Optimizing Data for Modeling Neuronal Responses

Peter Zeidman[1*†], Samira M. Kazan[1†], Nick Todd[2], Nikolaus Weiskopf[1,3], Karl J. Friston[1] and Martina F. Callaghan[1]

[1] Wellcome Centre for Human Neuroimaging, UCL Institute of Neurology, University College London, London, United Kingdom, [2] Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States, [3] Department of Neurophysics, Max Planck Institute for Human Cognition and Brain Sciences, Leipzig, Germany

In this technical note, we address an unresolved challenge in neuroimaging statistics: how to determine which of several datasets is the best for inferring neuronal responses. Comparisons of this kind are important for experimenters when choosing an imaging protocol—and for developers of new acquisition methods. However, the hypothesis that one dataset is better than another cannot be tested using conventional statistics (based on likelihood ratios), as these require the data to be the same under each hypothesis. Here we present Bayesian data comparison (BDC), a principled framework for evaluating the quality of functional imaging data, in terms of the precision with which neuronal connectivity parameters can be estimated and competing models can be disambiguated. For each of several candidate datasets, neuronal responses are modeled using Bayesian (probabilistic) forward models, such as General Linear Models (GLMs) or Dynamic Casual Models (DCMs). Next, the parameters from subject-specific models are summarized at the group level using a Bayesian GLM. A series of measures, which we introduce here, are then used to evaluate each dataset in terms of the precision of (group-level) parameter estimates and the ability of the data to distinguish similar models. To exemplify the approach, we compared four datasets that were acquired in a study evaluating multiband fMRI acquisition schemes, and we used simulations to establish the face validity of the comparison measures. To enable people to reproduce these analyses using their own data and experimental paradigms, we provide general-purpose Matlab code via the SPM software.

Keywords: dynamic causal modeling, DCM, fMRI, PEB, multiband

INTRODUCTION

Hypothesis testing involves comparing the evidence for different models or hypotheses, given some measured data. The key quantity of interest is the likelihood ratio—the probability of observing the data under one model relative to another—written $p\left(y|m_1\right) / p(y|m_2)$ for models m_1 and m_2 and dataset y. Likelihood ratios are ubiquitous in statistics, forming the basis of the F-test and the Bayes factor in classical and Bayesian statistics, respectively. They are the most powerful test for any given level of significance by the Neyman-Pearson lemma (Neyman and Pearson, 1933). However, the likelihood ratio test assumes that there is only one dataset y–and so cannot be used to compare different datasets. Therefore, an unresolved problem, especially pertinent to neuroimaging, is how to test the hypothesis that one dataset is better than another for making inferences.

Neuronal activity and circuitry cannot generally be observed directly, but rather are inferred from measured timeseries. In the case of fMRI, the data are mediated by neuro-vascular coupling, the BOLD response and noise. To estimate the underlying neuronal responses, models are specified which formalize the experimenter's understanding of how the data were generated. Hypotheses are then tested by making inferences about model parameters, or by comparing the evidence under different models. For example, an *F*-test can be performed on the General Linear Model (GLM) using classical statistics, or Bayesian Model Comparison can be used to select between probabilistic models. From the experimenter's perspective, the best dataset provides the most precise estimates of neuronal responses (enabling efficient inference about parameters) and provides the greatest discrimination among competing models (enabling efficient inference about models).

Here, we introduce Bayesian data comparison (BDC)—a set of information measures for evaluating a dataset's ability to support inferences about both parameters and models. While they are generic and can be used with any sort of probabilistic models, we illustrate their application using Dynamic Causal Modeling (DCM) for fMRI (Friston et al., 2003) because it offers several advantages. Compared to the GLM, DCM provides a richer characterization of neuroimaging data, through the use of biophysical models, based on differential equations that separate neuronal and hemodynamic parameters. This means one can evaluate which dataset is best for estimating neuronal parameters specifically. These models also include connections among regions, making DCM the usual approach for inferring effective (directed) connectivity from fMRI data.

By using the same methodology to select among datasets as experimenters use to select between connectivity models, feature selection and hypothesis testing can be brought into alignment for connectivity studies. Moreover, a particularly useful feature of DCM for comparing datasets is that it employs Bayesian statistics. The posterior probability over neuronal parameters forms a multivariate normal distribution, providing expected values for each parameter, as well as their covariance. The precision (inverse covariance) quantifies the confidence we place in the parameter estimates, given the data. After establishing that an experimental effect exists, for example by conducting an initial GLM analysis or by reference to previous studies, the precision of the parameters can be used to compare datasets. DCM also enables experimenters to distinguish among models, in terms of which model maximizes the log model evidence $\ln p(y|m)$. We cannot use this quantity to compare different datasets, but we can ask which of several datasets provides the most efficient discrimination among models. For these reasons, we used Bayesian methods as the basis for comparing datasets, both to provide estimates of neuronal responses and to distinguish among competing models.

This paper presents a methodology and associated software for evaluating which of several imaging acquisition protocols or data features affords the most sensitive inferences about neural architectures. We illustrate the framework by assessing the quality of fMRI time series acquired from 10 participants, who were each scanned four times with a different multiband

acceleration factor. Multiband is an approach for rapid acquisition of fMRI data, in which multiple slices are acquired simultaneously and subsequently unfolded using coil sensitivity information (Larkman et al., 2001; Xu et al., 2013). The rapid sampling enables sources of physiological noise to be separated from sources of experimental variance more efficiently, however, a penalty for this increased temporal resolution is a reduction of the signal-to-noise ratio (SNR) of each image. A detailed analysis of these data is published separately (Todd et al., 2017). We do not seek to draw any novel conclusions about multiband acceleration from this specific dataset, but rather we use it to illustrate a generic approach for comparing datasets. Additionally, we conducted simulations to establish the face validity of the outcome measures, using estimated effect sizes and SNR levels from the empirical multiband data.

The methodology we introduce here offers several novel contributions. First, it provides a sensitive comparison of data by evaluating their ability to reduce uncertainty about neuronal parameters, and to discriminate among competing models. Second, our procedure identifies the best dataset for hypothesis testing at the group level, reflecting the objectives of most cognitive neuroscience studies. Unlike a classical GLM analysis—where only the maximum likelihood estimates from each subject are taken to the group level—the (parametric empirical) Bayesian methods used here take into account the uncertainty of the parameters (the full covariance matrix), when modeling at the group level. Additionally, this methodology provides the necessary tools to evaluate which imaging protocol is optimal for effective connectivity analyses, although we anticipate many questions about data quality will not necessarily relate to connectivity. We provide a single Matlab function for conducting all the analyses described in this paper, which is available in the SPM (http://www.fil.ion.ucl.ac.uk/spm/) software package (spm_dcm_bdc.m). This function can be used to evaluate any type of imaging protocol, in terms of the precision with which model parameters are estimated and the complexity of the generative models that can be disambiguated.

METHODS

We begin by briefly reprising the theory behind DCM and introducing the set of outcome measures used to evaluate data quality. We then illustrate the analysis pipeline in the context of an exemplar fMRI dataset and evaluate the measures using simulations.

Dynamic Causal Modeling

DCM is a framework for evaluating generative models of time series data. At the most generic level, neural activity in region i of the brain at time t may be modeled by a lumped or neural mass quantity z_t^i. Generally, the experimenter is interested in the neuronal activity of a set of interconnected brain regions, the activity of which can be written as a vector z. The evolution of neural activity over time can then be written as:

$$\dot{z} = f\left(z, u, \theta^{(n)}\right) \qquad (1)$$

Where $\dot{z}$ is the derivative of the neural activity with respect to time, u are the time series of experimental or exogenous inputs and $\theta^{(n)}$ are the neural parameters controlling connectivity within and between regions. Neural activity cannot generally be directly observed. Therefore, the neural model is combined with an observation model g, with haemodynamic/observation parameters $\theta^{(h)}$, specifying how neural activity is transformed into a timeseries, y:

$$y = g\left(z, \theta^{(h)}\right) + \epsilon^{(1)} \qquad (2)$$

Where $\epsilon^{(1)}$ is zero-mean (I.I.D.) additive Gaussian noise, with log-precision specified by hyperparameters $\lambda = \lambda_1 \ldots \lambda_r$ for each region r. The I.I.D. assumption is licensed by automatic pre-whitening of the timeseries in SPM, prior to the DCM analysis. In practice, it is necessary to augment the vector of neuronal activity with hemodynamic parameters that enter the observation model above. The specific approximations of functions f and g depend on the imaging modality being used, and the procedures described in this paper are not contingent on any specific models. However, to briefly reprise the basic model for fMRI—which we use here for illustrative purposes—f is a Taylor approximation to any nonlinear neuronal dynamics:

$$\dot{z} = \left(A + \sum_j u_j(t)B^j\right) z + Cu(t) \qquad (3)$$

There are three sets of neural parameters $\theta^{(n)} = (A, B, C)$ and j experimental inputs. Matrix A represents the strength of connections within (i.e., intrinsic) and between (i.e., extrinsic) regions—their effective connectivity. Matrix B represents the effect of time-varying experimental inputs on each connection (these are referred to as modulatory or condition-specific effects) and the corresponding vector $u_j(t)$ is a timeseries encoding the timing of experimental condition j at time t. Matrix C specifies the influence of each experimental input on each region, which effectively drives the dynamics of the system, given $u(t)$ which is the vector of all experimental inputs at time t.

The hemodynamics (the observation model g above) are modeled with an extended Balloon model (Buxton et al., 2004; Stephan et al., 2007), which comprises a series of differential equations describing the process of neurovascular coupling by which activity ultimately manifests as a BOLD signal change. The majority of the parameters of this hemodynamic model are based on previous empirical measurements; however, three parameters are estimated on a region-specific basis: the transit time τ, the rate of signal decay κ, and the ratio of intra- to extra-vascular signal $\epsilon^{(h)}$.

The observation parameters $\theta^{(h)}$ are concatenated with the neural parameters $\theta^{(n)}$ and the hyperparameters λ and a prior multivariate normal density is defined (see **Table 1**). The parameters are then estimated using a standard variational Bayes scheme called variational Laplace (Friston et al., 2003; Friston, 2011). This provides a posterior probability density for the parameters, as well as an approximation of the log model evidence (i.e., the negative variational free energy), which scores

TABLE 1 | Priors on DCM parameters.

Parameter(s)	Prior expectation	Prior variance
A	0	1/64
B	0	1
C	0	1
τ	0	1/256
κ	0	1/256
$\epsilon^{(h)}$	0	1/256
λ_i	6	1/128

the quality of the model in terms of its accuracy minus its complexity.

