

# Towards a unified theory of efficient, predictive and sparse coding

Matthew Chalk<sup>\*</sup>, Olivier Marre<sup>†</sup>, and Gašper Tkačik<sup>\*</sup>

<sup>\*</sup>Institute of Science and Technology Austria, Am Campus 1, A-3400 Klosterneuburg, Austria, and <sup>†</sup>Institut de la Vision, INSERM UMRS 968, UPMC UM 80, CNRS UMR 7210, Paris, France

**A central goal in theoretical neuroscience is to predict the response properties of sensory neurons from first principles. Several theories have been proposed to this end. “Efficient coding” posits that neural circuits maximise information encoded about their inputs. “Sparse coding” posits that individual neurons respond selectively to specific, rarely occurring, features. Finally, “predictive coding” posits that neurons preferentially encode stimuli that are useful for making predictions. Except in special cases, it is unclear how these theories relate to each other, or what is expected if different coding objectives are combined. To address this question, we developed a unified framework that encompasses these previous theories and extends to new regimes, such as sparse predictive coding. We explore cases when different coding objectives exert conflicting or synergistic effects on neural response properties. We show that predictive coding can lead neurons to either correlate or decorrelate their inputs, depending on presented stimuli, while (at low-noise) efficient coding always predicts decorrelation. We compare predictive versus sparse coding of natural movies, showing that the two theories predict qualitatively different neural responses to visual motion. Our approach promises a way to explain the observed diversity of sensory neural responses, as due to a multiplicity of functional goals performed by different cell types and/or circuits.**

Sensory neural circuits perform a myriad of computations, which allow us to make sense of, and interact with, our environment. For example, neurons in the primary visual cortex encode information about local edges in an image [1], while neurons in higher-level areas encode more complex features, such as textures or faces [2, 3]. A central aim of sensory neuroscience is to develop a mathematical theory to explain the purpose and nature of such computations, and, ultimately, predict neural responses to stimuli from first principles.

Several theories have been proposed about the function that sensory systems have evolved to perform. The efficient coding hypothesis posits that sensory circuits transmit maximal information about their inputs, given internal constraints, such as metabolic costs and/or noise [4, 5, 6, 7]. Alternatively, the sparse coding hypothesis posits that individual neurons respond selectively to specific, rarely occurring, features in the environment [8, 9, 10]. Finally, the more recent predictive coding hypothesis<sup>1</sup> posits that sensory neurons transmit maximal information about stimuli that are predictive about the future while discarding non-predictive information [11, 12].

One may ask which, if any, of these objectives are fulfilled by sensory neural circuits. This question is all the more important given that, in many cases, different coding objectives appear to directly conflict with each other. For example, a classic result of efficient coding in the low-noise regime is that neurons should temporally decorrelate their inputs and preferentially encode fast stimulus features [13, 14, 15, 16]. In contrast, predictive coding favours the extraction of temporally-correlated, slow features [17, 18]. Likewise, sparse coding requires that neurons respond selectively to a single, preferred stimulus feature. It is unclear if this is compatible with predictive coding, which requires neurons to respond to stimuli as quickly as possible.

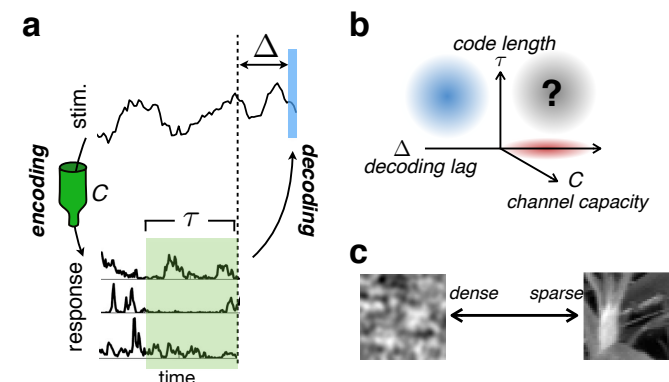
While a large body of theoretical work exists on efficient and sparse coding (reviewed in [19, 20]), there is little work on how neurons could optimally encode stimuli that are pre-

dictive about the future (with the exception of [21, 17]); in short, the general implications of predictive coding for neural circuits are still unknown. We also do not understand how different coding objectives relate to each other, or what happens when they are combined.

