Title Page

1. Manuscript Title

Practical Bayesian Inference in Neuroscience: Or How I Learned To Stop Worrying and Embrace the Distribution

2. Abbreviated Title

Practical Bayesian Inference in Neuroscience

3. List all Author Names and Affiliations in order as they would appear in the published article

Brandon S Coventry[1] , Edward L Bartlett[2]

[1]Department of Neurological Surgery and the Wisconsin Institute for Translational Neuroengineering, University of Wisconsin-Madison, Madison, WI USA 53705
[2]Weldon School of Biomedical Engineering, Department of Biological Sciences, and the Institute for Integrative Neuroscience, Purdue University, West Lafayette, IN USA 47907

4. Author Contributions:

BSC and ELB designed research, analyzed data, and wrote the manuscript. BSC performed research and provided analytic tools.

5. Correspondence should be addressed to

Edward L Bartlett

email: ebartle@purdue.edu

6. Number of Figures: 6
7. Number of Tables: 1
8. Number of Multimedia: 0
9. Number of Words in Abstract: 197
10. Number of Words for Significance Statement: 97
11. Number of Words for Introduction: 662
12. Number of Words for Discussion: 719
13. Acknowledgements:

Funding for experiments which generated young and aged responses to SAM stimuli was provided by the National Institutes of Health (NIDCD R01DC011580, PI: ELB). Funding for experiments generating INS dose-response data was provided by the Purdue Institute for Integrative Neuroscience collaborative training grant (PI: BSC).

14: Conflict of Interest

BSC and ELB hold a provisional patent (USPTO: 18/083,490) related to INS study data used in this tutorial.

15: Funding sources

ELB: NIDCD R01DC011580

BSC: Purdue Institute for Integrative Neuroscience Collaborative Training Grant.

16. Code and Data Repository

https://github.com/bscoventry/BayesianNeuralAnalysis

ABSTRACT

Typical statistical practices in biological sciences have been increasingly called into question due to difficulties in replication of an increasing number of studies, much of which is confounded by the relative difficulty of null significance hypothesis testing designs and interpretation of p-values. Bayesian inference, representing a fundamentally different approach to hypothesis testing, is receiving renewed interest as a potential alternative or complement to traditional null significance hypothesis testing due to its ease of interpretation and explicit declarations of prior assumptions. Bayesian models are more mathematically complex than equivalent frequentist approaches, which have historically limited applications to simplified analysis cases. However, the advent of probability distribution sampling tools with exponential increases in computational power now allows for quick and robust inference under any distribution of data. Here we present a practical tutorial on the use of Bayesian inference in the context of neuroscientific studies. We first start with an intuitive discussion of Bayes' rule and inference followed by the formulation of Bayesian-based regression and ANOVA models using data from a variety of neuroscientific studies. We show how Bayesian inference leads to easily interpretable analysis of data while providing an open-source toolbox to facilitate the use of Bayesian tools.

Significance Statement

Bayesian inference has received renewed interest as an alternative to null-significance hypothesis testing for its interpretability, ability to encapsulate prior knowledge into current inference, and robust model comparison paradigms. Despite this renewed interest, discussions of Bayesian inference are often obfuscated by undue mathematical complexity and misunderstandings underlying the Bayesian inference process. In this article, we aim to empower neuroscientists to adopt Bayesian statistical inference by providing a practical methodological walkthrough using single and multi-unit recordings from the rodent auditory circuit accompanied by a well-documented and user-friendly toolkit containing regression and ANOVA statistical models commonly encountered in neuroscience.

Keywords: Bayesian Inference, Neural Data Analysis, Statistical Inference

Introduction

Inference tools are foundational to all studies in neuroscience, providing the necessary machinery to make decisions and conclusions from data. Frequentist-based null significance hypothesis testing (NHST) has been the gold standard of inference in neuroscience and science at large in part due to the computational simplicity of frequentist models compared to permutation sampling or Bayesian-based methods. A significant problem present in the current practice of NHST, however, arises in the adoption of the p-value as the *de facto* metric of experimental "success", notorious for its difficulty in interpretation and correct usage (Krueger and Heck, 2019). The confluence of exponential increases in computational power with the wider discussion of problems with NHST usage has created renewed interest in Bayesian inference as an alternative to frequentist NHST while offering interpretability benefits over the p-value and NHST overall.

The use of p-values, the ubiquitous decision rule in frequentist methods, is fraught with problems due to fundamental misunderstandings of its use, interpretability, and most pathologically, its susceptibility to intentional and unintentional p-hacking(Nuzzo, 2014). Contrary to the initial intent of Ronald Fisher(Fisher, 1992), the p-value has often become the gatekeeper of significance in studies. In this role, it limits deeper observations into data, and it is often used without proper experimental design to ensure proper use and control. Methods of statistical inference require that one first define a statistical model with the power to adequately describe the data-generating process. Inference is then performed to estimate the population distribution from limited samples of observed data. Once estimates of population distributions are made, the determination of whether or not these distributions represent a significant effect is determined. NHST is somewhat a victim of its own success, where common practice has distilled the practice of NHST to chase the somewhat arbitrary $p<0.05$ measure of significance devoid of model or data considerations(Krueger and Heck, 2019). Furthermore, even in the best of experimental designs,

the p-value is a surrogate for arguably what a researcher is most interested in: how likely is it that observed data has some effect different from null(Kruschke, 2011; Gelman and Shalizi, 2013).

Bayesian methods offer a solution to the problem of the p-value, providing a direct measure of the probability that observations have some effect(Kruschke, 2011; Gelman and Shalizi, 2013). This is done by reallocation of probability of possibilities as parameters in a mathematical model of the data-generating process, leading to probabilistic estimates desired by but not attainable with p-value analyses. Bayesian methods are inherently data-driven; models are built with prior knowledge directly incorporated from parameters estimated directly from observed data.

Bayesian inference, though chronologically younger than frequentist approaches, was not adopted as the primary inference paradigm due to the computational demands necessary to solve inference problems outside of certain canonical forms(Bishop, 2006) and the adoption of frequentist interpretation of probability(Fienberg, 2006). Inference on arbitrary distributions required a deeper mathematical knowledge and computation of integrals which were potentially intractable without modern numerical integration techniques. Frequentist paradigms however were more easily adapted to computationally simple algorithms, allowing researchers to “do statistics” without extensive formal training. However, exponential increases in computational power with the development of powerful Markov chain Monte Carlo (MCMC) sampling methods now allow researchers to perform meaningful Bayesian inference on arbitrary distributions underlying observed data(Gilks et al., 1996).

The goal of this tutorial is to remedy the opacity that often accompanies discussions of Bayesian inference by providing simple, step-by-step walkthroughs of Bayesian inference with four common inference paradigms. We also aim to demonstrate the explanatory power of Bayesian inference in the context of neuroscience data. While the aim of this article is focused on application, this tutorial will begin

with a brief introduction to Bayes' rule and its constituent components necessary for inference. For more theoretical and mathematical considerations of Bayesian inference, see the following books and articles(Gerwinn et al., 2010; Colombo and Seriès, 2012; Bielza and Larranaga, 2014; Kruschke, 2014; Kruschke and Vanpaemel, 2015; Ma, 2019; Gelman et al., 2021; Van De Schoot et al., 2021).

Estimation of Spike Rates from Auditory Stimuli: A Motivating Example

To facilitate the discussion of Bayesian inference in neuroscience, consider an example found prominently in auditory neuroscience(Fig 1A-B). In our first experiment, single unit recordings were made from the inferior colliculus (IC) in response to applied sinusoidal amplitude-modulated tones (SAM, see SI Methods). The goal of this analysis is to create a linear model of SAM temporal auditory processing by quantifying increases in evoked single unit firing rates in response to decreased SAM modulation.

The linear regression model seeks to estimate a linear relationship between one (simple linear) or more (multilinear) predictor and measured variables. In this model, both the measured result and predictors are metric variables which map to a continuum of possible values. The simple linear regression model takes the form of:

$$y = \alpha + \beta x + \epsilon$$

where y is the measured (predicted) group, x is the predictor, β is the "slope" parameter dictating the relative increase or decrease in y per unit change in x, α is the intercept term which, in models of firing rate represents non-evoked, spontaneous firing rates, and ϵ is an error term which quantifies the difference between the expected value of y at a given x given a linear model versus the observed value of y at x. It should be noted that ϵ is not present in all regression models, but the authors suggest inclusion to quantify deviations from linear fit.

Linear regression thus forms a model in which AM depth predicts evoked firing rates in which the model parameters are estimated and used to draw conclusions about the relative dependency of y on x. To begin, an observation of the relative distribution of the measured data, in this case firing rates elicited from IC, will allow for robust inference model design. Inspection of the distribution of firing rates (Fig 1C) suggests that a log transform would allow for the data to be normally distributed, making model computations easier through use of canonical normal distributions. Before continuing to inference, it is important to describe the functional components of Bayesian inference's computational tool; Bayes rule.

Bayes' Rule

Foundational to Bayesian approaches is a complementary, but epistemically differing view of probability from that of frequentist approaches. While the frequentist perspective treats probability as the **relative frequency** of the occurrence of some event, the Bayesian perspective instead treats probability as the **expectation** of an event occurring which can be used to not only quantify the state of knowledge of an event, but also the uncertainty involved in measuring an event. Traditionally, the Bayesian perspective has been called 'belief', a perhaps unfortunate name which belies the fact that the Bayesian perspective of uncertainty of an event is fundamentally quantifiable. Perhaps a better description of Bayesian belief is instead quantification of the state of knowledge by accounting for uncertainty. The cornerstone of Bayesian inference is Bayes rule, defined as:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

where H is the quantification of the state of a hypothesis, and E is the quantification of observed evidence. In the context of inference, it is helpful to explicitly state the role of the model in Bayesian formulations:

$$P(\theta|E,M) = \frac{P(E|\theta,M)P(\theta|M)}{P(E|M)}$$

where M is the model of the data generating process and θ are the model parameters. The individual components of Bayes' rule are given names corresponding to the purpose they serve, with $P(\theta|E,M)$

called the posterior distribution, $P(E|\theta, M)$ the likelihood function, $P(\theta|M)$ the prior distribution, and $P(E|M)$ the evidence or marginal likelihood function. Taken together, Bayes' equation represents the quantification of observed data accounting for prior knowledge(Fig 2A). Each component plays a key role in Bayesian inference and each will be discussed briefly below.

The Model Evidence

The denominator term $P(E|M)$, called the model evidence (or just the evidence or marginal likelihood in Bayesian parlance) is the quantification of the probability of observing the data under a chosen model of the data generating function. At first glance, the calculation of the total evidence appears to be an insurmountable task. In reality this term is the weighted average of parameter values in a given model weighted by the relative probability of a given parameter value(Kruschke, 2014) and thus acts as a normalization term to ensure the numerator is a proper probability distribution. The structure of $P(E|M)$ will change based on whether the distributions represent probability mass functions (discrete case) or probability density functions (continuous case). In the discrete case, the evidence is

$$P(E|M) = \sum_{\theta} p(E|\theta, M)\, p(\theta|M)$$

and in the continuous case:

$$P(E|M) = \int p(E|\theta, M) p(\theta|M) d\theta$$

The evidence function thus represents an average of the likelihood function across all parameter values conditioned on the prior distribution. The marginal likelihood can also be utilized to assess the plausibility of two competing models(Johnson et al., 2023). The evidence, especially in the continuous case, is historically what made Bayesian inference difficult due to the need to evaluate a complex integral numerically. However, the advent of Markov-chain Monte Carlo (MCMC) methods with improvements in personal computer processing power has allowed for computationally efficient

integration without the need for supercomputing hardware. MCMC methods will be discussed in a subsequent section.

The Prior

The prior, $P(\theta|M)$, is often the major stumbling block for those entering into Bayesian inference, but this hurdle is less about the prior, and more about what the prior is perceived as. The prior, $P(H)$ describes the investigators prior beliefs on the state of knowledge of the study. Critics of Bayesian inference have described the prior as purely subjective, but we, and many others(Kruschke, 2010; Box and Tiao, 2011; Gelman and Shalizi, 2013), argue that the prior represents an explicit declaration of the investigators knowledge, assumptions, and the general state of a field which is implicit and often is present but not stated in frequentist approaches. Moreover, one is encouraged to perform prior predictive checks to compare the sensitivity of competing priors in a Bayesian inference model, as we will show subsequently. The practice of the design of experiments and their resulting publications are rife with implicit priors which are often not acknowledged or realized when reporting results. As an example, consider study of cortical extracellular single unit recordings(Paninski et al., 2004; Bartlett and Wang, 2007; De La Rocha et al., 2007; Coventry et al., 2023a as illustrative examples). The investigator could be leading a project with vast knowledge accumulated over years of study. Or the investigator is a trainee of a career researcher who draws a view of cortical physiology from their experienced mentor mixed with reading current literature. When designing an experiment, the investigator will have some intuition regarding likely and biologically feasible resting state and stimulus-evoked firing rates, cognitively assigning relatively low likelihood of seeing extremes of firing rates with higher likelihood assigned to moderate firing rates previously observed in literature or seen in experiments, and likely will discard or treat as outliers firing rates on the extremes or thought to be non-biological noise. The power of the prior distribution in Bayesian approaches is in part the need to explicitly quantify and report these prior beliefs, which can be analyzed and scrutinized as part of the peer review or post-publication process. Prior distributions

also require investigators to consider their biases and relative expectation on the importance of previously recorded and read data, promoting a deeper understanding of not only the data obtained within their lab, but also of the general state of the specific neuroscience field. As the name implies, prior beliefs are quantified as probability distributions by the investigators.

This begs the question as to what a prior might look like in newer avenues of study where a paucity of data exists. Or in situations where researchers and data analysts want the data to "speak for itself" outside any influence of the prior. In these cases, priors can be designed to be "non-informative" or "weakly-informative", assigning broad, non-committal distributions to the prior. One might assign a uniform distribution on the prior, effectively treating each parameter outcome as equally likely. Uniformly distributed priors do require some caution, however, as any parameter value outside of the bounds of the uniform distribution is automatically assigned probability 0 in the posterior, even if that value has been observed(Fig 2B). In many cases, it's better to allow small, but nonzero probabilities to extreme values, such as the tails of a normal distribution, such that evidence for unexpected events is represented in the posterior given strong data(Fig 2C). Conversely, priors can be made to be highly informative in situations where physiological bounds are well known and well-studied, where extreme values are known to be biophysically irrelevant or impossible or known to be due to instrument noise(e.g. large 50/60 Hz noise peak in power spectrum indicative of wall power noise).

The Likelihood

The likelihood function, $P(E|\theta, M)$ describes the probability that data is observed given parameter values θ in a data generating model M. In the context of inference, the likelihood function updates information given in a prior distribution to the posterior distribution given the observed data(Etz, 2018). The likelihood function is generally not a proper distribution, in that it is conditioned on yet unknown parameters and may not integrate to 1, but the evidence and prior terms ensures that resultant posterior distributions are

true probability densities. The idea of likelihood functions are present in both Bayesian and frequentist models, but has vastly different interpretations. The model parameters in a frequentist viewpoint converge upon singular values learned, usually though maximum likelihood estimation, from merging competing hypotheses of data. Bayesian approaches treat model parameters as ranges arising from distributions after observing the data at-hand.