Group Analyses With PEB

Having fitted a model of neuronal responses to each subject individually (a first level analysis), the parameters can be summarized at the group level (a second level analysis). We used a Bayesian GLM, implemented using the Parametric Empirical Bayes (PEB) framework for DCM (Friston et al., 2016). With N subjects and M connectivity parameters for each subject's DCM, the group-level GLM has the form:

$$\theta = X\beta + \epsilon^{(2)} \qquad (4)$$

The dimensions of this GLM are illustrated in **Figure 1**. Vector $\theta \in \mathbb{R}^{NM \times 1}$ are the neuronal parameters from all the subjects' DCMs, consisting of all parameters from subject 1, then all parameters from subject 2, etc. The design matrix $X \in \mathbb{R}^{NM \times M}$ was specified as:

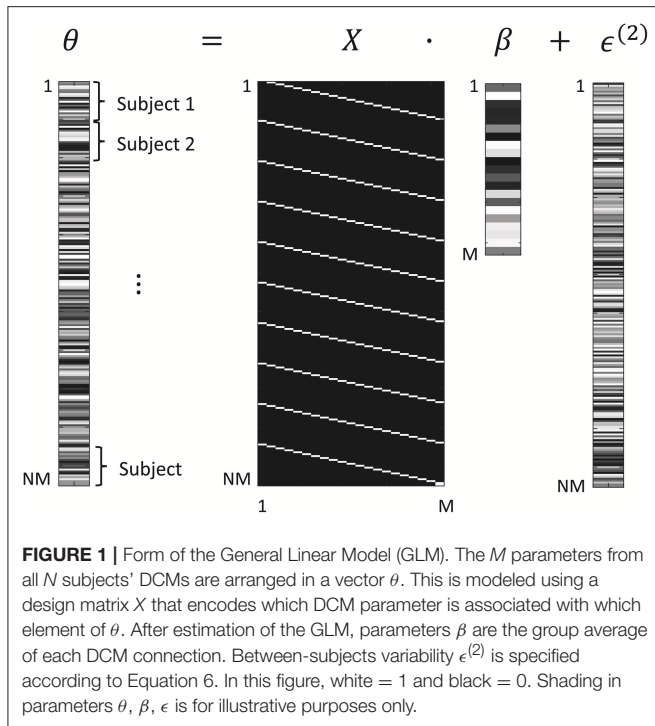$$X = \mathbf{1}_N \otimes I_M \qquad (5)$$

Where $\mathbf{1}_N$ is a column vector of 1 s of dimension N and I_M is the identity matrix of dimension M. The Kronecker product $\otimes$ replicates the identity matrix vertically for each subject. The use of a Kronecker product at the between subject level reflects the fact that between subject effects can be expressed at each and every connection. In this instance, we are just interested in the group mean and therefore there is only one between-subject effect. The resulting matrix X has one column (also called a covariate or regressor) for each connectivity parameter (**Figure 1**). The regressors are scaled by parameters $\beta \in \mathbb{R}^{M \times 1}$, which are estimated from the data and represent the group average strength of each connection. Finally, the errors $\epsilon^{(2)} \in \mathbb{R}^{NM \times 1}$ are modeled as zero-mean additive noise:

$$\epsilon^{(2)} = N(0, \ \Pi^{-1}) \qquad (6)$$

Where precision matrix $\Pi \in \mathbb{R}^{NM \times NM}$ is estimated from the data. This captures the between-subject variability in the connection strengths, parameterised using a single parameter γ:

$$\Pi = I_N \otimes \left(Q_0 + e^{-\gamma} Q_1\right) \qquad (7)$$

This is a multi-component covariance model. $Q_0 \in \mathbb{R}^{M \times M}$ is the lower bound on precision and ensures it is a positive number.

FIGURE 1 | Form of the General Linear Model (GLM). The M parameters from all N subjects' DCMs are arranged in a vector θ. This is modeled using a design matrix X that encodes which DCM parameter is associated with which element of θ. After estimation of the GLM, parameters β are the group average of each DCM connection. Between-subjects variability $\epsilon^{(2)}$ is specified according to Equation 6. In this figure, white = 1 and black = 0. Shading in parameters θ, β, ϵ is for illustrative purposes only.

Matrix $Q_1 \in \mathbb{R}^{M \times M}$ is the prior precision. When the parameter γ is zero, the precision is equal to $Q_0 + Q_1$. More negative values of γ equate to higher precision than the prior and vice versa. The Kronecker product $\otimes$ replicates the precision matrix for each subject, giving rise to the matrix Π of dimension $NM \times NM$ where the leading diagonal is the precision of each DCM parameter θ.

Prior (multivariate normal) densities are specified on the group-level parameters representing the average connection strengths across subjects β and the parameter controlling between-subject variability γ:

$$\beta = N\left(\mu_\beta, \Sigma_\beta\right)$$
$$\gamma = N(\mu_\gamma, \Sigma_\gamma) \qquad (8)$$

Where $\mu_\beta \in \mathbb{R}^{MN \times 1}$, $\Sigma_\beta \in \mathbb{R}^{MN \times MN}$, $\mu_\gamma \in \mathbb{R}^1$, $\sigma_\gamma \in \mathbb{R}^1$. The prior on β is set to be identical to the prior on the corresponding DCM parameter at the first (individual subject) level. In the example analysis presented here, parameters governing condition-specific neural effects (B) with prior $p\left(B_i\right) = N(0, 1)$ for each parameter i were taken to the group level. The prior on γ, the log precision of the between-subject random effects, was set to $p\left(\gamma\right) = N(0, 1/16)$. This expresses the prior belief that the between-subject variance is expected to be much smaller (16 times smaller) than the within-subject effect size.

To summarize, within and between-subject parameters are estimated using a hierarchical scheme, referred to as Parametric Empirical Bayes (PEB). Model estimation provides the approximate log model evidence (free energy) of the group-level Bayesian GLM—a statistic that enables different models to be compared (see **Appendix 1**). We take advantage of the free energy below to compare models of group-level data. For

full details on the priors, model specification, and estimation procedure in the PEB scheme, see Friston et al. (2016). Readers familiar with Random Effects Bayesian Model Selection (RFX BMS) (Stephan et al., 2009) will note the distinction with the PEB approach used here. Whereas, RFX BMS considers random effects over models, the PEB approach considers random effects at the group level to be expressed at the level of parameters; namely, parametric random effects. This means that uncertainty about parameters at the subject level is conveyed to the group level; licensing the measures described in the next section.

Outcome Measures
Parameter Certainty
To measure the information gain (or reduction in uncertainty) about parameters due to the data, we take advantage of the Laplace approximation used in the DCM framework, which means that the posterior and prior densities over parameters are Gaussian. In this case, the confidence of the parameter estimates can be quantified using the negative entropy of the posterior multivariate density over interesting parameters:

$$S_\theta = -0.5 ln \left|2\pi e \, \Sigma_\beta\right| \qquad (9)$$

Equation 9 uses the definition of the negative entropy for the multivariate normal distribution, applied to the neuronal parameter covariance matrix Σ_β. This has units of nats (i.e., natural units) and provides a summary of the precision associated with group level parameters—such as group means—having properly accounted for measurement noise and random effects at the between-subject level.

Datasets can be compared by treating the entropies as log Bayes factors (detailed in **Appendix 1**: Bayesian data comparison). In brief, this follows because the log Bayes factor can always be decomposed into two terms—a difference in accuracy minus a difference in complexity. The complexity is the KL-divergence between the posteriors Σ_β and the priors Σ_0, and it scores the reduction in uncertainty afforded by the data, in units of nats. Under flat or uninformative priors, the KL-divergence reduces to the negative entropy in Equation 9. A difference in entropy between 1.1 nats and 3 nats is referred to as "positive evidence" that one dataset is better than another, and corresponds to a difference in information gain between $e^{1.1} \approx 3$ fold and $e^3 \approx 20$ fold (Kass and Raftery, 1995). Similarly, a difference in entropy between 3 and 5 nats is referred to as "strong evidence," and differences in entropy beyond this are referred to as "very strong evidence." These same labels apply for the measures below.

Information Gain (Parameters)
The data quality afforded by a particular acquisition scheme can be scored in terms of the *relative entropy* or KL-divergence between posterior and prior distributions over parameters. This measure of salience is also known as Bayesian surprise, epistemic value or information gain and can be interpreted as the quantitative reduction of uncertainty after observing the data. In other words, it reflects the complexity of the model (the number of independent parameters) that can be supported by

the data. This takes into account both the posterior expectation and precision of the parameters relative to the priors, whereas the measure in part (a) considered only the posterior precision (relative to uninformative priors).

The KL-divergence for the multivariate normal distribution, between the posterior N_1 and the prior N_0, with mean μ_1 and μ_0 and covariance Σ_1 and Σ_0, respectively, is given by:

$$D_{KL}(N_1 \parallel N_0) = \frac{1}{2}\left(tr\left(\Sigma_0^{-1}\Sigma_1\right) + (\mu_0 - \mu_1)^T \Sigma_0^{-1} \right.$$
$$\left. (\mu_0 - \mu_1) - k + \ln\frac{\det\Sigma_0}{\det\Sigma_1} \right) \qquad (10)$$

Where $k = rank(\Sigma_0)$. This statistic increases when the posterior mean has moved away from the prior mean or when the precision of the parameters has increased relative to the precision of the priors. Note that this same quantity also plays an important role in the definition of the free energy approximation to log model evidence, which can be decomposed into accuracy minus complexity, the latter being the KL-divergence between posteriors and priors.

The measures described so far are based on posterior estimates of model parameters. We now turn to the equivalent measure of posterior beliefs about the models *per se*.

Information Gain (Models)

The quality of the data from a given acquisition scheme can also be assessed in terms of their ability to reduce uncertainty about models. This involves specifying a set of equally plausible, difficult to disambiguate models that vary in the presence or absence of experimental effects or parameters, and evaluating which dataset best enables these models to be distinguished.

Bayesian model comparison starts with defining a prior probability distribution over the models P_0. Here, we assume that all models are equally likely, therefore $P_0 = 1/p$ for each of p models. This prior is combined with the model evidence, to provide a posterior distribution over the models, P. To quantify the extent to which the competing models have been distinguished from one another, we measure the information gain from the prior P_0 to the posterior P. This is given by the KL-divergence used above for the parameters. After describing how we specified these models, we provide an example of this KL-divergence in practice.

Typically with Bayesian inference (e.g., DCM), the experimenter embodies each hypothesis as a model and compares the evidence for different models. In the example dataset presented here, we did not have strong hypotheses about the experimental effects, and so we adopted the following procedure. We first estimated a "full" group-level Bayesian GLM with all relevant free parameters from the subjects' DCMs. Next, we identified a set of reduced GLMs that only differed slightly in log evidence (i.e., they were difficult to discriminate). To do this we eliminated one connection or parameter (by fixing its prior variance to zero) and retained the model if the change in log evidence was >-3. This corresponds to a log odds ratio of approximately one in $e^3 \approx 20$, meaning that the model was retained if it was no more than 20 times less probable than the

full model. We repeated this procedure by eliminating another parameter (with replacement), ultimately obtaining the final model space. This procedure was performed rapidly by using Bayesian Model Reduction (BMR), which analytically computes the log evidence of reduced models from a full model (Friston et al., 2016).

