

# Statistical analysis and optimality of neural systems

## Highlights

- Optimality theories of neural function can be formulated as statistical priors
- This formulation enables inferences about optimality directly from experimental data
- The resulting framework is applicable to a broad range of biological datasets

## Authors

Wiktor Młynarski, Michal Hledík,  
Thomas R. Sokolowski, Gašper Tkačik

## Correspondence

wiktor.mlynarski@ist.ac.at

## In Brief

Młynarski et al. develop a new statistical methodology that brings together “top-down” theories of biological optimality and “bottom-up” data analysis. This new framework allows one to make inferences about the function and constraints of neural systems directly from the experimental data.



## Article

# Statistical analysis and optimality of neural systems

Wiktor Młynarski,<sup>1,2,4,\*</sup> Michal Hledik,<sup>1,2</sup> Thomas R. Sokolowski,<sup>1,3</sup> and Gašper Tkačik<sup>1</sup><sup>1</sup>Institute of Science and Technology Austria, Am Campus 1, 3400 Klosterneuburg, Austria<sup>2</sup>These authors contributed equally<sup>3</sup>Present address: Frankfurt Institute for Advanced Studies, Ruth-Moufang-Straße 1, 60438 Frankfurt am Main, Germany<sup>4</sup>Lead contact\*Correspondence: [wiktor.mlynarski@ist.ac.at](mailto:wiktor.mlynarski@ist.ac.at)<https://doi.org/10.1016/j.neuron.2021.01.020>

## SUMMARY

Normative theories and statistical inference provide complementary approaches for the study of biological systems. A normative theory postulates that organisms have adapted to efficiently solve essential tasks and proceeds to mathematically work out testable consequences of such optimality; parameters that maximize the hypothesized organismal function can be derived *ab initio*, without reference to experimental data. In contrast, statistical inference focuses on the efficient utilization of data to learn model parameters, without reference to any *a priori* notion of biological function. Traditionally, these two approaches were developed independently and applied separately. Here, we unify them in a coherent Bayesian framework that embeds a normative theory into a family of maximum-entropy “optimization priors.” This family defines a smooth interpolation between a data-rich inference regime and a data-limited prediction regime. Using three neuroscience datasets, we demonstrate that our framework allows one to address fundamental challenges relating to inference in high-dimensional, biological problems.

## INTRODUCTION

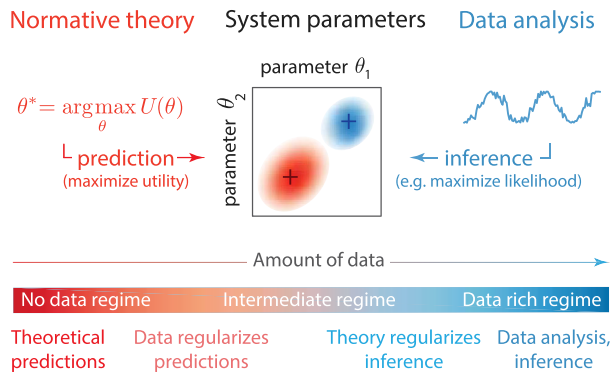
Ideas about optimization are at the core of how we approach biological complexity (Rosen, 2013; Bialek, 2012; Tkačik and Bialek, 2016). Quantitative predictions about biological systems have been successfully derived from first principles in the context of efficient coding (Laughlin, 1981; van Hateren, 1992); metabolic (Kacser and Burns, 1995; Ibarra et al., 2002), reaction (Savir et al., 2010; Tkačik et al., 2008), and transport (Tero et al., 2010) networks; evolution (Orzack, 2001); reinforcement learning (Alexander, 2003); and decision making (Geisler, 2011; Gold and Shadlen, 2007) by postulating that a system has evolved to optimize some utility function under biophysical constraints. Normative theories generate such predictions about living systems *ab initio*, with no (or minimal) appeal to experimental data. However, as such theories become increasingly high-dimensional and optimal solutions stop being unique, it gets progressively harder to judge whether theoretical predictions are consistent with data (Doi et al., 2012; Bittner et al., 2019) or to define rigorously what that even means (Wang et al., 2016; Park and Pillow, 2017; Eichhorn et al., 2009). Alternatively, data may be “close to” but not “at” optimality, and different instances of the system may show variation “around” optima (Pérez-Escudero et al., 2009; De Martino et al., 2018), but we lack a formal framework to address such scenarios. Lastly, normative theories typically make non-trivial predictions only under quantitative constraints, which ultimately must have an empirical origin, blurring the ideal-

ized distinction between a data-free normative prediction and a data-driven statistical inference.

In contrast to normative theories, which derive system parameters *ab initio*, the fundamental task of statistical inference is to reliably estimate model parameters from experimental observations. Here, too, biology has presented us with new challenges. While data are becoming increasingly high dimensional, they are not correspondingly more plentiful; the resulting curse of dimensionality that statistical models face is controlled by neither intrinsic symmetries nor the simplicity of disorder, as in statistical physics. To combat these issues and simultaneously deal with the noise and variability inherent to the experimental process, modern statistical methods often rely on prior assumptions about system parameters. These priors act as statistical regularizers to prevent overfitting or to capture low-level regularities such as smoothness, sparseness, or locality (Park and Pillow, 2011). Typically, however, their statistical structure is simple and does not reflect prior knowledge about system function.

Normative theories and inference share a fundamental similarity: they both make statements about parameters of biological systems. While these statements have traditionally been made in opposing “data regimes” (Figure 1), we observe that the two approaches are not exclusive and could in fact be combined with mutual benefit. To this end, we developed a Bayesian statistical framework that combines data likelihood with an “optimization prior” derived from a normative theory; contrary to simple, typically applied priors, optimization priors can induce a





**Figure 1. Normative theories and statistical inference**

Both approaches make statements about values of system parameters (center row; center panel). Normative theories predict which parameters would be of highest utility to the system (center row in red; left panel) without reference to experimental data. Data analysis infers parameter values from experimental observations (center row in blue; right panel). Large amounts of data support reliable inference of parameters. We consider a continuum of regimes that are applicable with different amounts of data (bottom row).

complex statistical structure on the space of parameters. This construction allows us to rigorously formulate and answer the following key questions: (1) Can one derive a statistical hypothesis test for the consistency of data with a proposed normative theory? (2) Can one define how close data are to the proposed optimal solution? (3) How can data be used to set the constraints in and resolve the degeneracies of a normative theory? (4) To what extent do optimization priors aid inference in high-dimensional statistical models?

The primary focus of this work is to develop conceptual and theoretical links between normative theories and statistical analyses. We illustrate the application of these developments to simple model systems and demonstrate their relevance to real-world data analysis on three diverse, yet still relatively tractable, examples. Applying similar methodology to large-scale high-dimensional data would necessitate the further development of sophisticated computational or approximative schemes. We recognize that as an outstanding and highly relevant challenge for future research.

## RESULTS

### Bayesian inference and optimization priors

Given a probabilistic model for a system of interest,  $P(x|\theta)$ , with parameters  $\theta$  and a set of  $T$  observations (or data)  $\mathcal{D} = \{x_t\}_{t=1}^T$ , Bayesian inference consists of formulating a (log) posterior over parameters given the data:

$$\log P(\theta|\mathcal{D}) = \log \mathcal{L}(\theta) + \log P(\theta) + \text{const}, \quad (\text{Equation 1})$$

where the constant term is independent of the parameters,  $\mathcal{L}(\theta) = \prod_{t=1}^T P(x_t|\theta)$  is the likelihood assuming independent and identically distributed observations  $x_t$ , and  $P(\theta)$  is the prior or the postulated distribution over the parameters in the absence of any observation. Much work has focused on how the prior

should be chosen to permit optimal inference, ranging from uninformative priors (Jeffreys, 1946), priors that regularize the inference and thus help models generalize to unseen data (Mackay, 2003; Murphy, 2012), or priors that can coarse-grain the model depending on the amount of data samples,  $T$  (Machta et al., 2013).

Our key intuition will lead us to a new class of priors that are fundamentally different from those considered previously. A normative theory for a system of interest with parameters  $\theta$  typically can be formalized through a notion of a (upper-bounded) utility function,  $U(\theta; \xi)$ , where  $\xi$  are optional parameters that specify the properties of the utility function itself. Optimality then amounts to the assumption that the real system operates at a point in parameter space,  $\theta^*$ , that maximizes utility,  $\theta^*(\xi) = \text{argmax}_{\theta} U(\theta; \xi)$ . Viewed in the Bayesian framework, the assertion that the system is optimal thus represents an infinitely strong prior where the parameters are concentrated at  $\theta^*$ —in other words,  $P(\theta|\xi) = \delta(\theta - \theta^*(\xi))$ . In this extreme case, no data are needed to determine system parameters; the prior fixes their values and typically no finite amount of data will suffice for the likelihood in Equation 1 to move the posterior away from  $\theta^*$ . This concentrated prior can, however, be interpreted as a limiting case of a softer prior that “prefers” solutions close to the optimum.

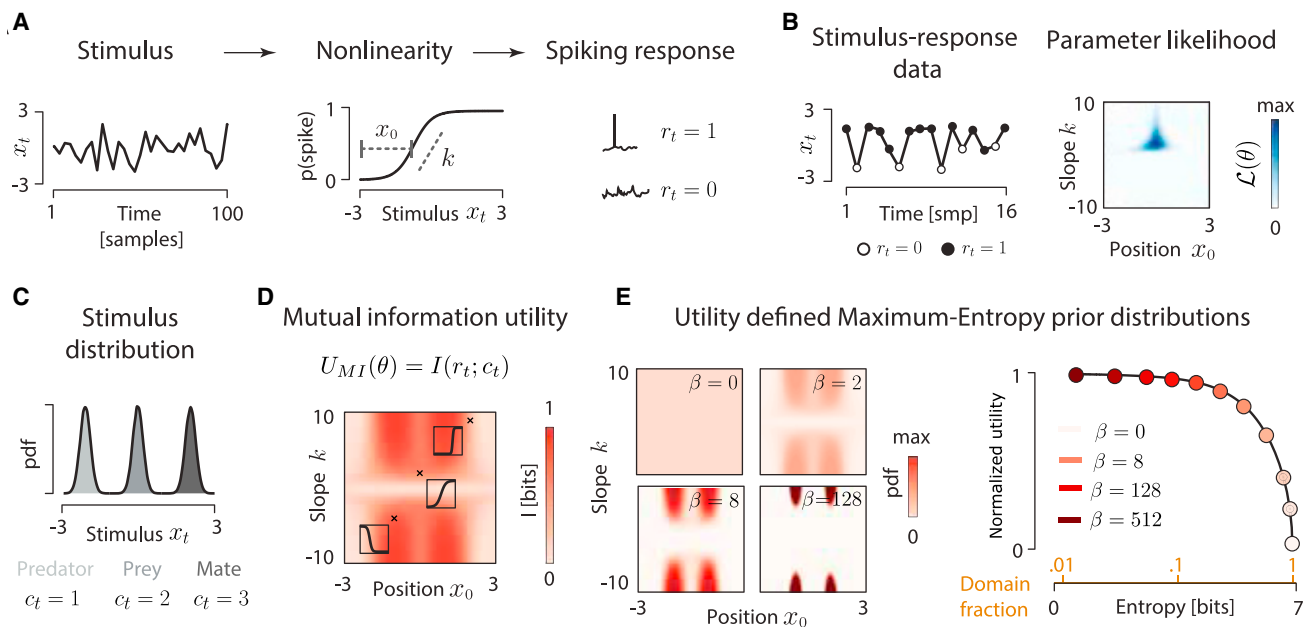
Consistent with the maximum entropy principle put forward by Jaynes (2003), we therefore consider for our prior distributions that are as random and unstructured as possible while attaining a prescribed average utility:

$$P(\theta|\beta, \xi) = \frac{1}{Z(\beta, \xi)} \exp[\beta U(\theta; \xi)]. \quad (\text{Equation 2})$$

This is in fact a family of priors, whose strength is parametrized by  $\beta$ ; when  $\beta = 0$ , parameters are distributed uniformly over their domain without any structure and in the absence of any optimization. As  $\beta \rightarrow \infty$ , parameter probability localizes at the point  $\theta^*(\xi)$  that maximizes the utility to  $U_{\max}(\xi)$  (if such a point is unique), irrespective of whether data support this or not. At finite  $\beta$ , however, the prior is “smeared” at  $\sim \theta^*(\xi)$  so that the average utility,  $\bar{U}(\beta, \xi) = \int d\theta P(\theta|\beta, \xi) U(\theta, \xi) < U_{\max}(\xi)$  increases monotonically with  $\beta$ . For this reason, we refer to  $\beta$  as the “optimization parameter,” and to the family of priors in Equation 2 as “optimization priors.”

The intermediate regime,  $0 < \beta < \infty$ , in the prior entering Equation 1 is interesting from an inference standpoint. It represents the belief that the system may be “close to” optimal with respect to the utility  $U(\theta; \xi)$  but this belief is not absolute and can be outweighed by the data: the log likelihood,  $\log \mathcal{L}$ , grows linearly with the number of observations,  $T$ , matching the roughly linear growth of the log prior with  $\beta$ . Varying  $\beta$  thus literally corresponds to the interpolation between an infinitely strong optimization prior and pure theoretical prediction in the “no data regime” and the uniform prior and pure statistical inference in the “data rich regime,” as schematized in Figure 1.

Additional parameters of the utility function,  $\xi$ , determine its shape in the domain of parameters  $\theta$ . Parameters  $\xi$  can be known and fixed for a specific theory or, if unknown *a priori*, inferred from the data in a Bayesian fashion. When there are no



**Figure 2. Efficient coding in a toy model neuron and the corresponding optimization prior**

(A) Model neuron uses a logistic nonlinearity (middle panel) to map continuous stimuli  $x_t$  (left panel) to a discrete spiking response  $r_t$  (right panel). The shape of the nonlinearity is described by 2 parameters: slope  $k$  and offset  $x_0$ .  
 (B) An example dataset (left panel) consisting of stimulus values (black line) and associated spiking responses (empty circles, no spike; full circles, spike). Likelihood function of the nonlinearity parameters defined by the observed data. Dark blue corresponds to most likely parameter values.  
 (C) Distribution of natural stimuli to which the neuron may be adapted. In this example, each mode corresponds to a behaviorally relevant state of the environment: presence of a predator, a prey, or a mate.  
 (D) Efficient coding utility function, here, the mutual information between neural response  $r_t$  and the state of the environment,  $c_t$ , with stimuli drawn from the distribution in (C). The amount of information conveyed by the neuron depends on the position and slope of the nonlinearity. Insets depict example nonlinearities corresponding to parameter values marked with black crosses.  
 (E) Four maximum-entropy optimization priors over parameters for the neural nonlinearity (left panel). Distributions are specified by the utility of each slope-offset combination. Increasing parameter  $\beta$  constrains the distribution (lowers its entropy) and increases the expected utility of the parameters (right panel). Here, we plot the normalized utility  $U(\theta)$ , see main text for explanation. Orange numbers on the horizontal axis specify the fraction of the entire domain effectively occupied by parameters at given  $\beta$ .

utility parameters  $\xi$  to consider, we suppress them for notational simplicity.

