# A General Method for Testing Bayesian Models using Neural Data

**Gabor Lengyel**
Center for Visual Science
Department of Brain and Cognitive Sciences
University of Rochester
Rochester, NY 14620
lengyel.gaabor@gmail.com

**Sabyasachi Shivkumar**
Zuckerman Institute
Columbia University
New York, NY 10027
Center for Visual Science
Department of Brain and Cognitive Sciences
University of Rochester
Rochester, NY 14620
sabyashiv@gmail.com

**Ralf Haefner**
Center for Visual Science
Department of Brain and Cognitive Sciences
University of Rochester
Rochester, NY 14620
ralf.haefner@rochester.edu

**Editors:** Marco Fumero, Emanuele Rodolà, Clementine Domine, Francesco Locatello, Gintare Karolina Dziugaite, Mathilde Caron

## Abstract

Bayesian models have been successful in explaining human and animal behavior, but the extent to which they can also explain neural activity is still an open question. A major obstacle to answering this question is that current methods for generating neural predictions require detailed and specific assumptions about the encoding of posterior beliefs in neural responses, with no consensus or decisive data about the nature of this encoding. Here, we present a new method and prove conditions for its validity, that overcomes these challenges for a wide class of probabilistic encodings – including the two major classes of neural sampling and distributed distributional codes. Our method tests whether the relationships between the model posteriors for different stimuli match the relationships between the corresponding neural responses – akin to representational similarity analysis (RSA), a widely used method for nonprobabilistic models. Finally, we present a new model comparison diagnostic for our method, based not on the agreement of the model with the data directly, but on the alignment of the model and data when injecting noise in our neural prediction generation method. We illustrate our method using simulated V1 data and compare two Bayesian models that are practically indistinguishable using behavior alone. Our results show a powerful new way to rigorously test Bayesian models on neural data.

## 1 Introduction

Bayesian computational models have been successfully used to explain animal and human behavior. Previous studies have shown that Bayesian models can capture not only perceptual processes (e.g.,

Proceedings of the I edition of the Workshop on Unifying Representations in Neural Models (UniReps 2023).

[11, 23, 24, 29]) and perceptual decision-making (e.g., [16, 32]), motor learning (e.g., [18, 26]), perceptual learning (e.g., [14]), and statistical learning (e.g., [41]) but also higher-level cognition such as abstract concept formation (e.g., [30, 46]), language acquisition (e.g., [51]), and rule learning (e.g., [13]). Thus, the Bayesian modeling framework has the potential to capture the full complexity of the computations that underlie human perception, cognition, and learning [31]. Testing the Bayesian Brain Hypothesis [25] using neural data has therefore been of great interest in systems neuroscience [12, 43]. There are two key elements to testing the Bayesian Brain Hypothesis using neural activity. First, what is the generative (Bayesian) model in which the brain performs inference and which defines the posterior over the latent variables of the brain? Second, how are posterior beliefs encoded in neural activity?

*Simplified assumptions about the brain's generative model:* Broadly speaking, there are two main ways in which generative models have been constructed. First, researchers have derived their models from a task (ideal observer model), assuming that the task-relevant variables can be decoded from neural responses (e.g., [36, 45, 50]). Second, researchers have derived their models from natural input statistics (e.g., [39, 44]) and linked posterior beliefs to neural activity by making very specific assumptions such as 'neural sampling' (e.g., [12, 20, 40]). Although the two approaches have diverged widely (see [34] for a summary), both use models that drastically simplify the rich internal model that the brain likely uses. Models derived from tasks that do not cover the range of natural behavior and consider only a few task-relevant variables and specific stimuli (e.g., orientation or motion). Even current image-computable models derived from natural input statistics ignore many stimulus dimensions, e.g., color or binocular disparity. Therefore, it is unclear under what conditions we can expect the neural predictions of a simplified Bayesian model to match the measured neural data, even if the assumed neural encoding of probabilities is correct.

*No consensus how probabilities are encoded neurally:* There are two major classes of theories of how the brain might encode probabilities mirroring the principal ways to perform Bayesian inference in machine learning: sampling-based codes (neural sampling, assuming that the neural activity represents samples from a posterior, e.g., [6, 12, 17, 20, 40]), and parametric codes (including Probabilistic Population Codes (PPCs), assuming that neural activities are linearly related to log probabilities, e.g., [1, 36, 45] and Distributed Distributional Codes (DDCs), assuming that neural activities are linearly related to probabilities, e.g., [48, 49]). However, despite multiple studies presenting evidence in favor of one of these codes, decisive evidence in favor of one and against the other codes is still missing ([2, 12, 34, 43], but see [47]). Therefore, it is unclear which code to assume when testing a specific Bayesian model, especially since each code usually requires several additional unknown parameters.

**Main contributions:** Our work (partially) overcomes the two challenges of model simplification and unknown encoding by providing a method for testing Bayesian models with neural data that is *independent* of the specific details of the encoding, as long as the encoding is part of two of the three major previously proposed encoding schemes (neural sampling and DDCs). We analytically derive the conditions under which our method allows for correct predictions of neural responses and illustrate them with ground truth simulations by comparing two qualitatively different probabilistic models of cortical area V1. Finally, we present a new model comparison diagnostic for our method, based not on the agreement of the model with the data directly, but on the alignment of model and data.

## 2 Results

### 2.1 Generating neural predictions

The Bayesian brain framework [12, 43] assumes that the brain infers the states of latent variables, $z$, given sensory observations, $o_z$, forming a posterior belief, $p\left(z \mid o_z\right)$. The neural activity is assumed to encode the brain's beliefs about the posterior probability distribution:

$$p\left(r \mid o_z\right) = \mathcal{R}_{p(z)\to p(r)}\left[p\left(z \mid o_z\right)\right] \tag{1}$$

where $r$ denotes the neural activity, and $\mathcal{R}_{p(z)\to p(r)}$ is the encoding mapping from $p\left(z\right)$ to $p\left(r\right)$. Importantly, only $p(r)$ is directly observable by the "scientist", while the brain's "true" latent variables, $z$, and its "true" internal model are unknown. Current models are highly simplified based on simplified