Having identified a set of plausible but difficult to disambiguate models (GLMs) for a given dataset, we then calculated the posterior probability of each model. Under flat priors, this is simply the softmax function of the log model evidence, as approximated by the free energy (see **Appendix 1**). We then computed the KL-divergence between the posterior and prior model probabilities, which is defined for discrete probability distributions as:

$$D_{KL}(P \parallel P_0) = \sum_{i=1\ldots k}\left(P_i \ln P_i\right) + \ln k \qquad (11)$$
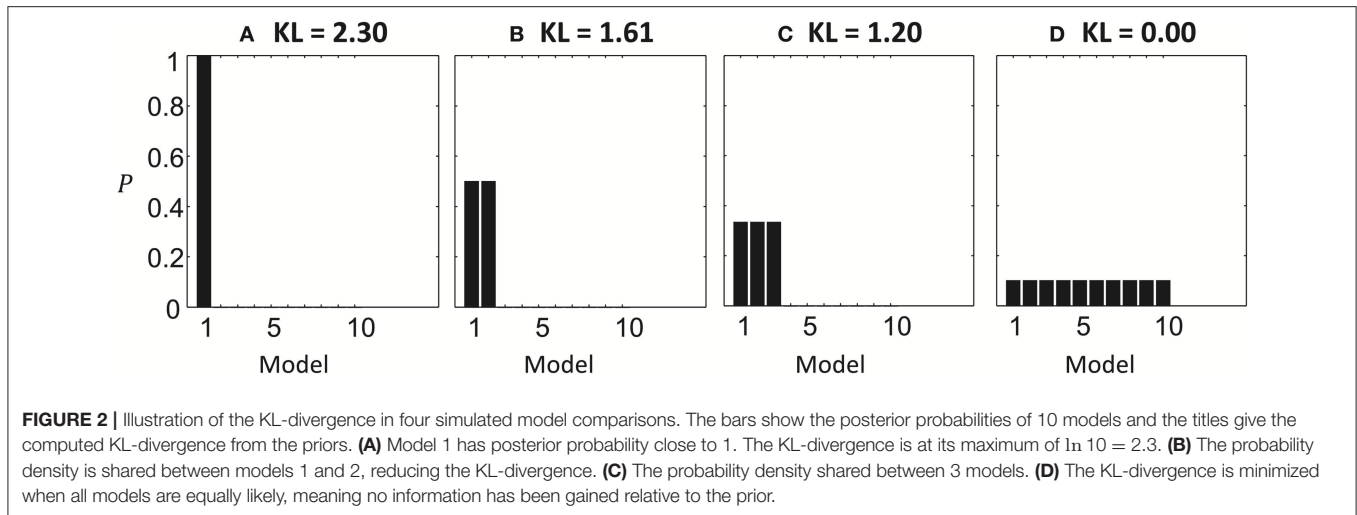
The behavior of the KL-divergence is illustrated in **Figure 2**, when comparing $k = 10$ simulated models. When one model has a posterior probability approaching one, and all other models have probability approaching zero, the KL-divergence is maximized and has the value $D_{KL} = \ln k = 2.30$ (**Figure 2A**). As the probability density is shared between more models, so the KL-divergence is reduced (**Figures 2B,C**). It reaches its minimum value of zero when all models are equally likely, meaning that no information has been gained by performing the model comparison (**Figure 2D**).

Summary of Measures and Analysis Pipeline

A key contribution of the measures introduced in this paper is the characterization of information gain in terms of both the parameters, and the models that entail those parameters. Together they provide a principled means by which to characterize the optimality of a given scheme for acquiring data. These are intended for use where the presence or absence of experimental effects in the data is already known—for example, based on previous studies and/or the results or an initial analysis collapsed across datasets (e.g., a mass-univariate GLM analysis).

We now suggest a pipeline for applying these measures to neuroimaging data. Step 1 provides estimates of neuronal parameters from each dataset. Steps 2 and 3 use the estimated parameters from all datasets to automatically identify a suitable model architecture (and could be skipped if the experimenter has strong priors as to what the model architecture should be). Steps 4 and 5 provide estimates of the group level parameters for each dataset and compare them using the measures described above. For convenience, steps 2–5 of this procedure can be run with a single Matlab function implemented in SPM (spm_dcm_bdc.m):

1) Model each subject's data using a Bayesian model (e.g., DCM). The objective is to obtain posterior estimates of neuronal parameters from each dataset. These estimates take the form of a multivariate probability density for each subject and dataset.
2) Identify the optimal group-level model structure. This step identifies a parsimonious model architecture, which can be

FIGURE 2 | Illustration of the KL-divergence in four simulated model comparisons. The bars show the posterior probabilities of 10 models and the titles give the computed KL-divergence from the priors. **(A)** Model 1 has posterior probability close to 1. The KL-divergence is at its maximum of $\ln 10 = 2.3$. **(B)** The probability density is shared between models 1 and 2, reducing the KL-divergence. **(C)** The probability density shared between 3 models. **(D)** The KL-divergence is minimized when all models are equally likely, meaning no information has been gained relative to the prior.

used to model all datasets (in the absence of strong hypotheses about the presence and absence of experimental effects). A Bayesian GLM is specified and fitted to the neuronal parameters from all subjects and datasets. To avoid bias, the GLM is not informed that the data derive from multiple datasets. The estimated GLM parameters represent the average connectivity across all datasets. This GLM is pruned to remove any redundant parameters (e.g., relating to the responses of specific brain regions) that do not contribute to the model evidence, using Bayesian Model Reduction (Friston et al., 2016). This gives the optimal reduced model structure at the group level, agnostic to the dataset.

3) Re-estimate each subject's individual DCM having switched off any parameters that were pruned in step 2. This step equips each subject with a parsimonious model to provide estimates of neuronal responses. This is known as "empirical Bayes," as the priors for the individual subjects have been updated based on the group level data. Again, this is performed analytically using Bayesian Model Reduction.

4) Fit separate Bayesian GLMs to the neuronal parameters of each dataset. This summarizes the estimated neuronal responses for each dataset, taking into account both the expected values and uncertainty of each subject's parameters.

5) Apply the measures outlined above to compare the quality or efficiency of inferences from each dataset's Bayesian GLM—in terms of parameters or models.

Collectively, the outcome measures that result from this procedure constitute an assessment of the goodness of different datasets in terms of inferences about connection parameters and models. Next, we provide an illustrative example using empirical data from an experiment comparing different fMRI multiband acceleration factors.

Multiband Example

For this example, we use fMRI data from a previously published study that evaluated the effect of multiband acceleration on fMRI data (Todd et al., 2017). We will briefly reprise the objectives

of that study. For a given effect size of interest, the statistical power of an fMRI experiment can be improved by acquiring a greater number of sample points (i.e., increasing the efficiency of the design) or by reducing measurement noise. This has the potential to enable more precise parameter estimates and provide support for more complex models of how the data were generated. Acquiring data with high temporal resolution both increases the number of samples per unit time and allows physiologically-driven fluctuations in the time series to be more fully sampled and subsequently removed or separated from the task-related BOLD signal (Todd et al., 2017). One approach to achieving rapid acquisitions is the use of the multiband or simultaneous multi-slice acquisition technique (Setsompop et al., 2012; Xu et al., 2013); in which multiple slices are acquired simultaneously and subsequently unfolded using coil sensitivity information (Setsompop et al., 2012; Cauley et al., 2014). The penalty for the increased temporal resolution is a reduction of the signal-to-noise ratio (SNR) of each image. This is caused by increased g-factor penalties, dependent on the coil sensitivity profiles, and reduced steady-state magnetization arising from the shorter repetition time (TR) and concomitant reduction in excitation flip angle. In addition, a shorter TR can be expected to increase the degree of temporal auto-correlation in the time series (Corbin et al., 2018). This raises the question of which MB acceleration factor offers the best trade-off between acquisition speed and image quality.

We do not seek to resolve the question of which multiband factor is optimal in general. Furthermore, there are many potential mechanisms by which multiband acquisitions could improve or limit data quality, including better sampling of physiological noise, and increasing the number of samples in the data. Rather than trying to address these questions here, we instead use these data to exemplify comparing datasets. In these data, physiologically-driven fluctuations—that are better sampled with higher multiband acceleration factor due to the higher Nyquist sampling frequency—were removed from the data by filtering. Subsequently, the data were down-sampled so as to have equivalent numbers of samples across multiband factors,

as described in Todd et al. (2017). The framework presented here could be used to test the datasets under many different acquisition and pre-processing procedures.

Data Acquisition

Ten healthy volunteers were scanned with local ethics committee approval on a Siemens 3T Tim Trio scanner. For each volunteer, fMRI task data with 3 mm isotropic resolution were acquired four times with a MB factor of either 1, 2, 4, or 8 using the gradient echo EPI sequence from the Center for Magnetic Resonance Research (R012 for VB17A, https://www.cmrr.umn. edu/multiband/). The TR was 2,800, 1,400, 700, and 350 ms for MB factor 1, 2, 4, and 8, respectively, resulting in 155, 310, 620, and 1,240 volumes, respectively, leading to a seven and a half min acquisition time per run. The data were acquired with the blipped-CAIPI scheme (Setsompop et al., 2012), without in-plane acceleration, and the leak-block kernel option for image reconstruction was enabled (Cauley et al., 2014).

fMRI Task

The fMRI task consisted of passive viewing of images, with image stimuli presented in 8 s blocks. Each block consisted of four images of naturalistic scenes or four images of single isolated objects, displayed successively for 2 s each. There were two experimental factors: stimulus type (images of scenes or objects) and novelty (2, 3, or 4 novel images per block, with the remainder repeated). This paradigm has previously been shown to induce activation in a well-established network of brain regions that respond to perceiving, imagining or recalling scenes (Spreng et al., 2009; Zeidman et al., 2015).

Preprocessing

All data were processed in SPM (Ashburner and Friston, 2005), version 12. This comprised the usual image realignment, co-registration to a T1-weighted anatomical image and spatial normalization to the Montreal Neurological Institute (MNI) template space using the unified segmentation algorithm, and smoothing with a $6 \times 6 \times 6$ mm full width at half maximum (FWHM) Gaussian kernel.

As described in Todd et al. (2016), all data were filtered using a 6th-order low pass Butterworth filter with a frequency cut-off of 0.18 Hz (corresponding to the Nyquist frequency of the MB = 1 data). This removed all frequency components between the cut-off frequency and the corresponding Nyquist frequency of the particular MB factor. In order to ensure equal numbers of samples per data set—regardless of MB factor used—the time series were decimated by down-sampling all datasets to the TR of the MB1 (TR = 2.8 s) data.

