A Appendix

A.1 The relationship between the error in the posterior matching and the error in the neural prediction for binary latents

The neural prediction error of a scientist model will reflect the extent to which the model can capture the brain's computations in the context of the stimuli only if, in the posterior matching step (Fig.2 A-C), we can perfectly match the test posterior as a mixture of the baseline posteriors. 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. For the perfect model of the brain (assuming an Linear Distributional Code, (LDC) [33]), we can write down the relationship between the error in the posterior matching and the error between the predicted and the true neural responses (Fig.5A). First, we assess the error in the posterior matching with the variational distance, V, between the approximated and true test posterior:

$$V = \frac{1}{2}\int dx \left|p(x|o_x^{\text{test,perfect}}) - p(x|o_x^{\text{test,jittered}})\right| \tag{9}$$

where $p(x|o_x^{\text{test,perfect}})$ and $p(x|o_x^{\text{test,jittered}})$ denote the approximated test posteriors using "perfect" weights, eliciting no error in the posterior matching step, and with jittered weights, eliciting an error in the posterior matching step, respectively.

If the scientist's model and the brain's model are the same, there is no mismatch between the two models, and the variational distance can be defined in terms of the difference in the posteriors for the brain's true model, $p(z|o_z)$:

$$V = \frac{1}{2}\int dx \left|p(z|o_z^{\text{test,perfect}}) - p(z|o_z^{\text{test,jittered}})\right| \tag{10}$$

In the case of binary latents in the brain's true model, the variational distance, V, can be expanded into a sum of two terms:

$$V = \frac{1}{2}\left|p(z=1|o_z^{\text{test,perfect}}) - p(z=1|o_z^{\text{test,jittered}})\right| + \frac{1}{2}\left|p(z=0|o_z^{\text{test,perfect}}) - p(z=0|o_z^{\text{test,jittered}})\right| \tag{11}$$

We use the fact that for binary latents $p(z=0|o) = 1 - p(z=0|o)$, thus

$$\left|p(z=0|o_z^{\text{test,perfect}}) - p(z=0|o_z^{\text{test,jittered}})\right| = \left|[1 - p(z=1|o_z^{\text{test,perfect}})] - [1 - p(z=1|o_z^{\text{test,jittered}})]\right| \tag{12}$$

which can be used to get the variational distance as

$$V = \left|p(z=1|o_z^{\text{test,perfect}}) - p(z=1|o_z^{\text{test,jittered}})\right| \tag{13}$$

Next, the posterior distribution over the brain's latents are mapped to the posterior over neural responses through the encoding function, $\mathcal{R}_{p(z)\to p(r)}$:

$$p(r|o_z^{\text{test,perfect}}) = \mathcal{R}_{p(z)\to p(r)}\left[p\left(z=1|o_z^{\text{test,perfect}}\right)\right] \tag{14}$$

For LDC, by definition, the following holds:

$$\mathcal{R}_{p(z)\to p(r)}\left[\sum_i \alpha_i p\left(z=1|o_z^{\text{test,perfect,i}}\right)\right] = \sum_i \alpha_i \mathcal{R}_{p(z)\to p(r)}\left[p\left(z=1|o_z^{\text{test,perfect,i}}\right)\right] \tag{15}$$

These also satisfy the necessary and sufficiency condition of an affine transformation [38], and therefore the neural responses can be related to the posterior over latents as follows:

$$\mathbf{p}(r|o_z^{\text{test,perfect}}) = \mathbf{A}p\left(z=1|o_z^{\text{test,perfect}}\right) + \mathbf{b} \tag{16}$$

where r are the spike counts in a counting window that lie in the set $[0,...,N]$ where N is the maximum spike count for the neuron. Let μ_r^{perfect} and μ_r^{jittered} be the predicted expected spike counts for the neural responses using the perfect and jittered weights from the posterior matching step, respectively. If $\mathbf{R}$ denotes the vector of spike counts from $[1, 2, \ldots, N]$, the expected spike counts become:

$$\mu_r^{\text{perfect}} = \mathbf{R}^\top \mathbf{p}(r|o_z^{\text{test,perfect}}) \tag{17}$$

$$\mu_r^{\text{jittered}} = \mathbf{R}^\top \mathbf{p}(r|o_z^{\text{test,jittered}}) \tag{18}$$

Substituting the affine form of the posterior

$$\mu_r^{\text{perfect}} = \mathbf{R}^\top \mathbf{A}p(z=1|o_z^{\text{test,perfect}}) + \mathbf{R}^\top \mathbf{b} \tag{19}$$

$$\mu_r^{\text{perfect}} - \mathbf{R}^\top \mathbf{b} = (\mathbf{R}^\top \mathbf{A} - \mathbf{1}^\top \mathbf{A})p(z=1|o_z^{\text{test,perfect}}) \tag{20}$$

For binary latents, $\mathbf{R}^\top \mathbf{A} - \mathbf{1}^\top \mathbf{A}$ is a scalar which we denote as d that we can substitute to relate the posterior over the latent to the expected spike counts:

$$(\frac{\mu_r^{\text{perfect}}}{d} - \frac{\mathbf{R}^\top \mathbf{b} - r_0(1 - \mathbf{1}^\top \mathbf{b})}{d}) = p(z=1|o_z^{\text{test,perfect}}) \tag{21}$$

$$(\frac{\mu_r^{\text{jittered}}}{d} - \frac{\mathbf{R}^\top \mathbf{b} - r_0(1 - \mathbf{1}^\top \mathbf{b})}{d}) = p(z=1|o_z^{\text{test,jittered}}) \tag{22}$$

Substituting these into the definition of the variational distance we get

$$V = \frac{1}{|d|}|\mu_r^{\text{perfect}} - \mu_r^{\text{jittered}}| \tag{23}$$

that could be rearranged to obtain

$$|\mu_r^{\text{perfect}} - \mu_r^{\text{jittered}}| = |d|V \tag{24}$$

or simply

$$|\mu_r^{\text{perfect}} - \mu_r^{\text{jittered}}| \propto V \tag{25}$$

A.2 The relationship between the error in the posterior matching and the error in the neural prediction for categorical latents

If the latents in the brain's model are categorical, the probabilities over all latent states can be parameterized as a column vector $\mathbf{p}(z|o_z)$ of size $n_{\text{latent}} - 1$ where n_{latent} is the number of states of the categorical latents. The probability of the remaining state ($z = 0$ without loss of generality) is $p(z=0|o_z) = 1 - \mathbf{1}^\top \mathbf{p}(z|o_z)$. Thus, the variational distance can be written as follows:

$$V = \frac{1}{2}\mathbf{1}^\top \left|\mathbf{p}(z|o_z^{\text{test,perfect}}) - \mathbf{p}(z|o_z^{\text{test,jittered}})\right| + \frac{1}{2}\left|p(z=0|o_z^{\text{test,perfect}}) - p(z=0|o_z^{\text{test,jittered}})\right| \tag{26}$$