In the following, we apply this framework to a toy model system, a single linear-nonlinear neuron, which is closely related to logistic regression. This example is simple, well understood across multiple fields, and low-dimensional so that all mathematical quantities can be constructed explicitly; the framework itself is, however, completely general. We then apply our framework to a more complex neuron model and to three experimental datasets. These examples demonstrate how the ability to encode the entire shape of the utility measure into the optimization prior opens up a more refined and richer set of optimality-related statistical analyses.

### Example: efficient coding in a simple model neuron

Let us consider a simple probabilistic model of a spiking neuron (Figure 2A), a broadly applied paradigm in sensory neuroscience (Sharpee and Bialek, 2007; Kastner et al., 2015; Paninski et al., 2007; Tkačik et al., 2010; Gjorgjieva et al., 2014). The neuron responds to one-dimensional continuous stimuli  $x_t$  either by eliciting a spike ( $r_t = 1$ ) or by remaining silent ( $r_t = 0$ ). The probability

of eliciting a spike in response to a particular stimulus value is determined by the nonlinear saturating stimulus-response function. The shape of this function is determined by two parameters: position  $x_0$  and slope  $k$  (see Method details).

Parameters  $\theta = \{x_0, k\}$  fully determine the function of the neuron yet remain unknown to the external observer. Statistical inference extracts parameter estimates  $\hat{\theta}$  using experimental data  $\mathcal{D}$  consisting of stimulus-response pairs (Figure 2B, left panel), by first summarizing the data with the likelihood  $\mathcal{L}(\theta)$  (Figure 2B, right panel), followed either by maximization of the likelihood  $\hat{\theta} = \text{argmax}_{\theta} \mathcal{L}(\theta)$  in the maximum-likelihood (ML) paradigm or by deriving  $\hat{\theta}$  from the posterior (Equation 1), in the Bayesian paradigm.

To apply our reasoning, we must propose a normative theory for neural function, for the optimization prior, and combine it with the likelihood in Figure 2B, as prescribed by the Bayes rule in Equation 1. An influential theory in neuroscience called efficient coding postulates that sensory neurons maximize the amount of information about natural stimuli that they encode into spikes given biophysical constraints (Barlow, 1961; van Hateren, 1992; Tkačik et al., 2010; Olshausen and Field, 1996; Smith and

Lewicki, 2006; Chalk et al., 2018). This information-theoretic optimization principle (Shannon, 1948) has correctly predicted neural parameters such as receptive field (RF) shapes (Olshausen and Field, 1996; Hyvärinen et al., 2009) and the distribution of tuning curves (Ganguli and Simoncelli, 2014; Wang et al., 2016), as well as other quantitative properties of sensory systems (Laughlin, 1981; Ratliff et al., 2010; Borghuis et al., 2008; Młynarski, 2015; Młynarski and McDermott, 2018; Carlson et al., 2012), *ab initio*, from the distribution of ecologically relevant stimuli (Olshausen and Field, 1996; Bialek, 2012).

To apply efficient coding, we need to specify a distribution from which the stimuli  $x_t$  are drawn. In reality, neurons would respond to complex and high-dimensional features of sensory inputs, such as a particular combination of odorants, timbre of a sound, or a visual texture, to help the animal discriminate between environmental states of very different behavioral relevance (e.g., a presence of a predator, prey, or a mate). To capture this intuition in our simplified setup, we imagine that the stimuli  $x_t$  are drawn from a multi-modal distribution, which is a mixture of three different environmental states, labeled by  $c_t$  (Figure 2C). Efficient coding then postulates that the neuron maximizes the mutual information,  $I(r_t; c_t)$ , between the environmental states,  $c_t$ , that gave rise to the corresponding stimuli,  $x_t$ , and the neural responses,  $r_t$ .

Mutual information, which can be evaluated for any choice of parameters  $k, x_0$  provides the utility function,  $U_{MI}(k, x_0) = I(r_t; c_t)$ , relevant to our case; in this simple example, the utility function has no extra parameters  $\xi$ . Figure 2D shows that  $U_{MI}$  is bounded between 0 and 1 bit (since the neuron is binary), but does not have a unique maximum. Instead, there are four combinations of parameters that define four degenerate maxima, corresponding to the neuron's nonlinearity being as steep as possible (high positive or negative  $k$ ) and located in any of the two "valleys" in the stimulus distribution (red peaks in Figure 2D). Moreover, the utility function forms broad ridges on the parameter surface, and small deviations from optimal points result only in weak decreases of utility. Consequently, formulating clear and unambiguous theoretical predictions is difficult, an issue that has been recurring in the analysis of real biological systems (Brinkman et al., 2016; Pitkow and Meister, 2012).

Given the utility function, the construction of the maximum-entropy optimization prior according to Equation 2 is straightforward. Explicit examples for different values of  $\beta$  are shown in Figure 2E (left panel). In general, the average utility of the prior monotonically increases as the prior becomes more localized around the optimal solutions, as measured by the decrease in entropy of the prior (Figure 2E, right panel). This can be interpreted as restricting the system into a smaller part of the parameter domain. If an increase in average utility requires a reduction in entropy by 1 bit, then this means that the parameters will be sampled from at most half the available domain.

Before proceeding, we note that our approach depends on several non-trivial choices. First, the fact that system parameterization and the size of the parameter domain can affect Bayesian inferences is well recognized (Gelman, 2004) and we discuss how it relates to our case in the Supplemental information (Methods S1 and S3; Figures S1 and S2). Second,  $\beta$  and the utility function enter the optimization prior of Equation 2 as a prod-

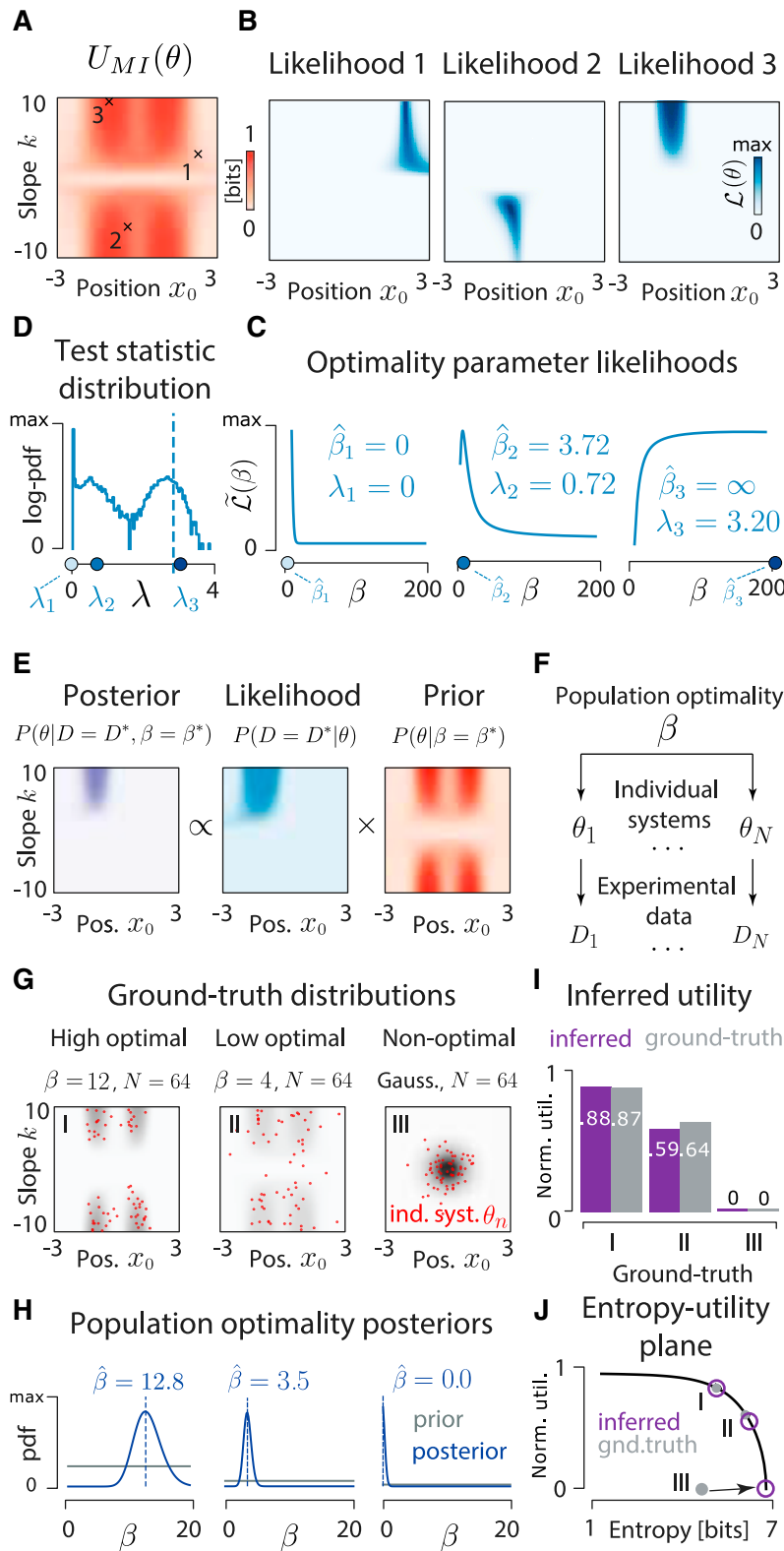
uct, leaving the scale of each quantity arbitrary. For interpretation purposes, we therefore define the normalized utility,  $\tilde{U} = (\bar{U}(\beta) - \bar{U}(\beta = 0)) / (U_{\max} - \bar{U}(\beta = 0))$ , which takes on values between 0 and 1 for non-negative  $\beta$  and is insensitive to linear scaling. We discuss the issue of  $\beta$  scaling in the Supplemental information (Method S4). Third, data and optimality theories could be combined in multiple ways. However, combining them via maxent optimization priors enjoys favorable theoretical guarantees that alternative approaches may lack, which we demonstrate in the Supplemental information (Method S5; Figures S4 and S5). These considerations complete our setup and allow us to address the four questions posed in the Introduction.

### Question 1: statistical test for the optimality hypothesis

Given a candidate normative theory and experimental data for a system of interest, a natural question arises: do the data support the postulated optimality? This question is non-trivial for two reasons. First, optimality theories typically do not specify a sharp boundary between optimal and non-optimal parameters, but rather a smooth utility function  $U(\theta)$  (Figure 3A); how should the test for optimality be defined in this case? Second, a finite dataset  $\mathcal{D}$  may be insufficient to infer a precise estimate of the parameters  $\theta$ , but will instead yield a (possibly broad) likelihood surface (Figure 3B); how should the test for optimality be formulated in the presence of such uncertainty?

Here, we devise an approach to address both issues. The basis of our test is a null hypothesis that the system is not optimized (i.e., that its parameters have been generated from a uniform random distribution on the biophysically accessible parameter domain). This distribution is exactly the optimization prior  $P(\theta|\beta=0)$ . The alternative hypothesis states that the parameters are drawn from a distribution  $P(\theta|\beta)$ , with  $\beta > 0$ . To discriminate between the two hypotheses, we use a likelihood ratio test with the statistic  $\lambda$ , which probes the overlap of high-likelihood and high-utility parameter regions. Specifically, we define the marginal likelihood of  $\beta$  given data,  $\tilde{\mathcal{L}}(\beta) = P(\mathcal{D}|\beta) = \int d\theta \mathcal{L}(\theta) P(\theta|\beta)$  (Figure 3C), and then define  $\lambda$  as the log ratio between the maximal marginal likelihood,  $\max_{\beta > 0} \tilde{\mathcal{L}}(\beta)$ , and the marginal likelihood under the null hypothesis,  $\tilde{\mathcal{L}}(\beta=0)$  (see Method details). Here, we assumed for simplicity that the utility function  $U$  does not depend on any additional parameters  $\xi$ ; this simplification is relaxed in the Supplemental information (Method S2; Figure S3).

The test statistic  $\lambda$  has a null distribution that can be estimated by sampling (Figure 3D), with large  $\lambda$  implying evidence against the null hypothesis; thus, given a significance threshold, we can declare the system to show a significant degree of optimization or to be consistent with no optimization. This is different from asking whether the system is "at" an optimum; such a narrow view seems too restrictive for complex biological systems (Barton and de Vladar, 2009; Wright, 1937). Evolution, for example, may not have pushed the system all the way to the biophysical optimum (e.g., due to mutational load or because the adaptation is still ongoing), or the system may be optimal under utility function or resource constraints slightly different from those postulated by our theory (De Martino et al., 2018). Instead, the proposed test asks whether the system has relatively high utility, compared to the utility distribution in the full parameter space.



**Figure 3. Statistical test and inference of the degree of optimality**

(A) Utility function  $U_{MI}(k, x_0)$ . Crosses and numbers show the locations of ground truth parameters.

(B) Likelihood of the nonlinearity parameters obtained from 20 stimulus-response  $(x_i, r_i)$  pairs. The 3 examples correspond to 3 ground truth parameter values (black crosses in A), and are ordered by increasing utility.

(C) Marginal likelihood of the optimality parameter  $\beta$ ,  $\tilde{\mathcal{L}}(\beta) = P(\mathcal{D}|\beta)$ , corresponding to data in (A). Maximum likelihood (ML) estimates  $\hat{\beta}_{1,2,3}$  (blue circles) indicate that the data would be most probable with no preference for high utility  $U_{MI}$  (left panel,  $\hat{\beta}_1 = 0$ ; note that we do not allow negative  $\hat{\beta}$ ), some preference for high  $U_{MI}$  (center panel,  $\hat{\beta}_2 > 0$  finite), and strong preference for high  $U_{MI}$  (right panel,  $\hat{\beta}_3 \rightarrow \infty$ ; blue circle displayed at  $\beta = 200$  for illustration purposes). Likelihood ratio statistic  $\lambda_{1,2,3}$  compares the marginal likelihood of  $\beta$  at  $\beta = 0$  versus  $\beta = \hat{\beta}_{1,2,3}$  (see Method details).