After initial processing and filtering, all data sets were modeled with a general linear model (GLM), with a high pass filter (cut-off period = 128 s) and regressors for motion, and physiological effects. In addition to these confounding effects, the stimulation blocks were modeled with boxcar functions convolved with the canonical hemodynamic response function. Temporal autocorrelations were accounted for with an autoregressive AR(1) model plus white noise. This was deemed sufficient given that after filtering and decimation each time series

had an effective TR of 2.8 s (Corbin et al., 2018). The contrast of scenes > objects was computed and used to select brain regions for the DCM analysis.

DCM Specification

We selected seven brain regions (**Figure 3A**) from the SPM analysis which are part of a "core network" that responds more to viewing images of scenes rather than images of isolated objects (Zeidman et al., 2015). These regions were: Early Occipital cortex (OCC), left Lateral Occipital cortex (lLOC), left Parahippocampal cortex (lPHC), left Retrosplenial cortex (lRSC), right Lateral Occipital cortex (rLOC), right Parahippocampal cortex (rPHC), and right Retrosplenial cortex (rRSC).
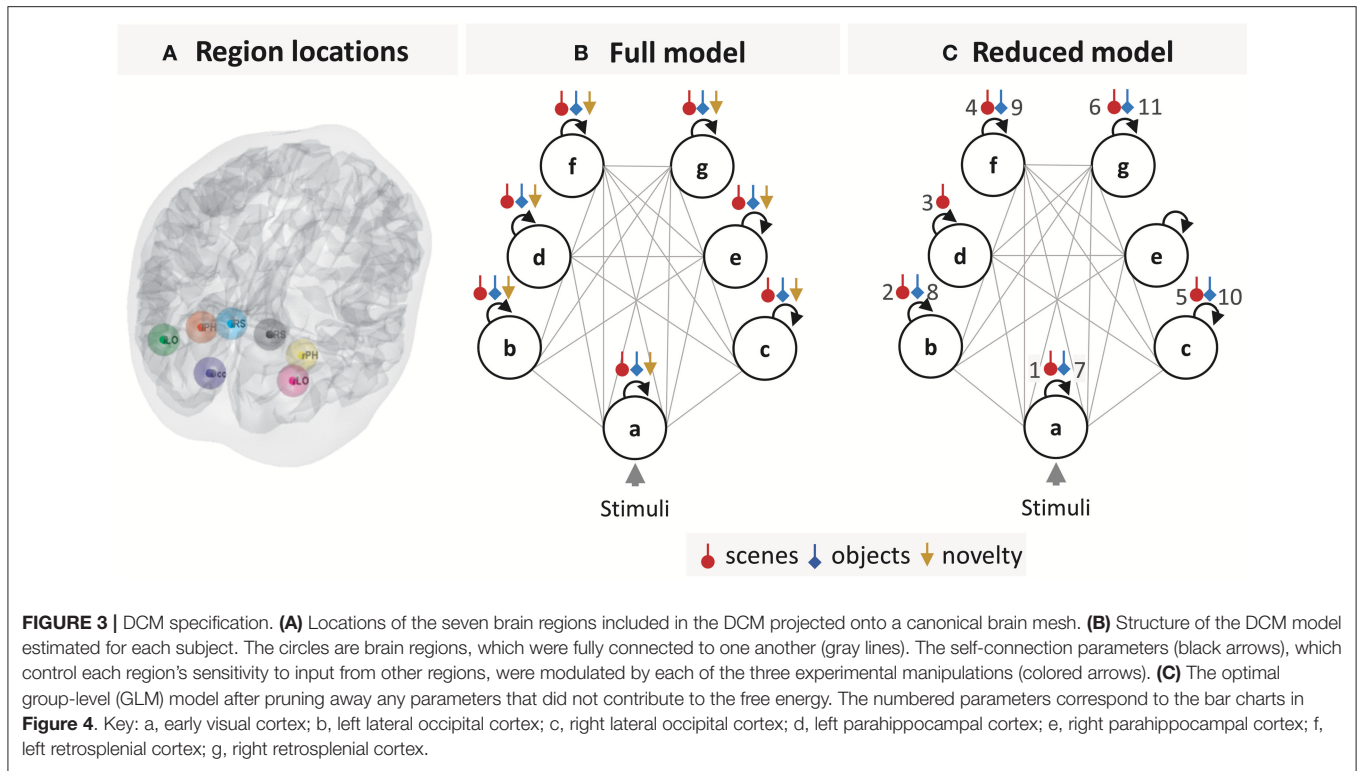
We extracted timeseries from each of these regions as follows. The group-level activation peak (collapsed across multiband factor to prevent bias) was identified from the contrast of scenes > objects (thresholded at $p < 0.05$ FWE-corrected) using a one-way ANOVA as implemented in SPM. Subsequently, a spherical region of interest (ROI) with 8 mm FWHM was centered on the peaks at the individual level that were closest to the group-level peaks. This size of the ROI sphere was arbitrary and provided a suitable trade-off between including a reasonable number of voxels and not crossing into neighboring anatomical areas. Voxels within each sphere surviving at least $p < 0.001$ uncorrected at the single-subject level were summarized by their first principal eigenvariate, which formed the data feature used for subsequent DCM analysis.

The neuronal model for each subject's DCM was specified as a fully connected network (**Figure 3B**). Dynamics within the network were driven by all trials, modeled as boxcar functions, driving occipital cortex (the circle labeled a in **Figure 3B**). The experimental manipulations (scene stimuli, object stimuli, and stimulus novelty) were modeled as modulating each region's self-inhibition (colored arrows in **Figure 3B**). These parameters control the sensitivity of each region to inputs from the rest of the network, in each experimental condition. Neurobiologically, they serve as simple proxies for context-specific changes in the excitatory-inhibitory balance of pyramidal cells and inhibitory interneurons within each region (Bastos et al., 2012). These parameters, which form the B-matrix in the DCM neuronal model (Equation 3), are usually the most interesting from the experimenters' perspective—and we focused on these parameters for our analyses.

Simulation

We also conducted simulations to confirm the face validity of the Bayesian data comparison approach presented here. Specifically, we wanted to ensure that subject-level differences in the precision of parameters across datasets were properly reflected in the group-level PEB parameters and the ensuing outcome measures. Note that these simulations were not intended to recapitulate the properties or behavior of multiband fMRI data. Rather, our intention was to conduct a simple and reasonably generic assessment of the outcome measures under varying levels of SNR.

We generated simulated neuroimaging data for 100 virtual experiments, each consisting of 100 datasets with differing SNR levels, and applied the Bayesian data comparison procedure to

FIGURE 3 | DCM specification. **(A)** Locations of the seven brain regions included in the DCM projected onto a canonical brain mesh. **(B)** Structure of the DCM model estimated for each subject. The circles are brain regions, which were fully connected to one another (gray lines). The self-connection parameters (black arrows), which control each region's sensitivity to input from other regions, were modulated by each of the three experimental manipulations (colored arrows). **(C)** The optimal group-level (GLM) model after pruning away any parameters that did not contribute to the free energy. The numbered parameters correspond to the bar charts in **Figure 4**. Key: a, early visual cortex; b, left lateral occipital cortex; c, right lateral occipital cortex; d, left parahippocampal cortex; e, right parahippocampal cortex; f, left retrosplenial cortex; g, right retrosplenial cortex.

each. These data were generated and modeled using General Linear Models (GLMs), which enabled precise control over the parameters and their covariance, as well as facilitating the inversion of large numbers of models in reasonable time (minutes on a desktop PC). For each simulation $i = 1 \ldots 100$, subject $j = 1 \ldots 16$, and level of observation noise $k = 1 \ldots 100$ we specified a GLM:

$$y^{(ijk)} = X \cdot \beta^{(ij)} + \epsilon^{(ijk)} \tag{12}$$

There were three regressors in the design matrix X matching the empirical multiband fMRI experiment reported in the previous section (corresponding to the scenes, objects, and novelty experimental conditions). For each virtual subject, the three corresponding parameters in vector $\beta^{(ij)}$ were sampled from a multivariate normal distribution:

$$\beta^{(ij)} \sim N\left(\mu_\beta, \Sigma_\beta\right)$$
$$\mu_\beta = [0.89, 0.89, 0.45]^T$$
$$\Sigma_\beta = I_3 \cdot \sigma_B^2 \tag{13}$$

Where I_3 is the identity matrix of dimension three. Vector μ_β were the "ground truth" parameters, chosen based on the empirical analysis (the first two parameters were set to 0.89, the mean of scene and object effects on occipital cortex, and the third parameter was set to half this value, to provide a smaller but still detectable effect). The between-subject variance σ_B^2 was set to 0.18, computed from the empirical PEB analyses and averaged over multiband datasets. Finally, we added I.I.D observation

noise $\epsilon^{(ijk)}$ with a different level of variance in each dataset, chosen to achieve SNRs ranging between 0.003 (most noisy) and 0.5 (least noisy) in steps of 0.005. Here, SNR was defined as the ratio of the variance of the modeled signal to the variance of the noise (residuals); the median SNR from the empirical data was 0.5 across subjects and datasets.

Having generated the simulated data, we then fitted GLMs using a variational Bayesian scheme (Friston et al., 2007) with priors on the parameters set to:

$$p\left(\beta^{(ij)}\right) = N(0, I_3 \cdot \sigma_W^2)$$

Where the within-subject prior variance was set to $\sigma_W^2 = 1$, to match the DCM parameters of interest in the empirical analysis above. To compute the information gain over models (the third outcome measure), we defined a model space with seven permutations of the parameters switched on or off (i.e., we specified a model for every possible permutation of the parameters, excluding the model with all three parameters switched off).

RESULTS

We followed the analysis pipeline described above (see Summary of measures and analysis) to compare data acquired under four levels of multiband acceleration. The group-level results below and the associated figures were generated using the Matlab function spm_dcm_bdc.m.

MB Leakage/Aliasing Investigation

While not the focus of this paper, we conducted an analysis to ensure that our DCM results were not influenced by a potential image acquisition confound. As with any accelerated imaging technique, the multiband acquisition scheme is vulnerable to potential aliased signals being unfolded incorrectly. This is important since activation aliasing between DCM regions of interest could potentially lead to artificial correlations between regions (Todd et al., 2016). This analysis, detailed in the **Supplementary Materials**, confirmed that the aliased location of any given region of interest used in the DCM analysis did not overlap with any other region of interest.

