A DATA

A.1 NEURAL DATA

Data from real neural activity is used in Fig. 1, in Fig. 3 panel **a** and panel **b** on the right side, and in Fig. 4. Responses were obtained via two-photon calcium imaging of layer L2/3 of the primary visual cortex (area V1) of the mouse. Recordings, experimental paradigm and pre-processing was similar to [16]. The data for Fig. 1 and Fig. 3 consists of the responses of 7672 neurons to 360 images where each image was presented 20 times. Since 7 trials were missing, this makes for a total of 7193 trials per neuron. The data for Fig. 4 consists of the responses of 5335 neurons to 4472 images in the training set, 522 images in the validation set and 100 images in the test set. The images of the test set were repeated 10 times which makes for 999 test trials per neuron since one trial was missing.

A.2 SIMULATED NEURAL DATA

Simulated neural data is used in Fig. 3 in the left part of panel **b** and both parts of panel **c**. We generated samples for 100 neurons, 360 stimuli, and 31 repeats per stimulus. Briefly, we simulated the data assuming a zero-inflated Log-Normal distribution where the parameters μ and σ^2 of the Log-Normal part are normal-gamma distributed. The complete description of the simulation is as follows:

$$
\begin{aligned}
y &\sim ZIL(\mu, \sigma^2, q, \tau) \\
\mu &\sim \mathcal{N}(\mu_\mu, \sigma^2/\nu) \\
\sigma^2 &\sim Gamma(\alpha_{\sigma^2}, \beta_{\sigma^2}) \\
q &\sim Beta(21, 117) \\
\tau &= \exp(-10) \\
\nu &\sim Gamma(8.29, 7.32) \\
\alpha_{\sigma^2} &\sim Gamma(27.81, 0.8) \\
\beta_{\sigma^2} &= \alpha_{\sigma^2}/\overline{\sigma}^2_{noise} \\
\begin{bmatrix} \mu_\mu \\ \overline{\sigma}^2_{noise} \end{bmatrix} &\sim \mathcal{N}\left(\begin{bmatrix} -3.13 \\ 0.36 \end{bmatrix}, \begin{bmatrix} 0.158 & -0.017 \\ -0.017 & 0.003 \end{bmatrix} \right)
\end{aligned}
$$

The parameter values were chosen such that the resulting simulated data resembles the real neural activity.

Simulating data with different SNRs In order to simulate data with different SNR values, we first generated samples y as described above and then transformed the data into the $\log$ space $z = \log(y)$. Next, to extract the noise we subtracted the mean per stimulus, scaled the noise and then added the mean back. This results in a set of samples where the mean stays the same while the noise level has been scaled. Finally, we transformed the data back into the original space by applying the $\exp$ function on the resulting samples:

$$
y = \exp\left((z - \overline{z}) * c + \overline{z}\right),
$$

where $\overline{z}$ is the average across repeats and c is the scaling factor for the noise across repeats. The SNR of the (simulated) responses is computed as $\frac{\text{Var}_x(\mathbb{E}_y[y|x])}{\mathbb{E}_x[\text{Var}_y(y|x)]}$ where $\text{Var}_x(\mathbb{E}_y[y|x])$ is the variance of averaged responses and $\mathbb{E}_x[\text{Var}_y(y|x)]$ is the average noise level.

Data for comparison of likelihood and correlation The artificial data used in Fig. 6 is not intended to simulate accurate patterns of neural activity. It is thus simply drawn from a normal and a gamma distribution, in the left and right plot respectively. We sampled 5000 train and 500 test repeats to three "stimuli" for two "neurons", resulting in three 2D distributions. In the case of gamma data (left plot):

$$\begin{bmatrix} y_{11} \\ y_{12} \end{bmatrix} \sim Gamma\left(\begin{bmatrix} .56 \\ .43 \end{bmatrix}, \begin{bmatrix} .28 \\ .34 \end{bmatrix}\right)$$
$$\begin{bmatrix} y_{21} \\ y_{22} \end{bmatrix} \sim Gamma\left(\begin{bmatrix} .32 \\ .27 \end{bmatrix}, \begin{bmatrix} .13 \\ .39 \end{bmatrix}\right)$$
$$\begin{bmatrix} y_{31} \\ y_{32} \end{bmatrix} \sim Gamma\left(\begin{bmatrix} .45 \\ .34 \end{bmatrix}, \begin{bmatrix} .35 \\ .37 \end{bmatrix}\right)$$

And in the case of normal data (right plot):

$$\begin{bmatrix} y_{11} \\ y_{12} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 15. \\ 21. \end{bmatrix}, \begin{bmatrix} 1. & 0. \\ 0. & 1. \end{bmatrix}\right)$$
$$\begin{bmatrix} y_{21} \\ y_{22} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 16.8 \\ 16.3 \end{bmatrix}, \begin{bmatrix} 4.2 & 0. \\ 0. & .5 \end{bmatrix}\right)$$
$$\begin{bmatrix} y_{31} \\ y_{32} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 21.0 \\ 10.1 \end{bmatrix}, \begin{bmatrix} 2.5 & 0. \\ 0. & 5.3 \end{bmatrix}\right)$$

The fitted likelihoods are a gamma and a χ^2 distribution, respectively.

B MOMENT MATCHING FOR ZERO-INFLATED LIKELIHOOD

In this section we demonstrate how to compute the moments of each component of a zero-inflated mixture model. Since the uniform zero part does not have any parameters, we express the moments of the non-zero part as a function of the moments of the entire data, under the assumption of a zero-inflated distribution. We then use those for moment-matching the parameters of the non-zero part. The detailed step-by-step derivation is as follows:

We first express the total mean μ_{total} and total variance σ^2_{total} in terms of the means and variances of each component of the mixture model. We then solve for mean μ_1 and variance σ^2_1 of the positive distribution:

$$\mu_{total} = \mathbb{E}_y[y] = (1-q)\cdot\mu_0 + q\cdot\mu_1$$

To compute the total variance we make use of the law of total variance $\text{Var}_y(y) = \mathbb{E}_m[\text{Var}_{y|m}(y)] + \text{Var}_m(\mathbb{E}_{y|m}[y])$ where

$$\begin{aligned}\mathbb{E}_m[\text{Var}_{y|m}(y)] &= \mathbb{E}_m[\{\text{Var}_{y|m=0}(y), \text{Var}_{y|m=1}(y)\}] \\ &= (1-q)\cdot\sigma_0^2 + q\cdot\sigma_1^2\end{aligned}$$