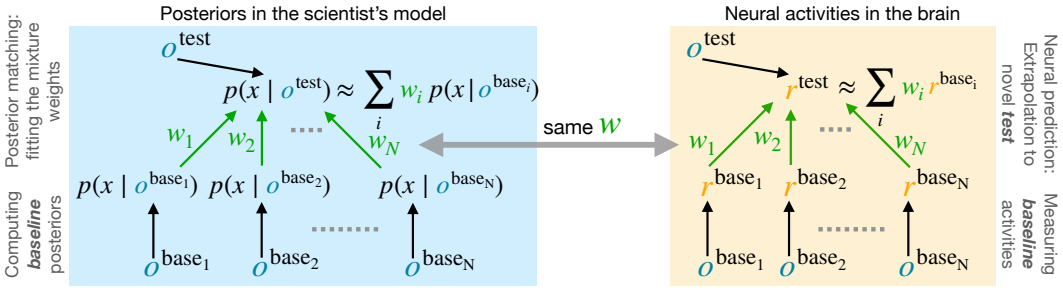


Figure 1: **Our method**: Comparing the relationship between the posteriors for the baseline and test stimuli (left, blue background) to the relationship between the neural responses to the baseline and test stimuli (right, yellow background). Using the scientist's Bayesian model, we fit the mixture weights, $w$, to match the posterior for a test stimulus, $p(x \mid o^{\text{test}})$, as a mixture of posteriors for some baseline stimuli, $p(x \mid o^{\text{base}_n})$. We then combine the neural responses to the baseline stimuli, $r^{\text{base}_n}$ with the same mixture weights, $w$, to predict the neural response to the test stimulus, $r^{\text{test}}$.

inputs and/or specific tasks. To understand the conditions under which such a simplified, "scientist" model, $p\left(x|o_x\right)$, can correctly predict neural responses, $r$, we first relate it to the brain's "true" model:

$$p\left(z \mid o_z\right) = \int \overbrace{p\left(z \mid x, o_{z \setminus x}\right)}^{\mathcal{M}_{x \to p(z)}} p\left(x \mid o_x\right) \, \mathrm{d}x \tag{2}$$

where $x$ and $o_x$ represent the latent variables and the observations in the scientist's model, respectively. $\mathcal{M}_{x \to p(z)}$ denotes the mapping from the scientist's to the brain's model and $o_{z \setminus x}$ denotes the subset of the brain's observations that are not shared with the "scientist model" (e.g., the color of a visual stimulus when the "scientist model" only assumes black and white observations). Thus, the probability over the neural activity given the observations can be written as follows:

$$p\left(r \mid o_z\right) = \mathcal{R}_{p(z) \to p(r)} \left[ \int \mathcal{M}_{x \to p(z)} \left(x, o_{z \setminus x}\right) p\left(x \mid o_x\right) \, \mathrm{d}x \right] \tag{3}$$

Previous studies had to make specific assumptions about the nature of $\mathcal{R}_{p(z) \to p(r)}$, e.g., neural sampling, PPC, or DDC, to make concrete neural predictions. They also ignored the difference between their assumed generative model and the brain's actual generative model without any guarantees that their model prediction would match the neural data even if the assumed neural code, $\mathcal{R}_{p(z) \to p(r)}$, were correct.

### 2.2 Our method

The idea behind our method is to compare the relationships between the posteriors computed for a set of stimuli using our Bayesian model to the relationship between the measured neural activities in responses to the same set of stimuli (Fig. 1). To this end, we formalize how to extrapolate from measured neural activities in responses to a set of simple 'baseline' stimuli using the relationships between the posteriors in the Bayesian model to predict the neural responses to complex 'test' stimuli (Figs. 1 & 2). Instead of making a specific (and probably incorrect) assumption about how the brain represents probabilities, we use the measured baseline responses (Fig. 2 D) together with the posteriors predicted by the Bayesian model for each baseline stimulus (Fig. 2 A) as a kind of look-up table. This table contains the link between posteriors in the scientist model and neural responses, combining $\mathcal{M}_{x \to p(z)}$ and $\mathcal{R}_{p(z) \to p(r)}$. Instead of making assumptions about the neural code, we measure it directly using the baseline stimuli.

This idea results in a simple linear extrapolation method for a wide class of encoding schemes called Linear distribution codes (LDCs) [33]. LDCs are defined by the property that the representation of a mixture of distributions corresponds to the weighted average of the representations of the individual mixture components [33]:

$$\mathcal{R}_{p(z) \to p(r)} \left[ \sum_i w_i p_i(z) \right] = \sum_i w_i \mathcal{R}_{p(z) \to p(r)} \left[ p_i(z) \right]. \tag{4}$$

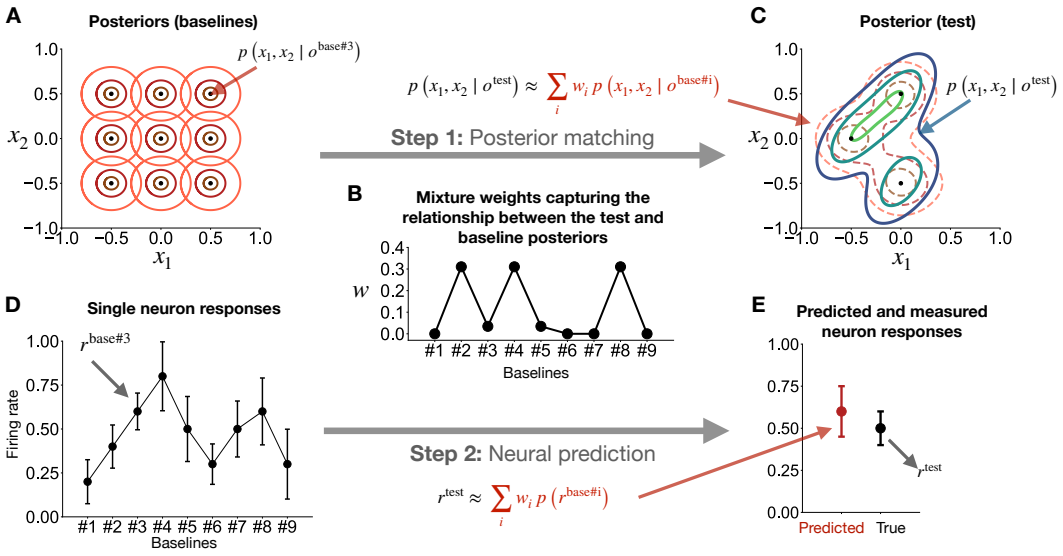


Figure 2: **Generating neural predictions from a Bayesian model using our method. A-C**: Step 1, fitting the weights, $w$ (B), to match the posterior for the test stimulus as a mixture of the posteriors for the baseline stimuli (C). $x_1$, $x_2$, $o^{\text{base}}$ and $o^{\text{test}}$ denote the latent variables and the observations when the baseline or the test stimulus is observed, respectively. **C**: Show the posterior for the test stimulus and its approximation by the mixture of the posteriors for the baselines. **D, B & E**: Step 2, using the weights ($w$) to extrapolate from the measured neural responses to the baseline stimuli (D) to predict the neural response to the test stimulus (E). $r^{\text{base}}$ and $r^{\text{test}}$ denote the neural activity in response to the baseline and the test stimuli, respectively.