Substituting $p(z=0|o_z) = 1 - \mathbf{1}^\top \mathbf{p}(z|o_z^{\text{test,perfect}})$

$$\left|p(z=0|o_z^{\text{test,perfect}}) - p(z=0|o_z^{\text{test,jittered}})\right| = \left|[1-\mathbf{1}^\top \mathbf{p}(z|o_z^{\text{test,perfect}})] - [1-\mathbf{1}^\top \mathbf{p}(z|o_z^{\text{test,jittered}})]\right| \tag{27}$$

which can be used to simplify the expression for the variational distance as follows:

$$V = \frac{1}{2}\mathbf{1}^\top \left|\mathbf{p}(z|o_z^{\text{test,perfect}}) - \mathbf{p}(z|o_z^{\text{test,jittered}})\right| + \frac{1}{2}\left|\mathbf{1}^\top \mathbf{p}(z|o_z^{\text{test,perfect}}) - \mathbf{1}^\top \mathbf{p}(z|o_z^{\text{test,jittered}})\right| \tag{28}$$

The approximated test posteriors can be written in the following ways:

$$\mathbf{p}(z|o_z^{\text{test,perfect}}) = \sum_i w_i^{\text{perfect}} \mathbf{p}(z|o_z^{\text{test,baseline,i}}) \tag{29}$$

$$\mathbf{p}(z|o_z^{\text{test,jittered}}) = \sum_i w_i^{\text{jittered}} \mathbf{p}(z|o_z^{\text{test,baseline,i}}) \tag{30}$$

We substitute these terms in eqs. 29 & 30 into in eq. 28:

$$V = \frac{1}{2}\mathbf{1}^\top \left|\sum_i (w_i^{\text{perfect}} - w_i^{\text{jittered}})\mathbf{p}(z|o_z^{\text{test,baseline,i}})\right| + \frac{1}{2}\left|\mathbf{1}^\top \left[\sum_i (w_i^{\text{perfect}} - w_i^{\text{jittered}})\mathbf{p}(z|o_z^{\text{test,baseline,i}})\right]\right| \tag{31}$$

and if we assume that $w_i^{\text{jittered}} \leq w_i^{\text{perfect}}$, then all terms inside the absolute value are positive thereby simplifying V to get the following:

$$V = \mathbf{1}^\top \left[\sum_i (w_i^{\text{perfect}} - w_i^{\text{jittered}})\mathbf{p}(z|o_z^{\text{test,baseline,i}})\right] \tag{32}$$

As in the binary case, the posterior distribution over spike counts and expected spike counts are

$$\mathbf{p}(r|o_z^{\text{test,perfect}}) = \mathbf{A}\mathbf{p}(z|o_z^{\text{test,perfect}}) + \mathbf{b} \tag{33}$$

$$\mu_r^{\text{perfect}} = \mathbf{R}^\top \mathbf{p}(r|o_z^{\text{test,perfect}}) \tag{34}$$

$$\mu_r^{\text{jittered}} = \mathbf{R}^\top \mathbf{p}(r|o_z^{\text{test,jittered}}) \tag{35}$$

Substituting the affine forms of the posteriors

$$\mu_r^{\text{perfect}} = \mathbf{R}^\top \mathbf{A}\mathbf{p}(z|o_z^{\text{test,perfect}}) + \mathbf{R}^\top \mathbf{b} \tag{36}$$

We can write the absolute difference in firing rates as

$$|\mu_r^{\text{perfect}} - \mu_r^{\text{jittered}}| = |\mathbf{R}^\top \mathbf{A}\mathbf{p}(z|o_z^{\text{test,perfect}}) + \mathbf{R}^\top \mathbf{b} - \mathbf{R}^\top \mathbf{A}\mathbf{p}(z|o_z^{\text{test,jittered}}) - \mathbf{R}^\top \mathbf{b}| \tag{37}$$

which can be simplified to

$$\left|\mu_r^{\text{perfect}} - \mu_r^{\text{jittered}}\right| = \left|\mathbf{R}^\top \mathbf{A}\left[\mathbf{p}(z|o_z^{\text{test,perfect}}) - \mathbf{p}(z|o_z^{\text{test,jittered}})\right]\right| \tag{38}$$

Since each column of $\mathbf{A}$ quantifies the influence of the corresponding latent state on neural responses, by assuming that each categorical state can be equivalently encoded in the spike counts, we have $\mathbf{R}^\top \mathbf{A} = d\mathbf{1}^\top$, that can be substituted into the previous equation to get

$$\left|\mu_r^{\text{perfect}} - \mu_r^{\text{jittered}}\right| = |d|\left|\mathbf{1}^\top \left[\mathbf{p}(z|o_z^{\text{test,perfect}}) - \mathbf{p}(z|o_z^{\text{test,jittered}})\right]\right| \tag{39}$$

For the case where $w_i^{\text{jittered}} \leq w_i^{\text{perfect}}$,

we can substitute

$$V = \left|\mathbf{1}^\top \left[\mathbf{p}(z|o_z^{\text{test,perfect}}) - \mathbf{p}(z|o_z^{\text{test,jittered}})\right]\right| \tag{40}$$

to get

$$|\mu_r^{\text{perfect}} - \mu_r^{\text{jittered}}| = |d|V \tag{41}$$

It is important to note that under this parameterization, the jittered weights need not sum to 1, thereby resulting in the jittered posterior not summing/integrating to 1. While this does not affect any of our derivations, for correctness, the variational distance (defined only for proper probability distributions) should be replaced by the integrated absolute value of the difference in density values when the weights don't add up to 1.

To provide further support for the proportional relationship between the posterior matching (V) and the neural prediction ($|\mu_r^{\text{perfect}} - \mu_r^{\text{jittered}}|$) errors in the case of categorical latent variables we tested it in a numerical simulation. We performed the weight perturbation analysis (Results 2.5) with a simulation that used the same Gaussian mixture ground truth model (App. A.3.2) except with categorical latents. In eq. 42, z became a vector of n categorical variables with three categories. Then, we used categorical Gibbs sampling for inferring z and converted the posterior into expected spike counts using expectation (i.e., $\mathbf{A}$ is the identity matrix and thus $d = 1$ in the above equations). We found that the correlations between the posterior matching (V) and the neural prediction ($|\mu_r^{\text{perfect}} - \mu_r^{\text{jittered}}|$) errors for the ground truth model were indeed perfect for all neurons and stimuli where there was sufficient variability in the errors to compute the correlation.