and

$$\begin{aligned}\text{Var}_m(\mathbb{E}_{y|m}[y]) &= \mathbb{E}_m[\mathbb{E}_{y|m}[y]^2] - \mathbb{E}_m[\mathbb{E}_{y|m}[y]]^2 \\ &= \mathbb{E}_m[\{\mathbb{E}_{y|m=0}[y]^2, \mathbb{E}_{y|m=1}[y]^2\}] - \mathbb{E}_m[\{\mathbb{E}_{y|m=0}[y], \mathbb{E}_{y|m=1}[y]\}]^2 \\ &= (1-q)\cdot\mu_0^2 + q\cdot\mu_1^2 - ((1-q)\cdot\mu_0 + q\cdot\mu_1)^2 \\ &= (1-q)\cdot\mu_0^2 + q\cdot\mu_1^2 - (1-q)^2\cdot\mu_0^2 - q^2\cdot\mu_1^2 - 2q(1-q)\cdot\mu_0\mu_1 \\ &= ((1-q) - (1-q)^2)\cdot\mu_0^2 + (q-q^2)\cdot\mu_1^2 - 2q(1-q)\cdot\mu_0\mu_1 \\ &= q(1-q)\cdot\mu_0^2 + q(1-q)\cdot\mu_1^2 - 2q(1-q)\cdot\mu_0\mu_1 \\ &= q(1-q)\cdot(\mu_0-\mu_1)^2.\end{aligned}$$

Notation $\mathbb{E}[\{a, b, ...\}]$ is used to denote that the expectation involves the terms in the set $\{a, b, ...\}$. The total variance can then be computed as

$$\begin{aligned}\sigma^2_{total} = \text{Var}_y[y] &= \mathbb{E}_m[\text{Var}_{y|m}(y) + \text{Var}_m(\mathbb{E}_{y|m}[y]) \\ &= (1-q)\cdot\sigma_0^2 + q\cdot\sigma_1^2 + q(1-q)\cdot(\mu_0-\mu_1)^2\end{aligned}$$

The mean and variance of the non-zero part can thus be computed as

$$\begin{aligned}\mu_1 &= \frac{\mu_{total} - (1-q)\cdot\mu_0}{q} \\ \sigma_1^2 &= \frac{\sigma^2_{total} - (1-q)\cdot\sigma_0^2 - q(1-q)(\mu_0-\mu_1)^2}{q}\end{aligned}$$

The parameters of the non-zero part can then be obtained by moment matching with μ_1 and σ^2_1. Note, however, that the mean of the non-zero distribution is not μ_1 itself but $\mu_1 - \tau$. In a case where there are no responses above the zero-threshold τ, μ_1 and σ^2_1 are not defined because of the denominator $q = 0$. In this case we assign a small value of 0.1 to the mean and 0.3 to the variance. We chose these values because they resulted in the best GS model performance for the PE approach.

B.1 ZERO-INFLATED LOG-NORMAL LIKELIHOOD

In the case of a Log-Normal non-zero part, the parameters μ_{LogN} and σ^2_{LogN} evaluate to

$$\begin{aligned}\mu_{LogN} &= \log\left(\frac{\mu_1 - \tau}{\sqrt{\frac{\sigma_1^2}{(\mu_1-\tau)^2} + 1}}\right) \\ \sigma^2_{LogN} &= \log\left(\frac{\sigma_1^2}{(\mu_1-\tau)^2} + 1\right)\end{aligned}$$

Note that μ_0, μ_1, σ_0^2 and σ_1^2 are the means and variances of the zero and non-zero part of the distribution, respectively. The parameters μ_{LogN} and σ^2_{LogN} are *not* the mean and variance of the Log-Normal distribution but of the underlying Normal distribution in log space.

C Posterior Predictive for zero-inflated likelihood

Our goal is to probabilistically infer the parameters of the distribution per image, in a leave-one-out manner. That is, to compute $p(y_i|\mathbf{y}_{\setminus i}, x)$. For brevity, we drop the conditioning on x in the following derivations. Following the graphical model in Fig. 2, let's define some of the density functions that will be used later on:

$$\begin{aligned} p(y, \theta, m, q) &= p(y|\theta, m)p(m|q)p(\theta)p(q) \\ p(m|q) &= q^m \cdot (1-q)^{1-m} \\ p(y|\theta, m) &= p(y|\theta_0)^{1-m} \cdot p(y|\theta_1)^m \\ p(q) &= q^{\alpha-1} \cdot (1-q)^{\beta-1} \cdot \frac{1}{\mathrm{B}(\alpha, \beta)} \end{aligned}$$

Marginalizing over m:

$$\begin{aligned} p(y, \theta, q) &= \sum_{m \in \{0,1\}} p(y, \theta, m, q) \\ &= p(\theta)p(q) \sum_{m \in \{0,1\}} p(y|\theta, m)p(m|q) \\ &= p(\theta)p(q) \left[p(y|\theta, m=0)p(m=0|q) + p(y|\theta, m=1)p(m=1|q)\right] \\ &= p(\theta)p(q) \left[p(y|\theta_0)(1-q) + p(y|\theta_1) \cdot q\right] \\ &= p(\theta)p(q)p(y|\theta, q) \end{aligned}$$

Our goal is to compute the posterior predictive distribution $p(y_i|\mathbf{y}_{\setminus i})$:

$$\begin{aligned} p(y_i|\mathbf{y}_{\setminus i}) &= \int_{\theta,q} p(y_i, \theta, q|\mathbf{y}_{\setminus i}) d\theta dq \\ &= \int_{\theta,q} \underbrace{p(y_i|\theta, q, \mathbf{y}_{\setminus i})}_{= p(y_i|\theta,q) \text{ since } y_i \perp\!\!\!\perp \mathbf{y}_{\setminus i}|\theta,q} p(\theta, q|\mathbf{y}_{\setminus i}) d\theta dq \\ &= \int_{\theta,q} \underbrace{p(y_i|\theta, q)}_{\text{likelihood}} \underbrace{p(\theta, q|\mathbf{y}_{\setminus i})}_{\text{posterior}} d\theta dq \end{aligned}$$

Let us now compute the quantities we need for the posterior predictive $p(y_i|\mathbf{y}_{\setminus i})$, for a single neuron and a single image.