Importantly, this property (which defines an affine mapping since $\sum_i w_i = 1$ [38]) holds for both neural sampling schemes and all DDCs, but not for PPCs or expectile codes.

The **central result** of our paper is the following. If a (test) *posterior* can be expressed as a mixture of (baseline) *posteriors* in the scientist model, then the *neural response* to the test stimulus (assuming an LDC) is a linear combination of the *neural responses* to the baseline stimuli using the posterior mixture weights as linear coefficients (Fig. 1).

To prove this, first, we write the neural activity in response to the test stimuli using eq. (3):

$$p(r^{\text{test}}) = \mathcal{R}_{p(z)\to p(r)} \left[ \int \mathcal{M}_{x\to p(z)} \left( x, o^{\text{test}}_{z\setminus x} \right) p\left( x \mid o^{\text{test}}_x \right) \mathrm{d}x \right] \tag{5}$$

Next, using mixture weights $w_i$, we approximate the posterior for the test stimulus, $p\left(x \mid o^{\text{test}}_x\right)$, as a mixture of baseline posteriors, computed for the baseline stimuli (Fig.2 A-C, Step 1: Posterior matching):

$$p\left(x \mid o^{\text{test}}_x\right) \approx \sum_i w_i p\left(x \mid o^{\text{base}_i}_x\right). \tag{6}$$

Assuming that the brain's observations that are not modeled by the "scientist model" ($o_{z\setminus x}$) are the same when responses to baseline and test stimuli are measured, we can substitute eq. (6) into (5):

$$p(r^{\text{test}}) \approx \mathcal{R}_{p(z)\to p(r)} \left[ \int \mathcal{M}_{x\to p(z)} \left( x, o^{\text{base}_i}_{z\setminus x} \right) \sum_i w_i p\left( x \mid o^{\text{base}_i}_x \right) \mathrm{d}x \right] \tag{7}$$

If the encoding, $\mathcal{R}_{p(z)\to p(r)}$, is an LDC, we can rewrite eq. (7) as:

$$p(r^{\text{test}}) \approx \sum_i w_i \overbrace{\mathcal{R}_{p(z)\to p(r)} \left[ \int \mathcal{M}_{x\to p(z)} \left( x, o^{\text{base}_i}_{z\setminus x} \right) p\left( x \mid o^{\text{base}_i}_x \right) \mathrm{d}x \right]}^{p\left(r^{\text{base}_i}\right)} = \sum_i w_i\, p\left(r^{\text{base}_i}\right) \tag{8}$$

This means we can measure the neural activity in response to some *baseline stimuli* and use a weighted mixture of those measurements to predict the neural activity in response to the desired *test stimuli*. This way, we use the brain's 'true' model with its encoding directly by measuring it in an experiment with the baseline stimuli, and we do not need to make any assumptions about $\mathcal{R}_{p(z)\to p(r)}$ or $\mathcal{M}_{x\to p(z)}$ other than that $\mathcal{R}_{p(z)\to p(r)}$ is an LDC and that $o_{z\setminus x}$ is common for the baseline and test stimuli. Note that this latter assumption is inherent to all laboratory measurements where all aspects of the stimulus apart from the parameter of interest are held fixed. In practice, it is not always feasible to keep all variables constant except those under study. Still, the hope in such experiments is that this variability will average out in those cases, and only add noise, but not bias to ones measurements. However, this issue is not unique to our method; instead, it is a challenge inherent in experimenting in general. Crucially, the weights $w_i$ are determined by the scientist model, $p(x|o_x)$, and reflect how the baseline posteriors and the test posterior are related. This relationship makes an empirically testable prediction about how the empirical responses to the baseline stimuli should be related to the empirical response to the test stimulus.

### 2.3 The two steps to generate neural predictions from Bayesian models

Step 1: Posterior matching in the scientist model: determine $w_i$ such that
$p\left(x \mid o_x^{\text{test}}\right) \approx \sum_i w_i \, p\left(x \mid o_x^{\text{base}_i}\right)$

Step 2: Predict response to test stimulus as the weighted average of baseline activities
$p\left(r^{\text{test}}\right) \approx \sum_i w_i \, p\left(r^{\text{base}_i}\right)$

The first step only requires us to compute the posteriors using our Bayesian model, while the second step requires us to measure neural activities in response to the baseline stimuli. Importantly, the quality of the neural prediction depends on the ability of the scientist's model to correctly capture the *relationship* between the brain's posteriors for the baseline and for the test stimuli. Critically, the relationships predicted by different models differ depending on the nature of their latents. This allows us to use neural data to compare different models, even if they make identical behavioral predictions. The quality of the prediction using this method depends only on the scientist's model, and it is agnostic to what the brain's true encoding is as long as it is an LDC.

### 2.4 The relationship between the posterior matching error and the neural prediction error

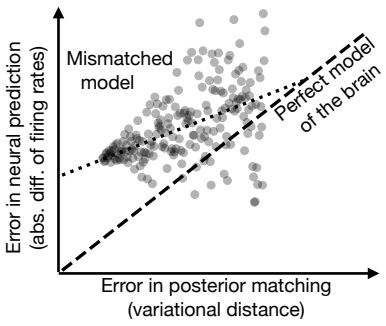


Figure 3: **The relationship between neural prediction error and posterior matching error.**

The error in the neural prediction will reflect the extent to which the scientist's model can capture the brain's computations in the context of the stimuli only if, we can perfectly match the test posterior as a mixture of the baseline posteriors in the posterior matching step (Fig.2 A-C). If we cannot perfectly match the test posterior in the weight-fitting step, this introduces an additional source of error in the neural prediction step. In general, one can expect that for a 'good' model, the better the mixture of baseline posteriors matches the test posterior, the smaller the error in the neural predictions will be. Interestingly, we could show (in general for binary and under reasonable assumptions for categorical latents, see App. A.2&A.2) that the relationship between the posterior matching error (quantified as variational distance) and neural response error is perfectly linear for the correct (true brain) model (Fig. 3). On the other hand, it does not hold for models that are mismatched: the neural error will not be zero for zero posterior matching error, and the correlation between both types of errors will be less than one since a set of mixture weights that yield a poor posterior match may produce a weighted average of neural responses that, by chance, match the empirically observed response to the test stimulus. Overall, this suggests that better models should imply a tighter correlation between posterior error and neural error, which we have also observed in our simulations (see Fig. 7 & Fig.S6 for a couple of example neurons).