Here, we incorporate the three existing theories—sparse coding, efficient coding, and predictive coding—into a unified framework. In this framework, a small set of optimisation parameters determines the functional goals and constraints faced by sensory neurons. Previous theories correspond to specific values of these optimisation parameters. As a result, we can investigate the conditions under which different coding objectives, such as encoding predictive information versus maximising efficiency, have conflicting or synergistic effects on neural responses. Further, we can explore qualitatively new coding regimes, such as neural codes that are both predictive and sparse. We end by hypothesizing that the observed diversity of sensory neural responses spans the space of coding tradeoffs accessible by varying the parameters of our new theory.

## A unified framework for predictive and efficient coding

We consider a stimulus,  $y_{-\infty:t} \equiv (\dots, y_{t-1}, y_t)$ , giving rise to a sensory input,  $x_t = y_t + n_t$ , where  $n_t$  represents input noise. We look for the optimal neural code,  $p(r_t|x_{-\infty:t})$ , such that neural responses  $r$  within a temporal window of length



**Fig. 1. Schematic of modeling framework. (a)** A stimulus (above) elicits a response in a population of neurons (below). We look for optimal codes, where the responses within a time window of length  $\tau$  maximise information encoded about the stimulus at lag  $\Delta$ , subject to a constraint on the information about past inputs,  $C$ . **(b)** For a given stimulus, the optimal code depends on three parameters:  $\tau$ ,  $\Delta$ , and  $C$ . Previous work on efficient temporal coding looked at  $\tau > 0$ , and  $\Delta < 0$  (blue shade). Previous work on predictive coding looked at  $\Delta > 0$  and  $\tau \sim 0$  (red shade). Our theory is valid in all regimes, but we focus in particular on  $\Delta > 0$  &  $\tau > 0$  (black shade). **(c)** We further explore how optimal codes change when there is a sparse latent structure in the stimulus (natural image patch, right) vs when there is none (filtered Gaussian noise, left).

<sup>1</sup>The term ‘predictive coding’ has been used previously to describe several different approaches. In our work, we use the definition given by [11], where neurons encode maximal information about the future, given information encoded about the past. Alternative definitions are described in the discussion.

$\tau$  encode maximal information about the stimulus at time lag  $\Delta$ , given a fixed amount of information encoded about past inputs,  $C$  (Fig a).

This problem can be formalised using the information bottleneck (IB) framework [22], in which one seeks a code,  $p(r_t|x_{-\infty:t})$ , that maximises the objective function:

$$L_{p(r_t|x_{-\infty:t})} = I(R_{t-\tau:t}; Y_{t+\Delta}) - \gamma I(R_{t-\tau:t}; X_{-\infty:t}), \quad [1]$$

where the first term is the mutual information between responses  $R_{t-\tau:t}$  and stimulus  $Y_{t+\Delta}$ , to be maximised, and the second term denotes the mutual information between  $R_{t-\tau:t}$  and  $X_{-\infty:t}$ , to be constrained (which we call the channel capacity,  $C$ ). A constant,  $\gamma$ , determines the strength of this constraint, and thus, the tradeoff between coding fidelity and compression. When  $\gamma = 0$  the optimal solution is to encode the input perfectly; when  $\gamma = 1$  the optimal solution is to encode zero information about the input.

In general, it is impossible to exactly maximise the objective function in Eq (1). We previously presented an approximate method that instead maximizes the information about the stimulus which can be recovered from neural responses using a specific type of decoder (e.g., a linear decoder); formally, this amounts to a variational approximation that maximizes a lower bound on  $L$  [23], as described in SI Section 1. To recover efficient coding in the limit where  $X \approx Y$ , we replaced  $R_{t-\tau:t}$  in the second term of Eq (1) with the instantaneous response,  $R_t$ . With this modification, maximising  $L$  for  $\Delta < 0$  is equivalent to minimising the redundancy in the responses, as in efficient coding (see Table 1 and SI Section 1).