The Posterior

The prior, likelihood, and evidence then form the posterior $P(\theta|E,M)$, the reallocation or mapping of probability from likelihood function, prior, and model evidence to an all-encompassing distribution. The posterior thus is the evidence for parameters θ conditioned on observed data and a model of the data generating function. The posterior forms the basis for inference, with all relevant information encoded in its distribution. Inference on the posterior distribution is covered in a section below.

Estimation of the Posterior

Despite Bayes' rule being formulated before Fisher's description of frequentist methods, a major reason that Bayesian inference was not been widely adopted was fundamentally a computational one, in that evaluation of Bayes' rule often requires solving non-trivial integrals. A subset of computationally tractable prior distributions and likelihood functions formed canonical posteriors in which the posterior is easily inferred. However, these cases are not generalizable to experimental data which can be noisy and not well behaved. Modern Markov-chain Monte-Carlo (MCMC) tools have been developed to quickly and easily estimate arbitrary distributions. MCMC involves the generation of random samples which converge to a target probability distribution, the details of which can be learned from the following reviews(Hoffman and Gelman, 2011; Betancourt, 2017).

Making Decisions on the Posterior

We define inference broadly as the process by which reasoning and decisions about a phenomena are made from a sample of observations of the phenomena. Classical NHST does not offer zero probability of error hypothesis testing(Blackwell, 1980). However, incorporation of prior knowledge in Bayesian inference allows for optimal decision making on observed data(Blackwell and Ramamoorthi, 1982). The posterior contains all necessary information to make inferences on experimental data incorporating prior knowledge. However, it is best to consider the specific goals of inference before performing statistics. Possible goals of inference are as follows(Kruschke, 2014):

- Infer the parameters of a model.
- Reject or confirm a null hypothesis
- Compare two or more competing models

In the case of neuroscientific studies, inferring model parameters occurs when an experiment aims to establish how neural firing rates change with changes in applied stimuli. Or one may want to confirm and reject a null hypothesis that a treatment has the desired effect or that there are differences between neural populations. Importantly, because the Bayesian inference operates solely on the posterior distribution, one can confirm or reject competing hypotheses and not simply reject the null as in frequentist NHST.

Regardless of the goal, inference always involves analyzing the posterior, which provides a complete representation of the distribution of a given parameter given the experimental data. Therefore, decisions about the data, the effect of model parameters, and/or which hypothesis has more evidence is performed with calculations on the posterior. There are a multiplicity of decision rules that can be used to assess the posterior. The most common, and in the author's opinion, the most intuitive is that of the Bayesian credible interval. The confidence interval calculates the probability that a population parameter lies in a certain interval. As credible intervals are not strictly unique, Bayesian inference convention is to fix the interval to the smallest interval which contains 95% of the posterior distribution density mass called the

highest density interval (HDI). Observations of posterior HDIs can then be used to assess the relative effect of a parameter. Regions of practical equivalence (ROPE) may be included in the posterior distribution that explicitly define a range of values that are effectively equivalent to a null value, with parameters considered significant if 95% of the posterior distribution (95% HDI) does not contain 0 or any values in the ROPE (Kruschke, 2018). Along with posterior HDIs, calculations of maximum *a posteriori* (MAP, distribution mode) estimates from the posterior are performed to quantify a most likely parameter value. While decision rules are important to assess the relative effect of statistical model parameters, we reiterate that simply passing a decision rule should not conclude the inference step. Inference should be made in context of the evidence presented in model quality checks, observed data posterior distributions, and decision metrics.

Error Quantification and Model Comparison

Critical to any statistical model and inference therein is its fit to observed data. While it is entirely possible to perform linear regression on data distributions which are highly nonlinear, predictions and inference made by the model will likely be inaccurate. Both Bayesian and frequentist inference offer robust model error quantification. Bayesian approaches, however, can utilize the posterior distribution to not only quantify and bound the distribution of model errors, but also include *post hoc* posterior predictive sampling as part of the inference paradigm. Posterior predictive sampling involves making random draws from the posterior and building a sampling distribution. This distribution is then compared to the observed data distribution to quantify the model's disparity from observed data. Along with posterior predictive checks, prior predictive checks act as a sensitivity measure of the influence of the prior distribution on the posterior distribution. Taken together, Bayesian inference thus allows for robust statistical inference on observed experimental data which appropriately includes prior knowledge of the state of the field.

Formulation of Models and Applied Bayesian Inference

There are a multiplicity of programs and programming languages that facilitate Bayesian analysis, such as standalone programs of Jasp(Love et al., 2019) and probabilistic programming language packages such as BUGS(Brooks, 2003) and STAN(Carpenter et al., 2017), we chose to use PyMC(Salvatier et al., 2016) for its ease in explicitly declaring probability distributions and its implementation in Python which is in common use in neuroscientific data analysis. Model formation is often conserved between frequentist and Bayesian approaches; it is only the mode of inference that differs. However, for clarity, we will discuss both model formation and performing inference in the subsequent sections.

Performing Bayesian Inference on the Linear Regression Model

Turning back to the example of IC single unit firing rates in response to SAM depth stimuli, the first step in inference is to place a prior distribution on the data. Previous studies and data can be used to inform the prior, but for this example we chose to demonstrate regression with moderately informative priors on $\alpha, \beta,$ and ϵ so as to let observed data drive posterior inference. Given that the data observed data is roughly normal, a good first pass is to place a normal distribution on the prior with mean equal to the mean of the observed data and a variance that is wide enough to capture all observed data. After inference is made, sensitivity analyses can be performed to assess the relative importance of the prior parameter values on posterior estimates. Larger prior variances allow for small, but non-zero probabilities on extreme values. This tends to be a more robust approach than setting a value of 0 on extreme events, as observed data with strong evidence for an extreme value can be adequately represented in the posterior. After observation of the underlying distribution of the observed data and decision on a prior distribution, a linear regression inference model can be easily described in code as follows:

Code Example 1: PyMC initialization of a simple linear regression model

```
    with pm.Model() as regression:        #Define a model that we call regression
        a = pm.Normal('a', mu=prMean, sigma = 5) #Normally distributed prior on a
        B = pm.Normal('B', mu=prMean, sigma = 5) #Normally distributed prior on B
        eps = pm.HalfCauchy("eps", 5)            #Model error prior
# Now we define our likelihood function, which for regression is our regression
function
        reg = pm.Deterministic('reg', a + (B*modDepth))
        likelihood = pm.Normal('Y',mu = reg, sigma = eps, observed = fir-
ingRate)
#Deterministic is for non probabilistic data. This is a modification to help sam-
pling, inference is still probabilistic.
```

The likelihood variable then translates our model to one of Bayesian inference by casting the model as a probability distribution, in this case

$$y \sim N(\alpha + \beta x + \epsilon)$$

noting that observed firing rates are incorporated by the 'observed' parameter in the likelihood distribution. To generate the posterior, all that needs to be done is to initialize and run MCMC as follows:

Code Example 2: Running the MCMC sampler

```
with regression:                #Access our defined model
        trace = pm.sample(numSamples, tune=numBurnIn, target_accept=0.90,chains = 4)
        #4 parallel, independent MCMC chains.
```

This routine then generates a trace variable containing the posterior distributions of all model parameters after sampling numSamples with numBurnIn samples to initialize chains. We also ran 4 chains in parallel with a target_accept probability of 90%. Acceptance probability is somewhat based on the statistics of observed data and model, with more difficult posteriors benefiting from higher accept probability values(Gilks et al., 1996). Improper acceptance probabilities can give rise to insufficient number of draws and malformation of posterior distributions. PyMC provides a helpful readout for when posterior draws are malignant and indicative of higher acceptance probabilities. In summary, in a few lines of code the researcher has observed distributions of the data and explicitly defined a model of the data generator

and likely now has a better intuition of the data and how it is distributed. All that's left to observe the posteriors with HDIs to infer significance from the model.

Plotting the 95% HDI estimation of the regression line (Fig 3a) on modulation depth vs natural log-transformed firing rates suggest a small but significant increase in firing rates with increases in modulation depth. Posterior distributions of model parameters (Fig. 3B) also show that there is an estimated basal firing rate above 0 (α MAP = 3.1) and a slope increase small but significantly above 0 (β MAP = 0.018) with model error terms considered small for being significantly smaller than intercept term (ε MAP = 0.74). The spread of the 95% HDI on inferred parameters is used as a measure of uncertainty of the parameter, with narrow HDIs representing more certainty in MAP estimated parameter. In our model, the α parameter has a spread between 3.02 to 3.13, with a difference of 0.11 containing 95% of its posterior distribution, suggesting strong certainty in the MAP estimate of 3.1. Similar narrow spread is see in the β parameter, with a difference of 0.007 containing 95% of the posterior. The model error term shows that observed data deviation from the model is constrained between 0.71 and 0.76 suggesting relative certainty in the magnitude of deviation of the data from the model.

Statistical conclusions should not end after making inferences on model parameters however. Critical to the validity of statistical inference is the quality of the model fit to observed data. This goodness of fit in Bayesian approaches can be analyzed by posterior predictive checks, in which sample draws are made from the posterior distribution, simulating observations of data generated from the experiment from which the statistical model was fit, and comparing sampled to observed data to assess deviation of model predictions from observed data distributions. In PyMC, posterior predictive checks can be easily performed using the following code:

Code Example 3: Performing posterior predictive checks

```
 """
Now let's do some posterior predictive checks. PyMC has some nice functions
that make this quite easy. We will also sample the posterior distribution for the
standard 16,000 samples, which for this posterior should be more than enough.
    """
    with regression:
      ppcRegression = pm.sample_posterior_predictive(trace, random_seed=Random-
      seed)
      #The above code envokes the regression mode, then uses the posterior from
      the trace, pulling synthetic samples to compare to observed. Random seed
      is set so that each run can be perfectly replicated
    az.plot_bpv(ppcRegression, hdi_prob=0.95,kind='p_value')
      #Bayes p-values, similar to frequentist,can be used to assess if posterior
      predictive is sufficiently close to observed density. Should be centered
      around 0.50.
    az.plot_ppc(ppcRegression)
    az.plot_trace(trace,var_names=['a', 'B','eps'])
    plt.show()
```

To illustrate how posterior predictive checks can be used, a competing model was made which performs Bayesian linear regression to the same data and priors except without log transformation of the data. In each case, random draws were made from each log transformed and non-log transformed posteriors to create empirical data distributions. Comparison of empirical distributions qualitatively show that log-transformed models present a better fit to observed data than non-log transformed models. The relative disparity between posterior predictive model fits and observed data can be quantified by use of Bayesian p-values, a distance measure between two distributions(for details of Bayesian p-values, see Kruschke, 2014). The closer the Bayesian p-value is to 0.5, the better data sampled from the posterior overlaps with the distribution of observed data. Plotting the resulting distributions and the Bayesian p-values indeed show the log-transformed model fits better to observed data than the non-transformed model. Similar analyses can be performed around model free parameters, such as prior variables, to form a sensitivity analysis of prior distributions on resulting posterior inferences.

A secondary and quick check of posterior sampling can be performed by qualitative evaluation of the MCMC sampling chains, often called traces. Traces represent the long term run of a Markov chain which represent the distribution of interest. As such, good traces show evidence of effective sampling and convergence to target probability distributions. PyMC offers easy ways to visualize posterior MCMC traces using the *plot_trace* function. Figure 3 shows traces obtained from our Bayesian regression example. Kernel density estimates of traces corresponding to the posterior distributions of regression parameters show good convergence of MCMC traces to a target distribution (Fig 3A). As MCMC chains are time series samples which form a distribution, evaluation of traces through sampling time can also be used as a diagnostic of sampling convergence. Traces should have a "fuzzy caterpillar" like appearance (Fig 3B) without any stark jump discontinuities from sample to sample. Quantitative trace evaluations are also available, with the Gelman-Rubin statistic ($\hat{r}$) being the most prominent. The Gelman-Rubin statistic measures the variance between MCMC chains to the within chain variance, effectively measuring chain stationarity and convergence(Gelman and Rubin, 1992). Heuristically, $\hat{r} < 1.05$ is considered good convergence of MCMC chains. This value can be calculated *post hoc* after sampling and PyMC will automatically flag if $\hat{r} \geq 1.05$ is detected.

While there are many reporting guidelines for Bayesian inference, we follow the Bayesian Analysis Reporting Guidelines as given by Kruscke(Kruschke, 2021) and provide an example reporting document including posterior predictive checks, Bayesian model comparisons, and sensitivity analysis as supplementary material.

Multilinear Regressions, Repeated Measures, and Hierarchical Models

In many experiments, inference across multiple possible data generating parameters must be analyzed and accounted for. These models, called multilinear regressions, are extensions of standard linear regression as follows:

$$y = X^T\beta + \epsilon \rightarrow y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_n x_n + \epsilon$$

where n is the total number predictors.

To illustrate the use of multilinear regressions, consider the case of thalamocortical infrared neural stimulation (INS)(Fig 5A). Auditory thalamic neurons in the medial geniculate body were excited by pulse trains of optical stimuli varying in pulse energy and time between pulses. The resulting auditory cortex single unit responses are recorded using a planar, Utah style array in layer 3/4. An important and understudied aspect of INS is the effect of laser energy and interstimulus interval changes on evoked firing rate responses; a so-called dose-response curve. We begin by specifying predicted and predictor values. Dose-response relationships were measured by predicting maximum firing rates in response to applied INS energy (E) and inter-pulse intervals (ISI). As we suspect an interaction between E and ISI, an interaction term of E*ISI was incorporated. Therefore, the model was defined as:

$$\max(FR) = \alpha + \beta_1 E + \beta_2 ISI + \beta_3 (E * ISI) + \epsilon$$

An important aspect of this study was that rats underwent chronic recordings through the duration of the lifetime of the implant. It almost a certainty that stimulation and recording quality will change over the lifetime of the devices due to neural adaptation to stimulation(Falowski et al., 2011) and glial response and encapsulation of the devices(Van Kuyck et al., 2007; Woolley et al., 2013). This experimental paradigm is thus complicated by potentially meaningful repeated measures within subject variability. Furthermore, slight differences in electrode and optrode placement between rodents could create a heterogeneity in the receptive fields of recorded neurons(Vasquez-Lopez et al., 2017), representing a potentially meaningful between-subject variance.