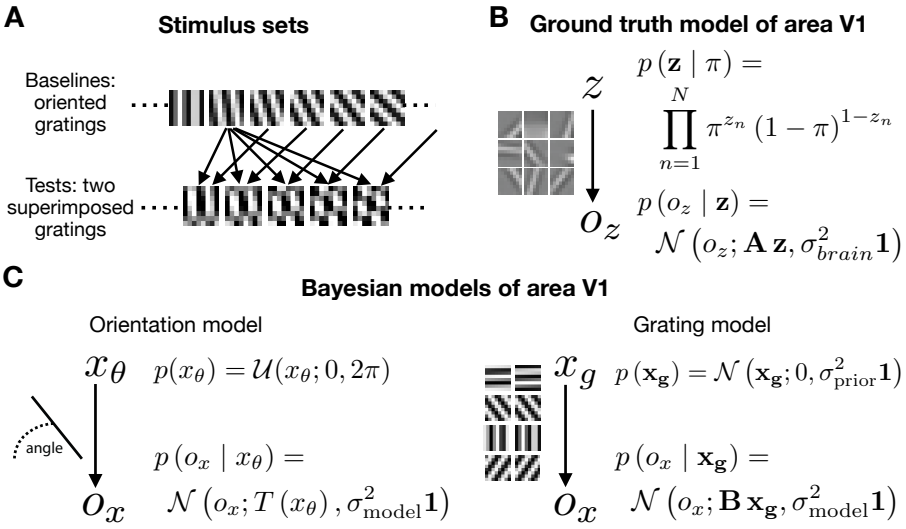


Figure 4: **Illustration of method using ground truth simulations. A**: Examples of baseline (top) and test stimuli (bottom). **B**: The ground truth model from which we generated simulated neural data. **C**: The two Bayesian models being compared.

### 2.5 A new diagnostic for model comparison

The relationship between the error in the posterior matching and the error in the neural prediction demonstrates the importance of controlling for the posterior matching error when comparing models. Without such control, we can't discern whether the difference between the models' prediction errors is due to different posterior matching errors or if one model genuinely aligns more closely with the brain than the other. Therefore, we developed a new diagnostic for model comparison based on how tightly the posterior matching error is coupled to the neural prediction error. This diagnostic doesn't depend on the absolute value in the posterior matching and it can be computed with the following weight perturbation analysis:

**For** each test stimulus

1. Generate neural predictions using Steps 1 & 2 above (section 2.3).
2. Generate many sets of perturbed weights ($w$) that deviate from the weights, fitted in the previous point, providing the best possible match for the test posterior. E.g., $w_i^{\text{jittered}} \sim Dirichlet(w_i^{\text{perfect}}\alpha)$.
3. Generate neural predictions using Step 2 (section 2.3) for the sets of perturbed weights.
4. Assess the strength of the relationship (e.g., using a correlation coefficient) between the error in the posterior matching and the error in the neural prediction for the perturbed weights. (Optionally, convert this to an error metric, e.g., 1-correlation)

Our numerical experiments using ground truth simulations confirm the value of this metric and show that it provides complementary information in many cases in which the traditional root mean squared error (RMSE) metric for model quality is the same for two competing models (see App. A.3.6 for more details and Fig. 7 & Fig.S6). Traditional metrics of model fit (here, RMSE) evaluate only one point of the relationship between the neural prediction and the posterior matching error when the posterior matching error is zero. Our new metric provides information about the strength of the relationship between the two errors when the posterior error is different from zero.

### 2.6 Validation & illustration of the method

Using a simulated ground truth model, we demonstrate the validity of our approach and its applicability in comparing two Bayesian models. Since validating a method requires data for which the ground truth is known, and no such neurophysiological data exists, we relied on synthetic data. We simulated

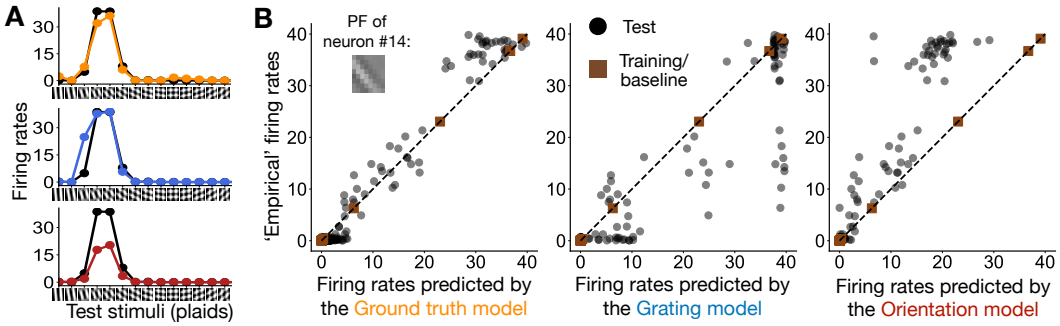


Figure 5: **The neural predictions for a representative neuron. A:** Empirical (black) and predicted tuning curves for the ground truth (orange), grating (blue), and orientation (red) models for 14 example test stimuli. **B:** The empirical as a function of the predicted firing rates for all test stimuli. Closer to the diagonal represents better predictions. Brown squares and black points represent the baseline and the test stimuli, respectively. Left inset: projective field of the neuron.

ground truth neural data from a binary sparse coding image model for V1 trained on natural images [5] (Fig.4B). The binary latent variables ($\mathbf{z}$, binary vector) represent 128 projective fields ($\mathbf{A}$, $8 \times 8$ (image pixels) by 128 (projective fields) matrix). We compare two toy Bayesian models using the simulated neural data (assuming an orientation discrimination task): 1) a scalar orientation model derived from the orientation task and 2) a grating Gaussian mixture model derived from the stimulus images. The orientation model (Fig.4C, left) assumes that the stimuli are generated from a single scalar variable, orientation ($x_\theta$, scalar), using a template function ($T(.)$) transforming an orientation into the image of a grating. The grating model (Fig.4C, right) assumes a Gaussian mixture model with 24 continuous latent variables ($\mathbf{x_g}$, vector) with gratings as projective fields ($\mathbf{B}$, $8 \times 8$ (image pixels) by 24 (projective fields) matrix). The spatial frequency of the grating projective fields was fixed to one, but the orientation $(0°, 45°, 90°, 135°)$ and the phase $(0°, 60°, 120°, 180°, 240°, 300°)$ varied. We used 20 single gratings of varying orientation, tiling the space of $0° - 180°$ for baseline stimuli. For test stimuli, we used 210 plaids generated by adding two baseline gratings pixel by pixel (all unique combinations) (Fig.4A). We chose single gratings and plaids for our stimuli to highlight the difference between the grating and the orientation models. Since the ground truth mixture model, similar to the grating mixture model, allows for multiple grating features to be present in the stimuli, we expect that the grating model will provide better fits to the ground truth model for the test plaids than the orientation model which tries to capture the plaid with only a single latent feature. Note that this simulation is meant to demonstrate the use and value of our method in comparing Bayesian models using neural data, *not* to provide a realistic model of the actual V1. (See App. A.3 & Fig. S1 for more details on the models and the stimuli.)