A.3 Illustration of the method

To demonstrate our method, we apply it to predict simulated neural responses in area V1, where we assume that the true internal model of the brain is a binary sparse coding image model [19]. Such a binary model was shown to explain the properties of V1 receptive fields as successfully as the original continuous model of [39] and was found to be biologically plausible [6]. Furthermore, we assume neural sampling [12, 20, 35] where samples from the binary latents can directly represent spikes. Consequently, the encoding function becomes a direct, one-to-one mapping satisfying the LCD [33] assumption of our method. The goal of the illustration is to predict the firing of a single, simulated neuron in response to superimposed grating stimuli using two competing scientist's models (Fig. 4 C).

A.3.1 The baseline and test stimuli

Just like in an experiment, we vary one parameter of the grating stimulus, the orientation, and keep the other parameters, such as frequency and phase, at a fixed value that elicits the maximum response from the neuron whose firing we want to predict. To highlight the main difference between the two competing scientist's models, we use superimposed grating stimuli, i.e., plaids (Fig. 4 A). We used 20 baseline grating images (Gabor filters, 8-by-8 pixels, with fixed parameters) with varying orientations that tiled the orientations space ($0 - \pi$) uniformly (Fig.S1 Top). The test stimuli were superimposed grating images (i.e., plaids) generated by adding the pixels of two baseline grating images with different orientations element-wise (Fig.S1 bottom). All unique combinations of the two superimposed gratings allowing to combine gratings with the same orientation resulted in 210 test stimuli in total.

A.3.2 The ground truth model of the brain

We generated the ground truth neural activity in response to the baseline and test stimuli using a binary sparse coding internal model of the brain [19]. For this illustration, we assume that this model is the "true" model of the brain in which the sensory observations are generated as follows (Fig. 4 B):

$$
\begin{aligned}
\mathbf{p}\left(\mathbf{z} \mid \pi\right) &= \prod_{n=1}^{N} \pi^{z_n}\left(1-\pi\right)^{1-z_n} \\
p\left(o_z \mid \mathbf{z}\right) &= \mathcal{N}\left(o_z; \mathbf{A}\,\mathbf{z}, \sigma_{\text{brain}}^2 \mathbf{1}\right)
\end{aligned}
\tag{42}
$$

where z_n represents the nth binary latent variable that activates the corresponding projective field, $\mathbf{A}_{:,n}$, and N denotes the number of binary latent variables. The 128 projective fields were learned on natural images in [7]. Each of these projective fields represents a column in the matrix, $\mathbf{A}$ (Fig.2 Bottom & Fig. 3 B). The prior distribution over $\mathbf{z}$ is sparse for small π, and the sensory observations, o_z, are generated by drawing from a Gaussian distribution around the linear combination of the projective fields given $\mathbf{z}$ with variance σ_{brain}^2.

We assumed that the brain infers the posterior probability over its latent variables, given its sensory observations. However, in the Bayesian framework, the sensory observations are also random latent variables, and the only observed variable, in this example, is the image of the grating stimulus, I. Therefore, to compute the posterior probability over the latent variables given the stimulus, the uncertainty in the sensory observations needs to be marginalized out:

$$
\begin{aligned}
p\left(\mathbf{z} \mid I\right) &= \int p\left(\mathbf{z} \mid o_z\right) p\left(o_z \mid I\right)\, \mathrm{d}o_z \\
&= \int \frac{p\left(o_z \mid \mathbf{z}\right) p\left(\mathbf{z}\right)}{p\left(o_z\right)} p\left(o_z \mid I\right)\, \mathrm{d}o_z \\
&= \int \frac{\mathcal{N}(o_z; \mathbf{A}\,\mathbf{z}, \sigma_{\text{brain}}^2 \mathbf{1}) \prod_{n=1}^{N} \pi^{z_n}(1-\pi)^{1-z_n}}{\int \mathcal{N}(o_z; \mathbf{A}\,\mathbf{z}, \sigma_{\text{brain}}^2 \mathbf{1}) \prod_{n=1}^{N} \pi^{z_n}(1-\pi)^{1-z_n}\, \mathrm{d}z} \mathcal{N}(o_z; I, \sigma_{\text{exp}}^2 \mathbf{1})\, \mathrm{d}o_z
\end{aligned}
\tag{43}
$$

In eq. 43, the posterior, $p\left(\mathbf{z} \mid o_z\right)$, will be approximated using Gibbs sampling.

$p\left(\mathbf{z} \mid I\right)$ will be approximated with Monte Carlo integration by drawing samples from $p\left(o_z \mid I\right)$:

$$
p\left(\mathbf{z} \mid I\right) \approx \frac{1}{M} \sum_{i \sim p(o_z \mid I)} p\left(\mathbf{z} \mid o_z = i\right) \tag{44}
$$

where M is the number of samples drawn from $p\left(o_z \mid I\right)$. We further assume that the probability of the observations given the stimulus can be represented by a multivariate Gaussian distribution centered at the pixel values of the grating image, I:

$$
p\left(o_z \mid I\right) = \mathcal{N}\left(o_z; I, \sigma_{\text{exp}}^2 \mathbf{1}\right) \tag{45}
$$

where $\sigma_{\text{exp}}^2 \mathbf{1}$ represents the variance of the Gaussian pixel noise added on the grating and $\mathbf{1}$ is the identity matrix.

Finally, we generate the neural activity from this model by drawing samples from the binary latent variable, z, which, being binary, can directly be considered as spikes:

$$
\mathbf{r} \mid I \sim p\left(\mathbf{z} \mid I\right) \tag{46}
$$

Note that this sampling-based encoding belongs to the LDC class [33] since $p\left(\mathbf{r} \mid I\right) = p\left(\mathbf{z} \mid I\right)$, thus the encoding is the identity function.

A.3.3 The two competing scientist's models

In the illustration, we compare two scientist's models. The first model assumes that the brain represents orientations in V1 (Fig. 4 C, left), and the other one hypothesizes that oriented gratings are represented in the brain (Fig. 4 C, right).

The orientation model assumes that the brain represents orientation, and the images can be generated from orientations using a template function that transforms an orientation into a grating image:

$$
p\left(o_x \mid x_\theta\right) = \mathcal{N}\left(o_x; T\left(x_\theta\right), \sigma_{\text{model}}^2 \mathbf{1}\right) \tag{47}
$$

where x_θ denotes the latent variable in the scientist's model representing the orientation of the grating stimulus, and $\sigma_{\text{model}}^2 \mathbf{1}$ captures the uncertainty in the sensory processing. $T(x_\theta)$ represents the template function, transforming the scalar quantity of orientation, x_θ, into an observation, which is an image of a grating, plus random Gaussian noise. We further assume uniform prior over the orientations, x_θ.