(D) Null distribution of the test statistic  $\lambda$ . Point mass at  $\lambda = 0$  corresponds to cases in which the ML optimality parameter is zero,  $\hat{\beta} = 0$ . High values of  $\lambda$  are evidence against the null hypothesis that  $\beta = 0$ , and hence support optimality. The dashed vertical line represents the  $p = 0.05$  significance threshold; blue circles show  $\lambda_{1,2,3}$ . Only  $\lambda_3$  crosses the threshold, indicating significant preference for high utility parameters.

(E) Posterior over nonlinearity parameters, inferred for a single system with a utility-derived prior at fixed optimality parameter,  $\beta = \beta^*$ .

(F) A hierarchical model of a population of optimized systems. Population optimality parameter  $\beta$  controls the distribution of parameters for individual systems ( $n = 1, \dots, N$ ),  $\theta_n$ , which give rise to observed data,  $\mathcal{D}_n$ .

(G) Nonlinearity parameters (64 red dots per distribution) sampled from 3 different ground truth distributions (denoted by roman numerals in G–J): a strongly optimized population ( $\beta = 12$ ; left), a weakly optimized population ( $\beta = 4$ ; center), and a non-optimal distribution (Gaussian distribution; right). For each model neuron  $\theta_n$ , data  $\mathcal{D}_n$  consists of 100 stimulus-response pairs.

(H) Results of hierarchical inference. Posteriors over  $\beta$  (purple lines) and MAP estimates,  $\hat{\beta}$  (dashed purple lines) were obtained using simulated data from (G). Priors (gray lines) were uniform on the  $[0, 20]$  interval.

(I) Normalized utility  $\bar{U}$ . Estimated values (purple bars) closely match ground truth (gray bars).

(J) Entropy and normalized utility of ground truth distributions (gray, filled circles) and inferred distributions parametrized by  $\hat{\beta}$  (purple, empty circles).

While principled, this hypothesis test is computationally expensive, since it entails an integration over the whole parameter space to compute the marginal likelihoods,  $\tilde{\mathcal{L}}(\beta)$ , as well as Monte Carlo sampling to generate the null distribution. The first difficulty can be resolved when the number of observations  $T$  is sufficient such that the likelihood of the data,  $\mathcal{L}(\theta)$ , is sharply localized in the parameter space; in this case, the value of the utility function at the peak of the likelihood itself becomes the test statistic and the costly integration can be avoided (see [Method details](#)). The second difficulty can be resolved when we can observe many systems and collectively test them for optimality; in this case, the distribution of the test statistic approaches the standard  $\chi^2$  distribution (see [Method details](#)).

### Question 2: inferring the degree of optimality

Hypothesis testing provides a way to resolve the question of whether the data provide evidence for system optimization (or to quantify this evidence with a p value). However, statistical significance does not necessarily imply biological significance: with sufficient data, rigorous hypothesis testing can support the optimality hypothesis, even if the associated utility increase is too small to be biologically relevant. Therefore, we formulate a more refined question: how strongly is the system optimized with respect to a given utility,  $U(\theta)$ ?

Methodologically, we are asking about the value of the optimization parameter,  $\beta$ , that is supported by the data  $\mathcal{D}$ . In the standard Bayesian approach, all of the parameters of the prior are considered fixed before performing the inference; the prior is then combined with likelihood to generate the posterior ([Figure 3E](#)). Our case corresponds to a hierarchical Bayesian scenario, where  $\beta$  is itself unknown and of interest. In the previous section, we chose it by maximizing the marginal likelihood,  $\tilde{\mathcal{L}}(\beta)$ , to devise a yes/no hypothesis test. Here, we consider a fully Bayesian treatment, which is particularly applicable when we observe many instances of the same system. In this case, we interpret different instances (e.g., multiple recorded neurons) as samples from a distribution determined by a single population optimality parameter  $\beta$  ([Figure 3F](#)) that is to be estimated. Stimulus-response data from multiple neurons are then used directly to estimate a posterior over  $\beta$  via hierarchical Bayesian inference.

To explore this possibility, we generate parameters  $\theta_n$  of  $n = 1, \dots, N$  model neurons from three different distributions: strongly optimized ( $\beta = 12$ ; [Figure 3G](#), left panel), weakly optimized ( $\beta = 4$ ; [Figure 3G](#), center panel), and non-optimal (Gaussian distribution of parameters; [Figure 3G](#), right panel). For each of the three examples, we simulate stimulus-response data for all of the neurons and use these data in a standard hierarchical Bayesian inference to compute posterior distributions over the population optimality parameter,  $\beta$  ([Figure 3H](#); see [Method details](#)).

Following hierarchical inference, we can interpret the inferred population optimality parameter  $\hat{\beta}$  by mapping it onto normalized utility (*cf.* [Figure 2E](#)). This reports optimality on a  $[0, 1]$  scale, with 1 corresponding to the maximum achievable utility  $U_{\max}$ , and thus a fully optimal system, and 0 corresponding to the average utility under random parameter sampling,  $\bar{U}(\beta = 0)$ . Normalized utility for the three examples is shown in [Figure 3I](#).

Our framework enables us to draw inferences about optimality that are not possible otherwise. For example, in addition to esti-

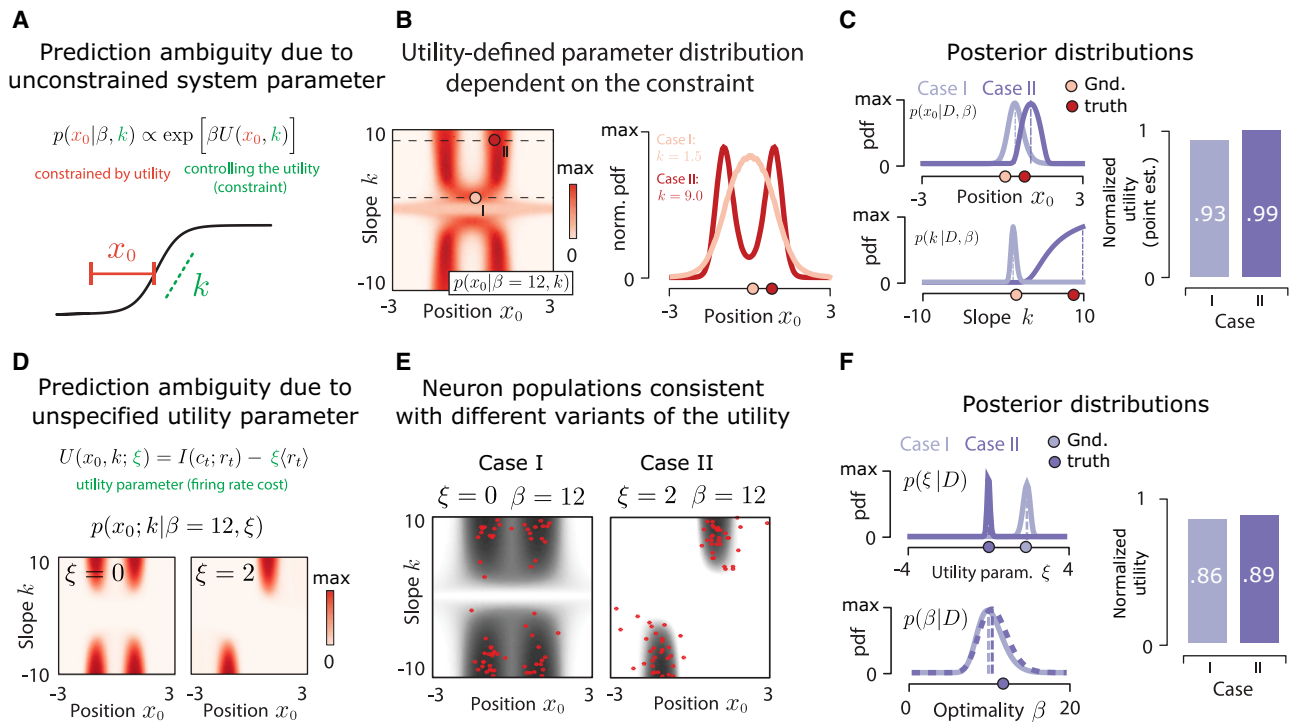
imating the normalized utility, we can also quantify how restrictive the optimization needs to be to achieve that level of utility. This restriction is measured by the entropy associated with  $\hat{\beta}$  ([Figure 3J](#)). In example I from [Figures 3G–3I](#),  $\hat{\beta} = 12.8$  is associated with a decrease in entropy of  $\sim 1.75$  bits compared to  $\beta = 0$ , meaning that nonlinearity parameters are effectively restricted to a fraction  $\sim 2^{-1.75} \approx 0.3$  of the parameter domain. Example III with  $\hat{\beta} = 0$  is consistent with a high-entropy optimization prior and indicates almost no parameter space restriction. This is despite the fact that the actual parameters were sampled from a Gaussian highly concentrated (i.e., with low entropy) in the parameter space, but not in a region of high utility. This mismatch suggests that such a system could be optimized for a different utility function or shaped by other constraints. The system could also be anti-optimized (i.e., prefer negative values of  $\tilde{U}$ ), which could easily be identified by permitting negative  $\beta$  values during inference. Another clear benefit of the probabilistic framework is the possibility of computing uncertainty estimates of  $\beta$  and the associated utility and entropy.

### Question 3: data resolve ambiguous theoretical predictions

Predictions derived from optimality theories can be non-unique and ambiguous. This ambiguity can manifest itself in different ways.

The first kind of ambiguity results from the existence of multiple maxima of the utility function. Before formulating statistical questions, it is important to pause and clarify the underlying biological context: could different observed instances of the system freely sample from all utility maxima (as in [Figure 3G](#), example I), or is a single maximum relevant, perhaps because it is the only one that nature realized by evolutionary adaptation? In the latter case, the first task of statistical analysis is to identify that single maximum. For low-dimensional systems, this ambiguity can be resolved trivially; in our toy model, for example, a few data points suffice to zero in on one of the four degenerate utility maxima ([Method S6](#); [Figure S6](#)). In contrast, in high-dimensional parameter spaces, the task of finding the “closest optimum” is non-trivial ([Doi et al., 2012](#)) and could be aided by sampling methods derived from optimization priors, which is a topic for further research.

The second kind of ambiguity results from system parameters that enter the utility function but are unconstrained by the optimization theory in question. Such parameters limit the performance of the whole system, with the utility typically achieving its global maximum when they take on extremal values (e.g.,  $\pm\infty, 0$ ); yet, these extremal values often correspond to physically implausible scenarios (e.g., infinite averaging time or energy consumption, zero noise, instantaneous response time). Optimization theory cannot make a non-trivial prediction about these parameters, so they must either be fixed *a priori* based on known external constraints or inferred from data simultaneously with the optimization of the remaining parameters. An additional subtlety comes into play when we analyze multiple instances of a system (e.g., neurons); either each individual neuron has its own value of the constraint parameter, to be determined from data (which we address in the following paragraph), or all neurons share a single value of the constraint that needs to be inferred jointly.



**Figure 4. Resolving ambiguities of theoretical predictions**

(A) Prediction ambiguity due to an unconstrained system parameter. Utility is evaluated over the position parameter  $x_0$  (red), with the slope parameter  $k$  (green) interpreted as an externally imposed biophysical constraint.  $k$  is inferred from data for each neuron separately; for different  $k$ , optimality may predict different optimal positions,  $x_0$ .

(B) Optimization priors for  $x_0$  are conditional maxent distributions over  $x_0$  parametrized by values of  $k$  (rows of the matrix, here at fixed  $\beta = 12$ ) (left). Distributions over  $x_0$  for 2 example values of  $k$  (dashed black lines at left) are displayed in the right panel, with optimal  $x_0$  values marked (pink and red circles for cases I and II, respectively).

(C) Posteriors over the position ( $x_0$ , left column, top) and the slope ( $k$ , left column, bottom) parameters, estimated for cases I and II (light and dark purple lines, respectively; dashed lines, MAP estimates), by marginalizing the joint posterior. Ground-truth values are marked with circles. Normalized utility of  $x_0$ , relative to the maximal utility for  $k$  inferred separately for cases I and II.

(D) Prediction ambiguity due to an unspecified utility function. Utility prefers high mutual information  $I$  at a low average firing rate ( $r$ ), with an unknown trade-off parameter  $\xi$ . Optimization prior with no firing rate constraint (left,  $\xi = 0$ ) shows 4 degenerate maxima; the constraint (right,  $\xi = 2$ ) partially lifts the degeneracy.

(E) Two ground truth distributions (gray) corresponding to different values of the firing rate constraint  $\xi$ . Red dots denote  $N = 64$  sample neurons.

(F) Posteriors over the firing rate constraint  $\xi$  (left column, top) and the optimality parameter  $\beta$  (left column, bottom), estimated for cases I and II (light and dark purple lines, respectively; dashed lines, MAP estimates), by marginalizing the joint posterior. Ground-truth values are marked with circles. Normalized utilities computed for  $\xi$  inferred separately for cases I and II.

In our model, the nonlinearity slope  $k$  is unconstrained by optimization; mutual information increases monotonically as  $|k| \rightarrow \infty$  (Figure 4A). This corresponds to vanishing noise in neural spiking. Since such noise cannot physically vanish, we must change the interpretation of the utility function,  $U_{MI}(\theta)$ , and evaluate it only over positions  $x_0$ , while treating the slope  $k$  as a constraint to be fit from data, which we indicate by writing  $U_{MI}(x_0; k)$ . Here, slope  $k$  determines the entire shape of the utility function (Figure 4B). Unreliable neurons with a small slope have a unique optimal position  $x_0 = 0$ , while for neurons with large  $|k|$  the utility is bimodal, with optimal positions separating peaks of the stimulus distribution. As before, we can infer both parameters for a “noisy” (case I) and “precise” (case II) simulated neuron (Figure 4C); this time, however, the optimization prior acts only on  $x_0$ , while the prior over slope  $k$  remains uniform. To properly assess optimality, we must normalize the utility by the maximal utility achievable at the estimated

value of  $k$ :  $\tilde{U}(\hat{x}_0; \hat{k}) = (U(\hat{x}_0; \hat{k}) - \bar{U}(\beta = 0; \hat{k})) / (U_{\max}(x_0; \hat{k}) - \bar{U}(\beta = 0; \hat{k}))$ . In both cases, the relative utility exceeds 0.9 (Figure 4C). Because theoretical predictions now depend on the biophysical constraint—which itself is a free parameter adjustable separately for each system instance—high values of normalized utility can be achieved by neurons with very different  $x_0$ .