Identifying the Optimal Group-Level Model

We obtained estimates of each subject's neuronal responses by fitting a DCM to their data, separately for each multiband factor (the structure of this DCM is illustrated in **Figure 3B**). Then we estimated a single group-level Bayesian GLM of the neuronal parameters and pruned any parameters that did not contribute to the model evidence (by fixing them at their prior expectations with zero mean and zero variance). This gave the optimal group-level model, the parameters of which are illustrated in **Figure 3C**. The main conditions of interest were scene and object stimuli. Redundant modulatory effects of object stimuli were pruned from bilateral PHC, while the effect of scene stimuli was pruned from right PHC only. Redundant effects of stimulus novelty were pruned from all regions. This was not surprising, as the experimental design was not optimized for this contrast—and the regions of interest were not selected on the basis of tests for novelty effects.

Modeling Each Dataset

Having identified a single group-level model architecture across all datasets (**Figure 3C**), we next updated each subject's DCMs to use this reduced architecture (by setting their priors to match the group level posteriors and obtaining updated estimates of the DCM parameters). We then estimated a group-level GLM for each dataset. The parameters of these four group-level GLMs are illustrated in **Figures 4A–D**. The numbered parameters, which correspond to those in **Figure 3C**, describe the change of sensitivity of each region to their inputs. More positive values signify more inhibition due to the task and more negative values signify dis-inhibition (excitation) due to the task. The results were largely consistent across multiband factors, with scene and object stimuli exciting most regions relative to baseline. Interestingly, modulation of early visual cortex by scenes and objects (parameters 1 and 7) were the largest effect sizes, so contributed the most to explaining the network-wide difference in scene and object stimuli.

Figure 4E shows the precision (inverse variance) of each parameter from **Figures 4A–D**. Each group of bars relates to a neuronal parameter, and each of the four bars relate to the four datasets (i.e., each of the four multiband factors). It is immediately apparent that all parameters (with the exception of parameter 7) achieved the highest precision with dataset MB4 (i.e., multiband acceleration factor 4). However,

examining each parameter separately in this way is limited, because we cannot see the covariance between the parameters. The covariance is important in determining the confidence with which we can make inferences about parameters or models. Next, we apply our novel series of measures to these data, which provide a simple summary of the qualities of each dataset while taking into account the full parameter covariance.

Comparing Datasets

In agreement with the analysis above, the dataset with multiband acceleration factor 4 (MB4) gave neuronal parameter estimates with the greatest precision or certainty (**Figure 5A**), followed by MB1 and MB2, and MB8 had the least precision. The difference between the best (MB4) and worst (MB8) performing datasets was 1.64 nats, equivalent to 84% probability of a difference (calculated by applying the softmax function to the plotted values). This may be classed as "positive evidence" for MB4 over MB8 (Kass and Raftery, 1995), however the evidence was not strong enough to confidently claim that MB4 was better than MB1 or MB2.

The information gain over parameters (see Outcome measures) is the extent to which the parameters were informed by the data. It reflects the number of independent parameters in the model (its complexity) that the data can accommodate. The best dataset was MB4 (**Figure 5B**), followed closely by MB2 and then MB1 and MB8. The difference between the best (MB4) and worst (MB8) datasets was 1.82 nats, or a 86.06% probability of a difference (positive evidence). There was also positive evidence that MB4 was better than MB1 (1.31 nats = 78.75%), but MB4 could not be distinguished from MB2 (0.33 nats = 58.18%). Thus, not only were the parameters most precise in dataset MB4 (**Figure 5A**), but they also gained the most information from the data, relative to the information available in the priors. This effect was most pronounced in comparison to MB8, and to a lesser extent, in comparison to MB1.

Next we computed the information gain over models (see Outcome measures, part c), which quantified the ability of the datasets to discriminate between similar models. Whereas, the previous measures were relative to the worst performing dataset, this measure was relative to the prior that all models were equally likely (zero nats). The automated procedure described in the Methods section identified eight similar candidate models that differed only in their priors (i.e., which parameters were switched on or off). Because there were eight models, the maximum possible information gain over models was $\ln 8 = 2.08$ nats. We found that MB4 afforded the best discrimination between models (**Figure 5C**), with an information gain of 1.34 nats relative to the prior that all models were equally likely (positive evidence, 79% probability). The other three datasets provided poorer discriminability: MB8 with 0.67 nats, MB2 with 0.62 nats, and MB1 with 0.34 nats.

To summarize, all three of the measures favored the dataset with multiband acceleration 4 (MB4). However, the magnitudes of these differences were generally not substantial, with "positive evidence" rather than "strong evidence" under all measures. MB4 consistently fared better than MB8—with
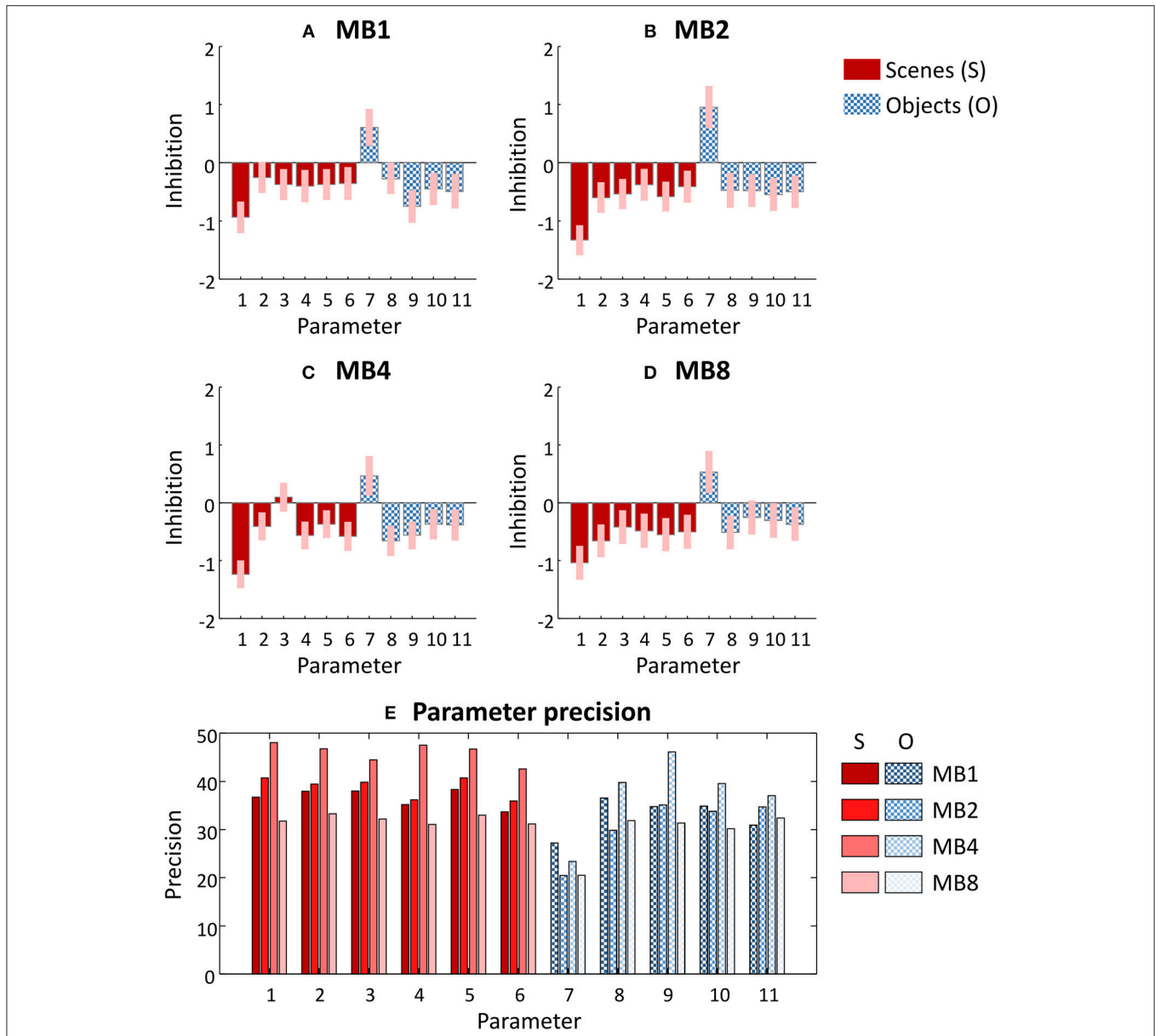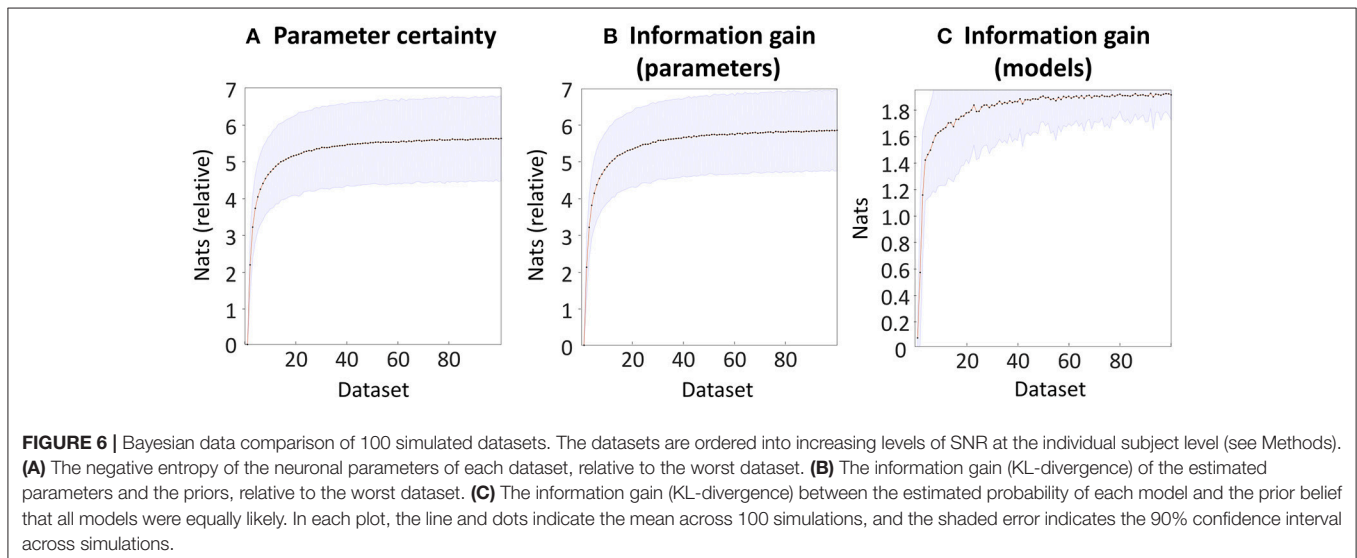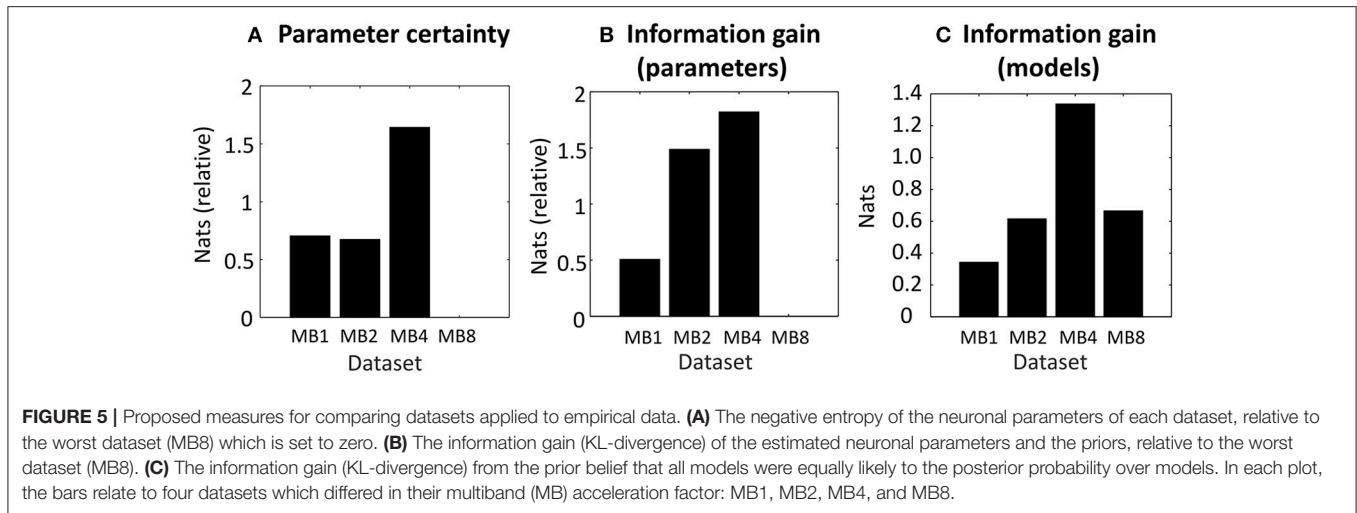
FIGURE 4 | Parameters of the group-level General Linear Model fitted to each dataset. **(A–D)** Posterior estimates of each parameter from each dataset. The bars correspond to the parameters labeled in **Figure 3C**, and for clarity these are divided into regional effects of scene stimuli (solid red) and of object stimuli (chequered blue). These parameters scale the prior self-connection of each region, and have no units. Positive values indicate greater inhibition due to the experimental condition and negative values indicate disinhibition (excitation). Pink error bars indicate 90% confidence intervals. MBx, multiband acceleration factor x. **(E)** The precision of each parameter—i.e., the inverse of the variance which was used to form the pink error bars in **(A–D)**. Parameters 1–6 relate to the effects of scene stimuli (S) and parameters 7–11 relate to the effects of object stimuli (O). Each group of four bars denote the four datasets in the order MB = 1, 2, 4, and 8 from left to right.