Although our method can be used for all types of neural data, including population activities and imaging data, as long as eq. (4) holds, for simplicity, we will illustrate how to use it to predict single-neuron activities. First, we assume that a single latent variable in the Bayesian models describes the response of a single neuron. Then, following the two steps in section 2.3, we generate neural prediction for a single test stimulus by fitting a set of weights to approximate the scientist's posterior for the test stimulus as a mixture of baseline posteriors (Fig.2A,B&C). Then, we used those weights to combine the measured baseline activities to predict the neuron's activity in response to the test stimulus (Fig.2D,C&E). (See these steps in Fig.S3). We repeated this procedure for every test stimulus. Since the orientation model has a single latent, it produces a single prediction for all neurons. However, the grating model has multiple latents, and we generated multiple predictions for the neurons using the marginal posterior over each latent. We then picked the best fitting latent for each neuron. (See App. A.3.4 for more details)

First, we demonstrate the validity of our method. We generated neural predictions from the ground truth model using our method and compared them to the neural predictions of the orientation and the grating models. We found that the neural predictions of the ground truth model provided significantly better fits to the simulated activity of the neurons across the test stimuli than the other two models (Wilcoxon signed-rank test, $P$s$< 0.05$) (see App. A.3.5 and Fig.5&Fig.S5). This demonstrates that with our method, we can correctly identify better models.

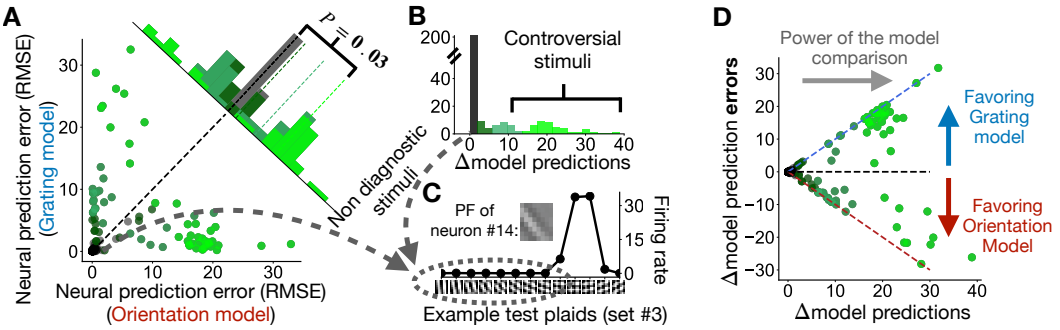


Figure 6: **Comparing models.** RMSEs (**A**), prediction differences (**B**), and RMSE differences between the two models (**D**) for diagnostic (green) and non-diagnostic (black) stimuli (**C**). $\Delta$ denotes the difference between the two models.

Next, we illustrate how to compare the neural predictions of the orientation and the grating models for a representative neuron. First, by looking at example tuning curves (Fig.5A), we see that both models capture the firing rates in response to the test stimuli qualitatively well. Overall, the orientation model seems to underestimate (Fig.5A&B in red) while the grating model seems to overestimate the firing rates (Fig.5B-middle). Interestingly, the two models generated almost identical predictions for most stimuli (Fig.6B, in black). Most of these test stimuli elicit minimal activity from the simulated neuron (Fig.6C), and both models accurately predict no activity for these stimuli. Responses to the diagnostic, controversial stimuli [15], on the other hand, provided substantial evidence in favor of either model (Fig.6B&D, in green). Considering all diagnostic stimuli, the grating model had substantially lower RMSE than the orientation model for this representative neuron (Wilcoxon signed-rank test, $T = 241$, $P = 0.03$, Fig.6A&D, in green, see other example neurons in Fig.S5) Overall, this example highlights that our method allows researchers to choose their stimuli before experimenting to maximize statistical power in their model comparison.

Finally, we performed the weight perturbation analysis for all the test stimuli for the same representative neuron. For each test stimulus, we generated 30 sets of perturbed weights. Then, across the 30 sets of weights, we computed the correlation coefficient between the posterior matching error and the error in the neural prediction. We found that for most of the test stimuli, the correlation coefficients were larger for the grating model than the orientation model (Wilcoxon signed-rank test, $T = 12471$, $P = 3 \cdot 10^{-12}$, Fig.7B). Importantly, many stimuli, which produced similar neural predictions across the two models and were found to be non-diagnostic with the RMSE metric, become diagnostic with this new correlational metric (Fig.6C, see other example neurons in Fig.S5). This implies that in the context of our method, the weight perturbation analysis provides complementary information to the neural prediction error, and suggests that it might be a more sensitive diagnostic than the conventional RMSE.

## 3 Discussion

Our work makes two main contributions: First, we present a method for generating neural predictions from Bayesian models that does not depend on the details of how probabilities are encoded in neural responses as long as the encoding is part of a large class of codes (LDCs), including neural sampling and DDCs. Second, we present a new metric tailored to our method that is complementary to classic prediction errors.

Methods for comparing nonprobabilistic models to neural data have yielded many insights into how the brain processes sensory information (e.g., [8, 21, 22, 27, 28, 37]). Our method is related to Representational Similarity Analysis (RSA) [9] in that it compares the relationship between model responses (in our case, posteriors) to a set of stimuli to the relationship between neural responses to the same set of stimuli. However, it differs from RSA in that RSA compares the relationship between many model (or neural) responses for pairs of stimuli, while our method compares the relationship between a single model (or neural) response across many stimuli. Further, RSA is based on 2nd order summary statistics, while our method makes specific testable predictions for individual neural responses. Our method is also similar to the main alternative to RSA, deterministic encoding models