We know the likelihood: $p(y|\theta, q) = (1-q) \cdot p(y|\theta_0) + q \cdot p(y|\theta_1)$. Since the two distributions of our mixture model are not overlapping we can re-write the likelihood as follows:

$$p(y|\theta, q) = \begin{cases} (1-q) \cdot p(y|\theta_0) & \text{if } y \leq \tau \ (m=0) \\ q \cdot p(y|\theta_1) & \text{otherwise } (m=1) \end{cases}$$

The posterior can be derived as follows:

$$\begin{aligned} p(\theta, q|\mathbf{y}_{\setminus i}) &\propto p(\mathbf{y}_{\setminus i}|\theta, q)p(\theta)p(q) \\ &\propto \left(p(\theta_0) \cdot \prod_{y_j \in \mathbf{y}^0_{\setminus i}} (1-q) \cdot p(y_j|\theta_0)\right) \cdot \left(p(\theta_1) \cdot \prod_{y_j \in \mathbf{y}^1_{\setminus i}} q \cdot p(y_j|\theta_1)\right) \cdot p(q) \\ &\propto \left(p(\theta_0) \cdot \prod_{y_j \in \mathbf{y}^0_{\setminus i}} p(y_j|\theta_0)\right) \cdot \left(p(\theta_1) \cdot \prod_{y_j \in \mathbf{y}^1_{\setminus i}} p(y_j|\theta_1)\right) \cdot (1-q)^{n_0} \cdot q^{n_1} \cdot p(q) \\ &\propto p(\theta_0)p(\mathbf{y}^0_{\setminus i}|\theta_0) \cdot p(\theta_1)p(\mathbf{y}^1_{\setminus i}|\theta_1) \cdot (1-q)^{n_0} \cdot q^{n_1} \cdot q^{\alpha-1} \cdot (1-q)^{\beta-1} \cdot \frac{1}{\mathrm{B}(\alpha, \beta)} \\ &\propto p(\theta_0)p(\mathbf{y}^0_{\setminus i}|\theta_0) \cdot p(\theta_1)p(\mathbf{y}^1_{\setminus i}|\theta_1) \cdot (1-q)^{n_0+\beta-1} \cdot q^{n_1+\alpha-1} \cdot \frac{1}{\mathrm{B}(\alpha, \beta)}, \end{aligned}$$

where $\mathbf{y}^0_{\setminus i}$ are the zero responses, $\mathbf{y}^1_{\setminus i}$ are the positive responses, and n_0 and n_1 are the number of zero and positive responses, respectively. Since the joint distribution factorizes, the whole posterior factorizes (because it is just a re-scaled version of the joint). Normalizing each factor by its own constant, respectively, we get:

$$\begin{aligned} p(\theta, q|\mathbf{y}_{\setminus i}) &= \frac{p(\theta_0)p(\mathbf{y}^0_{\setminus i}|\theta_0)}{Z_1} \cdot \frac{p(\theta_1)p(\mathbf{y}^1_{\setminus i}|\theta_1)}{Z_2} \cdot \frac{(1-q)^{n_0+\beta-1} \cdot q^{n_1+\alpha-1}}{\mathrm{B}(n_1+\alpha, n_0+\beta)} \\ &= p(\theta_0|\mathbf{y}^0_{\setminus i}) \cdot p(\theta_1|\mathbf{y}^1_{\setminus i}) \cdot \mathrm{Beta}(n_1+\alpha, n_0+\beta) \end{aligned} \tag{7}$$

Note that in the case of the posterior over q since the distribution takes the form of a beta distribution we can simply adjust the denominator to the appropriate normalization factor for a beta distribution $B(n_1+\alpha, n_0+\beta)$.

Let us now combine these two components of the posterior predictive to compute $p(y_i|\mathbf{y}_{\setminus i})$:

$$\begin{aligned} p(y_i|\mathbf{y}_{\setminus i}) &= \int_{\theta,q} p(y_i|\theta, q)p(\theta, q|\mathbf{y}_{\setminus i})\, d\theta dq \\ &= \int_{\theta,q} p(y_i|\theta, q)p(\theta_0|\mathbf{y}^0_{\setminus i})p(\theta_1|\mathbf{y}^1_{\setminus i})p(q|\mathbf{y}_{\setminus i})\, d\theta dq \\ &= \int_q \underbrace{\left(\int_\theta p(y_i|\theta, q)p(\theta_0|\mathbf{y}^0_{\setminus i})p(\theta_1|\mathbf{y}^1_{\setminus i})\, d\theta\right)}_{=p(y_i|q,\mathbf{y}_{\setminus i})} p(q|\mathbf{y}_{\setminus i})\, dq \\ &= \int_q p(y_i|q, \mathbf{y}_{\setminus i})p(q|\mathbf{y}_{\setminus i})\, dq \end{aligned} \tag{8}$$

The posterior predictive can then be evaluated depending on whether the target response y_i is below the zero-threshold τ or above it:

If $y_i < \tau$:

$$\begin{aligned} p(y_i|\mathbf{y}_{\setminus i}) &= \int_q \int_\theta p(y_i|\theta, q)p(\theta_0|\mathbf{y}^0_{\setminus i})p(\theta_1|\mathbf{y}^1_{\setminus i})\, d\theta\, p(q|\mathbf{y}_{\setminus i})\, dq \\ &= \int_q \int_\theta p(y_i|\theta_0, q)p(\theta_0|\mathbf{y}^0_{\setminus i})p(\theta_1|\mathbf{y}^1_{\setminus i})\, d\theta\, p(q|\mathbf{y}_{\setminus i})\, dq \\ &= \int_q \int_{\theta_0} p(y_i|\theta_0, q)p(\theta_0|\mathbf{y}^0_{\setminus i})\, d\theta_0 \underbrace{\int_{\theta_1} p(\theta_1|\mathbf{y}^1_{\setminus i})\, d\theta_1}_{=1} p(q|\mathbf{y}_{\setminus i})\, dq \\ &= \int_q \int_{\theta_0} p(y_i|\theta_0, q)p(\theta_0|\mathbf{y}^0_{\setminus i})\, d\theta_0\, p(q|\mathbf{y}_{\setminus i})\, dq \\ &= \int_q p(y_i|q, \mathbf{y}^0_{\setminus i})\, p(q|\mathbf{y}_{\setminus i})\, dq \\ &= \int_q (1-q) \cdot p(y_i|\mathbf{y}^0_{\setminus i})\, p(q|\mathbf{y}_{\setminus i})\, dq \\ &= p(y_i|\mathbf{y}^0_{\setminus i}) \int_q (1-q) \cdot p(q|\mathbf{y}_{\setminus i})\, dq \end{aligned}$$

And if $y_i \geq \tau$:

$$\begin{aligned} p(y_i|\mathbf{y}_{\setminus i}) &= \int_q \int_{\theta_1} p(y_i|\theta_1, q) p(\theta_1|\mathbf{y}^1_{\setminus i})\, d\theta_1\, p(q|\mathbf{y}_{\setminus i})\, dq \\ &= \int_q p(y_i|q, \mathbf{y}^1_{\setminus i})\, p(q|\mathbf{y}_{\setminus i})\, dq \\ &= \int_q q \cdot p(y_i|\mathbf{y}^1_{\setminus i})\, p(q|\mathbf{y}_{\setminus i})\, dq \\ &= p(y_i|\mathbf{y}^1_{\setminus i}) \int_q q \cdot p(q|\mathbf{y}_{\setminus i})\, dq \end{aligned}$$

This means that depending on the target response y_i we either need to compute the posterior predictive of the zero distribution (i.e., Uniform) or positive distribution (i.e., Log-Normal).