In the grating model, the brain represents grating templates, $\mathbf{B}$, with different orientations $(0°, 45°, 90°, 135°)$ and phases $(0°, 60°, 120°, 180°, 240°, 300°)$, and the observations are drawn from a Gaussian distribution around the linear combination of the grating templates $(\mathbf{B})$ weighted by their activation, $\mathbf{x_g}$:

$$p\left(o_x \mid \mathbf{x_g}\right) = \mathcal{N}\left(o_x; \mathbf{B}\,\mathbf{x_g}, \sigma^2_{\text{model}}\mathbf{1}\right) \tag{48}$$

where σ^2_{model} represents, again, the noise in the sensory processing. We used 24 gratings with different orientations and phases from the baseline stimuli set as grating templates in B (Fig.S2 top, and Fig. 4B, right). We further assumed a Gaussian prior over the latent variables, $\mathbf{x_g}$, centered at zero with variance equal to $\mathbf{C}$.

The inference in both scientist's models can be computed in the same way as in the case of the brain's model in eq. 43 by marginalizing out the sensory observations:

$$\begin{aligned} p\left(x \mid I\right) &= \int p\left(x \mid o_x\right) p\left(o_x \mid I\right)\, \mathrm{d}o_x \\ &= \int \frac{p\left(o_x \mid x\right) p\left(x\right)}{p\left(o_x\right)} p\left(o_x \mid I\right)\, \mathrm{d}o_x \end{aligned} \tag{49}$$

where x represents the represented angle, x_θ, in the orientation model and the activation weights of the grating features, $\mathbf{x_g}$, in the grating model.

There is an analytical solution for the posterior in eq. 49 in the case of the grating model:

$$\begin{aligned} p(\mathbf{x_g} \mid I) &= \int p(\mathbf{x_g} \mid o_x) p(o_x \mid I)\, \mathrm{d}o_x \\ &= \int \frac{p(o_x \mid x_g) p(\mathbf{x_g})}{p(o_x)} p(o_x \mid I)\, \mathrm{d}o_x \\ &= \int \mathcal{N}(\mathbf{x_g}; \mathbf{m}, \mathbf{v})\, \mathcal{N}(o_x; I, \sigma^2_{\text{exp}}\mathbf{1})\, \mathrm{d}o_x \end{aligned} \tag{50}$$

where

$$\begin{aligned} \mathbf{v} &= \left(\mathbf{C}^{-1} + \frac{1}{\sigma^2_{\text{model}}}\,\mathbf{B}^T\,\mathbf{B}\right)^{-1} \\ \mathbf{m} &= \frac{1}{\sigma^2_{\text{model}}}\,\mathbf{v}\,\mathbf{B}^T\, o_x \end{aligned} \tag{51}$$

$\mathbf{C}^{-1}$ represents the covariance of the prior distribution over, $\mathbf{x_g}$ (note that we assumed that the prior distribution is centered at zero). The derivation of $\mathbf{v}$ and $\mathbf{m}$ can be found in [4].

Next, we transform $p(\mathbf{x_g} \mid o_x)$ with the pseudo inverse matrix $\mathbf{H}$ to get a normal distribution in o_x:

$$p(\mathbf{x_g} \mid o_x) = \mathcal{N}(\mathbf{H}\mathbf{x_g}; \mathbf{H}\mathbf{m}, \mathbf{H}^T\mathbf{v}\mathbf{H}) \tag{52}$$

Since $\mathbf{Hm} = o_x$ the posterior, $p(x_g \mid I)$, can be written as:

$$\begin{aligned} p(\mathbf{x_g} \mid I) &= \int \mathcal{N}(o_x; \mathbf{H}\mathbf{x_g}, \mathbf{H}^T\mathbf{v}\mathbf{H})\, \mathcal{N}(o_x; \mathbf{I}, \sigma^2_{\text{exp}}\mathbf{1})\, \mathrm{d}o_x \\ &= \mathcal{N}\left(\mathbf{J}\,\mathbf{I}; \mathbf{x_g}, \mathbf{v} + \mathbf{J}^T\sigma^2_{\text{exp}}\mathbf{J}\right) \end{aligned} \tag{53}$$

where

$$\begin{aligned} \mathbf{H} &= (\frac{1}{\sigma^2_{\text{model}}}\,\mathbf{v}\,\mathbf{B}^T)^+ \\ \mathbf{J} = \mathbf{H}^+ &= \frac{1}{\sigma^2_{\text{model}}}\,\mathbf{v}\,\mathbf{B}^T \end{aligned} \tag{54}$$

where $+$, T, and -1 denote the pseudo inverse (we assume it exists), the transpose, and the inverse, respectively.

In the case of the orientation model, there is no analytical solution for the posterior in eq. 49, thus similar to eq. 44, we will use Monte Carlo approximation:

$$p\left(x_\theta \mid I\right) \approx \frac{1}{M} \sum_{i \sim p(o_x \mid I)} p\left(x_\theta \mid o_x = i\right) \tag{55}$$

where M represents the number of samples. The posterior, $p\left(x_\theta \mid o_x = i\right)$, can be computed by applying the Bayes' rule and using the generative equation in eq. 47 with a uniform prior over x_θ.

A.3.4 The simulated neural data and the neural predictions

The simulated neural data consisted of 128 simulated firing rates (one for each neuron) for each baseline and test stimulus. Only 17/128 neurons showed unimodal tuning to orientation (see the tuning curves in response to the baseline stimuli for a few example neurons in Fig.S4). The tuning curves for the test stimuli become 2-dimensional, showing how the neural response of a neuron changes as a function of the orientations of the two gratings in the plaids (see the tuning curves in response to the test stimuli for a few example neurons in Fig.S5A).

We generated neural predictions from the ground truth (for validation), the grating, and the orientation models using the two steps of our method (see Results 2.3). Since the orientation model only contains a single latent variable, it also provides a single prediction per test stimulus per neuron. The grating model has 24 latent variables. To predict single-neuron activity with the grating model, we assumed that one latent variable describes one neuron. Therefore, we generated 24 neural predictions for each test stimulus and neuron using the marginal posterior distribution over each latent variable. Thus, the two steps for generating neural predictions from the grating model become:

For each neuron, n

- Step 1: Posterior matching for each latent variable, j: determine $w_{i,j,n}$ such that $p\left(x_j \mid o_x^{\text{test}}\right) \approx \sum_i w_{i,j,n}\, p\left(x_j \mid o_x^{\text{base}_i}\right)$
- Step 2: Predict response to test stimulus as the weighted average of baseline activities for each latent variable, j: $p\left(r_{j,n}^{\text{test}}\right) \approx \sum_i w_{i,j,n}\, p\left(r_n^{\text{base}_i}\right)$

where $p\left(x_j \mid o_x^{\text{test}}\right)$ and $p\left(r_{j,n}^{\text{base}_i}\right)$ denote the marginal posteriors over the scientist jth latent variable, x_j, when the test and the baseline stimuli are observed, respectively. $p\left(r_n^{\text{base}_i}\right)$ represents the probability of the neural activity in response to the baseline stimuli for the nth neuron, while $p\left(r_{j,n}^{\text{test}}\right)$ represents the predicted neural activity from the grating model for the nth neuron using the jth latent variable. Finally, $w_{i,j,n}$ is the fitted mixture weight for the ith baseline stimulus, for the jth latent variable, and for the nth neuron (in step 1) which is then used as an extrapolation weight (in step 2). In step 1, during the posterior matching step, the marginal posteriors were evaluated on a fine grid, and the mixture weights were fitted using linear regression, constraining the weights to be positive. However, for large dimensional joint posteriors, one needs another method for fitting the mixture weights in step 1 (e.g., using an expectation-maximization algorithm for Gaussian mixture distributions).