The third kind of ambiguity arises when the utility function itself depends on additional parameters,  $\xi$ . The mutual information utility  $U_{MI}$  of our toy model can be extended by considering the cost of neural spiking, resulting in a new compound function,  $U(x_0, k; \xi) = U_{MI}(x_0, k) - \xi \langle r_t \rangle$ , with the trade-off parameter  $\xi$ . Increasing  $\xi$  changes the shape of the new utility function (Figure 4D). Given multiple instances of a biological system (Figure 4E), we can ask about the most likely form of  $U$  (i.e., the single value of  $\xi$  shared across all instances of the system), together with the most likely value of the optimization parameter,  $\beta$ .

Note that such joint determination of  $\beta$  and  $\xi$  corresponds to answering question 2 (“Inferring the degree of optimality”), in the presence of ambiguity. This problem is solved by hyperparameter inference, which generates joint posteriors and maximum a posteriori (MAP) estimates of  $\beta$  and  $\xi$  (Figure 4F). Here, too, the normalized utilities are defined relative to the inferred value of  $\xi$  and can thus be comparable, even when the underlying utility functions are substantially different.

The difference between ambiguities of the second and third kind is subtle, yet important. Broadly speaking, the second kind of ambiguity arises if only a subset of system parameters  $\theta$  depends on the optimality parameter  $\beta$ , while the remaining parameters act as constraints that must be inferred. In the third kind of ambiguity, all system parameters  $\theta$  depend on the optimality parameter  $\beta$  and on additional parameters of the utility function  $\xi$ . The corresponding differences in parameter dependency patterns are summarized graphically in Method S6 and Figure S7.

#### Question 4: Optimization priors improve inference for high-dimensional problems

Here, we extend our toy model neuron with two parameters to a more realistic case with hundreds of parameters. We focus on a linear-nonlinear-Poisson (LNP) model (Paninski et al., 2007), whose responses to natural image stimuli are determined by a linear filter (also referred to as an RF),  $\phi \in \mathbb{R}^{16 \times 16}$  (Figure 5A). The purpose of this exercise is to show the tractability of our approach and the power of optimization priors for high-dimensional inference problems. Inference of neural filters,  $\phi$ , from data is a central data analysis challenge in sensory neuroscience, making our example practically relevant.

Experimentally observed filters  $\phi$  in the visual cortex have been suggested to maximize the sparsity of responses  $s_i$  to natural stimuli (Olshausen and Field, 1996). A random variable is sparse when most of its mass is concentrated around 0 at fixed variance. These experimental observations have been reflected in the normative model of sparse coding, in which the maximization of sparsity has been hypothesized to be beneficial for energy efficiency, flexibility of neural representations, and noise robustness (Hyvärinen et al., 2009; Olshausen and Field, 2004). Filters optimized for sparse utility  $U_{SC}(\phi)$  (see Method details) are oriented and localized in space and frequency (Figure 5B, leftmost panel) and famously resemble RFs of simple cells in the primary visual cortex (V1). A significant fraction of neural RFs, however, differ from optimally sparse filters (Ringach, 2002), perhaps due to the existence of additional constraints. One possible constraint is spatial locality, which leads to suboptimally sparse filters that increasingly resemble localized blobs (Doi and Lewicki, 2014), as shown in Figure 5B.

In our framework, sparse coding utility  $U_{SC}$  and locality  $U_{LO}$  combine into a single utility function with a parameter  $\xi$  that specifies the strength of the locality constraint. We wondered whether an optimization prior based on sparsity, even in the presence of an additional constraint of unknown strength, could successfully regularize the inference of linear filters,  $\phi$ .

We first consider a scenario in which the locality constraint is known *a priori* to equal zero. We simulate spike trains of 100 model neurons optimized under sparse utility  $U_{SC}$  responding to a sequence of 2,000 natural image patches (see Method de-

tails). Using these simulated data we infer the filter estimates,  $\hat{\phi}$ , using spike triggered average (STA) (Sharpee, 2013; Park and Pillow, 2017), which under our assumptions are equivalent to the maximum likelihood (ML) estimates (Paninski et al., 2007) (see Method details). STAs computed from limited data recover noisy estimates of neural filters (Figure 5C; column second from the left).

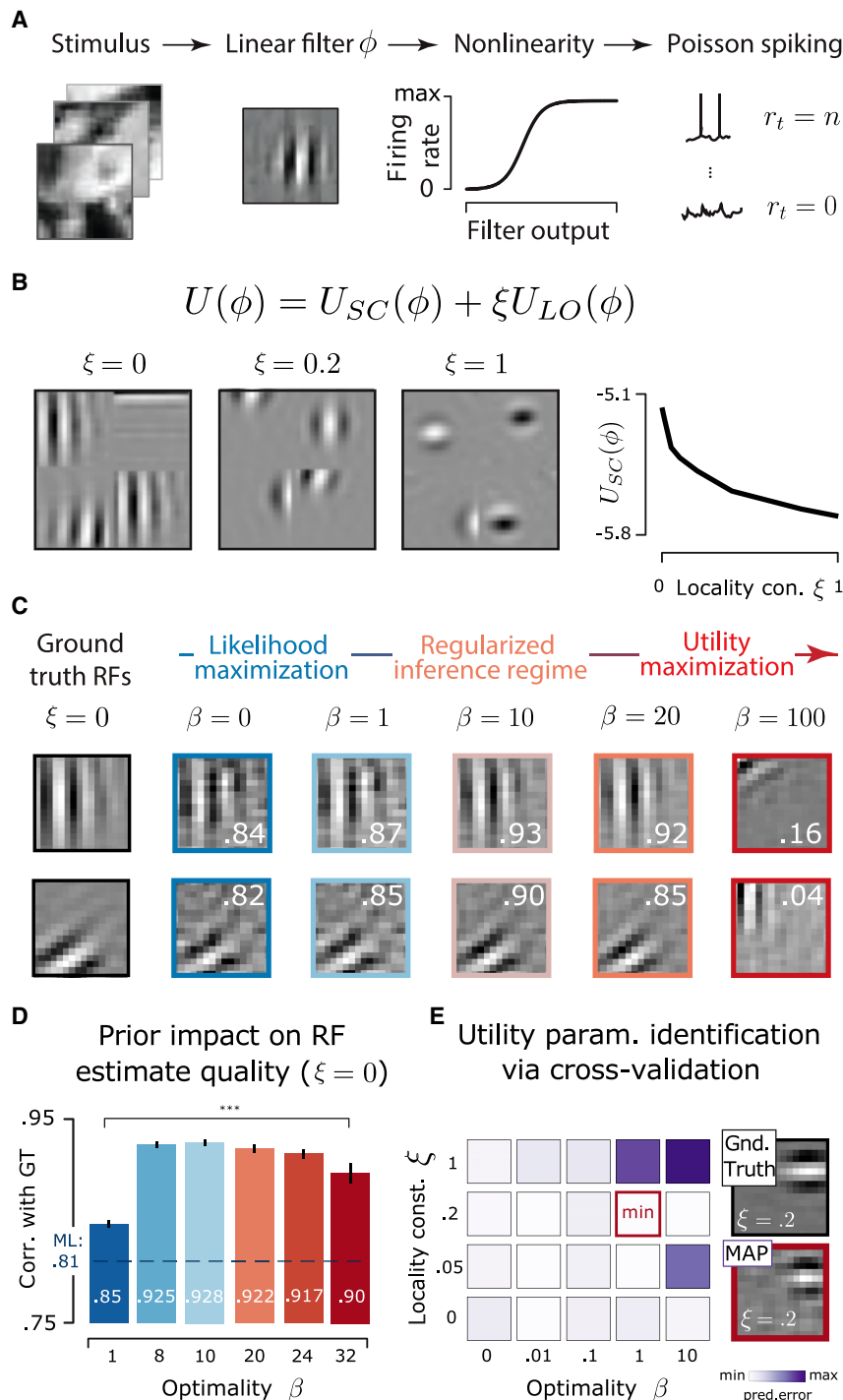
Can sparse coding provide a powerful prior to aid in the inference of high-dimensional filters? Using our sparse coding utility,  $U_{SC}(\phi)$ , we formulate optimization priors for various values of  $\beta$  and compute MAP filter estimates  $\hat{\phi}(\beta)$  from simulated data (Figure 5C; four rightmost columns; see Method details). Increasing values of  $\beta$  interpolate between pure data-driven ML estimation (Figure 5C, second column from the left) that ignores the utility and pure utility maximization (Figure 5C, right column) at very high  $\beta = 10^2$ , where the predicted filters become almost completely decoupled from data; these two regimes seem to be separated by a sharp transition. For intermediate  $\beta = 1, 10, 20$ , MAP filter estimates show a significant improvement in estimation performance relative to the ML estimate (Figure 5D).

We next consider a scenario in which the locality constraint is not known *a priori*, but can be identified together with the prior strength  $\beta$  using cross-validation (Kass et al., 2014), as described in question 3. To this end, we simulate responses of a single neuron whose filter was optimized with the locality constraint  $\xi = 0.2$  (Figure 5E, “Gnd. Truth”). We then use a subset of 1,800 of 2,000 stimulus-response pairs to compute the MAP estimate of the filter using a range of  $\beta$  and  $\xi$  values. Each MAP estimate of the filter is used to compute the prediction error for neural responses over withheld portion of the data. Cross-validation correctly identifies the true  $\xi$  and the optimal  $\beta$  values that minimize the prediction error (Figure 5E); the resulting filter estimate (Figure 5E, “MAP”) closely resembles the ground truth.

Optimization priors achieve a boost in performance because they quantitatively encode many characteristics we ascribe to the observed receptive fields (localization in space and bandwidth, orientation), which the typical regularizing priors (e.g., L2 or L1 regularization of  $\phi$  components) will fail to do. While hand-crafted priors designed for receptive field estimation can capture some of these characteristics (Park and Pillow, 2017; Savin and Tkacik, 2016), optimization priors grounded in the relevant normative theory represent the most succinct and complete way of summarizing our prior beliefs. For flexible optimization priors whose strength and additional parameters are set by cross-validation, one may expect that the postulated optimality theory need not be exactly correct to aid inference, so long as it captures some of the statistical regularities in the data.

#### Application 1: receptive fields in the visual cortex

Here, we analyze receptive fields of neurons in the primary visual cortex (V1) of the Macaque monkey (Ringach, 2002) (Figure 6A). This system is a good test case, for which multiple candidate optimality theories were developed and tested against data (Olshausen and Field, 1996; Wiskott and Sejnowski, 2002; Hyvärinen et al., 2009; Van Hateren and van der Schaaf, 1998). As in the example of Figure 5, we focus on sparse coding using utility  $U_{SC}$ , which prioritizes RFs localized in space and frequency



**Figure 5. Optimality priors improve inference of high-dimensional receptive fields**

(A) Linear-nonlinear-Poisson (LNP) neuron responding to  $16 \times 16$  pixel natural image patches,  $x_t$ . Stimuli are projected onto a linear filter  $\phi$ , which transforms them via logistic nonlinearity into an average firing rate of Poisson spiking,  $r_t$ .

(B) Receptive fields optimized for maximally sparse response to natural stimuli with a locality constraint  $\xi$ . First 3 panels on the left display  $2 \times 2$  example filters optimized at increasing  $\xi$ . The rightmost panel shows the decrease in average sparse utility of filters with increasing  $\xi$ .

(C) MAP estimates of 2 optimally sparse filters ( $\xi = 0$ ) obtained with optimality prior of increasing strength  $\beta$ . White digits denote correlation with the corresponding ground truth.

(D) Average correlations of  $N = 100$  filter estimates with the ground truth as a function of prior strength  $\beta$  for locality constraint  $\xi = 0$ . Dashed blue line denotes the average correlation for ML estimates. MAP estimate correlations are significantly higher than ML estimate correlations (t test; \*\*\* $p < 0.001$ ). Error bars denote standard errors of the mean.

(E) Identification of prior strength  $\beta$  and locality constraint  $\xi$  via cross-validation. Left panel, cross-validation errors in predicting withheld neural responses for a range of  $\beta$  and  $\xi$  values (heatmap). Parameter combination resulting in minimal error is marked with a red frame. Top right, a ground truth filter optimized with  $\xi = 0.2$ . Bottom right, MAP estimate of the filter, obtained with correctly identified values for  $\beta$  and  $\xi$ .

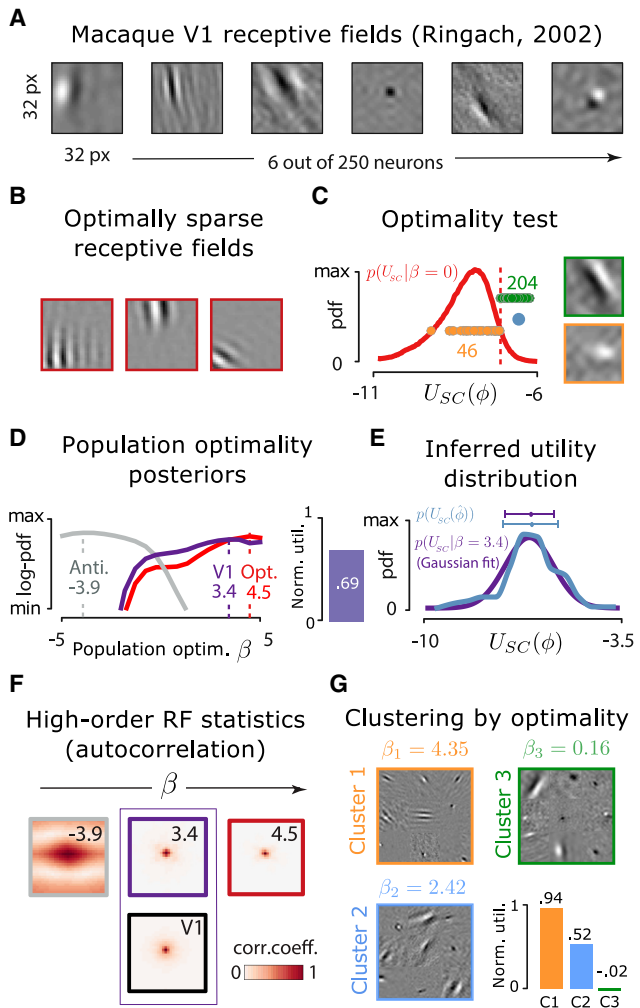
a test statistic. To construct the null distribution for the test, we sample  $10^6$  random filters consistent with optimization prior  $P(\phi|\beta = 0)$  and declare the 95<sup>th</sup> percentile to be the optimality threshold (Figure 6C). As expected, a large majority (204 neurons, green dots/example frame in Figure 6C) of V1 neurons pass the optimality threshold, with 46 neurons failing the test (orange dots/example frame in Figure 6C).