Finally, the complete posterior predictive distribution is estimated via numerical integration over q. Numerical integration in this particular case is feasible since q only takes values between 0 and 1.

C.1 Zero-inflated Log-Normal likelihood

We now apply the generic derivation in the previous section to zero-inflated Log-Normal distribution and derive the posterior predictive distribution for it. Let us start by assuming that the target response y_i is below the zero-threshold τ. In this case, the response falls into the Uniform distribution whose parameters are fixed and do not depend on the other zero responses. Therefore, the posterior predictive stays a uniform distribution: $p(y_i|\mathbf{y}_{\setminus i}) = 1/\tau$.

Alternatively, the target response y_i could be higher than the zero-threshold τ falling into the Log-Normal distribution. In this case, we first transform the responses via the $\log$ function into the Gaussian space, then compute the posterior predictive distribution, and finally normalize the resulting distribution to go back into the log space:

$$\begin{aligned} p(y_i|\mathbf{y}_{\setminus i}) &= p(\log(y_i)|\log(\mathbf{y}_{\setminus i})) \cdot |\det \nabla_{y_i} \exp(y_i)| \\ &= p(\log(y_i)|\log(\mathbf{y}_{\setminus i})) \cdot \frac{1}{y_i} \end{aligned} \tag{9}$$

We now focus on computing the posterior predictive in the Gaussian space. For brevity let us assign $\log(y)$ to a new variable $z = \log(y)$. To compute the posterior predictive distribution we need to specify a prior over our likelihood parameters, in this case μ and σ^2. For a Gaussian distribution with unknown μ and σ^2 the conjugate prior is the Normal-inverse gamma distribution with parameters μ_0, ν, α, and β. These parameters are estimated form the data. Once the prior parameters are known, we can then compute the posterior predictive distribution, which is a t-distribution in the case of a Gaussian likelihood:

$$p(z_i|z_{\setminus i}) = t_{2\alpha'}\left(z_i|\mu', \frac{\beta'(\nu'+1)}{\nu'+\alpha'}\right), \tag{10}$$

where

$$\begin{aligned} \mu' &= \frac{\nu\mu_0 + n\bar{z}_{\setminus i}}{\nu + n} \\ \nu' &= \nu + n \\ \alpha' &= \alpha + \frac{n}{2} \\ \beta' &= \beta + \frac{1}{2}\sum_{z_j \in z_{\setminus i}}^{n} (z_j - \bar{z}_{\setminus i})^2 + \frac{n\nu(\bar{z}_{\setminus i} - \mu_0)^2}{2(\nu + n)} \end{aligned}$$

with n being the number of left-out repeats $z_{\setminus i}$ and $\bar{z}_{\setminus i}$ being the mean of the left-out repeats. As the final step, to compute the posterior predictive in the original $\log$ space, we plug Eq. 10 back into Eq. 9:

$$p(y_i|q, \mathbf{y}_{\setminus i}) = t_{2\alpha'}\left(\log(y_i)|\mu', \frac{\beta'(\nu'+1)}{\nu'+\alpha'}\right) \cdot \frac{1}{y_i}$$

D IN WHICH CASES DOES THE BAYESIAN GS OUTPERFORM THE PE

The PE approach to obtain a Gold Standard model fails in many cases which is the reason behind using the Bayesian Posterior Predictive approach instead. Fig. S1 provides insight into why and in which cases the PE fails compared to the Bayesian GS. In summary: This is due to the combination of low-valued responses that fall into the positive distribution and the sparsity in the data:
Target responses that are slightly greater than the threshold value are not covered by the uniform distribution of the zero part but are an extreme value for the positive part (left panel). When this coincides with overfitting due to sparse data, i.e. low proportion q of positive responses (right panel), the Point Estimate results in a low log-likelihood. Note that the reason for this not being visible for the smallest value of q ($q = 0$) is that in this case no positive responses were available to estimate the PE parameters on. Since the target trial could still be positive, we needed to assign the PE parameters of the positive part of the distribution as a hyper parameter. This is equivalent to applying a delta-peak prior and results in a quasi-Bayesian approach for the PE in these rare cases.

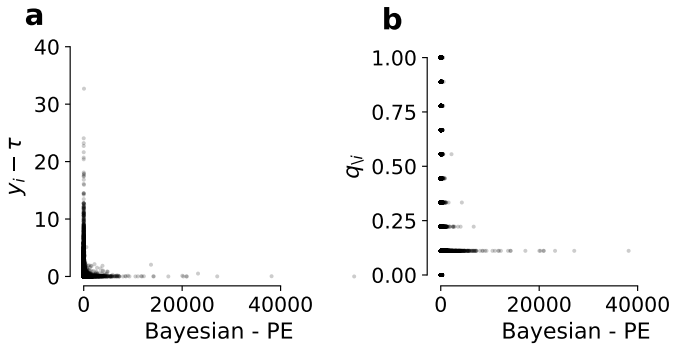

Fig. S1: Comparison between Bayesian and Point Estimate (PE) Gold Standard models. Since the two GS models share the same zero distribution, this analysis was only performed on the responses that fall into the positive distribution ($y \geq \tau$). **a:** Distance of the positive response from the zero-threshold τ as a function of the difference between Bayesian and the PE GS models. **b:** Fraction $q_{\setminus i}$ of positive leave-one-out responses $\mathbf{y}_{\setminus i}$ as a function of the difference between Bayesian and PE GS models. Data is per neuron, per repeat, and per stimulus).

E NORMALIZED INFORMATION GAIN IN TERMS OF KL DIVERGENCES

Here we provide the normalized information gain formulated in terms of KL divergence and derive the estimate presented in Eq.1:

$$
\begin{aligned}
\text{NInGa} =& \frac{\langle D_{KL}\left[p(y \mid x) \| p_0(y)\right]\rangle_x - \langle D_{KL}[p(y \mid x) \| \hat{p}(y \mid x)]\rangle_x}{\langle D_{KL}\left[p(y \mid x) \| p_0(y)\right]\rangle_x - \langle D_{KL}\left[p(y \mid x) \| p_*(y \mid x)\right]\rangle_x} \\
=& \frac{\left\langle \left\langle \log \frac{p(y|x)}{p_0(y)} \right\rangle_{y|x} \right\rangle_x - \left\langle \left\langle \log \frac{p(y|x)}{\hat{p}(y|x)} \right\rangle_{y|x} \right\rangle_x}{\left\langle \left\langle \log \frac{p(y|x)}{p_0(y)} \right\rangle_{y|x} \right\rangle_x - \left\langle \left\langle \log \frac{p(y|x)}{p_*(y|x)} \right\rangle_{y|x} \right\rangle_x} \\
=& \frac{\left\langle \langle \log \hat{p}(y \mid x) \rangle_{y|x} \right\rangle_x - \left\langle \langle \log p_0(y) \rangle_{y|x} \right\rangle_x}{\left\langle \langle \log p_*(y \mid x) \rangle_{y|x} \right\rangle_x - \left\langle \langle \log p_0(y) \rangle_{y|x} \right\rangle_x} \\
\approx& \frac{\sum_i \left(\log \hat{p}(y_i \mid x_i) - \log p_0(y_i)\right)}{\sum_i \left(\log p_*(y_i \mid x_i) - \log p_0(y_i)\right)}
\end{aligned}
$$