In the case of the ground truth model, we generated neural predictions following the method we described for the grating model. However, we only used the marginal posterior of the latent (i.e., neuron) corresponding to the neuron whose activity we want to predict. We could do that because, for the ground truth model, we know which latent variable (neuron) corresponds to which simulated activity. We also needed to assume that the marginal posterior of a binary latent in the ground truth model represents the firing rate. We used a Poisson distribution to convert the firing rates into posterior distributions, $p\left(x_j \mid o_x^{\text{test}}\right)$ and $p\left(r_{j,n}^{\text{base}_i}\right)$, that we used in the posterior matching step.

A.3.5 The validation of the method

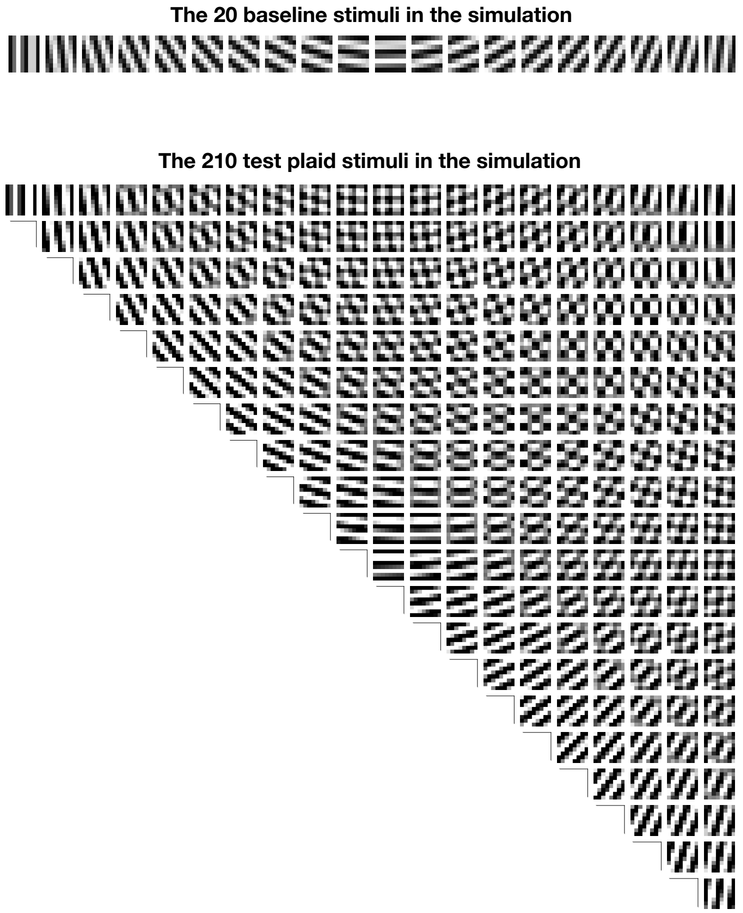
We used the Wilcoxon signed-rank test to compare the neural predictions of the ground truth and the orientation models for each neuron across all test stimuli. We computed the RMSEs for the fits across the models and found that the RMSEs of the ground truth model were significantly (*P*s < 0.05) lower than the RMSEs of the orientation model for all but eight neurons. For these eight neurons, the Wilcoxon signed-rank test was not significant (*P*s > 0.05). In the same way, next, we compared the RMSEs of the ground truth's and the grating model's neural predictions. We found that the RMSEs of the ground truth model were significantly (*P*s < 0.05) lower than the RMSEs of the grating model for all but nine neurons. For eight of these nine neurons, the Wilcoxon signed-rank test was not significant (*P*s > 0.09). In the case of that one neuron (#52), for which the grating model provided a significantly better neural prediction, the RMSE of the quantitatively better prediction of the grating model was still high (RMSE $= 8.49$) and similar to the RMSE of the ground truth model (RMSE $= 8.76$). When we look at the neurons for which we could not show that the ground truth model is better than the other two models, we can see that the neural activity in response to the baseline stimuli is much lower (essentially zero) than the neural response to the test stimuli. Therefore, in the ground truth model, we can't match the marginal posteriors for the test stimuli with the marginal posterior for the baseline stimuli because all the posteriors for the baselines peak far away (around zero firing rates) from the peaks of the test posteriors. These results show that, except for a few neurons, our method could identify that the ground truth model was indeed a better model for the simulated data than the orientation or the grating models. This demonstrates that our method can compare Bayesian models using neural data if our assumption about the encoding holds.

A.3.6 The evaluation of the new diagnostic

In the weight perturbation analysis, we used the Wilcoxon signed-rank test to compare the neural predictions of the ground truth and the orientation models using the new, correlation-based error metric for each neuron across all test stimuli. We found that the $1 - R$ errors of the ground truth model were significantly (*P*s < 0.05) lower than the $1 - R$ errors of the orientation model for all but 36 neurons. Please note that *R* is the correlation between the neural prediction and the posterior matching errors (see Results 2.5). We found exactly the same result when we compared the $1 - R$ errors of the ground truth and the grating models. 35 out of these 36 neurons were not activated by the baseline stimuli at all, therefore, computing the $1 - R$ metric was not possible. Consequently, the statistical test could not be computed. For one neuron (#69), the grating model had a significantly lower $1 - R$ error than the ground truth model. The grating model indeed provided a good prediction for this neuron (RMSE$= 2.01$). However, the difference in the $1 - R$ errors between the ground truth and the grating models was small ($1 - R = 0.45$ for the ground truth while $1 - R = 0.4$ for the grating models). The RMSE for this neuron was lower for the ground truth model than for the grating model but again, the difference was small (RMSE$= 1.48$ for the ground truth while RMSE$= 2.01$ for the grating models).