We next ask whether all RFs can be used together to quantify the degree of population optimality, as in question 2. We estimate approximate posteriors over parameter  $\beta$  via rejection sampling (see Method details), using all RFs in the population (Figure 6D, purple line). For comparison, we also compute posteriors using 250 utility-maximizing and 250 utility-mini-

mizing filters (Figure 6D, red and gray lines, respectively). MAP estimates of  $\beta$  obtained with simulated maximal and minimal utility RFs provide a reference for the interpretation of  $\beta$  estimated from real data. This estimate,  $\hat{\beta}_{V1}$ , is very close to the parameter value of the optimally sparse filters, implying a high degree of optimization. The normalized utility is 0.69, implying a significant yet not complete degree of optimization.

(Figure 6B; see Method details). An alternative utility prioritizing slow features is presented in the Supplemental information (Method S7; Figure S8).

We ask whether RFs of individual neurons support the optimality hypothesis, as in question 1. Given the high quality of RFs estimates, costly marginalization of the likelihood can be avoided, and the utility of estimated RFs can be used directly as



**Figure 6. Optimality of V1 receptive fields**

(A) Six example receptive fields (RFs) from Macaque visual cortex (courtesy of Dario Ringach; Ringach, 2002).

(B) Example simulated RFs optimized for sparsity.

(C) Null distribution of utility values used to test for optimality under sparse utility and the 95<sup>th</sup> percentile significance threshold (red dashed line). Significant (green) and non-significant (orange) receptive fields denoted with dots (x axis is truncated for visualization purposes); example RFs are shown in frames of matching colors. Blue dot shows the average RF utility (99.6<sup>th</sup> percentile of the null distribution).

(D) Approximate log-posteriors over population optimality parameter  $\beta$  derived from 250 RFs estimates (purple line), 250 maximum-utility filters (red line), and 250 minimal-utility filters (gray line). Dashed lines mark MAP estimates.

(E) Empirical distribution of RF utilities (blue line) compared with utility distribution consistent with the inferred  $\hat{\beta}_{V1}$  (purple line). Dots denote averages, and horizontal lines denote standard deviations.

(F) Spatial autocorrelation of RFs predicted for different  $\beta$  values (reported in top-right corner of each panel, cf. inferred values in D). Note a good match between data-derived RF autocorrelation (black frame) and the predicted autocorrelation at the inferred  $\hat{\beta}_{V1}$  (purple frame).

(G) Three clusters with different  $\beta$ , learned with a MaxEnt mixture model. For each cluster, 3 × 3 sample receptive fields are displayed, together with the corresponding normalized utility values in the bottom-right panel.

Since population optimality  $\beta$  parametrizes the entire distribution of receptive fields, inferring  $\beta$  allows us to make predictions inaccessible by other means. For example, given the inferred degree of optimality, we predict the entire distribution of utility values (not only its mean) across neurons. In principle, the predicted distribution (or its higher-order moments, for example, variance) could deviate from the empirically observed distribution, if the real system were adapted to a different utility or set of constraints. For V1 neurons, the predicted and empirical sparse utility distributions are very similar (Figure 6E).

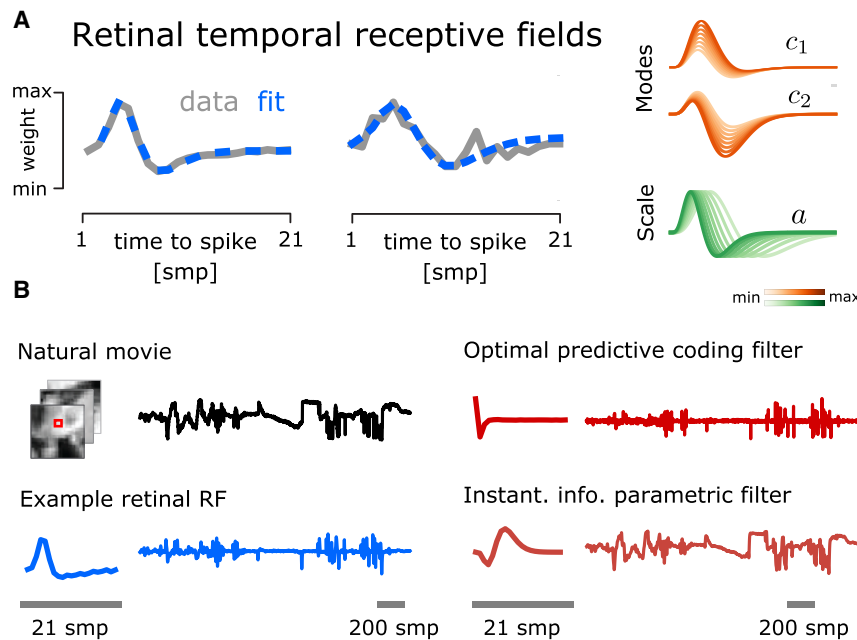
Another prediction concerns the correlation between system parameters, in our case, RF shapes. Different values of  $\beta$  predict very different spatial autocorrelation functions of RFs (Figure 6F), with the prediction at inferred  $\beta$  resembling the data-derived autocorrelation better than the alternative or extremal  $\beta$  values. These examples demonstrate that once the single parameter  $\beta$  is inferred, the optimality framework makes quantitative, rigorous, and parameter-free predictions of non-trivial statistics that can be directly tested against data.

Our framework can also be used to dissect sources of deviation from optimality. We fit a mixture model, in which each mixture component was parametrized by a separate value of  $\beta$  (Figure 6G; see Method details). This procedure clusters the RFs into three groups spanning a broad range of utility values. The largest cluster (135 RFs) achieves a nearly maximal normalized utility of 0.94; neurons in this cluster passed the significance test in Figure 6C. The existence of second- and third-largest clusters (95 RFs, normalized utility of 0.52; 20 RFs, normalized utility  $\sim$ 0, respectively) suggests that these cells may be subject to additional unknown constraints or may be optimizing a different utility. We emphasize that we analyze the optimality of individual neurons, whereas the optimization of complete populations could yield a more diverse set of RFs that are individually suboptimally sparse (Olshausen and Field, 1996; Zylberberg et al., 2011; Hyvärinen et al., 2009), accounting for the deviations we observe. Our analysis is intended as a demonstration of the applicability of our framework, rather than a definitive optimality claim about V1 neurons. Population-level analysis of optimality is a subject of future work.

### Application 2: receptive fields in the retina

Here, we analyze the temporal receptive fields of 117 retinal ganglion cells (RGCs) in the rat retina (Deny et al., 2017). Temporal RFs have a characteristic bimodal shape (Figure 7A, left), which can be captured well by a simple filter model with three parameters (Sun et al., 2017). Two parameters ( $c_1, c_2$ ) describe the amplitudes of both modes, while the third ( $a$ ) determines the temporal scale of the filter (Figure 7A, right). In what follows, we focus on the optimality of filter shapes in the space of these three parameters.

RGC receptive fields long have been hypothesized to instantiate predictive coding (PC), a canonical example of a normative theory in sensory neuroscience (Srinivasan et al., 1982). Temporal PC postulates that instead of tracking the exact stimulus value directly in their responses, neurons encode a difference between the stimulus and its linear prediction computed using past stimuli. Such a strategy has many potential benefits: it reduces the dynamic range of signals, minimizes the use of

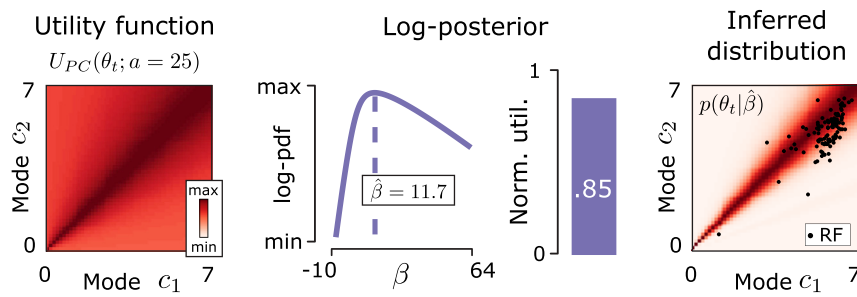


**Figure 7. Optimality of retinal receptive fields**

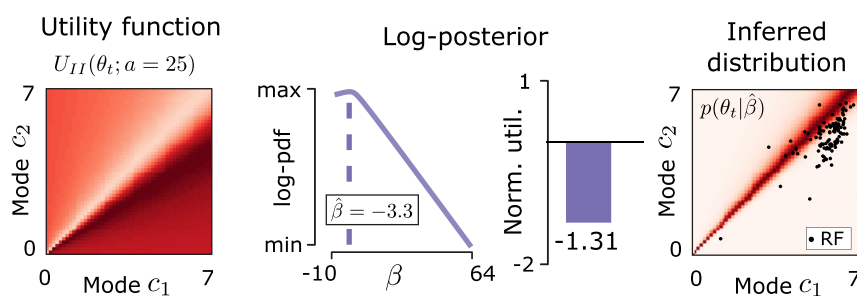
(A) Two example temporal receptive fields of rat retinal ganglion cells. Gray lines show RF estimates (courtesy of Olivier Marre; Deny et al., 2017), dashed blue lines show parametric fits. Fit parameters correspond to amplitudes of filter modes (parameters  $c_1, c_2$ , orange) and scale (parameter  $a$ , green). (B) Example natural stimulus: light intensity of a single pixel of a natural movie (top left, black). Representative retinal RF and its linear response to the natural stimulus (bottom left, blue line). Optimal predictive coding filter and its response to the same stimulus (top right, dark red line). Optimal instantaneous information transmission filter and its response (bottom right, pink line).

(C) Analysis of temporal RFs with the generalized predictive coding utility,  $U_{PC}$ . First panel: utility function of filter modes  $c_1, c_2$  constrained by time-scale  $a=25$ . Second panel: log-posterior (solid purple line) over population optimality parameter  $\beta$  (dashed vertical line, MAP estimate). Third panel: normalized utility of the RF population. Fourth panel: optimization prior distribution over  $(c_1, c_2)$  at the inferred  $\hat{\beta}$ , marginalized over all values of the time-scale parameter  $a$  (black dots, data-derived RFs). (D) Analysis of temporal RFs with the instantaneous information utility,  $U_{II}$ , analogous to (C).

**C** Predictive coding utility analysis



**D** Instantaneous information utility analysis



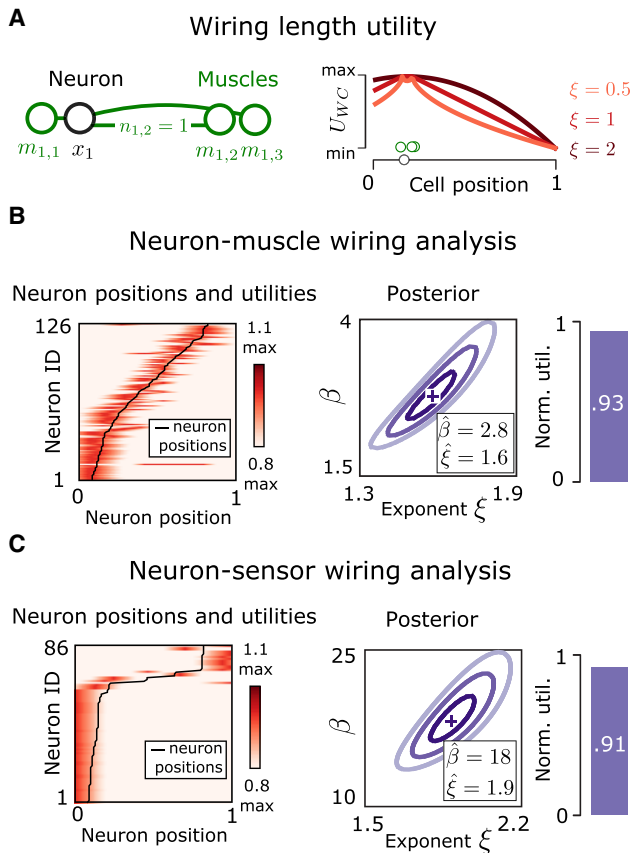
PC filters using natural light intensity time courses (see Method details). Optimal PC filter responses qualitatively resemble the responses of a representative retinal filter convolved with the same natural stimulus (Figure 7C). Both filters generate strong, spike-like transients to sudden changes in the stimulus mean, while their output remains close to 0 when the stimulus is not changing. This pattern is different from the response of a parametric bimodal filter (with parameters  $a, c_1, c_2$ ) optimized to track the stimulus, obtained by maximizing instantaneous information transmission in a low-noise regime ( $U_{II}$ , see Method details). More important, predicted responses can be very distinct despite the qualitative similarity between retinal, PC, and instantaneous information filters.

To evaluate the optimality of retinal RFs, we propose a new utility,  $U_{PC}$ , that mathematically generalizes the canonical formulation of predictive coding (Srinivasan et al., 1982). This utility prioritizes filters

that minimize power in their output, given a fixed filter norm, while allowing the filters to operate on timescales that are distinct from the stimulus frame rate (see Method details). We evaluate  $U_{PC}(\theta; a)$  as a function of the two filter mode parameters,  $\theta = (c_1, c_2)$ , but consider the timescale  $a$  to be an external constraint to be inferred from data for each neuron separately, as in question 3. Parameter  $a$  is a constraint because, much

metabolic resources, and can lead to efficient coding in the low noise limit, by performing stimulus decorrelation and response whitening (Srinivasan et al., 1982; Bialek, 2012; Dong and Atick, 1995; Van Hateren and van der Schaaf, 1998; Chalk et al., 2019).