positive evidence that it provided more confident parameter estimates (**Figure 5A**) and greater information gain (**Figure 5B**). There was also positive evidence that MB4 offered greater information gain than MB1 (**Figure 5B**). Finally, MB4 supported greater information gain over models than any other dataset (**Figure 5C**). Given these results, if we were to conduct this same experiment with a larger sample, we would select multiband acceleration factor MB4 as our preferred acquisition protocol.

Simulation Results

To validate the software implementing the outcome measures, we simulated 100 experiments, each of which compared 100 datasets with varying levels of SNR. As expected, increasing SNR was accompanied by an increase in the certainty (precision) of the group-level parameters (**Figure 6A**). This showed an initial rise and then plateaued, reaching 5.62 nats (very strong evidence) for the dataset with the highest SNR (dataset 100) compared to the dataset with the lowest SNR (**Dataset 1**). The information

FIGURE 5 | Proposed measures for comparing datasets applied to empirical data. **(A)** The negative entropy of the neuronal parameters of each dataset, relative to the worst dataset (MB8) which is set to zero. **(B)** The information gain (KL-divergence) of the estimated neuronal parameters and the priors, relative to the worst dataset (MB8). **(C)** The information gain (KL-divergence) from the prior belief that all models were equally likely to the posterior probability over models. In each plot, the bars relate to four datasets which differed in their multiband (MB) acceleration factor: MB1, MB2, MB4, and MB8.

FIGURE 6 | Bayesian data comparison of 100 simulated datasets. The datasets are ordered into increasing levels of SNR at the individual subject level (see Methods). **(A)** The negative entropy of the neuronal parameters of each dataset, relative to the worst dataset. **(B)** The information gain (KL-divergence) of the estimated parameters and the priors, relative to the worst dataset. **(C)** The information gain (KL-divergence) between the estimated probability of each model and the prior belief that all models were equally likely. In each plot, the line and dots indicate the mean across 100 simulations, and the shaded error indicates the 90% confidence interval across simulations.

gain over parameters, which quantified the KL-divergence from the priors to the posteriors relative to the first dataset, was very similar (**Figure 6B**). Finally, we compared the datasets in terms of their ability to distinguish seven models, one of which matched the model that generated the data. This measure could range from zero nats (the prior that all seven models were equally likely) to $\ln(7) = 1.95$ nats (a single model having a posterior probability of 100%). We found that the information gain increased with increasing SNR, to a ceiling of 1.91 nats (**Figure 6C**). Inspecting the results revealed that in the datasets with the lowest SNR, the probability mass was shared between model one (the "full" model which generated the data) and model three, in which the parameter quantifying the small novelty effect was disabled (fixed at zero). As SNR increased, model one was correctly selected with increasing confidence.

To summarize, these simulations provided a basic check that the within-subject parameters were conveyed to the group level and were captured by the outcome measures proposed in

this work. All three outcome measures showed a monotonic increase with SNR, which consisted of a large initial increase followed by diminishing returns as SNR further increased. Accompanying this paper, we provide a Matlab script for performing these simulations, which can be used for testing the impact of differing effect sizes or levels of between-subject variability, under any choice of (Bayesian) forward model.

DISCUSSION

This paper introduced Bayesian data comparison, a systematic approach for identifying the optimal data features for inferring neuronal architectures. We proposed a set of measures based on established Bayesian models of neuroimaging timeseries, in order to compare datasets for two types of analyses—inference about parameters and inference about models. We exemplified

these measures using data from a published experiment, which investigated the performance of multiband fMRI in a cohort of 10 healthy volunteers. This principled scheme, which can be applied by experimenters using a software function implemented in SPM (spm_dcm_bdc.m), can easily be applied to any experimental paradigm, any group of subjects (healthy or patient cohorts), and any acquisition scheme.

Comparing models based on their evidence is the most efficient procedure for testing hypotheses (Neyman and Pearson, 1933) and is employed in both classical and Bayesian statistics. The model evidence is the probability of the data y given the model m i.e., $p(y|m)$. Model comparison involves taking the ratio of the evidences for competing models. However, this ratio (known as the Bayes factor) assumes that each model has been fitted to the same data y. This means that when deciding which data to use (e.g., arbitrating between different multiband acceleration factors) it is not possible to fit models to each dataset and compare them based on their evidence. To address this, the measures we introduced here can be used in place of the model evidence to decide which of several datasets provides the best estimates of model parameters and best distinguishes among competing models.

The first step in our proposed analysis scheme is to quantify neuronal responses for each data acquisition. This necessarily requires the use of a model to partition the variance into neuronal, hemodynamic, and noise components. Any form of model and estimation scheme can be used, the only requirement being that it is probabilistic or Bayesian. In other words, it should furnish a probability density over the parameters. Here, we modeled each subject's neuronal activity using DCM for fMRI, in which the parameters form a multivariate normal distribution defined by the expected value of each parameter and their covariance (i.e., properly accounting for conditional dependencies among the parameters). Given that the main application of DCM is for investigating effective (causal) connectivity, the method offered in this paper is especially pertinent for asking which acquisition scheme will offer the most efficient estimates of connectivity parameters. Alternatively, the same analysis approach could be applied to the observation parameters rather than the neuronal parameters, to ask which dataset provides the best estimates of regional neurovascular coupling and the BOLD response. More broadly, any probabilistic model could have been used to obtain parameters relating to brain activity, one alternative being a Bayesian GLM at the single subject level (Penny et al., 2003).

Hypotheses in cognitive neuroscience are usually about effects that are conserved at the group level. However, the benefits of advanced acquisition schemes seen at the single subject level may not be preserved at the group level due to inter-subject variability (Kirilina et al., 2016). We were therefore motivated to develop a protocol to ask which acquisition scheme offers the best inferences at the group level, while appropriately modeling inter-subject variability. To facilitate this, the second step in our analysis procedure is to take the estimated neuronal parameters from every subject and summarize them using a group level model. Here we use a Bayesian GLM, estimated using a hierarchical (PEB) scheme. This provides the average (expected

value) of the connectivity parameters across subjects, as well as the uncertainty (covariance) of these parameters. It additionally provides the free energy approximation of the log model evidence of the GLM, which quantifies the relative goodness of the GLM in terms of accuracy minus complexity. The key advantage of this Bayesian approach, unlike the summary statistic approach used with the classical GLM in neuroimaging, is that it takes the full distribution over the parameters (both the expected values and covariance) from the single subject level to the group level. This is important in assessing the quality of datasets, where the subject-level uncertainty over the parameters is key to assessing their utility for parameter-level inference. Together, by fitting DCMs at the single subject level and then a Bayesian GLM at the group level, one can appropriately quantify neuronal responses at the group level.

Having obtained parameters and log model evidences of each dataset's group-level GLM, the final stage of our analysis procedure is to apply a set of measures to each dataset. These measures are derived from information theory and quantify the ability of the data to support two complementary types of inference. Firstly, inference about parameters involves testing hypotheses about the parameters of a model; e.g., assessing whether a particular neuronal response is positive or negative. A good dataset will support precise estimates of the parameters (where precision is the inverse variance) and will support the parameters being distinguished from one another (i.e., minimize conditional dependencies). We evaluated these features in each dataset by using the negative entropy of the parameters and the information gain. These provide a straightforward summary of the utility of each dataset for inference over parameters. A complementary form of inference involves embodying each hypothesis as a model and comparing these models based on their log evidence $\ln p(y|m)$. This forms the basis of most DCM studies, where models differ in terms of which connections are switched on and off, or which connections receive experimental inputs (specified by setting the priors of each model). We assessed each dataset in terms of its ability to distinguish similar, plausible and difficult-to-discriminate models from one another. This involved an automated procedure for defining a set of similar models, and the use of an information theoretic quantity—the information gain—to determine how well the models could be distinguished from one another in each dataset. This measure can be interpreted as the amount we have learnt about the models by performing the model comparison, relative to our prior belief that all models were equally likely.