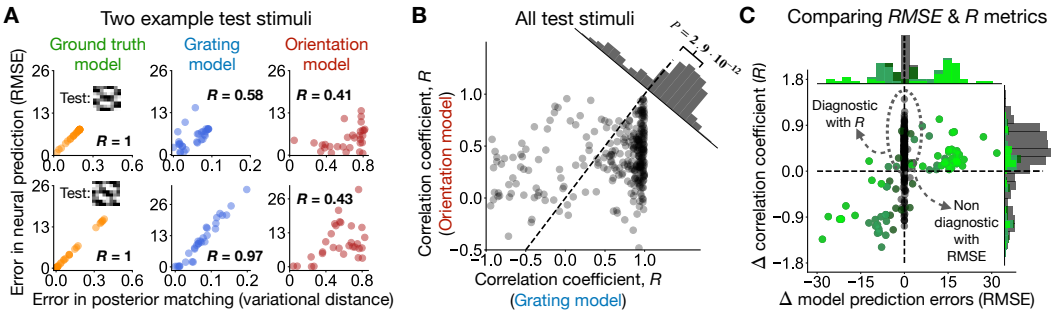


Figure 7: **Comparing the two models using the new, correlational diagnostic. A:** The correlations between the error in the neural prediction and the error in the posterior matching for the ground truth (orange), grating (blue), and orientation (red) models for two example test stimuli. **B:** Comparing the same correlations between the two models for all test stimuli. **C:** The difference of these correlations between the models plotted against the difference of the RMSEs between the models showing that most non-diagnostic stimuli using RMSE became diagnostic with the correlational metric.

(e.g., [8, 21, 22, 52, 53]), by fitting nuisance variables (the posterior weights in our case, the linear mapping from units/features to neurons in the case of deterministic encoding models) for training stimuli, and then testing how well they can predict neural responses to held-out stimuli. Unlike encoding models, our method does not require a specific mapping from model activations to neural responses as long as the mapping is part of the LDC family. In that sense, it also differs from prior studies of probabilistic models that have made specific assumptions about how probabilities are encoded in neural responses (e.g., PPCs, DDCs, or sampling).

Interestingly, our method can also be viewed as a generalization of [3], who compared average evoked activity in the visual cortex to spontaneous activity. Their prediction, derived from the model-based equality of average posterior being equal to the prior, is a special case of our posterior matching technique with all posterior weights being one, and base stimuli drawn from the distribution over natural images.

The main limitation of our method is the assumption of a Linear Distributional Code (LDC). While motivated by mathematical convenience, it enjoys empirical support both by studies that provide evidence for one of the various flavors of neural sampling, whether based on binary latents [3, 6, 42], or continuous latents [17, 20, 40], and those that compare LDC and non-LDC (e.g., PPCs) directly [47]. Most importantly, our method does not assume that the model we test is linear. The posteriors in the scientists' models can have arbitrarily nonlinear relationships to the input stimulus. It's only the encoding of the posterior in neural responses that we assume to be linear, akin to deterministic encoding models and RSA [10].

Although our illustration in the Results section only explicates how to test the model prediction for a single neuron, it is straightforward to apply it to populations of neurons in two complementary ways. First, one can apply our Method to all recorded neurons individually and compare population averages (or distributions) during the model comparison. Second, it is possible to apply our Method to joint population responses. All our derivations apply to the prediction of joint distributions, $p(r)$, where $r$ is a vector of responses, similar to marginals.

While we have only illustrated our method using the best-fit single latent from a high-dimensional model (i.e., the grating model), it is also straightforward to extend it to joint posteriors over multidimensional subspaces of $x$. In practice, the dimensionality of $x$ is primarily limited by one's ability to record neural responses to a large enough number of baseline stimuli to fill a multi-dimensional histogram. This number grows exponentially with the number of dimensions.

In this paper, we distinguished between the baseline and test stimuli for conceptual simplicity. However, in practice, there is no need for such a categorical distinction, and one can use a leave-one-out procedure where all stimuli except one are used for baselines, with the remaining one counting as a test. Choosing stimuli that yield posteriors that can be mixed to approximate the posteriors for a held-out test stimulus is a crucial step of our method. Implicitly, this is, required for any experiments: choosing the stimuli that can best test the model's predictions. E.g., to predict neural responses to the

motion of a particular object, we will need to show some set of moving stimuli, and we won't be able to test our predictions with stationary objects. One contribution of our work is to formalize this process and to take the guesswork out of which stimulus set to choose. Moreover, using our method, this can be done using the model's posteriors only, without collecting any neural data.

In future work, it may be possible to develop a more symmetrical method in which sets of weights are determined both within the model, and within the neural data, with the key comparison being between summary statistics applied to both sets of weights (similar to RSA).

Finally, the match of posteriors in our method will never be perfect, and the discrepancy may differ for different models using the same set of baseline stimuli. Therefore, in practice, it is essential to control for this discrepancy when comparing models. We therefore developed the correlational-based diagnostic that controls for the different errors in the posterior matching. Our new diagnostic based on weight perturbations only enjoys a firm analytical basis for binary latents, and for categorical latents under additional symmetry assumptions for the neural encoding. However, to what degree these assumptions can be justified empirically remains to be seen. In general, more work is required to compare the new diagnostic to classic model comparison metrics.

## Acknowledgments and Disclosure of Funding

We would like to thank Nikolaus Kriegeskorte and Richard D. Lange for their helpful comments on our manuscript. This work was supported by NIH/U19NS118246 (to GL, SS, and RH), NSF/CAREER IIS-2143440 (to RH), and NSF-1449828 (to SS).

## References

[1] J. Beck, A. Pouget, and K. A. Heller, "Complex inference in neural circuits with probabilistic population codes and topic models," in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25, Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/file/b7087c1f4f89e63af8d46f3b20271153-Paper.pdf.

[2] J. Beck, H. Ralf, P. Xaq, S. Cristina, and V. Eszter, "Competing theories of probabilistic computations in the brain," in *CCN 2020 Workshop GAC*, 2020. [Online]. Available: https://openreview.net/forum?id=uZMMO2obl50.

[3] P. Berkes, G. Orbán, M. Lengyel, and J. Fiser, "Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment," *Science*, vol. 331, no. 6013, pp. 83–87, 2011. DOI: 10.1126/science.1195870. eprint: https://www.science.org/doi/pdf/10.1126/science.1195870. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.1195870.

[4] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006, ISBN: 0387310738.

[5] J. Bornschein, M. Henniges, and J. Lücke, "Are v1 simple cells optimized for visual occlusions? a comparative study," *PLoS Computational Biology*, vol. 9, 2013.

[6] L. Buesing, J. Bill, B. Nessler, and W. Maass, "Neural dynamics as sampling: A model for stochastic computation in recurrent networks of spiking neurons," *PLOS Computational Biology*, vol. 7, no. 11, pp. 1–22, Nov. 2011. DOI: 10.1371/journal.pcbi.1002211. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1002211.