F RELATION BETWEEN NORMALIZED INFORMATION GAIN AND FEVE

Here we go through the complete derivation underlying Eq. 5. Let us start by defining FEVE:

$$FEVE = 1 - \frac{\left\langle (\mu_x - \hat{\mu}_x)^2 \right\rangle_x}{\sigma_s^2}, \tag{11}$$

where μ_x is the true conditional mean, $\hat{\mu}_x$ the estimated conditional mean by the model, and σ_s^2 is the signal variance. An estimator of FEVE was previously used by Cadena et al. [7] (which we also use to compute FEVE in Fig. 5**b**):

$$\begin{aligned}
FEVE &= 1 - \frac{\left\langle (\mu_x - \hat{\mu}_x)^2 \right\rangle_x}{\sigma_s^2} \\
&= 1 - \frac{\left\langle (\mu_x - \hat{\mu}_x)^2 \right\rangle_x}{\sigma_y^2 - \sigma_\epsilon^2} \\
&= 1 - \frac{\sigma_\epsilon^2 + \left\langle (\mu_x - \hat{\mu}_x)^2 \right\rangle_x - \sigma_\epsilon^2}{\sigma_y^2 - \sigma_\epsilon^2} \\
&= 1 - \frac{\left\langle (y - \mu_x)^2 \right\rangle_{x,y} + \left\langle (\mu_x - \hat{\mu}_x)^2 \right\rangle_x - \overbrace{2\left\langle (y - \mu_x)(\mu_x - \hat{\mu}_x) \right\rangle_{x,y}}^{=0} - \sigma_\epsilon^2}{\sigma_y^2 - \sigma_\epsilon^2} \\
&= 1 - \frac{\left\langle (y - \mu_x + \mu_x - \hat{\mu}_x)^2 \right\rangle_{x,y} - \sigma_\epsilon^2}{\sigma_y^2 - \sigma_\epsilon^2} \\
&= 1 - \frac{\left\langle (y - \hat{\mu}_x)^2 \right\rangle_{x,y} - \sigma_\epsilon^2}{\sigma_y^2 - \sigma_\epsilon^2},
\end{aligned} \tag{12}$$

where $\sigma_y^2 = \text{Var}(y)$ and $\sigma_\epsilon^2 = \mathbb{E}_x[\text{Var}(y|x)]$ are estimated from the data. Now let us expand $\langle D_{KL}[p(y|x)||\hat{p}(y|x)] \rangle_x$ in the case of a Gaussian likelihood:

$$\begin{aligned}
\langle D_{KL}[p(y|x)||\hat{p}(y|x)] \rangle_x &= \langle \log\left(p(y|x)\right) - \log(\hat{p}(y|x)) \rangle_{x,y} \\
&= \left\langle \log\left(\left(2\pi\sigma_x^2\right)^{-1/2}\right) - \frac{(y_x - \mu_x)^2}{2\sigma_x^2} - \log\left(\left(2\pi\hat{\sigma}_x^2\right)^{-1/2}\right) + \frac{(y_x - \hat{\mu}_x)^2}{2\hat{\sigma}_x^2} \right\rangle_{x,y} \\
&= \left\langle \log\left(\frac{\hat{\sigma}_x}{\sigma_x}\right) - \frac{(y_x - \mu_x)^2}{2\sigma_x^2} + \frac{(y_x - \hat{\mu}_x)^2}{2\hat{\sigma}_x^2} \right\rangle_{x,y} \\
&= \left\langle \log\left(\frac{\hat{\sigma}_x}{\sigma_x}\right) - \frac{(y_x - \mu_x)^2}{2\sigma_x^2} \right\rangle_{x,y} + \left\langle \frac{(y_x - \mu_x + \mu_x - \hat{\mu}_x)^2}{2\hat{\sigma}_x^2} \right\rangle_{x,y} \\
&= \left\langle \log\left(\frac{\hat{\sigma}_x}{\sigma_x}\right) - \frac{1}{2} \right\rangle_x + \left\langle \frac{(y_x - \mu_x)^2 + (\mu_x - \hat{\mu}_x)^2 + 2(y_x - \mu_x)(\mu_x - \hat{\mu}_x)}{2\hat{\sigma}_x^2} \right\rangle_{x,y} \\
&= \left\langle \log\left(\frac{\hat{\sigma}_x}{\sigma_x}\right) - \frac{1}{2} \right\rangle_x + \left\langle \frac{\left\langle (y_x - \mu_x)^2 \right\rangle_{y|x} + (\mu_x - \hat{\mu}_x)^2 + 2\overbrace{\left\langle (y_x - \mu_x)(\mu_x - \hat{\mu}_x) \right\rangle_{y|x}}^{=0}}{2\hat{\sigma}_x^2} \right\rangle_x \\
&= \left\langle \log\left(\frac{\hat{\sigma}_x}{\sigma_x}\right) - \frac{1}{2} \right\rangle_x + \left\langle \frac{\sigma_x^2 + (\mu_x - \hat{\mu}_x)^2}{2\hat{\sigma}_x^2} \right\rangle_x \\
&= \left\langle \log\left(\frac{\hat{\sigma}_x}{\sigma_x}\right) - \frac{1}{2} + \frac{\sigma_x^2 + (\mu_x - \hat{\mu}_x)^2}{2\hat{\sigma}_x^2} \right\rangle_x \\
&= \left\langle \log\left(\frac{\hat{\sigma}_x}{\sigma_x}\right) - \frac{1}{2} + \frac{\sigma_x^2}{2\hat{\sigma}_x^2} + \frac{(\mu_x - \hat{\mu}_x)^2}{2\hat{\sigma}_x^2} \right\rangle_x \\
&= \langle f(\hat{\sigma}_x) \rangle_x + \frac{1}{2} \left\langle \frac{(\mu_x - \hat{\mu}_x)^2}{\hat{\sigma}_x^2} \right\rangle_x
\end{aligned} \tag{13}$$

where $\hat{\sigma}_x$ is the noise estimated by the model, σ_x is the true noise, and $f(\hat{\sigma}_x) = \log\left(\frac{\hat{\sigma}_x}{\sigma_x}\right) - \frac{1}{2} + \frac{\sigma_x^2}{2\hat{\sigma}_x^2}$. Note that if $\hat{\sigma}_x = \sigma_x$ then $f(\hat{\sigma}_x) = 0$, and we would have:

$$\langle D_{KL}[p(y|x)||\hat{p}(y|x)]\rangle_x = \frac{1}{2}\left\langle \frac{(\mu_x - \hat{\mu}_x)^2}{\hat{\sigma}_x^2} \right\rangle_x$$