An optimal predictive coding filter must be adapted to the statistics of stimuli it encodes (Srinivasan et al., 1982). We optimize



**Figure 8. Optimality of neural wiring in *C. elegans***

(A) Left panel: connection schematic between example neuron at position  $x_1$  (black circle) and 3 muscles at positions  $m_{1,1}, m_{1,2}, m_{1,3}$  (green circles). Number of synapses between neuron  $x_1$  and muscle  $m_{1,j}$  is denoted  $n_{1,j}$ . The example neuron forms monosynaptic connections (green lines) only with the 3 muscles. Right panel: wiring cost utility,  $U_{WC}(x_1; \xi)$ , as a function of position  $x_1$ , corresponding to the scenario depicted at left. Position axis spans the entire *C. elegans* body length. Utility functions are shown for 3 exponent values  $\xi$ . (B) Neuron-muscle connection analysis. Left panel: utility  $U_{WC}(x; \xi = 2)$  (red, scaled to  $[0, 1]$  for each neuron) for all 126 neurons (rows), as a function of neuron positions  $x \in [0, 1]$ . The black line denotes positions of real neurons. Center panel: joint posterior over optimality parameter  $\beta$  and the exponent  $\xi$  (cross denotes MAP estimates reported in the legend). Right panel: normalized utility of neuron-muscle connectivity. (C) Neuron-sensor connection analysis, analogous to (B).

like  $k$  in the toy neuron example of Figure 2, its value is not set by optimality (which prefers  $a \rightarrow 0$ ) but by biophysical constraints or by the temporal horizon at which prediction is of highest use to the organism. For a broad range of  $a$  values,  $U_{PC}$  is highest close to the diagonal of the  $(c_1, c_2)$  plane, representing nearly balanced filters, as shown in Figure 7C (left).

We use all of the retinal RFs jointly to compute the posterior over the optimality parameter  $\beta$  (Figure 7C, second panel). The inferred  $\hat{\beta} \approx 11.7$  yields a normalized utility of 0.85, implying strong optimization for PC (Figure 7C, third panel from the left); even relative to the non-parametric optimal PC filter with no timescale constraint, the utility of retinal filters remains as high as 0.74. The high degree of optimization is visually evident in the  $(c_1, c_2)$

plane, where individual neurons fall onto high-utility regions of the maximum entropy distribution given inferred  $\hat{\beta}$  and marginalized over timescale  $a$  (Figure 7C, right). An analogous analysis performed using maximization of instantaneous information  $U_{II}$  (see Method details) reveals a negative  $\beta$  estimate and thus anti-optimization for this alternative utility, with real neurons avoiding high-utility regions of the maximum entropy distribution.

### Application 3: neural wiring in *Caenorhabditis elegans*

Here, we analyze neural wiring in *C. elegans*, which has been the subject of several normative studies (Chen et al., 2006; Chklovskii, 2004; Pérez-Escudero et al., 2009; Pérez-Escudero and de Polavieja, 2007). Relative positions of neurons could be partially predicted by minimizing the total wiring cost under the constraint that muscles and sensors need to be properly connected (Chen et al., 2006; Pérez-Escudero and de Polavieja, 2007). Instead of trying to predict individual neuron positions, we ask a different question: are the measured neuron positions optimized to minimize the wiring cost to muscles and sensors?

For each neuron  $i$ , the wiring cost is determined by the number of muscles it connects to, the distance between the neuron's position,  $x_i$ , the positions of muscles,  $m_{ij}$ , and the number of synapses formed by each connection,  $n_{ij}$  (Figure 8A). The resulting utility function for each neuron can be written as  $U_{WC}(x_i; m_i, n_i, \xi) = -\sum_{j=1}^{N_i} n_{ij} |x_i - m_{ij}|^\xi$ , where  $N_i$  is the number of muscles the neuron  $i$  connects with, and  $\xi$  is an exponent determining the form of the utility as a function of distance (Chen et al., 2006) (Figure 8B). The precise value of  $\xi$  is not specified by the theory and thus needs to be inferred from data, following the ambiguity of the third kind (*cf.* question 3).

Our analysis shows that a large proportion of 126 neurons that form connections with muscles align closely with the maxima of the utility function (Figure 8B, left panel). We estimate the joint posterior distribution over the optimality parameter  $\beta$  and the connection exponent  $\xi$ , for neuron-muscle and neuron-sensor connections separately (Figures 8B and 8C, center panels). In both cases, the normalized utility exceeds 90%, implying strong optimization. Interestingly, the estimates for the exponent  $\xi$  are relatively high: 1.6 for neuron-muscle connections and 1.9 for neuron-sensor connections, suggesting that neurons are only weakly penalized for small deviations from optimal positions. This is in contrast to a previously published analysis that focused instead on neuron-neuron connections (Pérez-Escudero et al., 2009), in which the authors find (and we confirm)  $\xi \approx 0.5$ . We interpret this discrepancy to imply that neuron-muscle and neuron-sensor connection costs are less important relative to neuron-neuron connections so far as the overall *C. elegans* body plan is concerned. One possible reason is that neuron-neuron wiring cost scales quadratically with the number of neurons, implying higher penalty (and thus lower  $\xi$ ) for deviations from optimality.

### DISCUSSION

Despite their theoretical appeal, the application of optimization principles to biological systems has been hindered by statistical issues that grow more pressing as the complexity and dimensionality of the models increases. These issues are not new. Instead of developing an *ad hoc* solution whenever called

for by a particular application, we decided to tackle these issues head on and flesh them out with simple examples. For instance, the issue of an unconstrained optimization parameter or a trade-off with unknown strength is well known to the practitioners but is often solved “by hand”; one manually adjusts the constraint until the optimality predictions are (visually) consistent with data. Such manual “fine-tuning” of constraints is clearly problematic from the statistical viewpoint, as it could easily amount to (over-)fitting that is not controlled for. In contrast, our framework performs inference and optimization jointly and provides a full posterior over constrained and unconstrained parameters alike. Another problematic issue arises from degenerate maxima of the utility functions. A frequent solution has been to postulate further constraints within the theory itself, which disambiguate the predictions (Doi et al., 2012). Our framework proposes a complementary mechanism: using a small amount of data to localize the theoretical predictions to the relevant optimum, against which further statistical tests can be carried out. As a last example, when fitting complex (e.g., nonlinear dynamical systems) models, one typically restricts parameters by hand to a domain that is thought to be “biologically relevant.” In contrast, optimization priors automatically suppress vast swaths of parameter space that lead to non-functioning systems, even if these systems are not fully optimized for the postulated utility. In this way, the statistical power of the data can be used with maximum effect in the parameter regime that is of actual biological relevance, without sacrificing statistical rigor.

The ability to exclude biologically irrelevant regions of the parameter space highlights a general advantage of optimality priors over simple, unstructured distributions. Frequently applied “regularization priors,” which penalize the norm of parameter values (e.g., Laplace, Gaussian; Park and Pillow, 2017; Sharpee, 2013) assign the highest probability when all parameters are equal to 0. Moreover, these priors are isotropic—they act with the same strength on each parameter and do not take into account interactions between them—which is an essential (and nontrivial) property of real systems. These two requirements enforced by the prior are often contradictory to the notion of a functioning biological system. For example, penalizing parameter magnitudes while inferring the shape of nonlinearity in our toy-model neuron would bias the inferences toward completely non-functional solutions (slope and offset equal to 0). Intuitively, the robustness against overfitting afforded by the regularization prior thus comes at a cost of biasing inferences away from functional solutions. Our approach, in contrast, attempts to avoid such a disastrous trade-off by incorporating knowledge about biological function directly into the structure of the prior.

While our framework provides a principled way to navigate a number of statistical issues in complex biological systems, important questions remain. A key challenge is to identify the relevant optimization criterion for a biological system and to express it in terms of experimentally measurable quantities. A candidate utility function that embodies an optimality criterion of interest could be selected from a possible discrete set of such functions (Wang et al., 2016; Mitynarski and Hermundstad, 2018; Chalk et al., 2019) or by inferring utility function parameters. Because we leverage the well-understood machinery of

Bayesian inference, one could perform model selection for the utility function that best explains the data. Such an approach could be used, for example, to rigorously verify whether entire neural populations in the visual cortex are jointly optimized for sparsity or a different utility, such as slowness (Wiskott and Sejnowski, 2002). An important caveat is that the more flexible our choice of the utility function becomes, the easier it is to claim an optimality for a system of interest. In principle, one could postulate a utility function with a fully unconstrained shape. In this limit, our framework would automatically recover the utility function shape from data (if these were sufficient), assuming the observed system is optimal, in a way reminiscent of inverse reinforcement learning (Chalk et al., 2019). This connection is an interesting topic for further research. In this article, however, we focused on optimization theories in which the number of adjustable utility parameters is smaller than the number of system parameters being predicted.

Our framework dovetails with other approaches that address the issues of ambiguity of theoretical predictions and model identifiability given the limited data in biology. “Sloppy-modeling” (O’Leary et al., 2015; Gutenkunst et al., 2007), grounded in dynamical systems theory, characterizes the dimensions of the parameter space that yields qualitatively similar behavior of the system. In our framework, these dimensions correspond to regions of the parameter space of equal or similar utility. Another important conceptual advance grounded in statistical inference has been the usage of limited data to coarse-grain probabilistic models (Bialek et al., 1996; Chen et al., 2018; Machta et al., 2013). In our framework, a related coarse-graining occurs when, instead of inferring all system parameters from data directly, optimization sets the values of most of these parameters, leaving only the unconstrained subset to be fitted. The resulting dimensionality reduction could be sizable (e.g., with optimization predicting high-dimensional RF shapes given inferred firing rate, locality, or neural noise constraints) and could efficiently parametrize neuronal heterogeneity in terms of a small number of constraints that vary from neuron to neuron or between neural populations. Another point of connection with recent work concerns the ability to instantiate high-dimensional maximum entropy distributions over parameters with complicated dependency structures (De Martino et al., 2018; Bittner et al., 2019; Gonçalves et al., 2020). Such computational innovations will be essential for statistical analyses of optimality that require sampling from maximum-entropy optimization priors.

## STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Model neuron and mutual information utility function

- Likelihood ratio test of optimality
- Data-rich-regime simplification
- Multiple system instances simplification
- Hierarchical inference of population optimality
- Inference of receptive fields with optimality priors
- Analysis of V1 receptive fields
- Analysis of retinal receptive fields
- Analysis of connectivity in *C. elegans*
- **QUANTIFICATION AND STATISTICAL ANALYSIS**

### SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at <https://doi.org/10.1016/j.neuron.2021.01.020>.

### ACKNOWLEDGMENTS

The authors thank Dario Ringach for providing the V1 receptive fields and Olivier Marre for providing the retinal receptive fields. W.M. was funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 754411. M.H. was funded in part by Human Frontiers Science grant no. HFSP RGP0032/2018.

### DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: February 25, 2020  
 Revised: September 10, 2020  
 Accepted: January 19, 2021  
 Published: February 15, 2021

### REFERENCES

- Alexander, R.M. (2003). *Principles of Animal Locomotion* (Princeton University Press).
- Barlow, H.B. (1961). Possible principles underlying the transformation of sensory messages. In *Sensory Communication*, W.A. Rosenblith, ed. (MIT Press).
- Barton, N.H., and de Vladar, H.P. (2009). Statistical mechanics and the evolution of polygenic quantitative traits. *Genetics* *181*, 997–1011.
- Bialek, W. (2012). *Biophysics: Searching for Principles* (Princeton University Press).
- Bialek, W., Callan, C.G., and Strong, S.P. (1996). Field theories for learning probability distributions. *Phys. Rev. Lett.* *77*, 4693–4697.
- Bittner, S.R., Palmigiano, A., Piet, A.T., Duan, C.A., Brody, C.D., Miller, K.D., and Cunningham, J.P. (2019). Interrogating theoretical models of neural computation with deep inference. *bioRxiv*, 837567.
- Borghuis, B.G., Ratliff, C.P., Smith, R.G., Sterling, P., and Balasubramanian, V. (2008). Design of a neuronal array. *J. Neurosci.* *28*, 3178–3189.
- Brinkman, B.A., Weber, A.I., Rieke, F., and Shea-Brown, E. (2016). How do efficient coding strategies depend on origins of noise in neural circuits? *PLoS Comput. Biol.* *12*, e1005150.
- Carlson, N.L., Ming, V.L., and Deweese, M.R. (2012). Sparse codes for speech predict spectrotemporal receptive fields in the inferior colliculus. *PLoS Comput. Biol.* *8*, e1002594.
- Chalk, M., Marre, O., and Tkačik, G. (2018). Toward a unified theory of efficient, predictive, and sparse coding. *Proc. Natl. Acad. Sci. USA* *115*, 186–191.
- Chalk, M., Tkacik, G., and Marre, O. (2019). Inferring the function performed by a recurrent neural network. *bioRxiv*, 598086.
- Chen, B.L., Hall, D.H., and Chklovskii, D.B. (2006). Wiring optimization can relate neuronal structure and function. *Proc. Natl. Acad. Sci. USA* *103*, 4723–4728.
- Chen, W.-C., Tareen, A., and Kinney, J.B. (2018). Density estimation on small data sets. *Phys. Rev. Lett.* *121*, 160605.
- Chklovskii, D.B. (2004). Exact solution for the optimal neuronal layout problem. *Neural Comput.* *16*, 2067–2078.
- De Martino, D., Mc Andersson, A., Bergmiller, T., Guet, C.C., and Tkačik, G. (2018). Statistical mechanics for metabolic networks during steady state growth. *Nat. Commun.* *9*, 2988.
- Deny, S., Ferrari, U., Macé, E., Yger, P., Caplette, R., Picaud, S., Tkačik, G., and Marre, O. (2017). Multiplexed computations in retinal ganglion cells of a single type. *Nat. Commun.* *8*, 1964.
- Doi, E., and Lewicki, M.S. (2014). A simple model of optimal population coding for sensory systems. *PLoS Comput. Biol.* *10*, e1003761.
- Doi, E., Gauthier, J.L., Field, G.D., Shlens, J., Sher, A., Greschner, M., Machado, T.A., Jepsen, L.H., Mathieson, K., Gunning, D.E., et al. (2012). Efficient coding of spatial information in the primate retina. *J. Neurosci.* *32*, 16256–16264.
- Dong, D.W., and Atick, J.J. (1995). Statistics of natural time-varying images. *Network* *6*, 345–358.
- Eichhorn, J., Sinz, F., and Bethge, M. (2009). Natural image coding in V1: how much use is orientation selectivity? *PLoS Comput. Biol.* *5*, e1000336.
- Ganguli, D., and Simoncelli, E.P. (2014). Efficient sensory encoding and Bayesian inference with heterogeneous neural populations. *Neural Comput.* *26*, 2103–2134.
- Geisler, W.S. (2011). Contributions of ideal observer theory to vision research. *Vision Res.* *51*, 771–781.
- Gelman, A. (2004). Parameterization and bayesian modeling. *J. Am. Stat. Assoc.* *99*, 537–545.
- Gjorgjieva, J., Sompolinsky, H., and Meister, M. (2014). Benefits of pathway splitting in sensory coding. *J. Neurosci.* *34*, 12127–12144.
- Gold, J.I., and Shadlen, M.N. (2007). The neural basis of decision making. *Annu. Rev. Neurosci.* *30*, 535–574.
- Gonçalves, Pedro, Lueckmann, J., Deistler, M., Nonnenmacher, M., Öcal, K., Bassetto, G., Chintaluri, C., Podlaski, w., Haddad, S., Vogels, T., Greenberg, D., et al. (2020). Training deep neural density estimators to identify mechanistic models of neural dynamics. *Elife* *9*, e56261.
- Gutenkunst, R.N., Waterfall, J.J., Casey, F.P., Brown, K.S., Myers, C.R., and Sethna, J.P. (2007). Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput. Biol.* *3*, 1871–1878.
- Hyvärinen, A., Hurri, J., and Hoyer, P.O. (2009). *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision* *Volume 39* (Springer Science & Business Media).
- Ibarra, R.U., Edwards, J.S., and Palsson, B.O. (2002). *Escherichia coli K-12* undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature* *420*, 186–189.
- Jaynes, E.T. (2003). *Probability Theory: The Logic of Science* (Cambridge University Press).
- Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond. A Math. Phys. Sci.* *186*, 453–461.
- Kacser, H., and Burns, J.A. (1995). The control of flux. *Biochem. Soc. Trans.* *23*, 341–366.
- Kass, R.E., Eden, U.T., and Brown, E.N. (2014). *Analysis of Neural Data* *Volume 491* (Springer).
- Kastner, D.B., Baccus, S.A., and Sharpee, T.O. (2015). Critical and maximally informative encoding between neural populations in the retina. *Proc. Natl. Acad. Sci. USA* *112*, 2533–2538.
- Laughlin, S. (1981). A simple coding procedure enhances a neuron's information capacity. *Z. Naturforsch. C Biosci.* *36*, 910–912.
- Machta, B.B., Chachra, R., Transtrum, M.K., and Sethna, J.P. (2013). Parameter space compression underlies emergent theories and predictive models. *Science* *342*, 604–607.
- MacKay, D.J. (2003). *Information Theory, Inference and Learning Algorithms* (Cambridge University Press).