Equation (1) clearly shows that the optimal coding strategy depends on three factors: the *decoding lag*,  $\Delta$ , the *code length*,  $\tau$ , and the *channel capacity*,  $C$  (determined by  $\gamma$ ). Previous theories of neural coding correspond to specific regions within the three-dimensional parameter space spanned by  $\Delta$ ,  $\tau$ , and  $C$  (Fig b). For example, efficient coding investigated how, at low noise, neurons transmit maximal information about past inputs ( $\Delta < 0$ ) by minimising temporal redundancy in their responses [13, 14]. This strategy is optimal when the stimulus can be read-out by integrating neural responses over time, i.e., when  $\tau > 0$  (blue region in Fig b). In contrast, predictive coding ( $\Delta > 0$ ) looked exclusively at near-instantaneous codes, where  $\tau \sim 0$  (red region in Fig b)<sup>2</sup> [12, 21, 17]. Below, we investigate the relation between these previous works and focus on the (previously unexplored) case of neural codes that are both predictive ( $\Delta > 0$ ) and temporal ( $\tau > 0$ ; grey region in Fig b). To specialize our theory to the biologically-relevant case, we further investigate predictive coding of natural stimuli. A hallmark of natural stimuli is their sparse structure [19, 20, 8, 24]: stimulus fragments can be constructed from a set of primitive features (e.g., image contours), each of which occurs rarely (Fig c). By incorporating sparsity into our information-theoretic framework, we explore the relationship between sparse and predictive coding.

## Results

**Dependence of neural code on coding objectives.** Our initial goal was to understand the influence of different coding objectives in the simplest scenario, where a single neuron linearly encodes a 1-d input. In this model, the neural response at time  $t$  is:

$$r_t = \sum_{k=0}^{\tau_w} w_k x_{t-k} + \eta_t, \quad [2]$$

where  $w = (w_0, \dots, w_{\tau_w})$  are the linear coding weights and  $\eta_t$  is a gaussian noise with unit variance.

With stimuli that have gaussian statistics, the objective function takes a very simple form:

$$L = -\frac{1}{2} \log \left\langle \left( y_{t+\Delta} - \sum_{k=0}^{\tau} u_k r_{t-k} \right)^2 \right\rangle - \gamma \frac{1}{2} \log \langle r_t^2 \rangle \quad [3]$$

where  $u = (u_0, \dots, u_{\tau})$  are the optimal linear read-out weights used to reconstruct the stimulus at time  $t + \Delta$  from the responses between  $t - \tau$  and  $t$ . Thus, the optimal code is the one that minimises the mean-squared reconstruction error at lag  $\Delta$ , constrained by the variance of the neural response (relative to the noise variance).

Initially, we investigated “instantaneous” predictive coding, where  $\tau = 0$ , so that the stimulus at time  $t + \Delta$  is estimated from the instantaneous neural response at time  $t$  (Fig 2a). We considered three different stimulus types, shown in Fig 2b. With a “Markov” stimulus, whose future trajectory depended on the current state,  $y_t$ , only (Fig 2b, top panel; see SI Section 2.1), to predict the stimulus at a future time,  $y_{t+\Delta}$ , neurons only needed to encode the current state  $y_t$ . Thus, when  $\tau = 0$ , we observed the trivial solution where  $r_t \propto y_t$ , irrespective of the decoding lag,  $\Delta$  (Fig 2c-d, top panels).