The term with the expectation can further be simplified. If the noise variance $\hat{\sigma}_x^2$ is not dependent on the stimulus x, which is the case for a Gaussian distribution, then $\hat{\sigma}_x^2 = \hat{\sigma}_\epsilon^2 = \left\langle \hat{\sigma}_x^2 \right\rangle_x$ and we can simply bring it outside the expectation:

$$\left\langle \frac{(\mu_x - \hat{\mu}_x)^2}{\hat{\sigma}_x^2} \right\rangle_x = \frac{\left\langle (\mu_x - \hat{\mu}_x)^2 \right\rangle_x}{\hat{\sigma}_\epsilon^2} \tag{14}$$

This would also mean that $f(\hat{\sigma}_x) = f(\hat{\sigma}_\epsilon) = \log\left(\frac{\hat{\sigma}_\epsilon}{\sigma_\epsilon}\right) - \frac{1}{2} + \frac{\sigma_\epsilon^2}{2\hat{\sigma}_\epsilon^2}$. However, if the noise variance is stimulus-dependent then the term with the expectation can be approximated via first-order Taylor expansion around the expected values of the numerator and denominator $\mathbf{c} = (\left\langle (\mu_x - \hat{\mu}_x)^2 \right\rangle_x, \left\langle \hat{\sigma}_x^2 \right\rangle_x)$:

$$\begin{aligned}
\left\langle \frac{(\mu_x - \hat{\mu}_x)^2}{\hat{\sigma}_x^2} \right\rangle_x &\approx \frac{\left\langle (\mu_x - \hat{\mu}_x)^2 \right\rangle_x}{\langle \hat{\sigma}_x^2 \rangle_x} \\
&\approx \frac{\left\langle (\mu_x - \hat{\mu}_x)^2 \right\rangle_x}{\hat{\sigma}_\epsilon^2}
\end{aligned}$$

Since we are dealing with a Gaussian distribution we will continue with the case where noise variance is not stimulus-dependent. However, the derivation applies to the approximate case too. Let us now relate Eq. 11 and Eq. 13:

$$\begin{aligned}
\langle D_{KL}[p(y|x)||\hat{p}(y|x)]\rangle_x &= f(\hat{\sigma}_\epsilon) + \frac{1}{2}\frac{\left\langle (\mu_x - \hat{\mu}_x)^2 \right\rangle_x}{\hat{\sigma}_\epsilon^2} \\
&= f(\hat{\sigma}_\epsilon) + \frac{1}{2}\frac{\left\langle (\mu_x - \hat{\mu}_x)^2 \right\rangle_x}{\sigma_\epsilon^2} \times \frac{\sigma_s^2}{\sigma_s^2} \\
&= f(\hat{\sigma}_\epsilon) + \frac{1}{2}\frac{\left\langle (\mu_x - \hat{\mu}_x)^2 \right\rangle_x}{\sigma_s^2} \times \frac{\sigma_s^2}{\hat{\sigma}_\epsilon^2} \\
&= f(\hat{\sigma}_\epsilon) + \frac{1}{2}(1 - FEVE) \times \frac{\sigma_s^2}{\hat{\sigma}_\epsilon^2}
\end{aligned} \tag{15}$$

Note that when estimated noise variance matches the true noise variance, then Eq. 15 becomes:

$$\langle D_{KL}[p(y|x)||\hat{p}(y|x)]\rangle_x = \frac{1}{2}(1 - FEVE) \times SNR$$

G RELATION BETWEEN NORMALIZED INFORMATION GAIN AND CORRELATION

Another commonly used metric is the trial-averaged correlation between the model prediction $\hat{\mu}_x = \langle \hat{y}|x \rangle_{\hat{y}|x}$ and true responses $\mu_x = \langle y|x \rangle_{y|x}$:

$$\rho(\hat{\mu}_x, \mu_x) = \frac{\text{Cov}(\hat{\mu}_x, \mu_x)}{\sqrt{\hat{\sigma}_s^2 \cdot \sigma_s^2}}$$

To relate this quantity to $\langle D_{KL}[p(y|x)||\hat{p}(y|x)] \rangle_x$, we start by expanding Eq. 14. Specifically, we add and subtract $\mu = \langle \mu_x \rangle = \langle \hat{\mu}_x \rangle$ in the numerator:

$$\begin{aligned}
\frac{\left\langle (\mu_x - \hat{\mu}_x)^2 \right\rangle_x}{\hat{\sigma}_\epsilon^2} &= \frac{\left\langle ((\mu_x - \mu) - (\hat{\mu}_x - \mu))^2 \right\rangle_x}{\hat{\sigma}_\epsilon^2} \\
&= \frac{\left\langle (\mu_x - \mu)^2 + (\hat{\mu}_x - \mu)^2 - 2(\mu_x - \mu)(\hat{\mu}_x - \mu) \right\rangle_x}{\hat{\sigma}_\epsilon^2} \\
&= \frac{\sigma_s^2 + \hat{\sigma}_s^2 - 2\text{Cov}(\hat{\mu}_x, \mu_x)}{\hat{\sigma}_\epsilon^2} \\
&= \frac{\sigma_s^2 + \hat{\sigma}_s^2 - 2\text{Cov}(\hat{\mu}_x, \mu_x)}{\hat{\sigma}_\epsilon^2} \times \frac{\sigma_s^2}{\sigma_s^2} \\
&= \frac{\sigma_s^2 + \hat{\sigma}_s^2 - 2\text{Cov}(\hat{\mu}_x, \mu_x)}{\sigma_s^2} \times \frac{\sigma_s^2}{\hat{\sigma}_\epsilon^2} \\
&= \left(1 + \frac{\hat{\sigma}_s^2}{\sigma_s^2} - \frac{2\hat{\sigma}_s}{\sigma_s}\rho(\hat{\mu}_x, \mu_x)\right) \times \frac{\sigma_s^2}{\hat{\sigma}_\epsilon^2}
\end{aligned}$$