We used the Wilcoxon signed-rank test to compare the effect sizes using the $1 - R$ to the effect sizes using the RMSEs in the comparisons. Using the $1 - R$ metric, the effect sizes and the statistical power increased when we compared the ground truth model to the other two models (Wilcoxon signed-rank test for comparing the effect sizes in the comparison between the ground truth and the grating models for all neurons, $T = 517$, $P = 5 \cdot 10^{-13}$, and the same statistical test in the comparison between the ground truth and the orientation models, $T = 309$, $P = 1 \cdot 10^{-15}$). Interestingly, when we used the new $1 - R$ metric to compare the grating and the orientation models, the effect sizes shifted toward supporting the grating model compared to when we used the RMSE for the comparison (Wilcoxon signed-rank test for comparing the effect sizes of the RMSEs and the new metric in the comparison between the grating and orientation models for all neurons, $T = 3182$, $P = 0.024$).

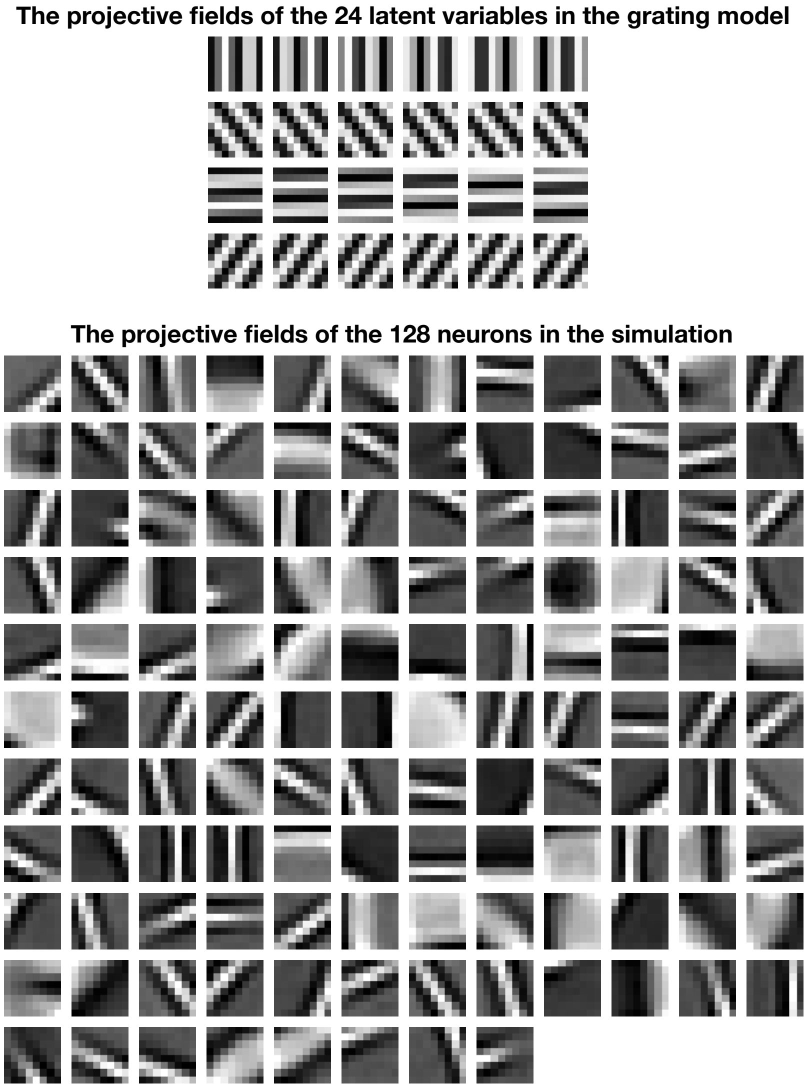
A.4 Supplementary figures



The 20 baseline stimuli in the simulation

The 210 test plaid stimuli in the simulation

Figure S1: **The baseline and test stimuli. Top:** Baselines: 20 gratings with orientations tiling the 0-180 deg. space. **Bottom:** Tests: 210 plaids generated from the combinations of the baseline gratings allowing to combine gratings with the same orientation (see the plaids along the diagonal).



The projective fields of the 24 latent variables in the grating model

The projective fields of the 128 neurons in the simulation

Figure S2: **The projective fields in the simulation. Top:** The projective fields used in the grating model. The four different orientations $(0°, 45°, 90°, 135°)$ are shown in the rows, while in the columns, the six different phases $(0°, 60°, 120°, 180°, 240°, 300°)$ can be observed. **Bottom:** The 128 projective fields in the binary sparse coding V1 model.