Hierarchical Structures Capture Latent Variables

Models in both Bayesian and frequentist paradigms capture these within and between subject variances by adding hierarchical structure to the model. From the Bayesian perspective, hierarchical models are

399 defined by allocating hyperparameters on the prior which encode within and between group variances in
400 the model, with each hyperparameter containing hyperprior distributions. Graphically, this is organized
401 in Fig 5B. Bayesian and frequentist hierarchical models share similar roots, with particular hyperprior
402 distributions in Bayesian paradigms becoming proportional to frequentist random effects models.
403
404 While this appears to be a herculean task in data modeling, PyMC allows for declarations of hierarchical
405 models, as shown in Code Snippet 4:
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421 *Code Example 4: Creating a hierarchical regression model*

```
animal_code_idx = data.animal_code.values       #Encodes within and between
subject variances as parameter indices
    with pm.Model() as Heirarchical_Regression:
        # Hyperpriors for group nodes
        mu_a = pm.Normal("mu_a", mu=0.0, sigma=1)
        sigma_a = pm.HalfNormal("sigma_a", 5)
        mu_b = pm.Normal("mu_b", mu=0.0, sigma=1)
        sigma_b = pm.HalfNormal("sigma_b", 5)
        mu_b2 = pm.Normal("mu_b2",mu=0.0, sigma=1)
        sigma_b2 = pm.HalfNormal("sigma_b2",5)
        mu_b3 = pm.Normal("mu_b3", 1)
        sigma_b3 = pm.HalfNormal("sigma_b3",5)

        sigma_nu = pm.Exponential("sigma_nu",5.0)
        #Base layer
        nu = pm.HalfCauchy('nu', sigma_nu)          #Nu for robust regression
        a_offset = pm.Normal('a_offset', mu=0, sigma=1, shape=(n_channels))
        a = pm.Deterministic("a", mu_a + a_offset * sigma_a)
        # Declare Regression parameters under a normal distribution
        b1_offset = pm.Normal('b1_offset', mu=0, sigma=1, shape=(n_channels))
        b1 = pm.Deterministic("b1", mu_b + b1_offset * sigma_b)

        b2_offset = pm.Normal("b2_offset",mu=0, sigma=1, shape=(n_channels))
        b2 = pm.Deterministic("b2", mu_b2 + b2_offset*sigma_b2)

        b3_offset = pm.Normal("b3_offset",mu=0, sigma=1, shape=(n_channels))
        b3 = pm.Deterministic("b3", mu_b3 + b3_offset*sigma_b3)
        #Add in the error term
        eps = pm.HalfCauchy("eps", 5,shape=(n_channels))
        #Declare regression model
         regression = a[animal_code_idx] + (b1[animal_code_idx] *
      XenergyPerPulse) + (b2[animal_code_idx] * XDist)
      +(b3[animal_code_idx]*XenergyPerPulse*XDist)
        #Encode model into likelihood function
         likelihood =
      pm.StudentT("MaxZ_like",nu=nu,mu=regression,sigma=eps[animal_code_idx],
      observed= MaxZ)
```

Owing to the scarcity of thalamocortical INS data, we assigned noninformative, wide spread normal distributions on the priors and hyperpriors so as to let the data speak for itself. We also utilized a student-T distribution as the likelihood function to accommodate outliers in a modification known as "robust

regression"(Kruschke, 2014). Student-T distributions have tails which are not bounded by the exponential function, meaning that extreme values have less impact or skew on the posterior distribution. Half-Cauchy distributions are placed on the error term and Student-T normality parameter ν. Half-Cauchy distributions are advantageous in learning scale parameters from the data in hierarchical models (Gelman, 2006; Polson and Scott, 2012).

It is important to validate that our model and data generating functions indeed represent the observed data. Sensitivity analyses and posterior predictive checks thus can be performed to ensure the model chosen is the one that best describes the observed data. Sensitivity analyses were performed by varying prior variance and comparing models which were nominal or natural log transformed with normal and student-T likelihood functions. Model comparisons can be performed in many ways, but a common paradigm is the leave-one-out cross validation (LOO)(Gelman et al., 2014). LOO consists of partitioning data into training and test sets and iteratively fitting the model under test with training data and testing out of sample fits with test data. Models are then ranked using the expected log pointwise predictive density (ELPD) measure:

$$ELPD = \sum_{i=1}^{k} \int dy_i p_t \bar{y_i} \log\left(p(\bar{y_i}|y)\right)$$

where p_t, y_i are unknown distributions representing the true data generating function for estimates of true posterior predictive function $(\bar{y}|y)$ from observed data y(Vehtari et al., 2017). In general, larger values of ELPD represent better out of sample fits indicative of a better model conditioned on observed data. We can then use standard errors between the model with the best ELPD (dse) and all competing models to rank all models to observed data. Importantly, these metrics should be understood only in the context of a model relative to other models, and not a global predictor of model validity. Observations of posterior fits to the data using posterior predictive fits and Bayesian p-values should be utilized on the final model

to determine model fit. This seemingly complex model comparison can be quickly and easily done in PyMC with the following commands:

Code Example 5: Model Comparisons

```
Var5_loo = az.loo(trace_Var5)
Var100_loo = az.loo(trace_Var100)
df_comp_loo = az.compare({"Var5": trace_Var5, "Var100": trace_100})
```

Model comparison results are given in Table 1. Similar to the simple regression above, the log transformed model provided much better fits to observed data than non-log transformed models. Interestingly and instructively, moderately informative priors (variance 5) outperformed noninformative priors (variance 100), suggesting that constraining prior variance can have predictive power in inference. Posterior predictive checks on the winning model show good fits to observed data with a Bayesian p-value near 0.5.

We can now perform inference on our multiregression model. It was found (Fig 5C) that α was significantly above 0 (MAP = 2.2, 95% HDI does not cross 0) suggesting that basal firing rates of recorded neurons were typically above 0 as expected. It was also seen that maximal firing rates were significantly dependent on applied INS energy (β_1MAP = 0.58, HDI does not cross 0) with increases in INS energy leading to larger evoked maximal firing rates. The relative spread of the 95% HDI on β_1 of 0.27-0.88 suggests a heterogeneity in neuron dose-response characteristics that can be explored more. Somewhat surprisingly, there was no significant effect of ISI on maximum firing rates (β_2 MAP = -0.055). The relative spread across 0 of -0.45 to 0.3 suggests that extreme values of ISI might potentially have an effect, with smaller ISIs causing neural integration of singular INS pulses into a singular, large pulse. However, that cannot be determined given the INS parameters used in this study. Also surprisingly, there was no

significant effect of Energy-ISI interactions (β_3 MAP = 0.028), suggesting that INS energy is the primary mediator of evoked firing rates.

Bayesian ANOVAs

Comparison of differences between groups is another routine statistical procedure used when predictor variables are nominal or categorical in nature or a mixture of metric and categorical predictors. The frequentist treatment of these experimental designs largely uses analysis of variances methods, namely ANOVA for categorical predictors and, more generally, ANCOVAs for categorical predictors with metric covariates. ANOVAs are models that take the form of:

$$y = \alpha + \sum_i \beta_i x_i$$

where β_i, x_i are the parameters corresponding to nominal predictor class i, α is the offset or bias parameter, and y is the metric dependent variable. ANOVA parameters and class values β_i, x_i are treated differently than the regression case, as x_i are categorical as opposed to continuous, metric values. As such x categories are recast into "one-hot" encoded vectors $\vec{x} = [x_0, x_1, \dots, x_i]$in which only a singular value in an array can have a value of 1 and all other elements are cast to 0, allowing for binary indication of a given class among a group of classes. If an individual value falls into group j, for example, $\vec{x}_{i \neq j} = 0, \vec{x}_{i=j} = 1$. The coefficients β_i then encodes the change in dependent variable y from inclusion of datapoint x in category i. Importantly, deflections from baseline are constrained such that $\sum_i \beta_i = 0$. Both Bayesian and frequentist ANOVA models treat β_i parameters as group deflections about the baseline level of the dependent variable.

ANCOVA is a modification to the ANOVA model to include a metric covariance term:

$$y = \alpha + \sum_i \beta_i x_i + \beta_{co}, x_{co}$$

where β_{co}, x_{co} are the parameters corresponding to metric predictors. Metric predictors terms are valuable in accounting for within group variance which is attributable to some other metric measurable variable, such as decreased firing rates in response to an applied stimulus found in a class of aged animals.

Bayesian analogues of ANOVA and ANCOVA can be easily defined in PyMC and are termed BANOVA and BANCOVA (Fig 5A) respectively to distinguish models from their frequentist counterparts. Traditional ANOVAs make two key assumptions; that underlying data is normally distributed and a homogeneity of variance among groups. To account for these assumptions, normal distributions are placed on prior parameter and observed data distributions and a uniform distribution prior is placed on observed data variance σ_y. Importantly, observed data distributions should be assessed to assure distributions are normally distributed. While not strictly an ANOVA-like structure, an advantage of Bayesian approaches is the ability to create models which handle arbitrary distributions. While traditional ANOVAs also assume independent group variances, the relative shared influence between groups can be learned from the data by imposing a hyperprior on group variance σ_β(Gelman, 2006). As with any prior distributions, selection of σ_β should be informed by prior inspection of the data. A Half-Cauchy distribution is once again chosen as it weakly informative and allows for extreme values if data dictates(Gelman, 2006; Polson and Scott, 2012). Setting σ_β to a large constant replicates a traditional ANOVA.

As a guiding example, consider a similar experiment to that done in simple linear regression. In this experiment, we aim to understand age-related changes in IC auditory processing of sinusoidal amplitude modulated sounds. This experiment consisted of two groups of young (animals < 6 months in age) and aged (animals > 22 months in age). SAM stimuli at increasing modulation depths were played to the

515 animals with evoked single unit responses recorded from IC. As seen in the regression experiment (Fig
516 2), there is a significant increase in evoked firing rate with increased modulation depth in young animals.
517 As such, it should be included in comparison between the two groups. Taken together, this suggests
518 BANCOVA will serve as an appropriate model. BANCOVAs are inherently hierarchical(Gelman, 2005;
519 Kruschke, 2014) (Fig 6A) to allow for between subject variances to be represented in the prior if these
520 variances mutually inform one another. Setting this hyperprior to a constant creates a model analogous
521 to a frequentist ANCOVA(Kruschke, 2014). The formation of the BANCOVA is again relatively
522 straightforward:

523 *Code Example 6: Creating a Bayesian ANCOVA*

```
with pm.Model() as BANCOVA:
        #Define hyperprior on sigma
        bSigma = pm.HalfCauchy('bSigma',2.0) #Recommended by Gelman, this
parameter doesn't overemphasize 0 on sigma.
        #Define Prior, likelihood distributions. Relatively noninformative
        a = pm.Normal('a',yMean,sigma = np.sqrt(yStDv))
        B = pm.Normal('B',0,sigma=bSigma,shape=numCategories)
        Bcov = pm.Normal('Bcov',yMean,sigma = np.sqrt(yStDv))
        sigmaLikelihood = pm.Uniform('sigmaLikelihood',yStDv/100,yStDv*10)
        BancovaModel = a + B[ClassAge] + (Bcov*(modDepth - modDepthMean))
        y = pm.Normal('y',mu=BancovaModel,sigma = yStDv,observed=firingRate)
        #Now, make sure model coefficients sum to 0 to create an ANOVA-like
structure
        aScaled = pm.Deterministic('aScaled',a+aesara.tensor.mean(B) + Bcov*(-
modDepthMean))
        bScaled = pm.Deterministic('bScaled',B - aesara.tensor.mean(B))
```

524

525 with inference made in the exact same way as the previous models.

526 After model sampling, posterior sampling checks were performed to ensure posterior distributions adhere
527 well to observed data. Posterior predictive distributions show good qualitative fit to observed firing rate
528 data with Bayesian p-values centered around 0.51, suggesting good model fits to observed data (Fig 6B).
529 Comparisons between groups is simple once posterior distributions are obtained. All that needs to be

done is to measure differences between aged and young group parameter posteriors (Fig 6C), encoding relative influence of young and age groups on firing rates. Aged and young contrasts show significantly elevated firing rates in young rats across all SAM stimuli (Young-aged difference MAP = 0.25, 95% HDI excludes 0). Another advantage of Bayesian inference is the ability to observe the distribution, and thus the most likely value and spread of effect size. In this analysis, the effect of age in SAM stimulus processing is significant but small (effect size MAP =0.058, 95% HDI excludes 0) but with a wide spread of effect (95% HDI between 0.025-0.64) suggesting variable temporal acuity between rodent subjects. Finally, firing rates vs SAM amplitude depth for each class are plotted with $y = \alpha + \beta_{young/age} x_{young/age} + \beta_{cov} x_{cov}$ superimposed.

Multiple Comparisons in Bayesian Inference

In traditional frequentist analyses, corrections for multiple comparisons are necessary in order to ensure that maximum Type I errors (false positives) are constrained to a maximum of 5% ($\alpha = 0.05$). With Bayesian inference, a posterior distribution across all parameters is obtained which remains unchanged no matter how many comparisons are made(Kruschke, 2014). Furthermore, frequentist type I errors are classically defined in the context of rejection of a null hypothesis. Bayesian inference is not strictly concerned with rejection of a null hypothesis, instead weighing competing hypotheses given observed data. Bayesian models are not immune to making false conclusions about data. These errors, called type M for errors in magnitude and type S for errors in sign occur when outliers in data exert too much influence on inference. These errors can be controlled by proper choice of priors or by building hierarchical models (Fig 5A, Fig 6A) which can account for outliers by pulling parameters towards group means when evidence is small and allowing parameters with good evidence to remain in a phenomenon called partial pooling implicit to hierarchical structures(Gelman et al., 2009).

Discussion

Bayesian inference approaches present a powerful statistical tool which encourages deep and meaningful exploration of data and allows for presentation of data in intuitive and transparent ways. In this tutorial, we demonstrate the ease by which Bayesian inference can be performed across a wide variety of experimental designs and provide source code which can be modified to accommodate neuroscientific experiments using all free and open source tools. We intentionally used the base PyMC toolchain in order to explicitly show Bayesian model creation. However, there are PyMC plugin tools such as Bambi (Capretto et al., 2022) which can facilitate creation of Bayesian models in single lines of code. An example of Bambi-enabled model creation is provided in our Bayesian inference toolbox.

Tempering Expectations of Bayesian Inference

Despite the enthusiasm of some Bayesian advocates, Bayesian inference is not a panacea. It is subject to similar problems as frequentist NHST, in that models can be used which do not adequately fit underlying data statistics or priors can be chosen which dominate model performance and deemphasize observed data. However, Bayesian approaches support and encourage model transparency, requiring researchers to declare model priors and posteriors while encouraging continued discussion of inference on data as opposed to stopping if a p-value is below an arbitrary threshold. A second caveat is that running MCMCs can be slower than frequentist approaches, with run times sometimes in minutes as opposed to seconds. However, time increases are not astronomical and can be further reduced to levels similar to frequentist approaches by using GPU computing or using programs such as JASP(Love et al., 2019) which utilize a C backend to speed up computation.

The Controversy of the Prior

The prior is arguably the most contentious aspect of Bayesian inference, with arguments that the prior unduly influences decisions on data. It is absolutely possible to have priors that distort posterior distributions into poor inference. Similar arguments can be levied at Frequentist approaches which

perform similar distortions on decision metrics, such as applying ANOVA tests when underlying data is not normal. Often times, these mistakes are not done out of malevolence, but due to the modern framework of how statistics is performed. We argue that having to consider what prior to use, and thus what one's assumptions are, what distributions are physiologically relevant, and the distributions of observed data will help to prevent errors in statistical modeling while creating greater transparency in how conclusions on data are drawn.