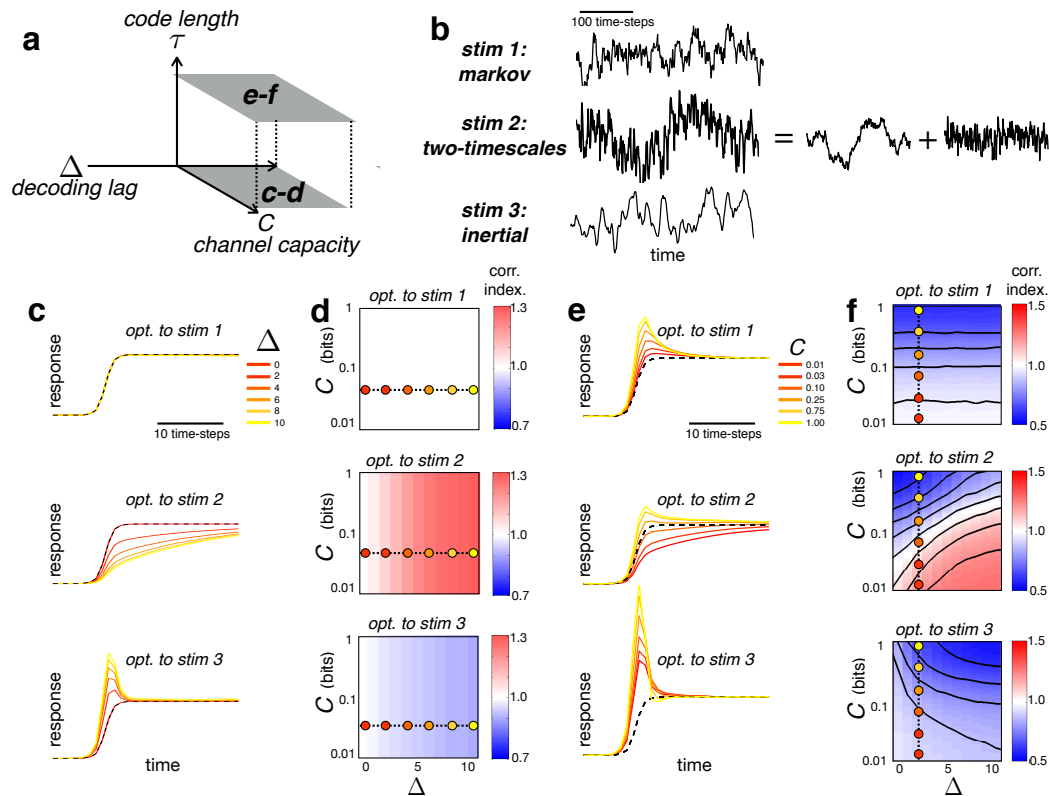
With a “two-timescale” stimulus, constructed from two Markov processes that vary over different timescales (Fig 2b, middle panel), the optimal solution was a low-pass filter, to selectively encode the predictive, slowly varying, part of the stimulus. The strength of the low-pass filter increased monotonically with the decoding lag,  $\Delta$  (Fig 2c-d, middle panels).

Finally, with an “inertial” stimulus, whose future trajectory depended on the previous two states,  $y_t$ , and  $y_{t-1}$  (Fig 2b, lower panel), the optimal solution was a high-pass filter, so as to transmit information about velocity. The strength of the high-pass filter also increased monotonically with the decoding lag,  $\Delta$  (Fig 2c-d, lower panels).

With an instantaneous code, varying the channel capacity,  $C$ , only rescales responses (relative to the noise amplitude), so as to alter their signal-to-noise ratio. However, the response shape is left unchanged (regardless of the stimulus statistics; Fig 2d). In contrast, with temporally extended codes, where  $\tau > 0$  (so the stimulus at time  $t + \Delta$  is estimated from the integrated responses between time  $t - \tau$  and  $t$ ; Fig 2a) the optimal neural code varies with the channel capacity,  $C$ . In common with classical efficient coding, at high  $C$  (i.e. high signal-to-noise ratio) neurons *always* decorrelated their input, regardless of both the stimulus statistics and decoding lag,  $\Delta$ . Also in common with classical efficient coding, decreasing  $C$  always led to more correlated responses [7]. However, *unlike* efficient coding, at low to intermediate values of  $C$  (i.e. intermediate to low signal-to-noise ratio) the optimal code was qualitatively altered by varying the decoding lag,  $\Delta$ . With the Markov stimulus, increasing  $\Delta$  had no effect; with the two-timescale stimulus it led to low-pass filtering; and with the inertial stimulus it led to stronger high-pass filtering.

Taken together, “phase diagrams” for optimal, temporally-extended codes show how regimes of decorrelation/whitening (high-pass filtering) and of smoothing (low-pass filtering) are preferred depending on channel capacity,  $C$ , and decoding lag,  $\Delta$ . We verified that a qualitatively similar transition from low- to high-pass filtering is also observed with higher dimensional stimuli, and/or more neurons. Importantly, we show that these phase diagrams depend in an essential way on the stimulus statistics already in the linear, gaussian case. We next examined what happens for non-gaussian, high-dimensional stimuli.

<sup>2</sup>In other words, previous efficient coding models maximised the encoded information *rate* at time  $t$ , while previous predictive coding models maximised the total encoded information at time  $t$ .