To exemplify the approach, we compared four fMRI datasets that differed in their multiband acceleration factor. The higher the acceleration factor, the faster the image acquisition. This affords the potential to better separate physiological noise from task-related variance—or to increase functional sensitivity by providing more samples per unit time. However, this comes with various costs, including reduced SNR and increased temporal auto-correlations. The datasets used were acquired in the context of an established fMRI paradigm, which elicited known effects in pre-defined regions of interest. The conclusion of the original study (Todd et al., 2017), which examined the datasets under the same pre-processing procedures used here, was that a multiband

acceleration factor between 4 (conservative) and 8 (aggressive) should be used. In the present analysis, the dataset acquired with multiband acceleration factor 4 (MB4) afforded the most precise estimates of neuronal parameters, and the largest information gain in terms of both parameters and models (**Figure 5**), although the differences between MB4 and MB2 were small. Our analysis of residual leakage artifact (**Supplementary Material**) showed this result was not confounded by aliasing, a potential issue with multiband acquisitions (Todd et al., 2016). Given that these data were decimated so as to have equivalent numbers of samples, regardless of MB factor, our results suggest that the improved sampling of physiological effects provided by multiband acceleration counterbalanced the loss of SNR. Speculatively, MB4 may have been optimal in terms of benefiting from physiological filtering (a sufficiently high Nyquist frequency to resolve breathing effects), despite any reduction in SNR. MB2 may have performed slightly less well because it suffered from the penalty of reduced SNR, without sufficient benefit from the filtering of physiological effects. Any advantage of MB8 in terms of physiological filtering may have been outweighed by the greater reduction in SNR.

One should exercise caution in generalizing this multiband result, which may not hold for different paradigms or image setups (e.g., RF coil types, field strength, resolution, etc.) or if using variable numbers of data points. Though congruent with a previous study (Todd et al., 2017), without these further investigations, the conclusions presented here should not be generalized. Going forward, the effect of each of these manipulations could be framed as a hypothesis, and tested using the procedures described here. One interesting future direction would be to investigate the contribution of the two pre-processing steps: filtering and decimation. Our data were filtered to provide improved sampling of physiological noise and were subsequently decimated in order to maintain a fixed number of data points for all multiband factors under investigation. This ensured a fair comparison of the datasets with equivalent handling of temporal auto-correlations. The protocol described here could be used to evaluate different filtering and decimation options. One might anticipate that the increased effective number of degrees of freedom within the data would be tempered by increased temporal auto-correlations arising from more rapid sampling.

A further specific consideration for the application of multiband fMRI to connectivity analyses is whether differences in slice timing across different acquisition speeds could influence estimates of effective connectivity in DCM. An approach for resolving this is slice timing correction—adjusting the model to account for acquisition time of each slice. DCM has an inbuilt slice timing model to facilitate this (Kiebel et al., 2007). Whether this is helpful for all application domains is uncertain. Following spatial realignment, coregistration, and normalization, the precise acquisition time of each slice is lost, so the modeled acquisition times can deviate from the actual acquisition times. On the other hand, if the modeled acquisition times are reasonably accurate, there may be some benefit—particularly for fast event-related designs. This uncertainty can be resolved using Bayesian model comparison—DCMs can be specified with

different slice timing options and their evidence compared. In the example dataset presented here, we had a slow block design which is unlikely to benefit significantly from slice timing correction, so for simplicity we used the default setting in DCM—aligning the onsets to the middle of each volume, thereby minimizing the error on average.

The procedure introduced here involves evaluating each dataset in terms of its ability to provide estimates of group-level experimental effects. An alternative approach would be to compare datasets at the individual subject level—for instance, by comparing the variance of each model's residuals, parameterized in DCM on a region-by-region basis (hyperparameters λ which control the log precision of the noise $\epsilon^{(1)}$, see Equation 2). However, this would only characterize the fit of each model as a whole, and would not evaluate the quality of inferences about neural parameters specifically, which are typically the quantities of interest in neuroimaging studies. Furthermore, neuroimaging studies typically evaluate hypotheses about groups of subjects rather than individuals, and thus assessing the quality of inferences is ideally performed using group-level models or parameters. For the example analysis presented here, we therefore chose to compare datasets in terms of the specific parameters of interest for the particular experiment (DCM B-matrix), summarized by the group level PEB model.

An important consideration—when introducing any novel modeling approach or procedure—is validation. Here, for our example analysis using empirical data, we used two extant models from the neuroimaging community—DCM for fMRI and the Bayesian GLM implemented in the PEB framework. The face validity of DCM for fMRI has been tested using simulations (Friston et al., 2003; Chumbley et al., 2007), its construct validity has been tested using extant modeling approaches (Penny et al., 2004; Lee et al., 2006), and its predictive validity has been tested using intracranial recordings (David et al., 2008a,b Reyt et al., 2010). The PEB model and the associated Bayesian Model Reduction scheme is more recent and so far has been validated in terms of its face validity using simulated data (Friston et al., 2016), its reproducibility with empirical data (Litvak et al., 2015) and its predictive validity in the context of individual differences in electrophysiological responses (Pinotsis et al., 2016).

The next validation consideration regards the novel contribution of this paper—the application of a set of outcome measures to the probabilistic models discussed above. These measures are simply descriptions or summary statistics of the Bayesian or probabilistic models to which they are applied. The measures themselves depend on two statistics from information theory—the negative entropy and the KL-divergence, which do not require validation in and of themselves, just as the t-statistic does not need validation when used to compare experimental conditions using the GLM. Rather, the implementation of the statistical pipeline needs validation, and we have assessed this using simulations. These confirmed that the measures behaved as expected, increasing monotonically with increasing SNR until they reached a saturation point, when further increases in SNR offered no additional benefit. It should be emphasized that although these simulations were based on the effect sizes from the empirical multiband fMRI data, they were not intended

to capture the detailed properties of multiband fMRI *per se*. For example, the simulated datasets differed in their level of additive Gaussian white noise, which cannot fully capture the complex noise properties specific to multiband fMRI. These simulations are therefore an illustration of any generic aspect of the imaging protocol that influences SNR. Additionally, we used general linear models (GLMs) to generate and model the 160,000 simulated datasets (100 repetitions × 100 datasets per repetition × 16 subjects), which would not have been tractable in reasonable time using DCMs. The use of GLMs did not recreate the nonlinearities and parameter covariance present in a more complex models (e.g., DCM), which may be expected to reduce parameter identifiablilty. Nevertheless, the use of GLMs was sufficient for establishing the face validity of the measures, while emphasizing that the outcome measures are not specific to the choice of forward model.

A complementary approach to comparing data would be to assess their predictive validity—i.e., whether effect sizes are detectable with sufficient confidence to predict left-out data (Strother et al., 2002). We haven't pursued this here because our objective is to select the data that maximizes the confidence of hypothesis testing (the precision of inferences over parameters or models). However, in contexts where the objective is to select the dataset which has the best predictive accuracy—such as when identifying biomarkers—this could be performed in the PEB software framework using tools provided for leave-one-out cross-validation (Friston et al., 2016).

Practically, we envisage that a comparison of datasets using the methods described here could be performed on small pilot groups of subjects, the results of which would inform decisions about which imaging protocol to use in a subsequent full-scale study. Regions of interest would be selected for inclusion in the model which are known to show experimental effects for the selected task—based on an initial analysis (e.g., SPM analysis) and/or based on previous studies. The pilot analysis would ideally have the same design—e.g., model structure—as intended for the full-scale study. This is because the quality measures depend on the neuronal parameters of the specific model(s) which will be used by the experimenter to test hypotheses. Following this, we do not expect there exists a "best" acquisition protocol in general for any imaging modality. Rather, the best dataset for a particular experiment will depend on the specific hypotheses (i.e., models) being tested, and the ideal dataset will maximize the precision of the parameters and maximize the difference in evidence between models. We anticipate that the protocol introduced here, implemented in the software accompanying this paper, will prove useful for experimenters when choosing their acquisition protocols.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of the Ethics Committee of University College London with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the Ethics Committee of University College London.

AUTHOR CONTRIBUTIONS

PZ and SK authored the manuscript, designed, and ran the analyses. NT, NW, and KF provided technical guidance and edited the manuscript. MC supervised the study and edited the manuscript.

FUNDING

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2018.00986/full#supplementary-material

REFERENCES

Ashburner, J., and Friston, K. J. (2005). Unified segmentation. *Neuroimage* 26, 839–851. doi: 10.1016/j.neuroimage.2005.02.018

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., and Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron* 76, 695–711. doi: 10.1016/j.neuron.2012.10.038

Buxton, R. B., Uludag, K., Dubowitz, D. J., and Liu, T. T. (2004). Modeling the hemodynamic response to brain activation. *Neuroimage* 23 (Suppl. 1):S220–233. doi: 10.1016/j.neuroimage.2004.07.013

Cauley, S. F., Polimeni, J. R., Bhat, H., Wald, L. L., and Setsompop, K. (2014). Interslice leakage artifact reduction technique for simultaneous multislice acquisitions. *Magn. Reson. Med.* 72, 93–102. doi: 10.1002/mrm.24898

Chumbley, J. R., Friston, K. J., Fearn, T., and Kiebel, S. J. (2007). A metropolis-hastings algorithm for dynamic causal models. *Neuroimage* 38, 478–487. doi: 10.1016/j.neuroimage.2007.07.028

Corbin, N., Todd, N., Friston, K. J., and Callaghan, M. F. (2018). *Accurate modeling of temporal correlations in rapidly sampled fMRI time series. Hum. Brain Mapp.* 39, 3884–3897. doi: 10.1002/hbm.24218

David, O., Guillemain, I., Saillet, S., Reyt, S., Deransart, C., Segebarth, C., et al. (2008b). Identifying neural drivers with functional MRI: an electrophysiological validation. *PLoS Biol.* 6, 2683–2697. doi: 10.1371/journal.pbio.0060315

David, O., Wozniak, A., Minotti, L., and Kahane, P. (2008a). Preictal short-term plasticity induced by intracerebral 1 Hz stimulation. *Neuroimage* 39, 1633–1646. doi: 10.1016/j.neuroimage.2007.11.005

Friston, K. J. (2011). Functional and effective connectivity: a review. *Brain Connect.* 1, 13–36. doi: 10.1089/brain.2011.0008

Friston, K. J., Harrison, L., and Penny, W. (2003). Dynamic causal modelling. *Neuroimage* 19, 1273–1302. doi: 10.1016/S1053-8119(03)00202-7