Decisions with Bayes Factors

Some studies which utilize Bayesian inference use a decision metric called a Bayes' factor, which is a measurement of the ratio of marginal likelihoods of two competing models providing log likelihood of evidence for one model over another(Johnson et al., 2023). We intentionally chose not to utilize Bayes' factor metrics because, in the authors' opinions, they reduce inference to evaluation of a single metric over an arbitrary threshold, as opposed to analysis over posterior distributions of observed data. Furthermore, certain prior declarations yield undefined Bayes' factors(Gelman and Rubin, 1995) potentially encouraging using suboptimum models in order to provide arbitrary decision metrics.

Bayesian and Frequentist Approaches: A Wholistic Approach to Inference

Following in the steps of Bayarri and Berger(Bayarri and Berger, 2004), data analysis should not consist solely of Bayesian or frequentist approaches devoid of the other. There are certainly cases where frequentist approaches should be used, such as clinical trials where preregistration and proper protocol design can provide bounds on false-positive and false negative rates necessary for translation of medical therapeutics. Hybrid frequentist and Bayesian approaches can also provide richer insight into analyses where posterior distributions are unidentifiable or difficult to sample(Raue et al., 2013) or in identifying when improper models have been chosen(Berger et al., 1997). Bayesian ideas of posterior predictive checks and model comparisons can also be applied to frequentist NHST, many of which would help

address problems of replication and data transparency. As frequentist approaches are often baked into the pedagogy of neuroscience and neural engineering, we aim for this tutorial to be a thorough introduction into the application of Bayesian statistics to help develop a toolkit which can be used for robust data analysis or in conjunction with previously established frequentist approaches. These models are also easily extendable into Bayesian analogs of logistic or multinomial regressions, gaussian mixture models, Bayesian time series analyses, among many more.

Code and Data Availability

The code/software described in the paper is freely available online at [URL redacted for double-blind review]. The code is available as Extended Data.

References

Bartlett EL, Wang X (2007) Neural Representations of Temporally Modulated Signals in the Auditory Thalamus of Awake Primates. J Neurophysiol 97:1005–1017.

Bayarri MJ, Berger JO (2004) The Interplay of Bayesian and Frequentist Analysis. Stat Sci 19 Available at: https://projecteuclid.org/journals/statistical-science/volume-19/issue-1/The-Interplay-of-Bayesian-and-Frequentist-Analysis/10.1214/088342304000000116.full.

Berger JO, Boukai B, Wang Y (1997) Unified frequentist and Bayesian testing of a precise hypothesis. Stat Sci 12 Available at: https://projecteuclid.org/journals/statistical-science/volume-12/issue-3/Unified-frequentist-and-Bayesian-testing-of-a-precise-hypothesis/10.1214/ss/1030037904.full.

Betancourt M (2017) A Conceptual Introduction to Hamiltonian Monte Carlo. Available at: https://arxiv.org/abs/1701.02434.

Bielza C, Larranaga P (2014) Bayesian networks in neuroscience: a survey. Front Comput Neurosci 8 Available at: http://journal.frontiersin.org/article/10.3389/fncom.2014.00131/abstract.

Bishop C (2006) Pattern Recognition and Machine Learning, 1st ed. New York, NY: Springer.

Blackwell DL (1980) There are no Borel SPLIFs. Ann Probab 8:1189–1190.

Blackwell DL, Ramamoorthi RV (1982) A Bayes but Not Classically Sufficient Statistic. Ann Stat 10:1025–1026.

Box GEP, Tiao GC (2011) Bayesian Inference in Statistical Analysis, 1st ed. Germany: Wiley.

631 Brooks SP (2003) Bayesian Computation: A Statistical Revolution. Philos Trans Math Phys Eng Sci
632 361:2681–2697.

633 Capretto T, Piho C, Kumar R, Westfall J, Yarkoni T, Martin OA (2022) **Bambi** : A Simple Interface for
634 Fitting Bayesian Linear Models in *Python*. J Stat Softw 103 Available at:
635 https://www.jstatsoft.org/v103/i15/.

636 Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Brubaker M, Guo J, Li P,
637 Riddell A (2017) *Stan* : A Probabilistic Programming Language. J Stat Softw 76 Available at:
638 http://www.jstatsoft.org/v76/i01/.

639 Colombo M, Seriès P (2012) Bayes in the Brain—On Bayesian Modelling in Neuroscience. Br J Philos
640 Sci 63:697–723.

641 Coventry BS, Lawlor GL, Bagnati CB, Krogmeier C, Bartlett EL (2023) Spatially specific, closed-loop
642 infrared thalamocortical deep brain stimulation. bioRxiv Available at:
643 http://biorxiv.org/lookup/doi/10.1101/2023.10.04.560859.

644 De La Rocha J, Doiron B, Shea-Brown E, Josić K, Reyes A (2007) Correlation between neural spike
645 trains increases with firing rate. Nature 448:802–806.

646 Etz A (2018) Introduction to the Concept of Likelihood and Its Applications. Adv Methods Pract Psychol
647 Sci 1:60–69.

648 Falowski SM, Sharan A, Reyes BAS, Sikkema C, Szot P, Van Bockstaele EJ (2011) An Evaluation of
649 Neuroplasticity and Behavior After Deep Brain Stimulation of the Nucleus Accumbens in an
650 Animal Model of Depression. Neurosurgery 69:1281–1290.

651 Fienberg SE (2006) When did Bayesian inference become "Bayesian"? Bayesian Anal 1 Available at:
652 https://projecteuclid.org/journals/bayesian-analysis/volume-1/issue-1/When-did-Bayesian-
653 inference-become-Bayesian/10.1214/06-BA101.full.

654 Fisher RA (1992) Statistical Methods for Research Workers. In: Breakthroughs in Statistics (Kotz S,
655 Johnson NL, eds), pp 66–70 Springer Series in Statistics. New York, NY: Springer New York.
656 Available at: http://link.springer.com/10.1007/978-1-4612-4380-9_6.

657 Gelman A (2005) Analysis of variance—why it is more important than ever. Ann Stat 33 Available at:
658 https://projecteuclid.org/journals/annals-of-statistics/volume-33/issue-1/Analysis-of-
659 variancewhy-it-is-more-important-than-ever/10.1214/009053604000001048.full.

660 Gelman A (2006) Prior distributions for variance parameters in hierarchical models (comment on article
661 by Browne and Draper). Bayesian Anal 1 Available at: https://projecteuclid.org/journals/bayesian-
662 analysis/volume-1/issue-3/Prior-distributions-for-variance-parameters-in-hierarchical-models-
663 comment-on/10.1214/06-BA117A.full.

664 Gelman A, Carlin J, Stern H, Dunson D, Vehtari A, Rubin D (2021) Bayesian Data Analysis, 3rd ed. Boca
665 Raton: Chapman and Hall/CRC.

Gelman A, Hill J, Yajima M (2009) Why we (usually) don't have to worry about multiple comparisons. Available at: https://arxiv.org/abs/0907.2478.

Gelman A, Hwang J, Vehtari A (2014) Understanding predictive information criteria for Bayesian models. Stat Comput 24:997–1016.

Gelman A, Rubin DB (1992) Inference from Iterative Simulation Using Multiple Sequences. Stat Sci 7:457–472.

Gelman A, Rubin DB (1995) Avoiding Model Selection in Bayesian Social Research. Sociol Methodol 25:165.

Gelman A, Shalizi CR (2013) Philosophy and the practice of Bayesian statistics: *Philosophy and the practice of Bayesian statistics*. Br J Math Stat Psychol 66:8–38.

Gerwinn S, Macke JH, Bethge M (2010) Bayesian inference for generalized linear models for spiking neurons. Front Comput Neurosci 4 Available at: http://journal.frontiersin.org/article/10.3389/fncom.2010.00012/abstract.

Gilks WR, Richardson S, Spiegelhalter DJ (1996) Markov Chain Monte Carlo in Practice. Boca Raton: Chapman and Hall/CRC.

Hoffman MD, Gelman A (2011) The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. Available at: https://arxiv.org/abs/1111.4246.

Johnson VE, Pramanik S, Shudde R (2023) Bayes factor functions for reporting outcomes of hypothesis tests. Proc Natl Acad Sci 120:e2217331120.

Krueger JI, Heck PR (2019) Putting the *P* -Value in its Place. Am Stat 73:122–128.

Kruschke JK (2010) What to believe: Bayesian methods for data analysis. Trends Cogn Sci 14:293–300.

Kruschke JK (2011) Bayesian Assessment of Null Values Via Parameter Estimation and Model Comparison. Perspect Psychol Sci 6:299–312.

Kruschke JK (2014) Doing Bayesian Data Analysis: A tutorial with R, JAGS, and stan, 2nd ed. Academic Press. Available at: http://www.indiana.edu/~kruschke/DoingBayesianDataAnalysis/.

Kruschke JK (2018) Rejecting or Accepting Parameter Values in Bayesian Estimation. Adv Methods Pract Psychol Sci 1:270–280.

Kruschke JK (2021) Bayesian Analysis Reporting Guidelines. Nat Hum Behav 5:1282–1291.

Kruschke JK, Vanpaemel W (2015) Bayesian Estimation in Hierarchical Models. In: The Oxford Handbook of Computational and Mathematical Psychology, pp 279–299. Oxford University Press.

Love J, Selker R, Marsman M, Jamil T, Dropmann D, Verhagen J, Ly A, Gronau QF, Smíra M, Epskamp S, Matzke D, Wild A, Knight P, Rouder JN, Morey RD, Wagenmakers E-J (2019) **JASP** :

698 Graphical Statistical Software for Common Statistical Designs. J Stat Softw 88 Available at:
699 http://www.jstatsoft.org/v88/i02/.

700 Ma WJ (2019) Bayesian Decision Models: A Primer. Neuron 104:164–175.

701 Nuzzo R (2014) Scientific method: Statistical errors. Nature 506:150–152.

702 Paninski L, Fellows MR, Hatsopoulos NG, Donoghue JP (2004) Spatiotemporal Tuning of Motor Cortical
703 Neurons for Hand Position and Velocity. J Neurophysiol 91:515–532.

704 Polson NG, Scott JG (2012) On the Half-Cauchy Prior for a Global Scale Parameter. Bayesian Anal 7
705 Available at: https://projecteuclid.org/journals/bayesian-analysis/volume-7/issue-4/On-the-Half-
706 Cauchy-Prior-for-a-Global-Scale-Parameter/10.1214/12-BA730.full.

707 Raue A, Kreutz C, Theis FJ, Timmer J (2013) Joining forces of Bayesian and frequentist methodology: a
708 study for inference in the presence of non-identifiability. Philos Trans R Soc Math Phys Eng Sci
709 371:20110544.

710 Salvatier J, Wiecki TV, Fonnesbeck C (2016) Probabilistic programming in Python using PyMC3. PeerJ
711 Comput Sci 2:e55.

712 Van De Schoot R, Depaoli S, King R, Kramer B, Märtens K, Tadesse MG, Vannucci M, Gelman A, Veen
713 D, Willemsen J, Yau C (2021) Bayesian statistics and modelling. Nat Rev Methods Primer 1:1.

714 Van Kuyck K, Welkenhuysen M, Arckens L, Sciot R, Nuttin B (2007) Histological Alterations Induced
715 by Electrode Implantation and Electrical Stimulation in the Human Brain: A Review.
716 Neuromodulation Technol Neural Interface 10:244–261.

717 Vasquez-Lopez SA, Weissenberger Y, Lohse M, Keating P, King AJ, Dahmen JC (2017) Thalamic input
718 to auditory cortex is locally heterogeneous but globally tonotopic. eLife 6:e25141.

719 Vehtari A, Gelman A, Gabry J (2017) Practical Bayesian model evaluation using leave-one-out cross-
720 validation and WAIC. Stat Comput 27:1413–1432.

721 Woolley AJ, Desai HA, Otto KJ (2013) Chronic intracortical microelectrode arrays induce non-uniform,
722 depth-related tissue responses. J Neural Eng 10:026007.

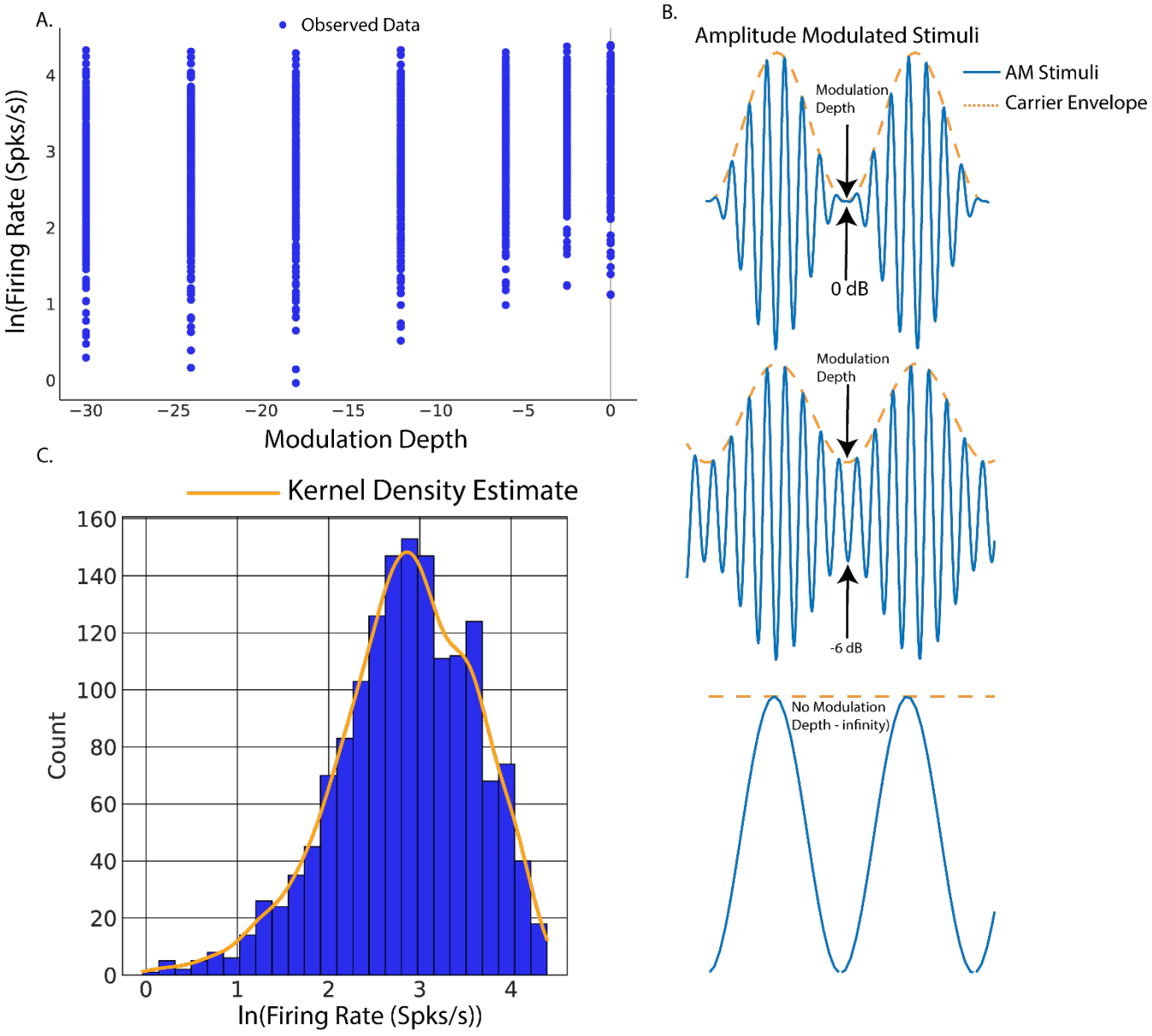
723

724

725

726

727 **Figures**



728

729 Figure 1: Example of Bayesian simple linear regression on population estimates of firing rate vs
730 amplitude modulation depth stimuli. This model was applied to population single unit firing rates
731 elicited from inferior colliculus with sinusoidal amplitude modulated (SAM) tones. The goal of this
732 model was to predict evoked firing rates from increases in SAM modulation depths. A. Scatterplot of
733 observed firing rates vs SAM modulation depth and fitted regression estimates. B. Schematic of

734 amplitude modulated stimuli. C. Kernel density estimates of the observed log transformed data
735 probability distribution function. C. An example of Bayesian model comparison. Left: Regression
736 model with untransformed data. Right: natural log transformed firing rate model. Posterior predictive
737 checks reveal that natural log transformed firing rate models better match observed data.