**Fig. 2. Dependence of optimal code on decoding lag,  $\Delta$ , code length,  $\tau$ , and channel capacity,  $C$ .** (a) We investigated two types of code: instantaneous codes, where  $\tau = 0$  (panels c-d), and temporally extended codes, where  $\tau > 0$  (panels e-f). (b) Training stimuli used in our simulations. Markov stimulus: the future only depends on the present state. Two-timescale stimulus: sum of two Markov processes that vary over different timescales (shown at right). Inertial stimulus: future depends on stimulus at previous two time-steps. (c) Neural responses to probe stimulus (dashed line) after optimising code with varying  $\Delta$ , and  $\tau = 0$ . Responses are normalised by the final, steady state value. (d) Correlation index after optimisation with varying  $\Delta$  &  $C$ . Correlation index measures the correlation between neural responses at adjacent timesteps, normalized by the stimulus correlation at adjacent timesteps, i.e.,  $\langle r_t r_{t+1} \rangle / \langle r_t^2 \rangle$  divided by  $\langle x_t x_{t+1} \rangle / \langle x_t^2 \rangle$ . Values greater / smaller than 1 indicate that neurons temporally correlate (red) / decorrelate (blue) their input. Filled circles show the parameter values used in panel c. (e-f) Same as c-d, but with code optimised for  $\tau \gg 0$ . Plots in panel e correspond to responses to probe stimulus (dashed line) at varying channel capacity & fixed decoding lag (i.e.,  $\Delta = 3$ , indicated by dashed line in panel f).

**Predictive versus efficient coding of naturalistic stimuli.** Natural stimuli exhibit a strongly non-gaussian statistical structure which is essential for human perception [25, 24]. A large body of work has investigated how neurons could efficiently represent such stimuli by encoding their non-redundant, or independent, components [19]. Under fairly general conditions, this is equivalent to finding a sparse code, where each neuron responds selectively to a single, rarely occurring, stimulus feature. For natural images this leads to neurons selective for spatially localised image contours, qualitatively similar to the receptive fields (RFs) of V1 simple cells [8, 26]. For natural movies this leads to neurons selective for a particular motion direction, again similar to observations in area V1 [27].

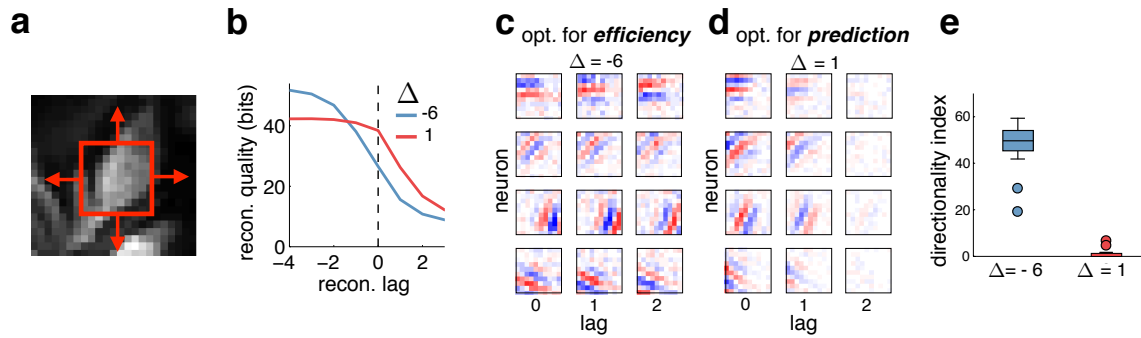
However, an independent (sparse) temporal code has only been shown to be optimal when: (i) the goal is to maximise information about *past* inputs, i.e.,  $\Delta < 0$ ; (ii) at low noise, i.e., at high capacity,  $C \gg 0$ . We were interested, therefore, in what happens when these two criteria are violated; for example when neural responses are optimised to encode predictive information, i.e., for  $\Delta \geq 0$ .

To explore these questions we modified the objective function of Eq (3) to deal with multi-dimensional stimuli and non-gaussian statistics of natural images. To achieve this, we generalized the second term of our objective function to allow optimization of the neural code with respect to higher-order

(i.e., beyond covariance) response statistics. Crucially, this modification, described in SI Section 1 and [23], permits—but does not enforce by hand—the sparsity of neural responses. For non-sparse, gaussian stimuli the modification automatically recovers the results of the previous section; for natural stimuli it replicates previous sparse coding results in the limit  $\Delta < 0$  and  $C \gg 0$  (see SI Fig 3), without introducing any new tuneable parameters.