- Młynarski, W. (2015). The opponent channel population code of sound location is an efficient representation of natural binaural sounds. *PLoS Comput. Biol.* *11*, e1004294.
- Młynarski, W.F., and Hermundstad, A.M. (2018). Adaptive coding for dynamic sensory inference. *eLife* *7*, e32055.
- Młynarski, W., and McDermott, J.H. (2018). Learning midlevel auditory codes from natural sound statistics. *Neural Comput.* *30*, 631–669.
- Murphy, K.P. (2012). *Machine Learning: A Probabilistic Perspective* (MIT Press).
- O’Leary, T., Sutton, A.C., and Marder, E. (2015). Computational models in the age of large datasets. *Curr. Opin. Neurobiol.* *32*, 87–94.
- Olshausen, B.A., and Field, D.J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* *381*, 607–609.
- Olshausen, B.A., and Field, D.J. (1997). Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vision Res.* *37*, 3311–3325.
- Olshausen, B.A., and Field, D.J. (2004). Sparse coding of sensory inputs. *Curr. Opin. Neurobiol.* *14*, 481–487.
- Orzack, S.H. (2001). *Adaptation and Optimality* (Cambridge University Press).
- Paninski, L., Pillow, J., and Lewi, J. (2007). Statistical models for neural encoding, decoding, and optimal stimulus design. *Prog. Brain Res.* *165*, 493–507.
- Park, M., and Pillow, J.W. (2011). Receptive field inference with localized priors. *PLoS Comput. Biol.* *7*, e1002219.
- Park, I.M., and Pillow, J.W. (2017). Bayesian efficient coding. *bioRxiv*, 178418.
- Pérez-Escudero, A., and de Polavieja, G.G. (2007). Optimally wired subnetwork determines neuroanatomy of *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. USA* *104*, 17180–17185.
- Pérez-Escudero, A., Rivera-Alba, M., and de Polavieja, G.G. (2009). Structure of deviations from optimality in biological systems. *Proc. Natl. Acad. Sci. USA* *106*, 20544–20549.
- Pitkow, X., and Meister, M. (2012). Decorrelation and efficient coding by retinal ganglion cells. *Nat. Neurosci.* *15*, 628–635.
- Ratliff, C.P., Borghuis, B.G., Kao, Y.-H., Sterling, P., and Balasubramanian, V. (2010). Retina is structured to process an excess of darkness in natural scenes. *Proc. Natl. Acad. Sci. USA* *107*, 17368–17373.
- Ringach, D.L. (2002). Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *J. Neurophysiol.* *88*, 455–463.
- Rosen, R. (2013). *Optimality Principles in Biology* (Springer).
- Savin, C., and Tkacik, G. (2016). Estimating nonlinear neural response functions using gp priors and kronecker methods. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds. (Curran Associates), pp. 3603–3611.
- Savir, Y., Noor, E., Milo, R., and Tlusty, T. (2010). Cross-species analysis traces adaptation of Rubisco toward optimality in a low-dimensional landscape. *Proc. Natl. Acad. Sci. USA* *107*, 3475–3480.
- Shannon, C.E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* *27*, 379–423.
- Sharpee, T.O. (2013). Computational identification of receptive fields. *Annu. Rev. Neurosci.* *36*, 103–120.
- Sharpee, T., and Bialek, W. (2007). Neural decision boundaries for maximal information transmission. *PLoS ONE* *2*, e646.
- Smith, E.C., and Lewicki, M.S. (2006). Efficient auditory coding. *Nature* *439*, 978–982.
- Srinivasan, M.V., Laughlin, S.B., and Dubs, A. (1982). Predictive coding: a fresh view of inhibition in the retina. *Proc. R. Soc. Lond. B Biol. Sci.* *216*, 427–459.
- Sun, Y., Nern, A., Franconville, R., Dana, H., Schreier, E.R., Looger, L.L., Svoboda, K., Kim, D.S., Hermundstad, A.M., and Jayaraman, V. (2017). Neural signatures of dynamic stimulus selection in *Drosophila*. *Nat. Neurosci.* *20*, 1104–1113.
- Tero, A., Takagi, S., Saigusa, T., Ito, K., Bebber, D.P., Fricker, M.D., Yumiki, K., Kobayashi, R., and Nakagaki, T. (2010). Rules for biologically inspired adaptive network design. *Science* *327*, 439–442.
- Tkačik, G., and Bialek, W. (2016). Information processing in biological systems. *Annu. Rev. Condens. Matter Phys.* *7*, 89–117.
- Tkačik, G., Callan, C.G., Jr., and Bialek, W. (2008). Information flow and optimization in transcriptional regulation. *Proc. Natl. Acad. Sci. USA* *105*, 12265–12270.
- Tkačik, G., Prentice, J.S., Balasubramanian, V., and Schneidman, E. (2010). Optimal population coding by noisy spiking neurons. *Proc. Natl. Acad. Sci. USA* *107*, 14419–14424.
- van Hateren, J.H. (1992). Theoretical predictions of spatiotemporal receptive fields of fly Imcs, and experimental validation. *J. Comp. Physiol. A Neuroethol. Sens. Neural Behav. Physiol.* *171*, 157–170.
- van Hateren, J.H., and Ruderman, D.L. (1998). Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proc. Biol. Sci.* *265*, 2315–2320.
- van Hateren, J.H., and van der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. Biol. Sci.* *265*, 359–366.
- Wang, Z., Stocker, A.A., and Lee, D.D. (2016). Efficient neural codes that minimize lp reconstruction error. *Neural Comput.* *28*, 2656–2686.
- Wiskott, L., and Sejnowski, T.J. (2002). Slow feature analysis: unsupervised learning of invariances. *Neural Comput.* *14*, 715–770.
- Wright, S. (1937). The Distribution of Gene Frequencies in Populations. *Proc. Natl. Acad. Sci. USA* *23*, 307–320.
- Zylberberg, J., Murphy, J.T., and DeWeese, M.R. (2011). A sparse coding model with synaptically local plasticity and spiking neurons can account for the diverse shapes of V1 simple cell receptive fields. *PLoS Comput. Biol.* *7*, e1002250.

## STAR★METHODS

## KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
MATLAB 2014	Mathworks	<a href="https://www.mathworks.com">https://www.mathworks.com</a>
Other		
V1 Receptive fields of a macaque (provided by Dario Ringach)	<a href="#">Ringach 2002</a>	N/A
Retinal receptive fields of a rat (provided by Olivier Marre)	<a href="#">Deny et al., 2017</a>	N/A
Neural connectivity of <i>C. Elegans</i>	<a href="#">Chen et al., 2006</a>	<a href="https://www.wormatlas.org/neuronalwiring.html">https://www.wormatlas.org/neuronalwiring.html</a>

## RESOURCE AVAILABILITY

## Lead contact

Further information and requests for code should be directed to Wiktor Młynarski ([wiktor.mlynarski@ist.ac.at](mailto:wiktor.mlynarski@ist.ac.at)).

## Materials availability

This study did not generate any new materials.

## Data and code availability

This study did not generate any new datasets. The code supporting the current study is freely available from the corresponding author upon request.

## METHOD DETAILS

## Model neuron and mutual information utility function

A model neuron elicits a spike at time  $t$  ( $r_t = 1$ ) with a probability:

$$P(r_t = 1|x_t) = \frac{1}{1 + \exp[-k(x_t - x_0)]}; \quad (\text{Equation 3})$$

the stimuli  $x_t$  were distributed according to a Gaussian Mixture Model,  $P(x_t) = \sum_{i=1}^3 w_i \mathcal{N}(\mu_i, \sigma_i^2)$ , where  $w_i = 1/3$  are weights of the mixture components,  $\mu_{1,2,3} = -2, 0, 2$  are the means, and  $\sigma_i = 0.2$  are standard deviations.

To estimate mutual information between class labels and neural responses, we generated  $5 \cdot 10^4$  stimulus samples  $x_t$  from the stimulus distribution. Each sample was associated with a class label  $c_t \in \{1, 2, 3\}$ , corresponding to a mixture component. We created a discrete grid of logistic-nonlinearity parameters by uniformly discretizing ranges of slope  $k \in [-10, 10]$  and position  $x_0 \in [-3, 3]$  into 128 values each. For each pair of parameters on the grid, we simulated responses of the model neuron to the stimulus dataset and estimated the mutual information directly from a joint histogram of responses  $r_t$  and class labels  $c_t$ .

## Likelihood ratio test of optimality

The proposed test uses the likelihood ratio statistic,

$$\lambda = 2 \log \frac{\max_{\beta > 0} P(\mathcal{D}|\beta)}{P(\mathcal{D}|\beta = 0)}. \quad (\text{Equation 4})$$

The null hypothesis is rejected for high values of  $\lambda$ . The marginal likelihood of  $\beta$ ,  $\tilde{\mathcal{L}}(\beta) = P(\mathcal{D}|\beta)$ , depends on the overlap of parameter likelihood and the optimization prior,  $P(\mathcal{D}|\beta) = \int_{\Theta} P(\mathcal{D}|\theta) P(\theta|\beta) d\theta$ , where  $\Theta$  is the region of biophysically feasible parameter combinations.

The null distribution of  $\lambda$  is obtained by sampling in three steps: (i) sample a parameter combination  $\theta$  from a uniform distribution on  $\theta$ , i.e.,  $P(\theta|\beta = 0)$ ; (ii) sample a dataset  $\mathcal{D}$  according to the likelihood  $P(\mathcal{D}|\theta)$ ; (iii) compute the test statistic  $\lambda$  according to (4). This computationally expensive process simplifies in two situations described below.

### Data-rich-regime simplification

In the data-rich regime, when the parameter likelihood  $P(\mathcal{D}|\theta)$  is concentrated at a sharp peak positioned at  $\hat{\theta}_{ML}$ , likelihood ratio depends only on the value of utility at  $\hat{\theta}_{ML}$ :

$$\lambda = 2 \log \frac{\max_{\beta > 0} \int_{\Theta} P(\mathcal{D}|\theta) P(\theta|\beta) d\theta}{\int_{\Theta} P(\mathcal{D}|\theta) P(\theta|\beta=0) d\theta} \quad (\text{Equation 5})$$

$$= 2 \log \frac{\max_{\beta > 0} P(\hat{\theta}_{ML}|\beta)}{P(\hat{\theta}_{ML}|\beta=0)} \quad (\text{Equation 6})$$

$$= 2 \log \left( Z(0) \max_{\beta > 0} \frac{e^{\beta U(\hat{\theta}_{ML})}}{Z(\beta)} \right), \quad (\text{Equation 7})$$

which is a non-decreasing function of the utility  $U(\hat{\theta}_{ML})$ . Thus, this test is equivalent to a test that uses the utility estimate itself,  $U(\hat{\theta}_{ML})$ , as the test statistic, making it possible to avoid the costly integration over  $\Theta$ . The null distribution can then be obtained by computing  $U(\theta)$  at uniformly sampled  $\theta$ .

### Multiple system instances simplification

If multiple instances of the system are available and we can assume that their parameters  $\theta_1, \theta_2, \dots, \theta_N$  are i.i.d. samples from the same distribution  $P(\theta|\beta)$ , then the datasets  $\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_N$  are also i.i.d.,  $P(\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_N|\beta) = \prod_{n=1}^N P(\mathcal{D}_n|\beta)$ . We test the hypotheses  $\beta = 0$  versus  $\beta > 0$  with the likelihood ratio statistic

$$\lambda = 2 \log \frac{\max_{\beta > 0} \prod_{n=1}^N P(\mathcal{D}_n|\beta)}{\prod_{n=1}^N P(\mathcal{D}_n|\beta=0)}. \quad (\text{Equation 8})$$

By Wilks' theorem, for large  $N$  the null distribution of  $\lambda$  approaches the  $\chi_1^2$  distribution (with a point mass of weight 1/2 at  $\lambda = 0$ , because we only consider  $\beta \geq 0$ ). This removes the need for sampling in order to obtain the null distribution.