738

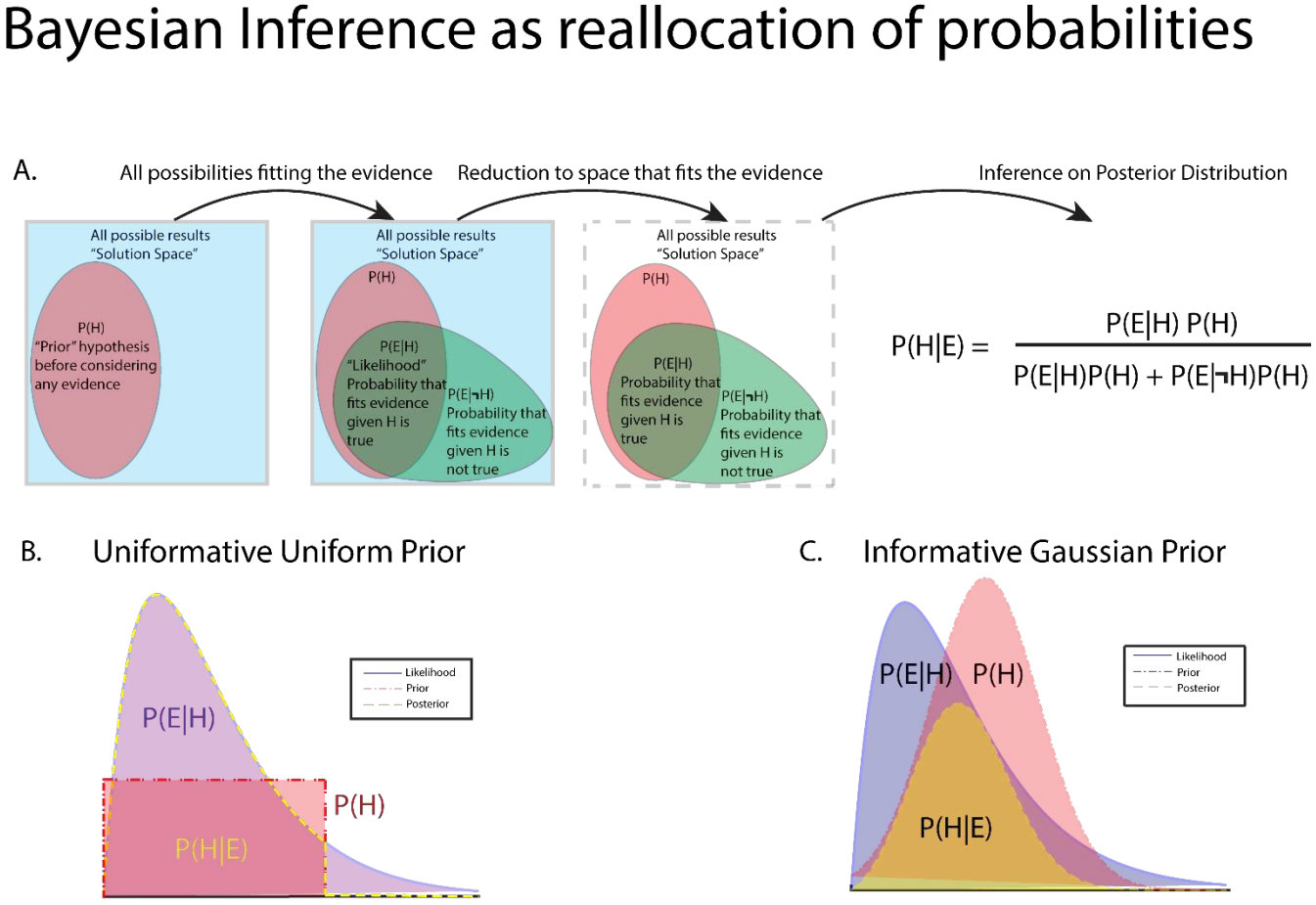
739

740

741

742

743



744

745 **Figure 2:** Graphical description of Bayes rule and the interaction between prior distributions and
746 likelihood functions leading to the final posterior distribution. A. Bayes rule can be thought of as a
747 reallocation of probability to the posterior after accounting for prior distributions and observed
748 evidence. B. An example of posterior generated from an inverse-Gamma distributed likelihood and a
749 uniformly distributed prior. Uniform priors reflect the likelihood function, and thus the observed data
750 with no redistribution probability, making uniform distributions uninformative priors. However, care
751 must be taken in using uniform distributions as observed data outside of prior bounds is mapped to 0
752 probability. C. An example of a posterior generated from an inverse-Gamma distributed likelihood and a
753 gaussian distributed prior. This prior is considered informative as it shapes the posterior distribution to a
754 greater extent than a uniform distribution. Prior distributions with longer tails can handle extremes of

observed data by mapping extreme events to low, but non-zero representation in the posterior. Examples B and C represent extremes of prior choices, with minimally informative priors often chosen to let the data "speak for itself" with little change to posterior from prior influence.

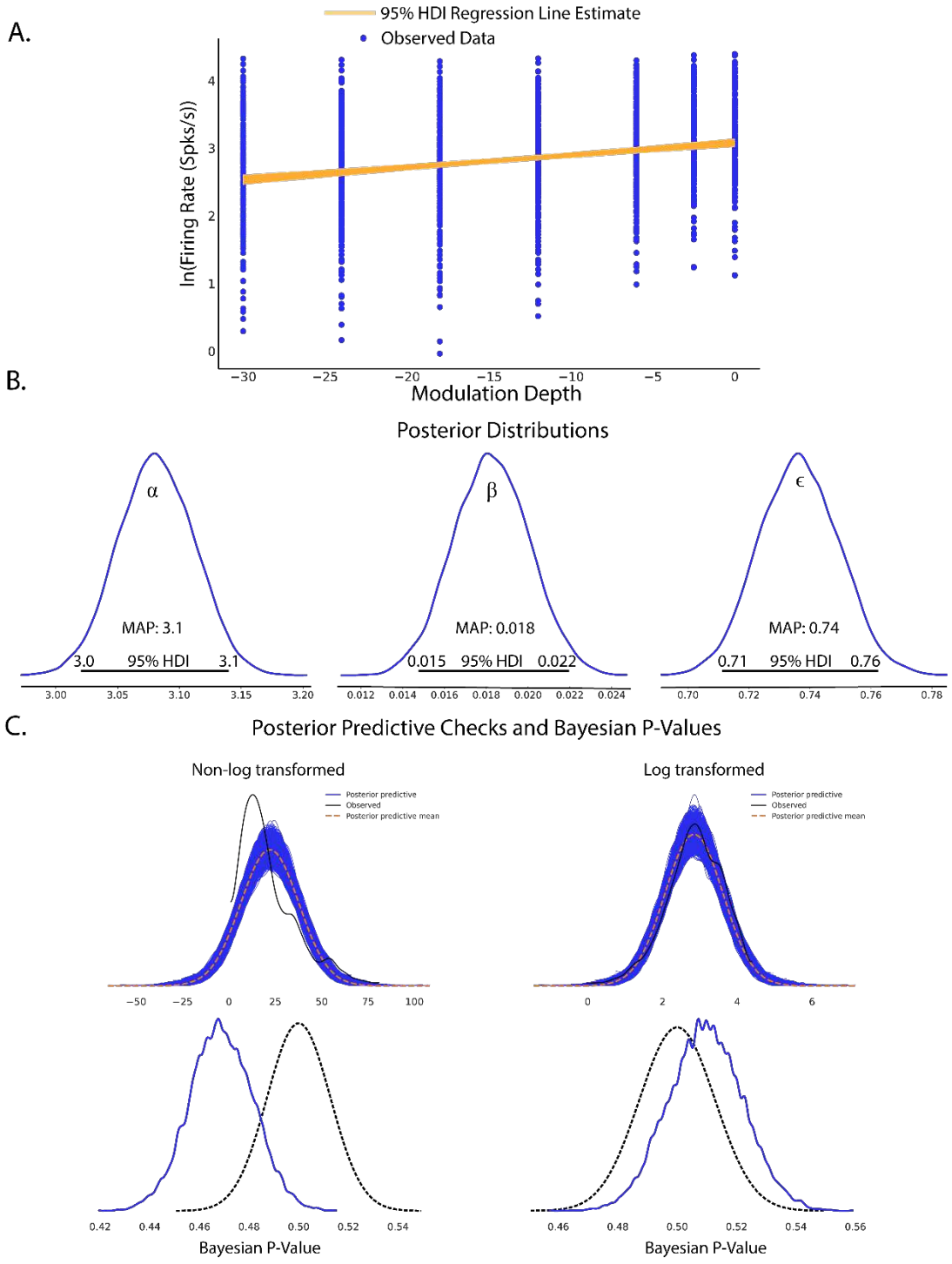


Figure 3: Completed Bayesian inference quantifying linear relationships in evoked firing rate from increases in modulation depth. A. Scatterplot of observed firing rates vs SAM depth stimuli with fitted regression line estimates superimposed. 95% HDI estimates of regression slopes are shown in orange,

776 with the spread of lines encoding the 95th percentile of most likely slope values. B. Estimates of
777 Bayesian linear regression parameters. Intercept term α was significantly above 0 (MAP = 3.1, 95%HDI
778 does not overlap 0) which indicates basal firing rates above 0. Regression slope was small but
779 significantly above 0 (MAP = 0.018, 95% HDI does not overlap 0) suggesting an increase in evoked
780 firing rates with increased modulation depth. Error term ϵ was significantly above 0 (MAP = 0.74, 95%
781 HDI does not overlap 0) suggesting some model deviation from observed data. However, error terms
782 were considered small as ϵ MAP $<$ α basal firing rate MAPs. C. Posterior predictive checks of linear
783 (left) and log linear (right) regression models show that log transformed firing rate models produce
784 posterior predictions most inline with observed data. Disparity of empirical posterior predictive
785 distributions from observed data as quantified through Bayesian P-values also suggest log transformed
786 firing rates creates a superior model fit.

787

788

789

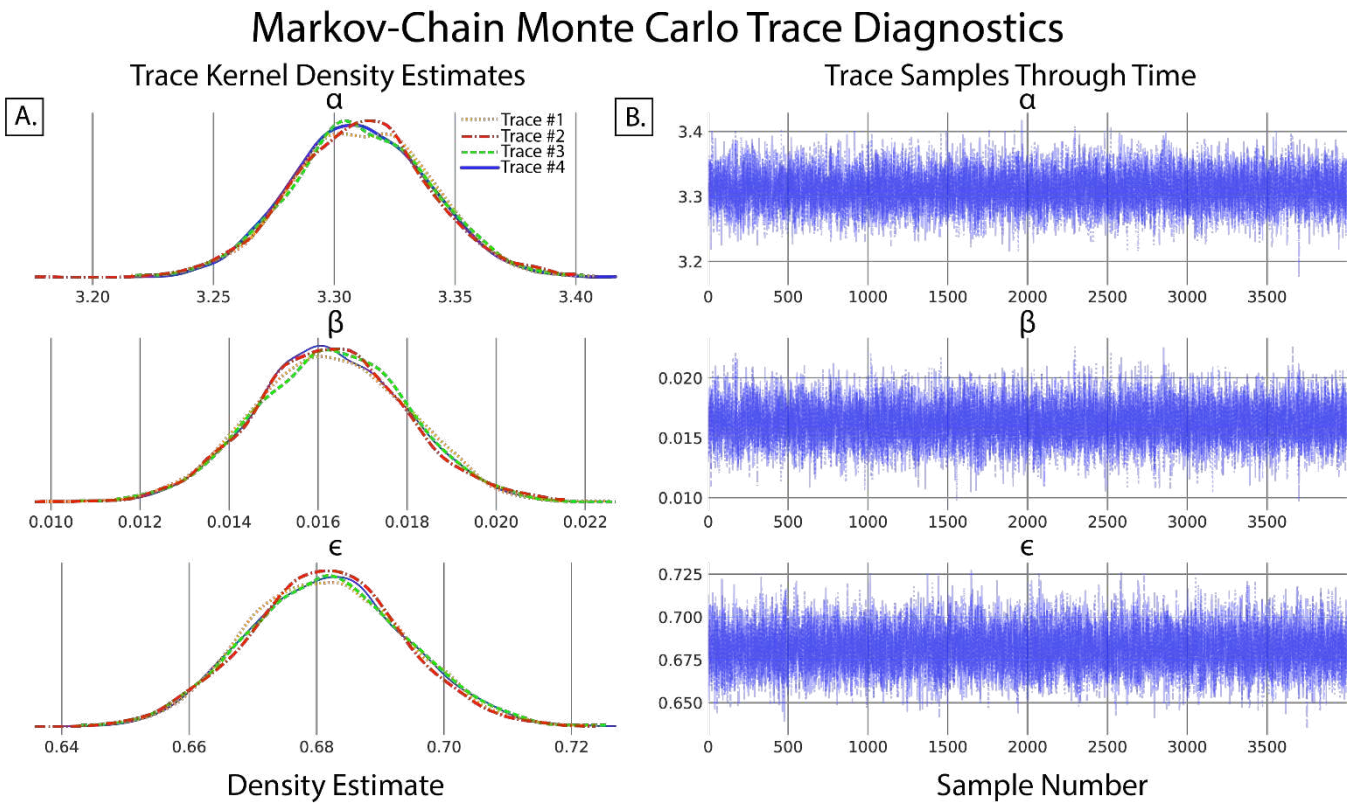
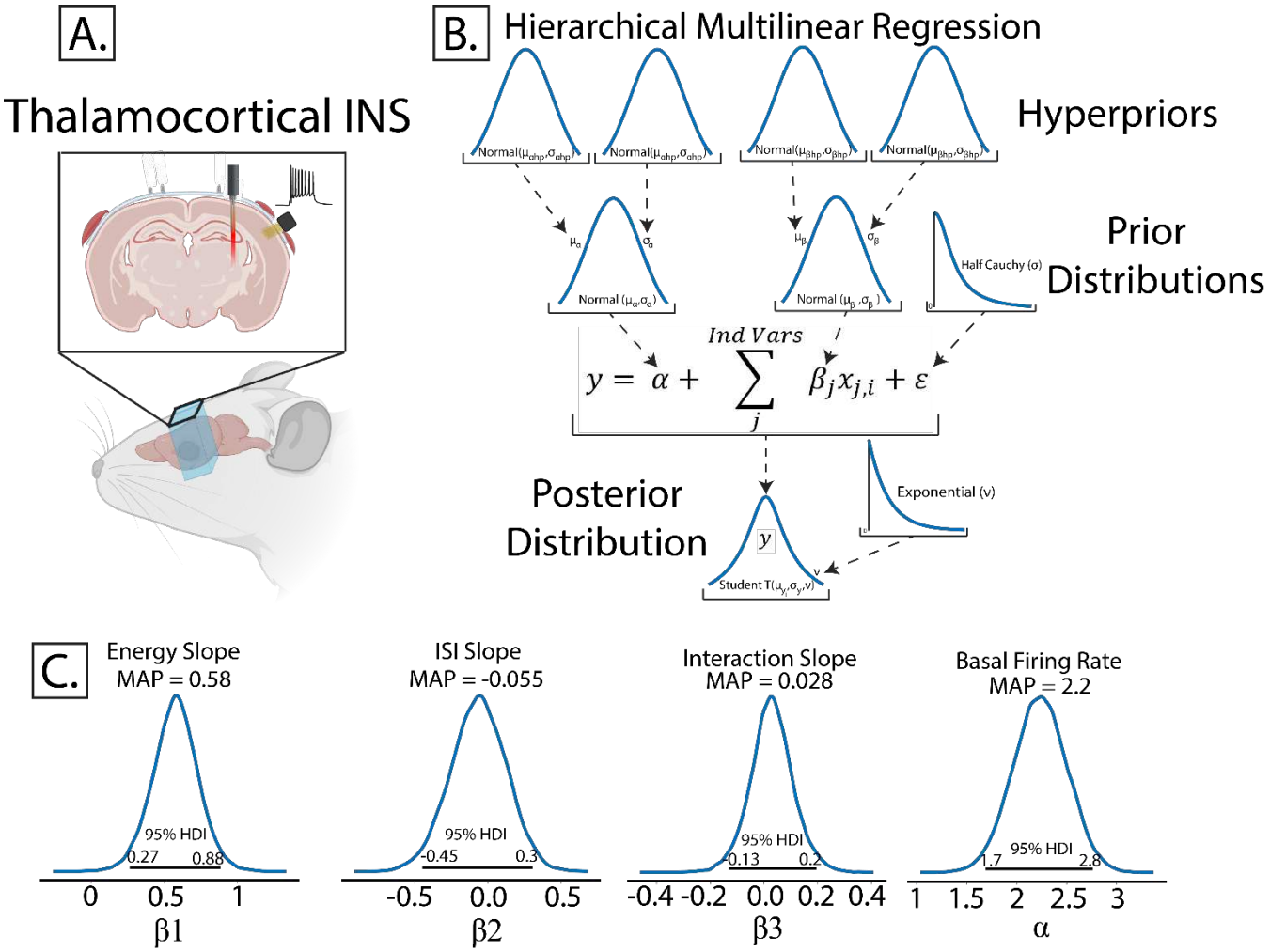