Friston, K. J., Litvak, V., Oswal, A., Razi, A., Stephan, K. E., van Wijk, B. C. M., et al. (2016). Bayesian model reduction and empirical Bayes for group

(DCM) studies. *Neuroimage* 128, 413–431. doi: 10.1016/j.neuroimage.2015.
11.015

Friston, K. J., Mattout, J., Trujillo-Barreto, N., Ashburner, J., and Penny, W.
(2007). Variational free energy and the Laplace approximation. *Neuroimage* 34,
220–234. doi: 10.1016/j.neuroimage.2006.08.035

Kass, R. E., and Raftery, A. E. (1995). Bayes Factors. *J. Am. Stat. Assoc.* 90, 773–795.
doi: 10.1080/01621459.1995.10476572

Kiebel, S. J., Kloppel, S., Weiskopf, N., and Friston, K. J. (2007).
Dynamic causal modeling: a generative model of slice timing in
fMRI. *Neuroimage* 34, 1487–1496. doi: 10.1016/j.neuroimage.2006.
10.026

Kirilina, E., Lutti, A., Poser, B. A., Blankenburg, F., and Weiskopf, N. (2016).
The quest for the best: The impact of different EPI sequences on the
sensitivity of random effect fMRI group analyses. *Neuroimage* 126, 49–59.
doi: 10.1016/j.neuroimage.2015.10.071

Larkman, D. J., Hajnal, J. V., Herlihy, A. H., Coutts, G. A., Young, I. R.,
and Ehnholm, G. (2001). Use of multicoil arrays for separation of signal
from multiple slices simultaneously excited. *J. Magn. Reson. Imaging* 13,
313–317. doi: 10.1002/1522-2586(200102)13:2<313::AID-JMRI1045>3.0.
CO;2-W

Lee, L., Friston, K., and Horwitz, B. (2006). Large-scale neural
models and dynamic causal modelling. *Neuroimage* 30, 1243–1254.
doi: 10.1016/j.neuroimage.2005.11.007

Litvak, V., Garrido, M., Zeidman, P., and Friston, K. (2015). Empirical bayes for
group (DCM) studies: a reproducibility study. *Front. Hum. Neurosci.* 9:670.
doi: 10.3389/fnhum.2015.00670

Neyman, J. P., and Pearson, E. S. (1933). On the problem of the most
efficient tests of statistical hypotheses. *Philos. Trans. R. Soc.* 231, 289–337.
doi: 10.1098/rsta.1933.0009

Penny, W., Kiebel, S., and Friston, K. (2003). Variational bayesian
inference for fMRI time series. *Neuroimage* 19, 727–741.
doi: 10.1016/S1053-8119(03)00071-5

Penny, W. D., Stephan, K. E., Mechelli, A., and Friston, K. J. (2004).
Modelling functional integration: a comparison of structural equation
and dynamic causal models. *Neuroimage* 23 (Suppl. 1), S264–274.
doi: 10.1016/j.neuroimage.2004.07.041

Pinotsis, D. A., Perry, G., Litvak, V., Singh, K. D., and Friston, K. J. (2016).
Intersubject variability and induced gamma in the visual cortex: DCM
with empirical B ayes and neural fields. *Hum. Brain Mapp.* 37, 4597–4614.
doi: 10.1002/hbm.23331

Reyt, S., Picq, C., Sinniger, V., Clarencon, D., Bonaz, B., and David,
O. (2010). Dynamic causal modelling and physiological confounds: a
functional MRI study of vagus nerve stimulation. *Neuroimage* 52, 1456–1464.
doi: 10.1016/j.neuroimage.2010.05.021

Setsompop, K., Gagoski, B. A., Polimeni, J. R., Witzel, T., Wedeen, V. J., and Wald,
L. L. (2012). Blipped-controlled aliasing in parallel imaging for simultaneous
multislice echo planar imaging with reduced g-factor penalty. *Magn. Reson.
Med.* 67, 1210–1224. doi: 10.1002/mrm.23097

Spreng, R. N., Mar, R. A., and Kim, A. S. (2009). The common neural basis
of autobiographical memory, prospection, navigation, theory of mind, and
the default mode: a quantitative meta-analysis. *J. Cogn. Neur.* 21, 489–510.
doi: 10.1162/jocn.2008.21029

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., and Friston, K. J.
(2009). Bayesian model selection for group studies. *Neuroimage* 46, 1004–1017.
doi: 10.1016/j.neuroimage.2009.03.025

Stephan, K. E., Weiskopf, N., Drysdale, P. M., Robinson, P. A., and Friston, K. J.
(2007). Comparing hemodynamic models with DCM. *Neuroimage* 38, 387–401.
doi: 10.1016/j.neuroimage.2007.07.040

Strother, S. C., Anderson, J., Hansen, L. K., Kjems, U., Kustra, R., Sidtis,
J., et al. (2002). The quantitative evaluation of functional neuroimaging
experiments: the NPAIRS data analysis framework. *Neuroimage* 15, 747–771.
doi: 10.1006/nimg.2001.1034

Todd, N., Josephs, O., Zeidman, P., Flandin, G., Moeller, S., and Weiskopf,
N. (2017). Functional sensitivity of 2D simultaneous multi-slice echo-planar
imaging: effects of acceleration on g-factor and physiological noise. *Front.
Neurosci.* 11, 158. doi: 10.3389/fnins.2017.00158

Todd, N., Moeller, S., Auerbach, E. J., Yacoub, E., Flandin, G., and Weiskopf, N.
(2016). Evaluation of 2D multiband EPI imaging for high-resolution, whole-
brain, task-based fMRI studies at 3T: Sensitivity and slice leakage artifacts.
Neuroimage 124, 32–42. doi: 10.1016/j.neuroimage.2015.08.056

Xu, J., Moeller, S., Auerbach, E. J., Strupp, J., Smith, S. M., Feinberg, D. A., et al.
(2013). Evaluation of slice accelerations using multiband echo planar imaging
at 3 T. *Neuroimage* 83, 991–1001. doi: 10.1016/j.neuroimage.2013.07.055

Zeidman, P., Mullally, S. L., and Maguire, E. A. (2015). Constructing, perceiving,
and maintaining scenes: hippocampal activity and connectivity. *Cereb. Cortex*
25, 3836–3855. doi: 10.1093/cercor/bhu266

APPENDIX

APPENDIX 1: BAYESIAN DATA COMPARISON

In this appendix we illustrate why the entropy of the posterior over model parameters obtained from two different datasets, can be compared as if they were log Bayes factors. We first consider the comparison of different models of the same data. We then turn to the complementary application of Bayes theorem to the comparison of different data under the same model, in terms of their relative ability to reduce uncertainty about model parameters.

Bayesian Model Comparison

To clarify model comparison in a conventional setting, consider two models m_1 and m_2 fitted to the *same* data y. The Bayes factor in favor of model 1 is the ratio of the model evidences under each model:

$$BF_1 = \frac{p\left(y|m_1\right)}{p\left(y|m_2\right)} \tag{A1}$$

In place of the model evidence $p(y|m)$ we estimate an approximation (lower bound)—the negative variational free energy [this usually calls on variational Bayes, although other approximations can be used such as harmonic means from sampling approximations and other (Akaike or Bayesian) information criteria]:

$$F \approx \ln p(y|m) \tag{A2}$$

As we are working with the log of the model evidence, it is more convenient to also work with the log of the Bayes factor:

$$\ln BF_1 = \ln p\left(y|m_1\right) - \ln p\left(y|m_2\right) \approx F_1 - F_2 \tag{A3}$$

The log Bayes factor for two models is simply the difference in their free energies, in units of nats. Kass and Raftery (1995) assigned labels to describe the strength of evidence for one model over another. For example, "strong evidence" requires a Bayes factor of 20 or a log Bayes factor of $\ln(20) \approx 3$. We can also transform the log Bayes factor to a posterior probability for one model relative to the other under uninformative priors over models: $p\left(m_1\right) = p\left(m_2\right) = p\left(m\right)$. By using Bayes rule and re-arranging, we find this probability is a sigmoid (i.e., logistic) function of the log Bayes factor:

$$p\left(m_1|y\right) = \frac{p\left(y|m_1\right)p(m)}{p\left(y\right)} = \frac{p\left(y|m_1\right)}{p\left(y|m_1\right) + p\left(y|m_2\right)}$$
$$= \frac{1}{1 + \exp(-lnBF_1)} \tag{A4}$$

The final equality shows that the logistic function of the log Bayes factor in favor of model 1 gives the posterior probability for model 1, relative to model 2. When dealing with multiple models, the logistic function becomes a softmax function (i.e., normalized exponential function).

Accuracy and Complexity

The log evidence or free energy can be decomposed into accuracy and complexity terms:

$$F = \underbrace{\langle \ln p(y|m)\rangle_q}_{accuracy} - \underbrace{KL\left[q(\theta) \parallel p(\theta)\right]}_{complexity} \tag{A5}$$

Where $q\left(\theta\right) \approx p\left(\theta|y\right)$ is the posterior over parameters, $p(\theta)$ are the priors on the parameters and $\langle \cdot \rangle_q$ denotes the expectation under $q\left(\theta\right)$. This says that the accuracy is the expected log likelihood of the data under a particular model. Conversely, the complexity scores how far the parameters had to move from their priors to explain the data, as measured by the KL-divergence between the posterior and prior distributions. This divergence is also known as a relative entropy or information gain. In other words, the complexity scores the reduction in uncertainty afforded by the data. We can now use this interpretation of the complexity to compare the ability of different data sets to reduce uncertainty about the model parameters.

Comparing Across Datasets

Consider one model fitted to two datasets y_1 and y_2 with approximate posteriors over the parameters $q_1(\theta)$ and $q_2(\theta)$ for each data set. The corresponding complexity difference is:

$$\begin{aligned}
&\left(KL\left[q_1(\theta) \parallel p\left(\theta\right)\right] - KL\left[q_2(\theta) \parallel p\left(\theta\right)\right]\right) \\
&= \langle \ln q_1(\theta)\rangle_q - \langle \ln q_2(\theta)\rangle_q \\
&= H[q_2(\theta)] - H[q_1(\theta)] \\
&= \left\langle \ln \frac{q_1(\theta)}{q_2(\theta)} \right\rangle_q
\end{aligned} \tag{A6}$$

Where $H[q(\theta)]$ is the entropy of $q\left(\theta\right)$. Therefore, the reduction in conditional uncertainty (i.e., the difference in the entropies of the approximate posteriors) corresponds to the difference in information gain afforded by the two sets of data. Because this difference is measured in *nats*, it has the same interpretation as a difference in log evidence—or a log odds ratio (i.e., Bayes factor). The last equality above shows that the difference in entropies (or complexities) corresponds to the expected log odds ratio of posterior beliefs about the parameters. The entropies are defined for the multivariate normal distribution in Equation 9 of the main text.