### Hierarchical inference of population optimality

Assuming that experimental datasets  $\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_N$  are i.i.d., the posterior over population optimality parameter  $\beta$  takes the form:

$$P(\beta|\mathcal{D}_1, \dots, \mathcal{D}_N) \propto P(\beta) \prod_{n=1}^N \int_{\theta_n} P(\mathcal{D}_n|\theta_n) P(\theta_n|\beta) d\theta_n, \quad (\text{Equation 9})$$

where  $\theta = (k_n, x_{0,n})$  is a vector of neural parameters (slope and position), and  $P(\beta)$  is a prior over  $\beta$ . We approximated integrals numerically via the method of squares. Neural parameter values were sampled from ground-truth distributions via rejection sampling.

### Inference of receptive fields with optimality priors

We randomly sampled  $16 \times 16$  pixel image patches from the van Hateren natural image database (van Hateren and van der Schaaf, 1998) and standardized them to zero mean and unit standard deviation. Neural responses were simulated using a Linear-Nonlinear Poisson (LNP) model:

$$P(r_t|x_t, \phi, k, x_0) = \frac{\lambda_t^{r_t} e^{-\lambda_t}}{r_t!}, \quad (\text{Equation 10})$$

where  $\lambda_t$  is the rate parameter equal to:

$$\lambda_t = \frac{L}{1 + \exp[-\phi^T x_t]}, \quad (\text{Equation 11})$$

where  $L = 20$  was the maximal firing rate.

Given a linear filter  $\phi$ , we quantified sparsity of its responses to natural images using the following function:

$$U_{SC}(\phi) = - \langle |\phi^T x_t| \rangle. \quad (\text{Equation 12})$$

Filter sparsity was averaged across the natural image dataset consisting of  $5 \cdot 10^4$  standardized image patches randomly drawn from the van Hateren image database. The mean and standard deviation of filters  $\phi$  was set to be 0 and 1 respectively. We optimized filters

which either maximize or minimize the sparse utility measure via gradient descent. Different random initializations led to different filter shapes.

The locality utility of neural filters was defined as follows:

$$U_{LO}(\phi) = - \sum_{i,j} \left( (i - i_{max})^2 + (j - j_{max})^2 \right) \phi_{ij}^2, \quad (\text{Equation 13})$$

where  $i_{max}, j_{max}$  are positions of the RF pixel with the largest absolute value. This definition of locality was introduced in Doi et al. (2012).

Sparsity and locality utilities were combined into a single utility:

$$U(\phi; \xi) = U_{SC}(\phi) + \xi U_{LO}(\phi). \quad (\text{Equation 14})$$

To estimate receptive fields (neural filters), we first simulated the responses of the model population to 2000 natural image patches. We estimated linear receptive fields from simulated data by computing the spike-triggered average (STA), a widely applied estimator of neural receptive fields (Sharpee, 2013). In the STA model, response of neuron  $n$  at time  $t$  is assumed to follow the normal distribution (Park and Pillow, 2017):

$$P(r_{t,n} | \mathbf{s}_{t,n}, \phi_n) = \mathcal{N}(\phi_n^T \mathbf{s}_t; \sigma^2) \quad (\text{Equation 15})$$

where  $\phi_n$  is the linear receptive field of the  $n$ -th neuron, and  $\sigma^2$  is the noise variance.

To infer the receptive fields from simulated neural responses using our framework, we assumed the following optimization prior over receptive fields derived from the sparsity utility in Equation 12:

$$P(\phi_n | \beta) \propto \exp[\beta U_{SC}(z(\phi_n))], \quad (\text{Equation 16})$$

where  $z(\phi_n)$  denotes normalization of the receptive field to zero mean and unit variance. The sparse utility was evaluated over  $10^4$  randomly sampled image patches. The resulting log-posterior took the following form:

$$E(\phi_n | D, S, \beta) \propto - \frac{1}{\sigma^2} \sum_{t=1}^T (\phi_n^T \mathbf{s}_t - r_{t,n})^2 + \beta U_{SC}(z(\phi_n)). \quad (\text{Equation 17})$$

MAP inference was performed via gradient ascent on the log-posterior. Receptive fields were inferred with different priors corresponding to following values of the  $\beta$  parameter: 0, 1, 10, 20, 100. Receptive fields were estimated after reducing the dimensionality of stimuli with Principal Component Analysis to 64 dimensions. Estimation via gradient ascent on the log-posterior was performed in the PCA domain. PCA preprocessing is equivalent to low-pass filtering the stimuli.

To estimate value of the locality constraint  $\xi$  as well as the prior strength  $\beta$  via cross-validation, we split the data into the training and testing datasets comprising of 80% and 20% of data respectively. We estimated receptive fields for a range of  $\beta$  and  $\xi$  values ([0, 0.01, 0.1, 1, 10] and [0, 0.05, 0.2, 1] respectively). For each MAP RF estimate, we predicted neural responses  $\hat{r}_t$  using stimuli from the test dataset. We then computed the average error  $\langle (\hat{r}_t - r_t)^2 \rangle$  using neural responses in the test dataset. Combination of hyperparameters  $\xi, \beta$  which resulted in the smallest error value was taken to be the estimate of the correct one.

### Analysis of V1 receptive fields

Receptive fields of 250 neurons in the Macaque V1 were published and analyzed in Ringach (2002). All receptive fields were down-sampled to 32x32 pixels size and normalized to have zero mean and unit variance.

To evaluate sparseness of V1 receptive fields, we relied on the following sparse utility:

$$U_{SC}(\phi) = \langle \log \left( 1 + (z(\phi^T) x_t)^2 \right) \rangle_t, \quad (\text{Equation 18})$$

where  $x_t$  are individual image patches and  $z(\phi_n)$  denotes normalization of the receptive field to zero mean and unit variance. The sparse utility was evaluated over  $5 \times 10^4$  randomly sampled image patches. This form of the sparse utility was proposed in Olshausen and Field (1997), and together with the measure specified in Equation 12 it belongs to a broad class of equivalent sparsity measures defined by convex functions (Hyvärinen et al., 2009).

To test individual RFs for optimality, we generated the null distribution of utility values by bootstrapping  $10^6$  random filters as follows: (i) draw a random integer  $K$  between 1 and 128; (ii) superimpose  $K$  randomly selected principal components of natural image patches; each component is multiplied by a random coefficient  $v \sim \mathcal{N}(0, 1)$ ; (iii) generate a 2D Gaussian spatial mask centered at a random position on the image patch; lengths of horizontal and vertical axes of the Gaussian ellipse were drawn independently; (iv) multiply the random filter and the Gaussian mask. This procedure ensures that a range of filters of different sparsity and slowness will be randomly generated. Filters were standardized to zero mean and unit standard deviation.

To establish a measure of optimality at a population level, we needed to simplify the integration over all receptive field parameters, which was intractable due to their high-dimensionality. Computation of posteriors over  $\beta$  in Equation 9 was therefore approximated as follows:

$$P(\beta|\mathcal{D}_1, \dots, \mathcal{D}_N) \approx P(\beta) \prod_{n=1}^N \frac{1}{Z(\beta)} P(\hat{\theta}_n|\beta). \quad (\text{Equation 19})$$

where  $\hat{\theta}$  are receptive fields estimates computed in Ringach (2002).

We approximated  $P(\hat{\theta}_n|\beta)$  via rejection sampling, noting that  $P(\hat{\theta}_n|\beta) = P(U(\hat{\theta}_n)|\beta)$ , i.e., the probability of a high dimensional receptive field is determined solely by a one-dimensional utility function.

For each  $\beta$  we randomly sampled  $10^6$  filters from the proposal distribution, as described above, and retained only those consistent with  $P(U_{SC}(\theta)|\beta)$  via rejection sampling. Obtained utility values were fitted with a Gaussian distribution, used to evaluate posteriors over  $\beta$ , with point estimates being posterior maxima; the prior over  $\beta$  was uniform over the range displayed in the figures. For sparse utility, we discretized  $\beta$  values into 20 values equally spaced on the  $[-5, 5]$  interval. Filters accepted for each  $\beta$  value were used to compute the average spatial autocorrelation.

For comparison we used optimally sparse receptive fields learned from natural image patches preprocessed with PCA. We note that this preprocessing step might not have a direct biological counterpart. To compare optimal solutions and neural data, we therefore evaluated sparsity of model and real V1 RFs in the domain of natural images without PCA preprocessing.

To cluster receptive fields according to optimality, we defined a mixture model:

$$P\left(\theta_n|\{w_1, \dots, w_K\}, \{\beta_1, \dots, \beta_K\}\right) = \sum_{k=1}^K w_k P(U_{SC}(\theta_n)|\beta_k) \quad (\text{Equation 20})$$

where  $w_k$  is the weight of the  $k$ th mixture component and  $\beta_k$  is the optimality of that component. To approximate utility-defined distributions, we used the Gaussian approximation described above i.e.:  $P(\theta_n|\beta) = P(U_{SC}(\theta_n)|\beta) = \mathcal{N}(U_{SC}(\theta_n); \mu_\beta, \sigma_\beta^2)$

Parameters of the model were learned via the standard expectation-maximization algorithm (EM).

### Analysis of retinal receptive fields

Temporal receptive fields of retinal ganglion cells were published and analyzed in Deny et al. (2017). We analyzed RFs of 117 neurons selected by temporal smoothness. Each RF was normalized to unit norm and fitted with a parametric biphasic filter model described in Sun et al. (2017).

We considered two different utility functions. First one was a generalization of the predictive coding objective introduced in Srinivasan et al. (1982). The predictive coding objective minimizes the squared difference between the stimulus value  $s_t$  at time  $t$  and the linear prediction of that stimulus value computed from  $N$  past values:  $E(\phi) = [\sum_{\tau=0}^N \phi_\tau s_{t-\tau}]^2$ , where  $\phi_\tau$  are the weights of the linear filter. In the classical approach it has been assumed that the linear weight of the current stimulus  $s_t$  is equal to 1 i.e.,  $\phi_0 = 1$ . We note that such form makes it difficult to evaluate predictive coding filters adapted to stimuli of unknown temporal scale. In particular, we optimize and evaluate our filters on natural movies whose frame rate might be mismatched with the timescale of the retina. We therefore relax the assumption that the predictive coding filter reduces the dynamic range by subtracting only the current stimulus from its prediction, and assume that what is being predicted is itself a linear combination of stimulus values (e.g., integrating stimulus value over some recent period of time). In practice this means that we allow all values of the filter including  $\phi_0$  to vary freely. To avoid trivial solutions, where the residue  $E(\phi)$  is minimized by setting all weights to 0, we impose a unit norm constraint on the filter  $\phi$ . The utility function of a filter  $\phi$  is then equal to:

$$U_{PC}(\phi) = - \left\langle \left[ \sum_{\tau=0}^N z(\phi)_\tau s_{n,t-\tau} \right]^2 \right\rangle_n \quad (\text{Equation 21})$$

where  $z$  denotes the unit norm operator, and  $n$  indexes stimulus epochs  $s_n$ .

We evaluated the utility  $U_{PC}$  using 50000, 21-sample long excerpts of single-pixel luminance extracted from natural movies of scenes in the African savanna (van Hateren and Ruderman, 1998).

We used these natural stimulus data to learn the optimal predictive-coding filter, as described in Srinivasan et al. (1982) via gradient descent.

The second considered utility was measuring the amount of information between the stimulus and the instantaneous filter output in a low-noise regime. Under the Gaussian approximation of stimulus and output distribution this utility takes the form:

$$U_{II}(\phi) = -\frac{1}{2} \log(1 - \rho^2), \quad (\text{Equation 22})$$

where  $\rho$  is the Pearson correlation coefficient between the stimulus  $s_t$  and the filter output  $r_t$ . This utility is high when the neural responses track the stimulus with high fidelity. Note that this is not the general solution to an efficient coding (infomax) problem, where the *full response trajectory*, not the instantaneous response, should encode high information about the stimulus, which leads to decorrelation / whitening in the low-noise regime. We evaluated  $U_{II}$  using a trajectory of 20000 samples of pixel intensity values extracted from the natural movie dataset.

To compute utility-defined distributions of the filter mode amplitude parameters  $c_1, c_2$ , we first discretized values of these parameters into 100 values uniformly spaced on the  $[0.01, 13]$  interval, where 13 was the maximum amplitude parameter value among fits to

normalized retinal RFs. For each filter we evaluated utility for each pair of discretized amplitude parameter values and a fixed value of the scale parameter  $a$  fitted to that filter. We used such utility surfaces to estimate the normalization constant of the utility-defined distribution parametrized by  $\beta$  and the scale parameter  $a$ .

We discretized the parameter  $\beta$  into 100 values uniformly spaced on the  $[-10, 64]$  interval. We estimated the posterior over  $\beta$  by numerically integrating over filter parameters  $c_1, c_2, a$ . We assumed a uniform prior over  $\beta$ .

### Analysis of connectivity in *C. elegans*

For our analysis we used the *C. elegans* neural wiring dataset available on Worm Atlas (<https://www.wormatlas.org>). This dataset has been published and analyzed before in Chen et al. (2006) as well as Pérez-Escudero et al. (2009; Pérez-Escudero and de Polavieja (2007) – for details about the dataset please refer to this prior work.

For the analyses depicted in Figure 8 we selected two sets of neurons. The first set consisted of 126 neurons connected to at least one muscle, and the second set consisted of 86 neurons connected to at least one sensor. "i-th" neuron was therefore characterized by its position,  $x_i$ , number of landmark cells (muscles or sensors) it was connected to,  $N_i$ , vectors of positions of the landmark cells,  $m_i$  (muscles), and  $s_i$  (sensors), and vectors of the number of synapses in each neuron-to-landmark connection,  $n_i$ . For each neuron the utility of its position was defined as:

$$U_{WC}(x_i; N_i, l_i, n_i, \xi) = - \sum_{j=1}^{N_i} n_{ij} |x_i - l_{ij}|^{\xi}. \quad (\text{Equation 23})$$

where  $l_i \in \{m_i, s_i\}$ , denotes the vector of landmark cell positions. We evaluated the utility function on the  $[0, 1]$  interval representing the linear extent of the worm body axis, discretized into 100 linearly spaced values. To compute the posterior distribution over parameters  $\beta$  and  $\xi$  we discretized them into 64 linearly spaced values. For neuron-muscle connections,  $\beta$  was defined over a  $[1.5, 4]$  interval and  $\xi$  over a  $[1.3, 1.9]$  interval. For neuron-sensor connections,  $\beta$  was defined over a  $[10, 25]$  interval and  $\xi$  over a  $[1.5, 2.2]$  interval. We assumed a uniform prior over parameters  $\beta, \xi$ .

### QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical test performed in Figure 5D was a two-tailed t test. Stars denote p values lower than 0.001. Error bars in the figure denote standard errors of the mean.