Figure 4: Evaluation of Markov-chain Monte Carlo (MCMC) chains can help diagnose ill fitting distributions. A. Kernel density estimates of the marginal posteriors corresponding to each of the regression parameters of each MCMC trace. Qualitatively, chain distributions should appear similar to each other, suggesting good convergence to target distributions. B. Time series plot of trace value vs sample number of marginal posteriors corresponding to each regression parameter. Qualitatively good traces should have a "fuzzy caterpillar" like shape, evident in all parameters of this model, indicative of good integration over the joint posterior distribution and effective sampling of the posterior.



801

802 **Figure5:** Example of Bayesian multilinear regression incorporating a hierarchical structure. A. In this
803 experiment, rodents were implanted with fiber optic arrays into auditory thalamus and planar recording
804 arrays into auditory cortex. Single unit responses were recorded from INS stimuli with applied energy
805 and interstimulus intervals varied to derive dose-response curves. Figure was drawn using BioRender
806 under publication license (www.biorender.com). B. Hierarchical schematic of Bayesian multilinear
807 regression. Hierarchical structures are advantageous in accounting for within and between subject
808 variability or for repeated measures designs. C. Resulting parameter distributions from dose-response
809 models. Energy was a significant contributor to maximum firing rate, with increasing laser energy
810 resulting in increased maximum firing rate, as determined by 95% HDI of the laser energy term β_1

excluding 0 (MAP = 0.58). Laser pulse interstimulus interval did not significantly contribute to changes in max firing rate as indicated by ISI parameter β_2 overlapping 0 in its 95% HDI with a MAP value near 0 (MAP = 0.028). The relatively wide spread about zero does suggest that there may be a subset of ISIs which contribute more strongly to firing rates and warrants further study. Laser energy-ISI interactions also did not significantly contribute to max firing rate as evidenced by interaction parameter β_3 including 0 in its 95% HDI. The intercept term α, correspondint to basal firing rates, were significantly above 0 (MAP = 2.2, 95% HDI excludes 0).

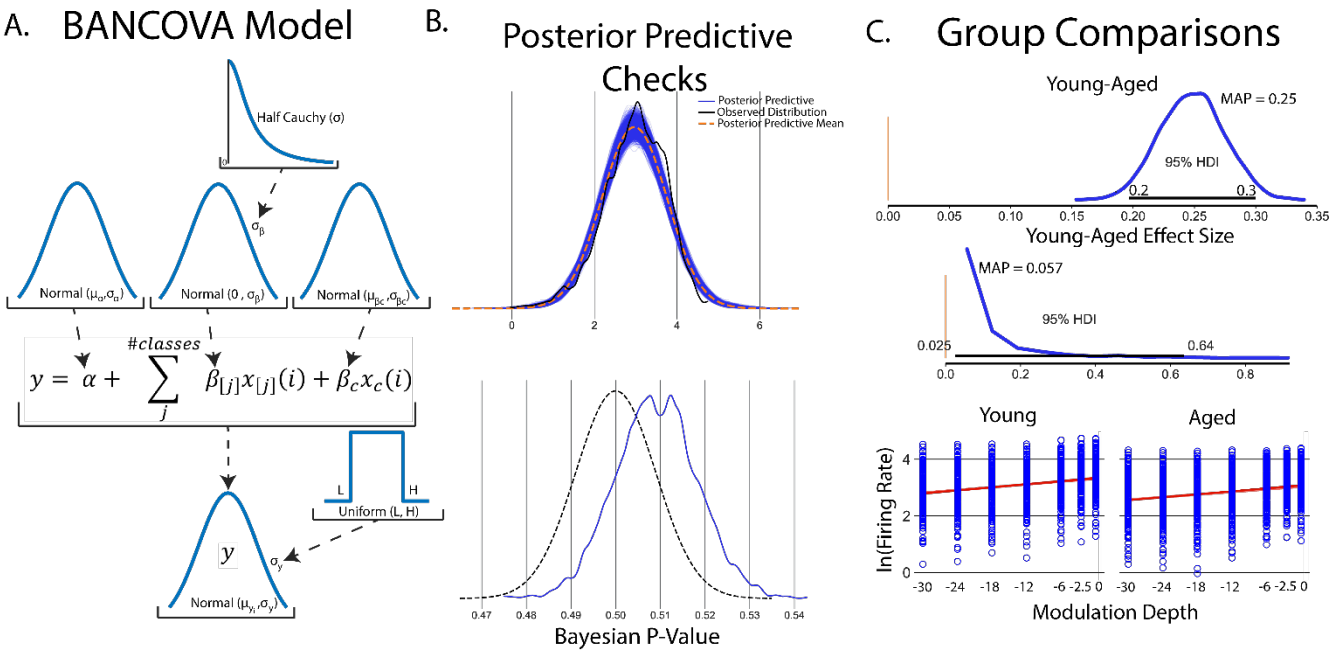


Figure 6: An example of Bayesian inference using ANOVA-like models. A. General schematic of BANOVA/BANCOVA models. Traditional ANOVAs have two key assumptions; normality of group data and homogeneity of variance. Normality of group data is imposed in BANOVA-like models as normal distributions around group parameters with homogeneity of variance encoded as a uniform distribution around posterior variance term σ_y. Traditional ANOVAs assume a fixed variance on group parameter values σ_β, imposing the constraint that each group is estimated independently from each other group. A uniquely Bayesian approach is to instead learn σ_β values from the data itself by placing a distribution on σ_β. B. Posterior predictive checks suggest posterior distributions show good fit in mean and variance to observed data. C. Once posterior distributions are calculated, group comparisons can be easily done by subtracting young and aged posteriors to yield a contrast distribution. It is found that firing rates across all modulation depths are significantly higher in aged vs young rodents (contrast MAP = 0.25, 95% HDI does not overlap 0). Another unique feature of Bayesian approaches is the ability to assess distributions on effect size. In this BANCOVA, while group differences are significant,

their relative effective size is small but significant (effect size MAP = 0.057, 95% HDI does not cross 0) suggesting marginal impact of age on firing rates elicited from SAM stimuli. Finally, metric covariates of firing rate in response to varying SAM depth in young and aged groups can be plotted as regressions superimposed on raw data.

Tables

Table 1: LOO Model comparisons and sensitivity analyses

Model	R	ELPD	DSE
St Log Var 5	1	-5337.48	0.00
ST Log Var 100	2	-5337.62	0.420867
ST Log Var 0.5	3	-5337.76	0.409773
St Log Var 25	4	-5338.15	0.492297
ST Log Var 10	5	-5338.18	0.300197
ST Log Var 1	6	-5338.26	0.331152
N Log Var 10	7	-5340.60	3.308668
N Log var 1	8	-5341.09	3.293779
N log var 5	9	-5341.16	3.296273
N log var 0.5	10	-5342.46	3.300550
ST Semilog Var 1	11	-5466.76	15.845916
St Semilog var 5	12	-5467.12	15.856552
ST semilog var 10	13	-5467.15	15.895646
ST semilog var 0.5	14	-5467.18	15.866405
ST Var 1	15	-15336.31	79.406629
ST var 0.5	16	-15355.67	80.415787
St var 5	17	-15355.67	80.415787
N var 10	18	-16119.11	82.384329

858

859

860

861

N var 1	19	-16132.23	83.549811
N var 0.5	20	-16154.55	84.262219

Supplementary Material Outline

1. Materials and Methods

Bayesian inference was performed on a range of data typical to neuroscience experiments. Regression models, ANOVA models, and group comparisons are performed on single-unit activity recorded from inferior colliculus (IC) neurons in response to auditory stimuli in young and aged rats(Palombi et al., 2001; Simon et al., 2004; C.F et al., 2012; Herrmann et al., 2017). Random-effects regression models are performed on single units recorded in the auditory cortex (A1) using high-density recording arrays in response to infrared neural stimulation(Izzo et al., 2007; Cayce et al., 2011, 2014; Coventry et al., 2023) of the medial geniculate body (MGB). To underscore that meaningful Bayesian inference does not require cluster computing or extensive computational resources, all computations were performed on an MSI GS-66 laptop with an Intel i7 processor with an Nvidia RTX2070 GPU. Our inference programs are CPU-bound, not requiring any GPU resources. Computations can be performed on most modern CPUs, but accelerate with more CPU threads and cores and parallelization on GPUs. All surgical procedures used in this study were approved by [redacted for double-blind review].

1.1 Disruption of Temporal Processing in the Inferior Colliculus Due to Aging

The inferior colliculus (IC) is the major integrative center of the auditory pathway, receiving excitatory inputs from ventral and dorsal cochlear nuclei, excitatory and inhibitory inputs from the lateral and medial superior olivary complex(Kelly and Caspary, 2005) and inhibitory inputs from superior paraolivary nucleus and the dorsal and ventral nuclei of the lateral lemniscus(Cant and Benson, 2006; Loftus et al., 2010). The IC encodes auditory information through hierarchical processing of input synaptics with local IC circuitry(Caspary et al., 2002; Rabang et al., 2012; Grimsley et al., 2013; Coventry et al., 2017). Age-related changes in auditory processing primarily arise as deficits in temporal processing(Frisina and Frisina, 1997; Parthasarathy et al., 2010; Parthasarathy and Bartlett, 2012; Herrmann et al., 2017). This dataset is composed of single unit responses recorded from young (Age$\leq$ 6 months) and aged (age $\geq$ 22 months) Fisher 344 rats. Auditory brainstem responses were recorded from animal subjects a few days prior to surgery to ensure hearing thresholds were typical of the rodent's age. Single unit recordings were performed in a 9'x9' double-walled, electrically isolated anechoic chamber (Industrial Acoustics Corporation). Animals were initially anesthetized via a bolus injection of ketamine (VetaKet, 60-80 mg/kg) and medetomidine (0.1-0.2 mg/kg) mixture via intramuscular injection. Oxygen was maintained via a manifold and pulse rate and blood oxygenation monitored through pulse oximetry. Supplemental doses of ketamine/medetomidine (20 mg/kg ketamine, 0.05 mg/kg medetomidine) were administered intramuscularly as required to maintain surgical plane of anesthesia. An incision was made down midline and the skull exposed. Periosteum was resected and a stainless steel headpost was secured anterior to bregma via 3 stainless steel bone screws. A craniectomy was made above inferior colliculus (-8.5 anterior/posterior, 1 mm medial/lateral from bregma). A single tungsten electrode was advanced dorsally towards the central nucleus of the inferior colliculus (ICC) during which bandpass noise (200 ms, center frequencies 1-36kHz in five steps per octave, 0.5 octave bandwidth) was delivered. ICC was identified based on short-latency driven responses to bandpass noise search stimuli with ascending tonotopy and narrowly tuned responses to pure tones of varying frequencies. Once neurons were identified, responses from 5-10 repetitions of sinusoidal amplitude-modulated tones (750 ms tone length, modulation depth between -30 to 0 dB) were recorded using a preamplifying headstage (RA4PA, Tucker-Davis Technologies) and discretized at a sampling rate of 24.41 kHz (RZ-5, TDT). Sinusoidal amplitude-modulated tones were defined as:

$$s(t) = A[1 + m * cos(2\pi f\_m\, t + \varphi)] * n(t)$$

where m is modulation depth ranging between 0.032-1 (-30 – 0 dB), f_m the modulation frequency, φ the reference phase of the modulator, A the scaling factor for stimulus sound level, and $n(t)$ the broadband noise stimulus. Single units were filtered between 0.3 and 5 kHz. Offline spike sorting was performed using OpenExplorer (TDT).