Putting this back into Eq. 13:

$$\langle D_{KL}[p(y|x)||\hat{p}(y|x)] \rangle_x = f(\hat{\sigma}_\epsilon) + \frac{1}{2}\left(1 + \frac{\hat{\sigma}_s^2}{\sigma_s^2} - \frac{2\hat{\sigma}_s}{\sigma_s}\rho(\hat{\mu}_x, \mu_x)\right) \times \frac{\sigma_s^2}{\hat{\sigma}_\epsilon^2} \tag{16}$$

Again, if the model's noise variance matches the true noise variance ($\sigma_\epsilon^2 = \hat{\sigma}_\epsilon^2$), we have:

$$\langle D_{KL}[p(y|x)||\hat{p}(y|x)] \rangle_x = \frac{1}{2}\left(1 + \frac{\hat{\sigma}_s^2}{\sigma_s^2} - \frac{2\hat{\sigma}_s}{\sigma_s}\rho(\hat{\mu}_x, \mu_x)\right) \times SNR$$

If we further assume that the model's signal variance matches the true signal variance, $\hat{\sigma}_s^2 = \sigma_s^2$, we get:

$$\langle D_{KL}[p(y|x)||\hat{p}(y|x)] \rangle_x = (1 - \rho(\hat{\mu}_x, \mu_x)) \times SNR$$

H OPTIMIZING CORRELATION ONLY FOCUSES ON MATCHING TRIAL-AVERAGED RESPONSES

In addition to trial-averaged correlation, neural encoding models are also evaluated via single-trial correlation [23]. While in the trial-averaged case the correlation obviously only focuses on conditional means, here we show this is the case even for single-trial correlation. That is, optimizing single-trial correlation only focuses on matching the conditional means:

$$\begin{aligned}
\rho_{st}(\hat{\mu}_x, y) &= \frac{\text{Cov}(\hat{\mu}_x, y)}{\sqrt{\hat{\sigma}_s^2 \sigma_y^2}} \\
&= \frac{\text{Cov}(\hat{\mu}_x, y)}{\sqrt{\hat{\sigma}_s^2 \sigma_y^2}} \\
&= \frac{\text{Cov}(\mathbb{E}[\hat{\mu}_x|x], \mathbb{E}[y|x]) + \overbrace{\mathbb{E}[\text{Cov}(\hat{\mu}_x, y|x)]}^{=0}}{\sqrt{\hat{\sigma}_s^2 \sigma_y^2}} \\
&= \frac{\text{Cov}(\hat{\mu}_x, \mu_x)}{\sqrt{\hat{\sigma}_s^2 \sigma_y^2}}
\end{aligned}$$

where $\hat{\mu}_x$ is the predicted conditional mean, μ_x is the trial-averaged response, $\hat{\sigma}_s^2$ is the model signal variance, and σ_y^2 is the total data variance computed across all trials. Note that this quantity is invariant to affine transformations of the predicted conditional mean.

I OTHER APPROACHES FOR A GS ESTIMATE

I.1 MAXIMUM A POSTERIORI ESTIMATE

Instead of using the full posterior predictive to obtain a good GS model, one can use the maximum a posteriori (MAP) estimate of the distribution parameters. Here, we show the derivation of the MAP estimate for the zero-inflated Log-Normal distribution. However, it does not perform as well as the posterior predictive approach, see Fig. S2.

The maximum a posteriori estimator of a parameter $\phi \in \{\theta_0, \theta_1, q\}$ of a zero inflated likelihood can be computed as

$$\begin{aligned}
\hat{\phi}_{MAP} &= \arg\max_{\phi} p(\mathbf{y}_{\setminus i}|\theta, q)p(\theta)p(q) \\
&= \arg\max_{\phi} p(\mathbf{y}^0_{\setminus i}|\theta_0)p(\theta_0)p(\mathbf{y}^1_{\setminus i}|\theta_1)p(\theta_1) \cdot q^{n_1}(1-q)^{n_0}p(q) \\
&= \arg\max_{\phi} \log\left(p(\mathbf{y}^0_{\setminus i}|\theta_0)p(\theta_0)\right) + \log\left(p(\mathbf{y}^1_{\setminus i}|\theta_1)p(\theta_1)\right) + \log\left(q^{n_1}(1-q)^{n_0}p(q)\right)
\end{aligned}$$

where the second step is analogous to Eq. 7.

MAP estimate for q. In order to obtain the maximum a posteriori estimator for q we set the derivative with respect to q to zero. As a prior for q we choose a Beta distribution $p(q) = \mathrm{Beta}(q; \alpha'', \beta'')$:

$$\begin{aligned}
\hat{q}_{MAP} &= \arg\max_{q} \log\left(q^{n_1}(1-q)^{n_0}p(q)\right) \\
&= \arg\max_{q} \log\left(q^{n_1}(1-q)^{n_0}\frac{q^{\alpha''-1}(1-q)^{\beta''-1}}{B(\alpha'', \beta'')}\right) \\
&= \arg\max_{q} \underbrace{\log\left(q^{n_1+\alpha''-1}(1-q)^{n_0+\beta''-1}\right)}_{:=f(q)}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial f(q)}{\partial q} &= \frac{\partial}{\partial q}\left[\log\left(q^{n_1+\alpha''-1}(1-q)^{n_0+\beta''-1}\right)\right] \\
&= \frac{\partial}{\partial q}\left[(n_1+\alpha''-1)\log(q) + (n_0+\beta''-1)\log(1-q)\right] \\
&= (n_1+\alpha''-1)\frac{1}{q} - (n_0+\beta''-1)\frac{1}{1-q} \\
&\stackrel{!}{=} 0 \\
\hat{q}_{MAP} &= \frac{n_1+\alpha''-1}{n_0+n_1+\alpha''+\beta''-2}
\end{aligned}$$

MAP estimate for θ_1. The parameters θ_1 are all parameters of the non-zero part of the distribution. In the case of a LogNormal distributions, this is $\theta_1 \in \{\mu, \sigma^2\}$ and we assume a Normal-Inverse-Gamma prior $p(\theta_1) = \mathcal{NG}^{-1}(\mu'', \lambda'', \alpha'', \beta'')$. The posterior then follows a Normal-Inverse-Gamma distribution as well

$$\begin{aligned}
p(\mathbf{y}^1_{\setminus i}|\theta_1)p(\theta_1) &\approx p(\theta_1|\mathbf{y}^1_{\setminus i}) \\
&= \mathcal{NG}^{-1}(\mu', \lambda', \alpha', \beta') \\
&= \frac{\sqrt{\lambda'}\beta'^{\alpha'}}{\sqrt{2\pi}\Gamma(\alpha')}\frac{1}{\sigma}\left(\frac{1}{\sigma^2}\right)^{\alpha'+1}\exp\left(-\frac{2\beta'+\lambda'(\mu-\mu')^2}{2\sigma^2}\right)
\end{aligned}$$