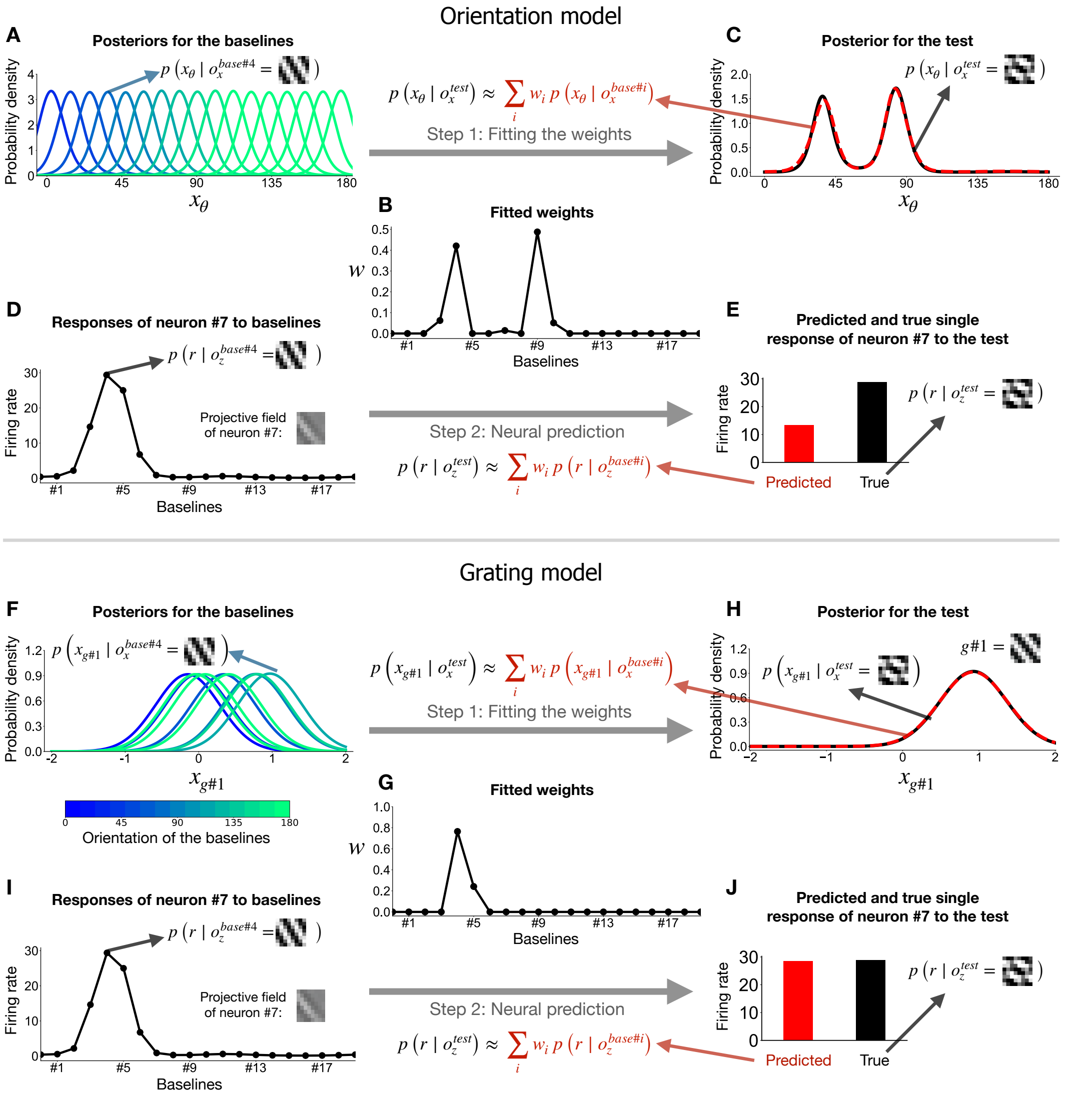


Figure S3: **Applying our method in the simulation for a single test plaid stimulus. A-C:** Weight fitting in the orientation model for a single test stimulus. **A:** Posteriors (colored lines) for the baseline stimuli. x_θ and $o_x^{\text{base}\#4}$ denote the latent variable (which is orientation) and the observation for baseline stimulus #4 in the orientation model, respectively. **B:** The fitted weights for matching the posterior to the test as a mixture of the baseline posteriors. w and o_x^{test} denote the weights and the observation for the test stimulus in the orientation model, respectively. **C:** The approximated (red, dashed line) and the true, computed (solid, black line) posteriors for the test stimulus. **D-E:** Neural prediction in the orientation model for a single synthetic neuron (the neuron's projective field is shown in D). **D:** The simulated firing rates in response to the baseline stimuli. r denotes the neural activity, and $o_z^{\text{base}\#4}$ represents the observation of the brain, which in this example is assumed to be equal to the observation in the scientist's model (note that the image is the same for both $o_x^{\text{base}\#4}$ and $o_z^{\text{base}\#4}$). **E:** The predicted (red) and the simulated (black) firing rates in response to the test stimuli. o_z^{test} denotes the observation for the test, which, again, is the same as in the scientist's model (note that the grating image is the same for both o_x^{test} and o_z^{test}). **F -J:** Same as **A - E** but for the grating model. We only use the marginal posteriors of one of the latent grating variables, $g\#1$. The projective field of the grating variable is shown in H.

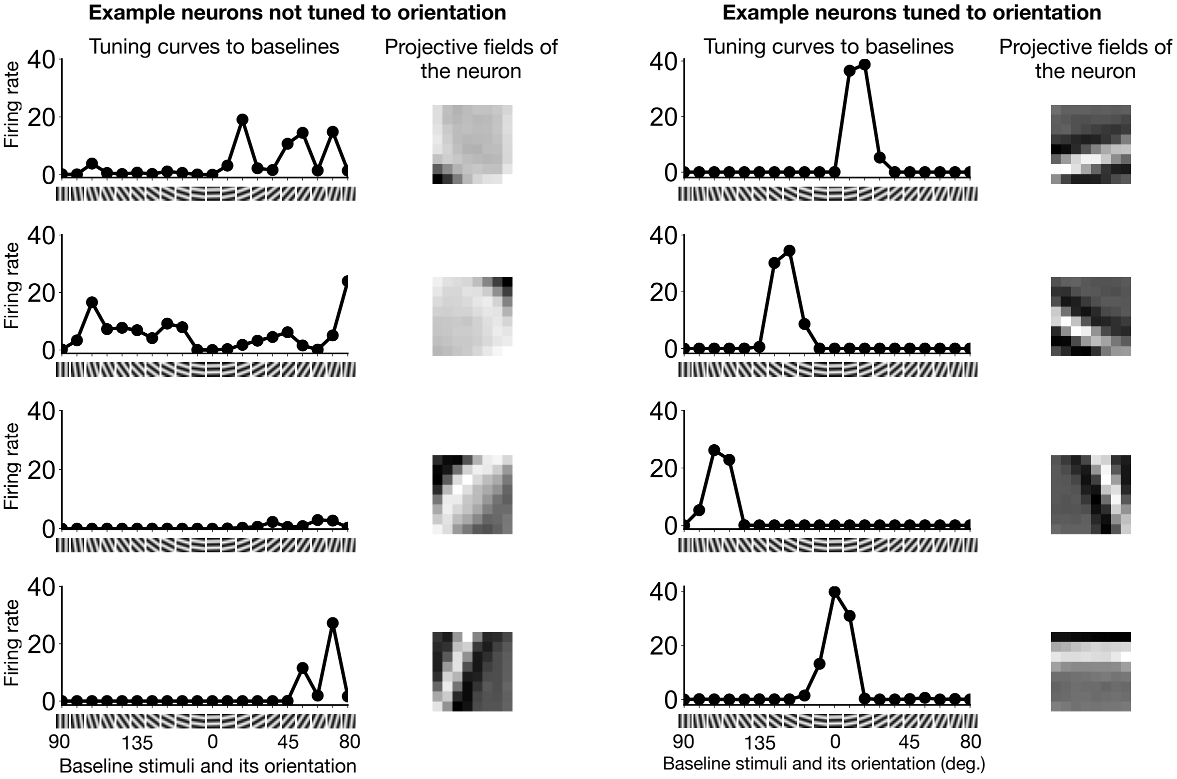


Figure S4: **Tuning curves in response to the baseline stimuli. Left column:** Example tuning curves for neurons not tuned to orientation. **Right column:** Example tuning curves for neurons tuned to orientation.