1.2 Thalamocortical Infrared Neural Stimulation

Infrared neural stimulation (INS) is an optical technique using coherent infrared light to stimulate nerves and neurons without the need for genetic modification of the target or direct contact with tissue that offers spatially constrained activation above electrical stimulation(Wells et al., 2005; Izzo et al., 2007; Cayce et al., 2011, 2014; Coventry et al., 2020, 2023). In this study, rats were chronically implanted in A1 with 16 channel planar Utah-style arrays (TDT, Alacua FL) and stimulating optrodes in the medial geniculate body of auditory thalamus (Thor Labs, Newton NJ). Rodents were initially anesthetized with a bolus injection of a ketamine (80 mg/kg) and medetomidine (0.2 mg/kg) cocktail. Oxygen was maintained via a manifold and pulse rate and blood oxygenation monitored through pulse oximetry. Supplemental doses of ketamine/medetomidine (20 mg/kg ketamine, 0.05 mg/kg medetomidine) were administered intramuscularly as required to maintain surgical plane of anesthesia. An incision was made down midline and the skull exposed. The periosteum was removed via blunt dissection and 3 stainless steel bone screws were placed in skull for headcap stability. An additional titanium bones crew was placed in skull to serve as a chronic ground and reference point for recording electrodes. Craniectomies were made above medial geniculate body (-6 anterior/posterior, -3.5 medial/lateral from bregma) and auditory cortex (-6 anterior/posterior, -5 medial/lateral from bregma). Fiber optic stimulating optrodes were placed in the midpoint of MGB (-6 dorsal/ventral from dura) and affixed to the skull using UV-curable dental acrylic (MidWest Dental). A 16 recording channel planar array was putatively placed in layers 3/4 of auditory cortex, with placement confirmed by short-latency high amplitude multiunit activity elicited from band pass noise (200 ms, center frequencies 1-36kHz in five steps per octave, 0.5 octave bandwidth) test stimuli. Recording electrodes were sealed onto the headcap. Animals were allowed to recover for 72 hours prior to the beginning of the recording regime. All recordings were performed in a 9'x9' electrically isolated anechoic chamber. During recording periods, animals received a intramuscular injection of medetomidine(0.2 mg/kg) for sedation. Optical stimuli were delivered from a 1907 nm diode laser (INSight open source optical stimulation system) coupled to the optrode with a 200 μm, 0.22 NA fiber (Thor Labs FG200LCC). Laser stimuli were controlled via a RX-7 stimulator (TDT) and consisted of train stimuli with pulse widths between 0.2-10 ms, interstimulus intervals between 0.2-100 ms and energy per pulse between 0-4 mJ. Applied laser energies were randomized to limit effects from neural adaptation with 30-60 repetitions per pulse width/interstimulus interval combinations. Signals from recording electrodes were amplified via a Medusa 32 channel preamplifier and discretized and sampled at 24.414 kHz with a RZ-2 biosignal processor and visualized using Open-Ex software (TDT). Action potentials were extracted from raw waveforms via real-time digital band-pass filtering with cutoff frequencies of 300-5000 Hz. Single units were extracted offline via superparamagnetic clustering in WaveClus (Quiroga et al., 2004). Studies were performed to assess the dose-response profiles of optically-based deep brain stimulation over the span of several months. As each electrode recorded diverse populations of neurons which are potentially subject to change due to electrode healing in, age of the device, and adaptation to the stimulus, a within subjects, repeated measures regression model was warranted. Bayesian hierarchical regressions can easily deal with complex models such as these.

2. An Example of Bayesian Analysis Reporting Guidelines

Bayesian Analysis Reporting Guidelines (BARG)(Kruschke, 2021) was initially proposed to promote transparent and reproducible Bayesian statistics reporting. While initially devised for social and psychological sciences, we adapted the BARG to suit neuroscientific data.

1. **Bayesian Model Descriptions and Sensitivity Analyses.** This report follows the guidelines for reporting of Bayesian Analysis (BARG) (Kruschke, 2021) consisting of:

- Necessary software and source code directory
- Goals of the analysis
- Model descriptions and decision criterion
- Prior and hyperprior descriptions
- Sensitivity analyses for varying prior distributions
- Posterior and MCMC diagnostics

1.1 Necessary software and source code directory

BARG: Step 2A, 6

This section describes the computational tools used for statistical analyses, including CPU and GPU use. For example:

Bayesian modeling was performed using Python 3.6.8 on a Razer Blade 15 Laptop with an Intel Core i7 processor (6 cores) and an Nvidia RTX2070 GPU. Models were implemented in PyMC version 4.11.5 (Salvatier et al., 2016), a probabilistic programming module in the Python environment. All source code is available at this paper's github repository (Link to software). All source data is available at this article's open science framework repository (Link to Data).

1.2 Goals of the Analyses

This section serves to establish goals of the analyses, brief description of the statistical models used and validation of Bayesian approaches.

BARG: Preamble

The goal of Bayesian regression analyses is to infer a linear relationship within inferior colliculus single unit firing rates resulting from changes in depth of sinusoidal amplitude modulated stimuli. While this is normally established using frequentist linear regression methods, Bayesian approaches allow for flexible and explicit model descriptions which provide rich and descriptive inference and quantification of uncertainty in measurement of single unit activity. Inference is completed using direct probability measures on posterior distributions as opposed to less intuitive and difficult to interpret p-values. Bayesian approaches are also data driven and account for previous knowledge to be encoded as prior distributions (see section 1.3).
The regression model utilized is:

$$\ln(FR) = \alpha + \beta * m + \epsilon$$

where FR is the mean evoked firing rate. Firing rate functions were calculated from recorded peristimulus time histograms. Parameter β quantifies the effect of modulation depth (m) on evoked firing rates respectively. The α parameter describes the model intercept and quantifies subthreshold spontaneous activity and the ϵ quantifies model error.

1.3 Prior Selection

Priors and rational for prior choice is described in this section

There is significant data detailing inferior colliculus responses to SAM stimuli from our lab and the auditory neuroscience community writ large(Citations redacted for double blind review). However,

the role of modulation depth on IC firing rates is understudied. As observations of single unit IC activity tends towards normal distributions, normal likelihood and prior distributions were chosen. Normal distributions also have the advantage of being moderately informative, refraining from undue influence on the posterior from the prior, allowing data to "speak for itself."

1.4 Posterior Decision Rules

This section details the decision rules used in inference (ROPE+HDI, Bayes factors, etc).

Inference was performed on posterior distributions with credible regions (analogous to frequentist confidence intervals) defined as a highest density interval (HDI) of 95% of parameter maximal *a posteriori* density (MAP) parameter estimates which represent the most probable value of the coefficient. MAP estimates are analogous to maximum likelihood estimation found in frequentist approaches. This allows for the quantification of parameter uncertainty as variance observed in posterior parameter distributions, with narrow HDIs representing more certain estimates. It is customary to define a region of practical equivalence (ROPE) if prior information dictates that incremental parameter changes are effectively the same. As we lack prior knowledge to inform the choice of a prior rope, we take an agnostic approach that any change seen is worth investigating and thus ROPEs are not presented. An effect was deemed significant if it's 95% HDI did not overlap with 0, in line with proposed decision rules typical of Bayesian inference(Kruschke, 2011, 2014).

1.5 Final Model

This section details the final model after prior and posterior sensitivity analyses. Helpful to include a descriptive figure of the inference model

Posterior predictive checks and sensitivity analysis were performed to titrate the best performing models as measured against observed data (Section 3). The final regression model is schematized in figure S1. Final models included deterministic nodes at outputs of prior nodes to prevent NUTS from becoming stuck in regions of the sampling space which are difficult to explore [1].

1.6 Model Sensitivity analyses

BARG: Step 3A,C

This section details the methods and results of any model sensitivity analyses. As an example, Model sensitivity analyses from hierarchical linear regression are given below.

To evaluate the dependance of hyperprior and prior parameters on Bayesian hierarchical linear regression, leave one out (LOO) cross validation(Gelman et al., 2014). A series of models were evaluated with model variances varied to test sensitivity of each model. Initial data analyses suggested that natural-log transformations of the dependent variable (firing rate) produced distributions which are better modeled as normal distributions. To this end, hierarchical models under test were as follows:

MODEL NAME	MODEL
REGRESSION	$FR = \alpha + \beta_1 * Energy + \beta_2 * ISI + \beta_3 * Energy * ISI + \epsilon$
SEMILOG REGRESSION	$ln(FR) = \alpha + \beta_1 * Energy + \beta_2 * ISI + \beta_3 * Energy * ISI + \epsilon$
NATURAL LOG REGRESSION	$ln(FR) = \alpha + \beta_1 * ln(Energy) + \beta_2 * ln(ISI) + \beta_3 * ln(Energy) * ln(ISI) + \epsilon$

Table S1: Regression models under test

For each model, the variance hyperprior was varied to assess the impact of prior parameters on posterior predictions. Prior classes were defined as: informative (variance ≤ 1), moderately informative (variance $= 5$), and weakly informative (variance ≥ 10). Primary metrics for model comparison were expected log pointwise predictive density (ELPD), defined as(Vehtari et al., 2017):

$$elpd = \sum_{i=1}^{k} \int dy_i p_t \bar{y_i} \log\left(p(\bar{y_i}|y)\right)$$

where p_t, y_i are unknown distributions representing the true data generating function for estimates of true posterior predictive function $(\bar{y}|y)$ from observed data y. Estimated p_t, y_i distributions are obtained via cross validation during LOO analysis. In general, higher values of ELPD are a result of higher out of sample predictive fit indicative of a better model. Weight values generated by LOO cross validation were also analyzed and predict the probability of each model given observed data. Finally, we observed the standard error of the ELPD estimate (SE), and the difference between the model with highest ELPD and every other model (dSE) with dSE of the top model set to 0.00 by definition. All LOO calculations were performed *post hoc* with the python package arviz, a plugin for PyMC.

MODEL	R	ELPD	WEIGHT	SE	DSE
ST LOG VAR 5	1	-5337.48	2.046623e-01	46.220458	0.00
ST LOG VAR 100	2	-5337.62	1.763745e-01	46.227682	0.420867
ST LOG VAR 0.5	3	-5337.76	1.552051e-01	49.173347	0.409773
ST LOG VAR 25	4	-5338.15	1.058393e-01	46.358847	0.492297
ST LOG VAR 10	5	-5338.18	9.996540e-02	46.141502	0.300197
ST LOG VAR 1	6	-5338.26	9.238175e-02	49.030330	0.331152
N LOG VAR 10	7	-5340.60	7.103823e-02	49.024680	3.308668
N LOG VAR 1	8	-5341.09	4.291607e-02	48.985814	3.293779
N LOG VAR 5	9	-5341.16	3.978488e-02	89.737613	3.296273
N LOG VAR 0.5	10	-5342.46	1.183257e-02	89.930943	3.300550
ST SEMILOG VAR 1	11	-5466.76	4.359604e-37	84.933022	15.845916
ST SEMILOG VAR 5	12	-5467.12	3.535240e-37	89.043113	15.856552
ST SEMILOG VAR 10	13	-5467.15	5.622764e-37	85.266895	15.895646

ST SEMILOG VAR 0.5	14	-5467.18	3.483572e-37	85.266895	15.866405
ST VAR 1	15	-15336.31	0.000000e+00	49.465018	79.406629
ST VAR 0.5	16	-15355.67	0.000000e+00	49.509352	80.415787
ST VAR 5	17	-15355.67	0.000000e+00	49.487001	80.415787
N VAR 10	18	-16119.11	0.000000e+00	49.510419	82.384329
N VAR 1	19	-16132.23	0.000000e+00	49.524316	83.549811
N VAR 0.5	20	-16154.55	0.000000e+00	49.514661	84.262219

Table S2: LOO model comparison results for the Bayesian hierarchical regression models. Var: Prior variance parameter, log: log predictor and predicted variable model. semilog: semilog predictor model. ST: Student T Likelihood models. N: Normal likelihood models

1.7 Posterior and MCMC Diagnostics

This section details model diagnostics surrounding the Bayesian Inference procedure.

BARG: Step 1E, 2A-D, 3A,C

1.7.1 Choice of MCMC method

This section details the choice of Markov-chain Monte Carlo model used. Many are available, but NUTS is the most common.

For sampling, the Hamiltonian-based MCMC method no U-turn sampling (NUTS)**(Hoffman and Gelman, 2011)** was used. NUTS presents a modification of general Hamiltonian Monte Carlo samplers and presents an efficient sampler for hierarchical and high-dimensional models at the cost of slower sampling times. Regression models ran 4 simultaneous chains with 2000 burn in samples and 4000 iterations with a 90% target inclusion probability.

MCMC Diagnostics

Energy transition plots were used to assess how well MCMC sampled the target posterior distribution of the best performing model as assessed by PSIS-LOO metrics which were compared between models**(Betancourt, 2017)**. NUTS sampling is based off dynamical systems modeling (Hamiltonian Monte Carlo) movement through the high entropy distributions towards a target distribution. MCMCs are modeled as dynamical systems with “position” and “momentum” associated with transition between states. This allows for a measurement of kinetic energy associated with the sampler. Efficiency in MCMC trajectory towards the target distribution can be assessed by comparing energy associated with the marginal energy distribution. The regression model displayed overlapping marginal energy and energy transition distributions (Fig S2) suggesting that sample to sample movement was nearly independent and indicative of efficient sampling of the target posterior distribution.

Furthermore, traces of sampled prior parameters in regression models suggest effective sampling of the posterior distribution (Fig S3). Furthermore, the Gelman-Rubin statistic, quantifying within

and between chain estimates and correlation was $\hat{r} < 1.05$, indicative of convergence of marginal posterior parameter values**(Gelman and Rubin, 1992)**.

1.7.2 Posterior Predictive Checks

This section describes any posterior predictive checks performed and a description of the posterior predictive decision rule.

A key advantage of Bayesian-based inference approaches is the ability to directly and explicitly compare model fits to observed data. During model development, posterior predictive checks were performed by sampling from the posterior distribution (16,000 draws). Kernel density estimates of posterior predictive draws from the posterior distribution were compared to kernel densities of observed data. Goodness of fit was quantified using the Bayesian p-value**(Gelman et al., 2021)**. Similar to the frequentist p-value, the Bayesian p-value is also a measure of discrepancy, quantifying the probability that posterior predictive-based draws are more extreme than observed data. The Bayesian p-value is defined as:

$$p_B = \iint dy^r d\theta \; I_{T(y^r,\theta)\geq T(y^r,\theta)} p(y^r|\theta)p(\theta|y)$$

where I is the indicator function, y^r is the posterior predictive distribution and y is the posterior distribution. Similar to the posterior distribution, posterior predictive distribution and Bayesian p-values were estimated using NUTS. The closer the Bayesian p-value is to 0.5, the better the data sampled from the posterior distribute around the observed data.

Posterior predictive fits and Bayesian p-values for the hierarchical linear and multinominal regression models suggest excellent posterior predictive fits with $\bar{p} = 0.51$ for the hierarchical linear regression model (Main article, Fig 4).

Prior and Posterior Trace plots

This section presents prior and posterior trace plots which are useful for diagnosing model fits.