with

$$\begin{aligned}
\mu' &= \frac{\mu''\nu'' + n_1\overline{y}}{\nu'' + n_1} \\
\nu' &= \nu'' + n_1 \\
\alpha' &= \alpha'' + \frac{n_1}{2} \\
\beta' &= \beta'' + \frac{1}{2}\sum_{y_j \in y^1_{\setminus i}}^{n_1} (y_j - \overline{y})^2 + \frac{n_1\nu''(\overline{y} - \mu'')^2}{2(\nu'' + n_1)}
\end{aligned}$$

where $\overline{y} := 1/n_1 \sum_{y_j \in \mathbf{y}^1_{\setminus i}} y_j$

The maximum a posteriori estimator of μ can then be obtained as follows:

$$\begin{aligned}
\hat{\mu}_{MAP} &= \arg\max_{\mu} \log\Big(p(\mathbf{y}^1_{\setminus i}|\theta_1)p(\theta_1)\Big) \\
&= \arg\max_{\mu} \log\left(\exp\left(-\frac{2\beta' + \lambda'(\mu - \mu')^2}{2\sigma^2}\right)\right) \\
&= \arg\max_{\mu} -\frac{\lambda'(\mu - \mu')^2}{2\sigma^2} \\
&= \mu'
\end{aligned}$$

And for σ^2:

$$\hat{\sigma}^2_{MAP} = \arg\max_{\sigma^2} \underbrace{\log\left(\frac{1}{\sigma}\left(\frac{1}{\sigma^2}\right)^{\alpha'+1}\exp\left(-\frac{2\beta' + \lambda'(\mu - \mu')^2}{2\sigma^2}\right)\right)}_{:= f(\sigma^2)}$$

$$\begin{aligned}
\frac{\partial f(\sigma^2)}{\partial \sigma^2} &= \frac{\partial}{\partial \sigma^2}\left(-\frac{1}{2}\log\sigma^2 - (\alpha' + 1)\log\sigma^2 - \frac{2\beta' + \lambda'(\mu - \mu')^2}{2}\frac{1}{\sigma^2}\right) \\
&= (-\alpha' - \frac{3}{2})\frac{1}{\sigma^2} + \frac{2\beta' + \lambda'(\mu - \mu')^2}{2}\frac{1}{\sigma^4} \\
&\stackrel{!}{=} 0 \\
\hat{\sigma}^2_{MAP} &= \frac{2\beta' + \lambda'\left(\hat{\mu}_{MAP} - \mu'\right)^2}{2\alpha' + 3}
\end{aligned}$$

MAP estimate for θ_0 In general, the maximum a posteriori estimator for θ_0 can be obtained analogously. In our case we model the zero part of the response distribution with a uniform distribution which does not have a parameter θ_0.

I.2 GOLD STANDARD MODEL AS A MIXTURE OF NULL AND POSTERIOR PREDICTIVE DISTRIBUTIONS

For some individual neurons and images the null model performs better than the Gold Standard model because the prior of the GS model is fitted per neuron but not per image. In cases with few positive responses where the GS model has to rely heavily on the prior, the performance can thus be sub-optimal for individual images. One idea, suggested by one of the reviewers, to circumvent this is to build a mixture model p_{**} between the GS p_* and Null p_0 model:

$$p_{**}(y_i|\mathbf{y}_{\setminus i}, y) = w_i \cdot p_*(y_i|\mathbf{y}_{\setminus i}) + (1 - w_i) \cdot p_0(y),$$

where $w \in [0, 1]$. We optimized w in a leave-one-out manner, just like p_* itself is obtained in a leave-one-out-manner: we obtained a w for each target repeat (per neuron per image) by optimizing p_{**} with respect to w on the other repeats. However, the resulting GS mixture model does not outperform the Bayesian GS model, see Fig. S2.

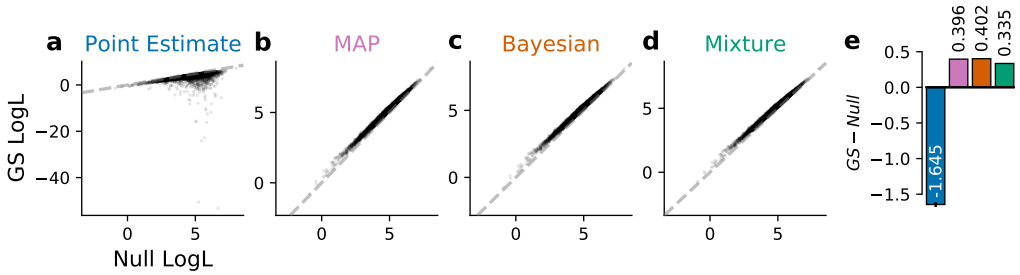

Fig. S2: Comparison of different methods to obtain an upper bound (GS): **a–d:** Various GS model log-likelihood vs. the Null model log-likelihood. Data is per neuron (averaged over repeats and stimuli). **e:** The full Bayesian Posterior Predictive outperforms the Point Estimate, the Maximum a posteriori (MAP), and the Mixture model. Each bar is the difference between the corresponding GS model and the Null model, averaged over repeats, stimuli, and neurons. Error bars correspond to the SEM and evaluate to ± 0.03 for the PE and ± 0.002 for the other GS models.

J NInGa across different datasets

We performed an analysis similar to Fig. 4c (blue bar) but for multiple datasets. We trained the same model described in section 3.2 on five different additional datasets from [23]. Our results show that using NInGa allows a better comparison of models that are trained on different datasets which can exhibit different levels of achievable performance (Fig. S3, compare left vs. right) . When models with the same architecture are trained on different datasets the resulting performances are more similar in NInGa (Eq. 1) than in the unnormalized IG (i.e. the numerator of Eq. 1), because the performance of the model is reported relative to the Null and Gold Standard model.

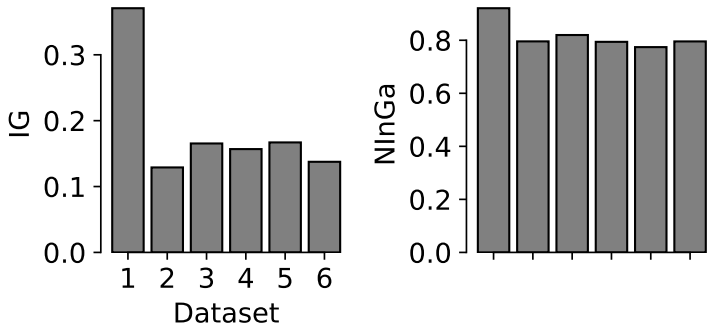

Fig. S3: Comparison of Trained Model performance on different datasets. **Left:** Models evaluated on simple Information Gain (IG), i.e. the numerator of Eq. 1. **Right:** Models evaluated on Normalized Information Gain (NInGa), i.e. the full Eq. 1.