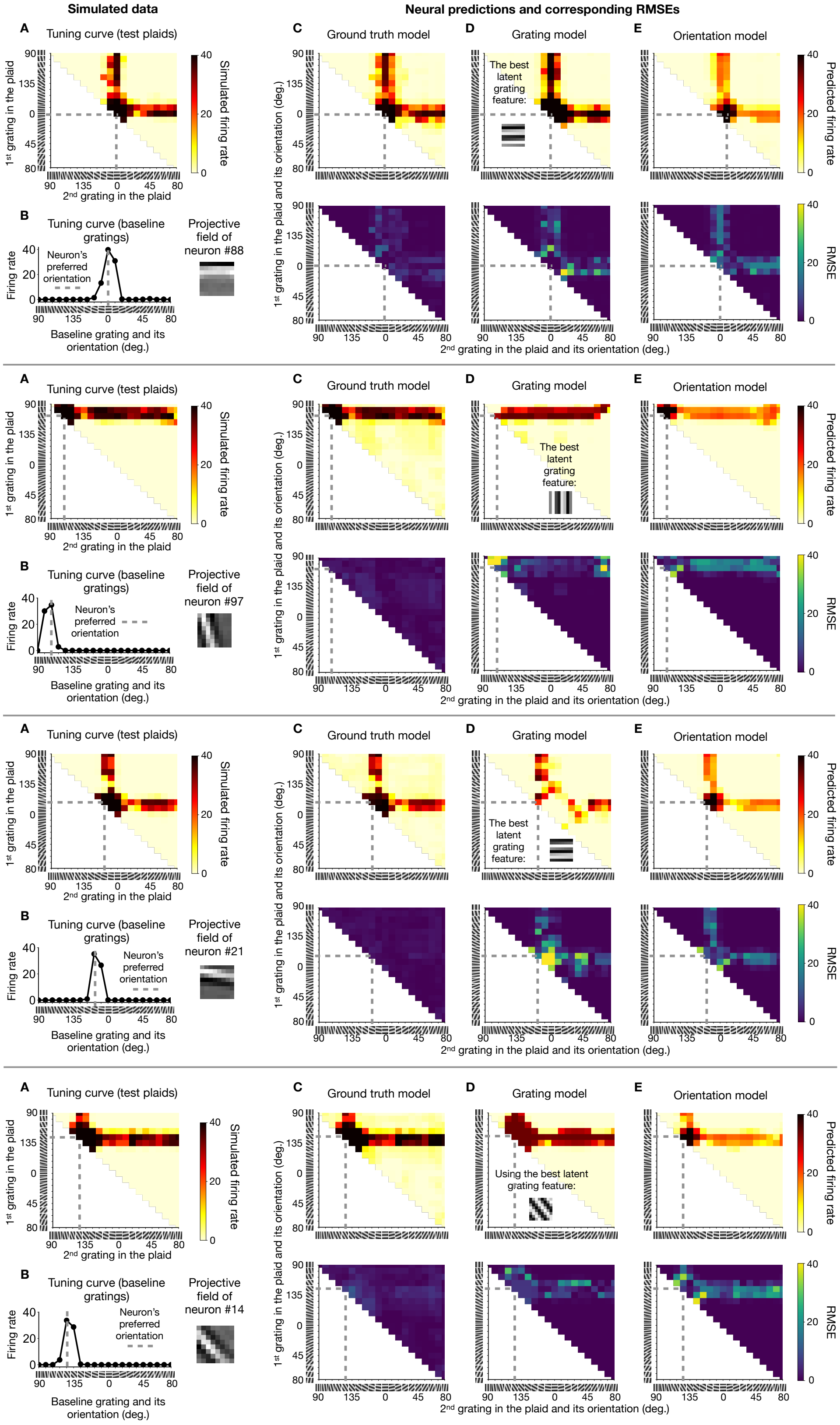


Figure S5: **Simulated and predicted firing rates for four example neurons. A:** The tuning curves for the test plaid stimuli. **B:** The tuning curves for the baseline stimuli. (**C-E**) Predicted firing rates (top row, low rates in dim yellow and high rates in black) and root mean squared errors (bottom row, RMSE, low in blue and high in yellow) for the ground truth (**C**), the grating (**D**), and the orientation (**E**) models. Note that the tuning curves for all test stimuli can be best plotted on these 2-dimensional heat maps, showing how the neural response of a neuron changes as a function of the orientations of the two gratings in the plaids

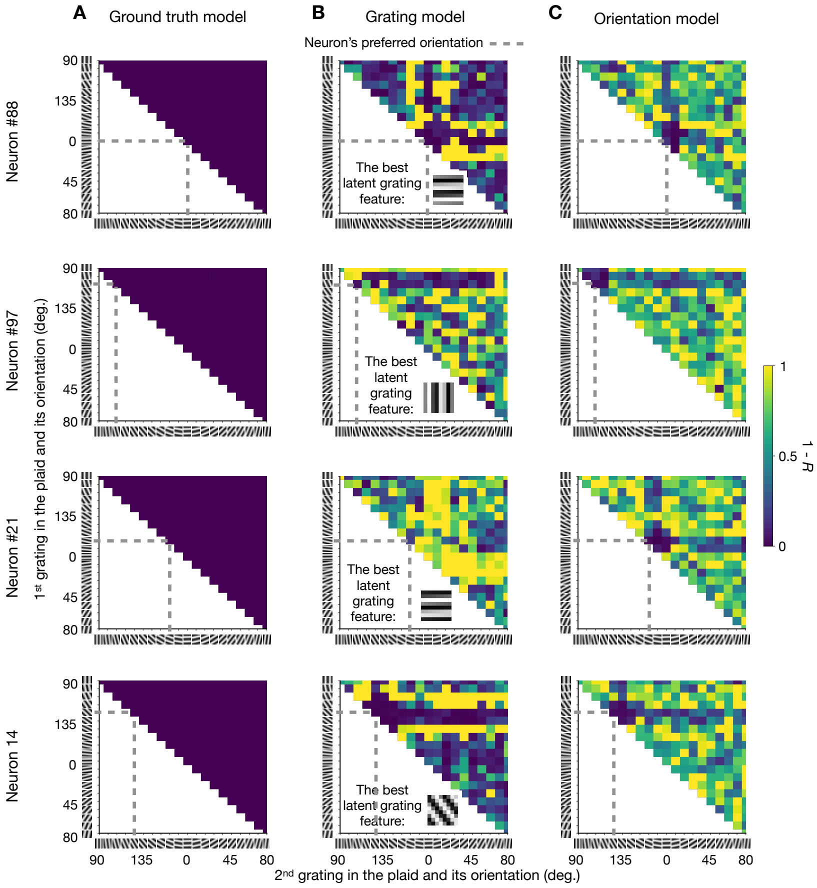


Figure S6: **The new, correlation-based diagnostic for model comparison. A-C:** 1-*R* (where *R* is the correlation between the neural prediction and the posterior matching errors) across the test plaids for the ground truth, grating, and orientation models for the four example neurons. See Fig. S5 for additional details about the fits for these neurons.