BARG: Step 2B,C

Critical to the performance of HMC MCMC sampling is the convergence of sampling traces. Output trace plots display the chain of sampled values and the resulting kernel density estimates of sampled prior distributions. All sampled traces showed no divergences in sampling, suggesting that sampled traces were “well behaved”, providing accuate and effective sampling of the distribution. The Gelman-Rubin statistic, quantifying within and between chain estimates and correlation was $\hat{r} < 1.5$ for all sampled traces thus showing good MCMC convergence. For clarity, traces are available on open science framework, with traces for the posterior presented in Figure S3. Traces were checked for characteristic sampling behavior**(Hoffman and Gelman, 2011)** with no pathological traces found in models.

3. Supplementary Figures

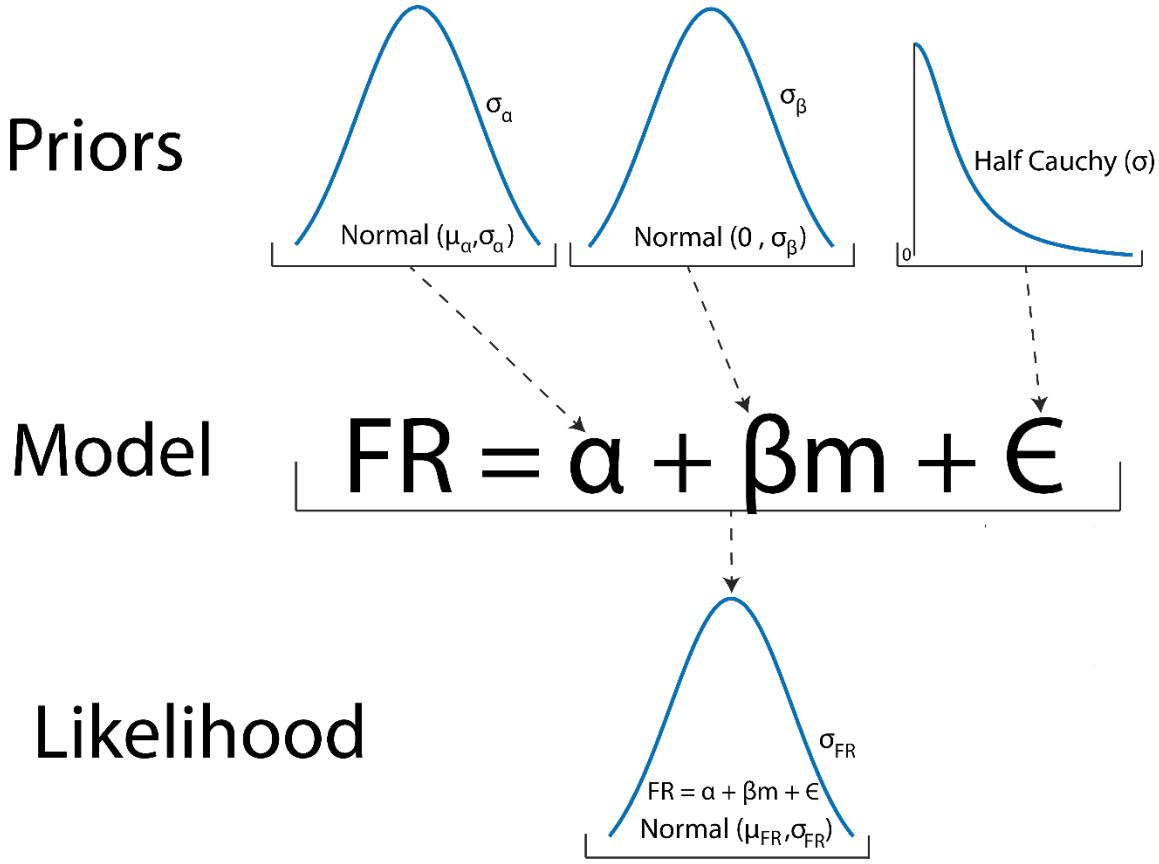


Figure S1: Schematic diagram of Bayesian Linear Regression

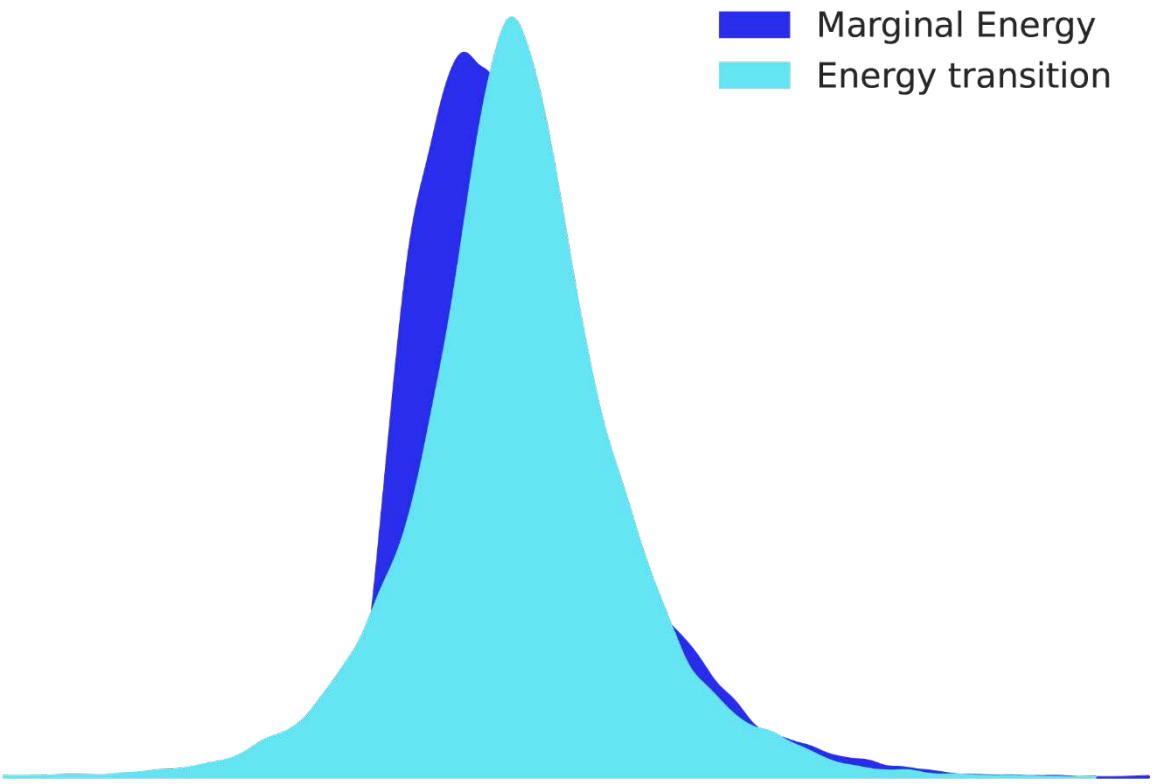


Figure S2: Energy transition diagram for Bayesian linear regression. Overlap of marginal energy with energy transition distributions suggests effective sampling of target distributions.

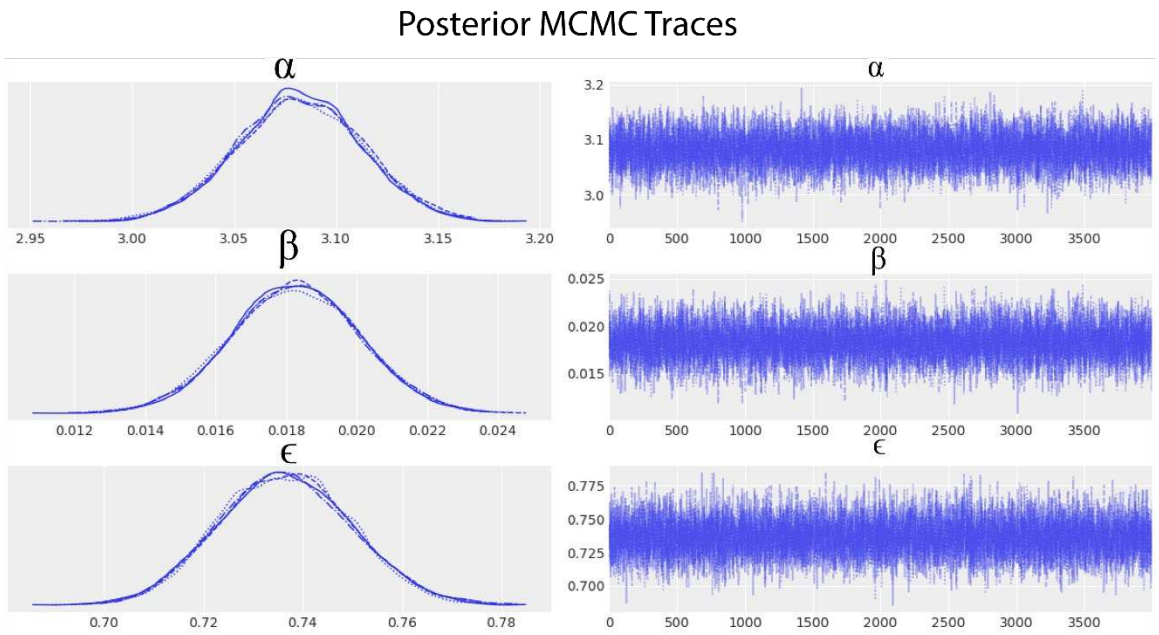


Figure S3: Bayesian regression posterior MCMC traces. MCMC chain distributions (left) built from HMC traces (right).

4. Supplementary References

Betancourt M (2017) A Conceptual Introduction to Hamiltonian Monte Carlo. Available at: https://arxiv.org/abs/1701.02434 [Accessed July 26, 2023].

Cant NB, Benson CG (2006) Organization of the inferior colliculus of the gerbil (Meriones unguiculatus): Differences in distribution of projections from the cochlear nuclei and the superior olivary complex. J Comp Neurol 495:511–528.

Caspary DM, Palombi PS, Hughes LF (2002) GABAergic inputs shape responses to amplitude modulated stimuli in the inferior colliculus. Hearing Research 168:163–173.

Cayce JM, Friedman RM, Chen G, Jansen ED, Mahadevan-Jansen A, Roe AW (2014) Infrared neural stimulation of primary visual cortex in non-human primates. NeuroImage 84:181–190.

Cayce JM, Friedman RM, Jansen ED, Mahavaden-Jansen A, Roe AW (2011) Pulsed infrared light alters neural activity in rat somatosensory cortex in vivo. NeuroImage 57:155–166.

C.F R, Parthasarathy A, Venkataraman Y, Fisher ZL, Gardner SM, Bartlett EL (2012) A computational model of inferior colliculus responses to amplitude modulated sounds in young and aged rats. Frontiers in Neural Circuits 6.

Coventry BS, Lawlor GL, Bagnati CB, Krogmeier C, Bartlett EL (2023) Spatially specific, closed-loop infrared thalamocortical deep brain stimulation. bioRxiv Available at: http://biorxiv.org/lookup/doi/10.1101/2023.10.04.560859 [Accessed October 12, 2023].

Coventry BS, Parthasarathy A, Sommer AL, Bartlett EL (2017) Hierarchical winner-take-all particle swarm optimization social network for neural model fitting. Journal of Computational Neuroscience 42:72–85.

Coventry BS, Sick JT, Talavage TM, Stantz KM, Bartlett EL (2020) Short-wave Infrared Neural Stimulation Drives Graded Sciatic Nerve Activation Across A Continuum of Wavelengths. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE.

Frisina DR, Frisina RD (1997) Speech recognition in noise and presbycusis: relations to possible neural mechanisms. Hearing Research 106:95–104.

Gelman A, Carlin J, Stern H, Dunson D, Vehtari A, Rubin D (2021) Bayesian Data Analysis, 3rd ed. Boca Raton: Chapman and Hall/CRC.

Gelman A, Hwang J, Vehtari A (2014) Understanding predictive information criteria for Bayesian models. Stat Comput 24:997–1016.

Gelman A, Rubin DB (1992) Inference from Iterative Simulation Using Multiple Sequences. Statistical Science 7:457–472.

Grimsley CA, Sanchez JT, Sivaramakrishnan S (2013) Midbrain local circuits shape sound intensity codes. Front Neural Circuits 7 Available at: http://journal.frontiersin.org/article/10.3389/fncir.2013.00174/abstract [Accessed August 27, 2023].

Herrmann B, Parthasarathy A, Bartlett EL (2017) Ageing affects dual encoding of periodicity and envelope shape in rat inferior colliculus neurons Foxe J, ed. Eur J Neurosci 45:299–311.

Hoffman MD, Gelman A (2011) The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. Available at: https://arxiv.org/abs/1111.4246 [Accessed July 26, 2023].

Izzo AD, Suh E, Pathria J, Walsh JT, Whitlon DS, Richter C-P (2007) Selectivity of neural stimulation in the auditory system: a comparison of optic and electric stimuli. Journal of biomedical optics 12:021008.

Kelly JB, Caspary DM (2005) Pharmacology of the inferior colliculus. In: The inferior colliculus., 1st ed. New York, NY: Springer.

Kruschke JK (2011) Bayesian Assessment of Null Values Via Parameter Estimation and Model Comparison. Perspectives on Psychological Science 6:299–312.

Kruschke JK (2014) Doing Bayesian Data Analysis: A tutorial with R, JAGS, and stan, 2nd ed. Academic Press. Available at: http://www.indiana.edu/~kruschke/DoingBayesianDataAnalysis/.

Kruschke JK (2021) Bayesian Analysis Reporting Guidelines. Nat Hum Behav 5:1282–1291.

Loftus WC, Bishop DC, Oliver DL (2010) Differential Patterns of Inputs Create Functional Zones in Central Nucleus of Inferior Colliculus. J Neurosci 30:13396–13408.

Palombi PS, Backoff PM, Caspary DM (2001) Responses of young and aged rat inferior colliculus neurons to sinusoidally amplitude modulated stimuli. Hearing Research 153:174–180.

Parthasarathy A, Bartlett E (2012) Two-channel recording of auditory-evoked potentials to detect age-related deficits in temporal processing. Hearing Research 289:52–62.

Parthasarathy A, Cunningham PA, Bartlett EL (2010) Age-Related Differences in Auditory Processing as Assessed by Amplitude-Modulation Following Responses in Quiet and in Noise. Front Ag Neurosci 2 Available at: http://journal.frontiersin.org/article/10.3389/fnagi.2010.00152/abstract [Accessed March 13, 2023].

Quiroga RQ, Nadasdy Z, Ben-Shaul Y (2004) Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. Neural computation 16:1661–1687.

Rabang CF, Parthasarathy A, Venkataraman Y, Fisher ZL, Gardner SM, Bartlett EL (2012) A computational model of inferior colliculus responses to amplitude modulated sounds in young and aged rats. Frontiers in neural circuits 6:77.

Salvatier J, Wiecki TV, Fonnesbeck C (2016) Probabilistic programming in Python using PyMC3. PeerJ Computer Science 2:e55.

Simon H, Frisina RD, Walton JP (2004) Age reduces response latency of mouse inferior colliculus neurons to AM sounds. The Journal of the Acoustical Society of America 116:469–477.

Vehtari A, Gelman A, Gabry J (2017) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat Comput 27:1413–1432.

Wells J, Kao C, Mariappan K, Albea J, Jansen ED, Konrad P, Mahadevan-Jansen A (2005) Optical stimulation of neural tissue in vivo. Optics Letters 30:504.