[7] A. Chattoraj, *Models of Approximate Inference in Vision*. University of Rochester, 2022.

[8] J. J. DiCarlo and D. D. Cox, "Untangling invariant object recognition," *Trends in Cognitive Sciences*, vol. 11, no. 8, pp. 333–341, 2007, ISSN: 1364-6613. DOI: 10.1016/j.tics.2007.06.010. [Online]. Available: https://doi.org/10.1016/j.tics.2007.06.010.

[9] J. Diedrichsen and N. Kriegeskorte, "Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis," *PLOS Computational Biology*, vol. 13, no. 4, pp. 1–33, Apr. 2017. DOI: 10.1371/journal.pcbi.1005508. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1005508.

[10] J. Diedrichsen and N. Kriegeskorte, "Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis," *PLOS Computational Biology*, vol. 13, no. 4, pp. 1–33, Apr. 2017. DOI: 10.1371/journal.pcbi.1005508. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1005508.

[11] M. O. Ernst and M. S. Banks, "Humans integrate visual and haptic information in a statistically optimal fashion," *Nature*, vol. 415, no. 6870, pp. 429–433, 2002, ISSN: 1476-4687. DOI: 10.1038/415429a. [Online]. Available: https://doi.org/10.1038/415429a.

[12] J. Fiser, P. Berkes, G. Orbán, and M. Lengyel, "Statistically optimal perception and learning: From behavior to neural representations," en, *Trends Cogn. Sci.*, vol. 14, no. 3, pp. 119–130, Mar. 2010.

[13] J. Fiser and G. Lengyel, "Statistical learning in vision," *Annual Review of Vision Science*, vol. 8, no. 1, pp. 265–290, 2022, PMID: 35727961. DOI: 10.1146/annurev-vision-100720-103343. eprint: https://doi.org/10.1146/annurev-vision-100720-103343. [Online]. Available: https://doi.org/10.1146/annurev-vision-100720-103343.

[14] J. Fiser and G. Lengyel, "A common probabilistic framework for perceptual and statistical learning," *Current Opinion in Neurobiology*, vol. 58, pp. 218–228, 2019, Computational Neuroscience, ISSN: 0959-4388. DOI: https://doi.org/10.1016/j.conb.2019.09.007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0959438819300352.

[15] T. Golan, P. C. Raju, and N. Kriegeskorte, "Controversial stimuli: Pitting neural networks against each other as models of human cognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 47, pp. 29 330–29 337, 2020. DOI: 10.1073/pnas.1912334117. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1912334117. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.1912334117.

[16] J. I. Gold and M. N. Shadlen, "The neural basis of decision making," *Annual Review of Neuroscience*, vol. 30, no. 1, pp. 535–574, 2007, PMID: 17600525. DOI: 10.1146/annurev.neuro.29.051605.113038. eprint: https://doi.org/10.1146/annurev.neuro.29.051605.113038. [Online]. Available: https://doi.org/10.1146/annurev.neuro.29.051605.113038.

[17] R. M. Haefner, P. Berkes, and J. Fiser, "Perceptual Decision-Making as probabilistic inference by neural sampling," *Neuron*, vol. 90, no. 3, pp. 649–660, 2016.

[18] J. B. Heald, M. Lengyel, and D. M. Wolpert, "Contextual inference underlies the learning of sensorimotor repertoires," *Nature*, vol. 600, no. 7889, pp. 489–493, 2021, ISSN: 1476-4687. DOI: 10.1038/s41586-021-04129-3. [Online]. Available: https://doi.org/10.1038/s41586-021-04129-3.

[19] M. Henniges, G. Puertas, J. Bornschein, J. Eggert, and J. Lücke, "Binary sparse coding," in *Proceedings of the 9th International Conference on Latent Variable Analysis and Signal Separation*, ser. LVA/ICA'10, St. Malo, France: Springer-Verlag, 2010, 450–457, ISBN: 364215994X.

[20] P. O. Hoyer and A. Hyvärinen, "Interpreting neural response variability as monte carlo sampling of the posterior," in *Proceedings of the 15th International Conference on Neural Information Processing Systems*, ser. NIPS'02, Cambridge, MA, USA: MIT Press, 2002, 293–300.

[21] A. G. Huth, W. A. de Heer, T. L. Griffiths, F. E. Theunissen, and J. L. Gallant, "Natural speech reveals the semantic maps that tile human cerebral cortex," *Nature*, vol. 532, no. 7600, pp. 453–458, 2016, ISSN: 1476-4687. DOI: 10.1038/nature17637. [Online]. Available: https://doi.org/10.1038/nature17637.

[22] K. N. Kay, T. Naselaris, R. J. Prenger, and J. L. Gallant, "Identifying natural images from human brain activity," *Nature*, vol. 452, no. 7185, pp. 352–355, 2008, ISSN: 1476-4687. DOI: 10.1038/nature06713. [Online]. Available: https://doi.org/10.1038/nature06713.

[23] D. Kersten, P. Mamassian, and A. Yuille, "Object perception as bayesian inference," *Annu. Rev. Psychol.*, vol. 55, no. 1, pp. 271–304, 2004.

[24] D. C. Knill and W. Richards, "Perception as bayesian inference," in Cambridge University Press, 1996, pp. v–vi.

[25] D. C. Knill and A. Pouget, "The bayesian brain: The role of uncertainty in neural coding and computation," en, *Trends Neurosci*, vol. 27, no. 12, pp. 712–719, Dec. 2004.

[26] K. P. Körding and D. M. Wolpert, "Bayesian integration in sensorimotor learning," *Nature*, vol. 427, no. 6971, pp. 244–247, 2004, ISSN: 1476-4687. DOI: 10.1038/nature02169. [Online]. Available: https://doi.org/10.1038/nature02169.

[27] N. Kriegeskorte, "Pattern-information analysis: From stimulus decoding to computational-model testing," *NeuroImage*, vol. 56, no. 2, pp. 411–421, 2011, Multivariate Decoding and Brain Reading, ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2011.01.061. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811911000978.

[28] N. Kriegeskorte *et al.*, "Matching categorical object representations in inferior temporal cortex of man and monkey," *Neuron*, vol. 60, no. 6, pp. 1126–1141, 2008, ISSN: 0896-6273. DOI: 10.1016/j.neuron.2008.10.043. [Online]. Available: https://doi.org/10.1016/j.neuron.2008.10.043.

[29] K. P. Körding, U. Beierholm, W. J. Ma, S. Quartz, J. B. Tenenbaum, and L. Shams, "Causal inference in multisensory perception," *PLOS ONE*, vol. 2, no. 9, pp. 1–10, Sep. 2007. DOI: 10.1371/journal.pone.0000943. [Online]. Available: https://doi.org/10.1371/journal.pone.0000943.

[30] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, no. 6266, pp. 1332–1338, 2015.

[31] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building machines that learn and think like people," en, *Behav. Brain Sci.*, vol. 40, e253, Jan. 2017.

[32] R. D. Lange, A. Chattoraj, J. M. Beck, J. L. Yates, and R. M. Haefner, "A confirmation bias in perceptual decision-making due to hierarchical approximate inference," *PLOS Computational Biology*, vol. 17, no. 11, pp. 1–30, Nov. 2021. DOI: 10.1371/journal.pcbi.1009517. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1009517.

[33] R. D. Lange and R. M. Haefner, "Task-induced neural covariability as a signature of approximate bayesian learning and inference," *PLOS Computational Biology*, vol. 18, no. 3, pp. 1–39, Mar. 2022. DOI: 10.1371/journal.pcbi.1009557. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1009557.

[34] R. D. Lange, S. Shivkumar, A. Chattoraj, and R. M. Haefner, "Bayesian encoding and decoding as distinct perspectives on neural coding," *bioRxiv*, 2022. DOI: 10.1101/2020.10.14.339770. eprint: https://www.biorxiv.org/content/early/2022/09/18/2020.10.14.339770.full.pdf. [Online]. Available: https://www.biorxiv.org/content/early/2022/09/18/2020.10.14.339770.

[35] T. S. Lee and D. Mumford, "Hierarchical bayesian inference in the visual cortex," *Journal of the Optical Society of America A: Optics and Image Science, and Vision*, vol. 20, no. 7, 1434 – 1448, 2003, Cited by: 897; All Open Access, Green Open Access. DOI: 10.1364/JOSAA.20.001434. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-0042565834&doi=10.1364%2fJOSAA.20.001434&partnerID=40&md5=9fcbb947dbfa64938d1a4315f39d1086.

[36] W. J. Ma, J. M. Beck, P. E. Latham, and A. Pouget, "Bayesian inference with probabilistic population codes," *Nature Neuroscience*, vol. 9, no. 11, pp. 1432–1438, 2006, ISSN: 1546-1726. DOI: 10.1038/nn1790. [Online]. Available: https://doi.org/10.1038/nn1790.

[37] T. M. Mitchell *et al.*, "Predicting human brain activity associated with the meanings of nouns," *Science*, vol. 320, no. 5880, pp. 1191–1195, 2008. DOI: 10.1126/science.1152876. eprint: https://www.science.org/doi/pdf/10.1126/science.1152876. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.1152876.

[38] K. Nomizu and T. Sasaki, *Affine Differential Geometry: Geometry of Affine Immersions*. Cambridge, UK: Cambridge University Press, 1994, ISBN: 978-0-521-06439-2.

[39] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1?" *Vision Research*, vol. 37, no. 23, pp. 3311–3325, 1997, ISSN: 0042-6989. DOI: https://doi.org/10.1016/S0042-6989(97)00169-7. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0042698997001697.

[40] G. Orbán, P. Berkes, J. Fiser, and M. Lengyel, "Neural variability and Sampling-Based probabilistic representations in the visual cortex," en, *Neuron*, vol. 92, no. 2, pp. 530–543, Oct. 2016.

[41] G. Orbán, J. Fiser, R. N. Aslin, and M. Lengyel, "Bayesian learning of visual chunks by human observers," en, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 7, pp. 2745–2750, Feb. 2008.

[42] D. Pecevski, L. Buesing, and W. Maass, "Probabilistic inference in general graphical models through sampling in stochastic networks of spiking neurons," *PLOS Computational Biology*, vol. 7, no. 12, pp. 1–25, Dec. 2011. DOI: 10.1371/journal.pcbi.1002294. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1002294.

[43] A. Pouget, J. M. Beck, W. J. Ma, and P. E. Latham, "Probabilistic brains: Knowns and unknowns," en, *Nat. Neurosci.*, vol. 16, no. 9, pp. 1170–1178, Sep. 2013.

[44] O. Schwartz and E. P. Simoncelli, "Natural signal statistics and sensory gain control," *Nature Neuroscience*, vol. 4, no. 8, pp. 819–825, Aug. 2001.

[45] C. I. Tajima, S. Tajima, K. Koida, H. Komatsu, K. Aihara, and H. Suzuki, "Population code dynamics in categorical perception," *Scientific Reports*, vol. 6, no. 1, p. 22 536, Mar. 2016.

[46] J. B. Tenenbaum, T. L. Griffiths, and C. Kemp, "Theory-based bayesian models of inductive learning and reasoning," en, *Trends Cogn. Sci.*, vol. 10, no. 7, pp. 309–318, Jul. 2006.

[47] B. B. Ujfalussy and G. Orbán, "Sampling motion trajectories during hippocampal theta sequences," *eLife*, vol. 11, A. Peyrache and L. L. Colgin, Eds., e74058, 2022, ISSN: 2050-084X. DOI: 10.7554/eLife.74058. [Online]. Available: https://doi.org/10.7554/eLife.74058.

[48] E. Vértes and M. Sahani, "Flexible and accurate inference and learning for deep generative models," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18, Montréal, Canada: Curran Associates Inc., 2018, 4170–4179.

[49] E. Vértes and M. Sahani, "A neurally plausible model learns successor representations in partially observable environments," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/dea184826614d3f4c608731389ed0c74-Paper.pdf.

[50] E. Y. Walker, R. J. Cotton, W. J. Ma, and A. S. Tolias, "A neural basis of probabilistic computation in visual cortex," *Nature Neuroscience*, vol. 23, no. 1, pp. 122–129, Jan. 2020.

[51] F. Xu and J. B. Tenenbaum, "Word learning as bayesian inference," *Psychological review*, vol. 114, no. 2, 245–272, 2007. DOI: 10.1037/0033-295X.114.2.245.

[52] D. L. K. Yamins and J. J. DiCarlo, "Using goal-driven deep learning models to understand sensory cortex," *Nature Neuroscience*, vol. 19, no. 3, pp. 356–365, 2016, ISSN: 1546-1726. DOI: 10.1038/nn.4244. [Online]. Available: https://doi.org/10.1038/nn.4244.

[53] D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo, "Performance-optimized hierarchical models predict neural responses in higher visual cortex," *Proceedings of the National Academy of Sciences*, vol. 111, no. 23, pp. 8619–8624, 2014. DOI: 10.1073/pnas.1403112111, eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1403112111. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.1403112111.