We investigated how the optimal neural code for natural stimuli varied with the decoding lag,  $\Delta$ , while keeping channel capacity,  $C$ , and code length,  $\tau$ , constant. Stimuli were constructed from  $10 \times 10$  pixel patches drifting stochastically across static natural images (Fig 3a & SI Fig 1; see SI Section 2.2). Neural encoding weights were optimised with two different decoding lags: for  $\Delta = -6$  the goal was to efficiently encode the past, while for  $\Delta = 1$  the goal was to predict the near future. Figure 3b confirms that the codes indeed are optimal either for efficiency ( $\Delta = -6$ ) or prediction ( $\Delta = 1$ ), as desired.

After optimisation at both values of  $\Delta$ , individual neurons were selective to local oriented edge features (Fig 3c-d) [8]. Varying  $\Delta$  qualitatively altered the temporal features encoded by each neuron, while having little effect on their spatial selectivity. Consistent with previous results in the efficient coding regime [27], single cells at  $\Delta = -6$  were responsive to stimuli moving in a preferred direction, as evidenced by spatially



**Fig. 3. Efficient versus predictive coding of natural stimuli.** (a) Movies were constructed from a  $10 \times 10$  pixel patch (red square) which drifted stochastically across static natural images. (b) Information encoded by neural responses about the stimulus at varying lag, after optimization with  $\Delta = -6$  (blue) and  $\Delta = 1$  (red). (c) Spatio-temporal encoding filters, for 4 example neurons, after optimisation with  $\Delta = -6$ . (d) Same as panel c, for  $\Delta = 1$ . (e) Directionality index of neural responses, after optimisation with  $\Delta = -6$  and  $\Delta = 1$ . The directionality index measures the percentage change in response to a grating stimulus moving in a neuron’s preferred direction, versus the same stimulus moving in the opposite direction.

displaced encoding filters at different lags (Fig 3c & SI Fig 4a-c), and a high “directionality index” (Fig 3e). In contrast, for predictive coding setup at  $\Delta = 1$ , cells responded equally to stimuli moving in either direction perpendicular to their encoded stimulus orientation. This was evidenced by spatio-temporally separable receptive fields (SI Fig 4d-f) and directionality indexes near zero. This qualitative difference between the efficient and predictive code for natural movies was highly surprising, and we sought to understand its origins.

**Trade-off between sparsity and predictive power.** To gain an intuitive understanding of how the optimal code varies with decoding lag  $\Delta$ , we constructed artificial stimuli from overlapping “gaussian bumps” which drifted stochastically along a single spatial dimension (Fig 4a; SI Section 2.3). While simple, this stimulus captured two key aspects of the natural movies: first, the gaussian bumps drifted smoothly in space, resembling the stochastic global motion over the image patches; second, the stimulus also had a sparse latent structure.

We optimised the neural code with  $\Delta$  ranging from  $-2$  to  $2$ , holding the channel capacity,  $C$ , and code length,  $\tau$ , constant. Fig 4b confirms that highest performance was achieved when the reconstruction performance was evaluated at the same lag for which each model was trained. This simpler setup recapitulated the surprising result we obtained with naturalistic stimuli: namely, that when  $\Delta < 0$  neurons were selective to a single preferred motion direction, while when  $\Delta \geq 0$  neurons responded equally to stimuli moving from either direction into their receptive field (Fig 4c-d).

Predicting the future state of the stimulus requires estimating its current motion direction and speed. How is it possible, then, that an optimal predictive code ( $\Delta > 0$ ) results in neurons being unselective to motion direction? This paradox is resolved by realising that it is the information encoded by the *entire neural population* that counts, not the information encoded by individual neurons. Indeed, when we looked at the information encoded by the neural population, we did find what we had originally expected: when optimised with  $\Delta > 0$ , the neural population as a whole encoded significantly more information about the stimulus velocity than its position (relative to when  $\Delta < 0$ ), despite the fact that individual neurons were unselective to motion direction (Fig 4e-f).

The change in coding strategy that is observed as one goes from efficient ( $\Delta < 0$ ) to predictive coding ( $\Delta \geq 0$ ) is in part due to a tradeoff between cells maintaining sparse responses (which is efficient) and responding quickly to stimuli

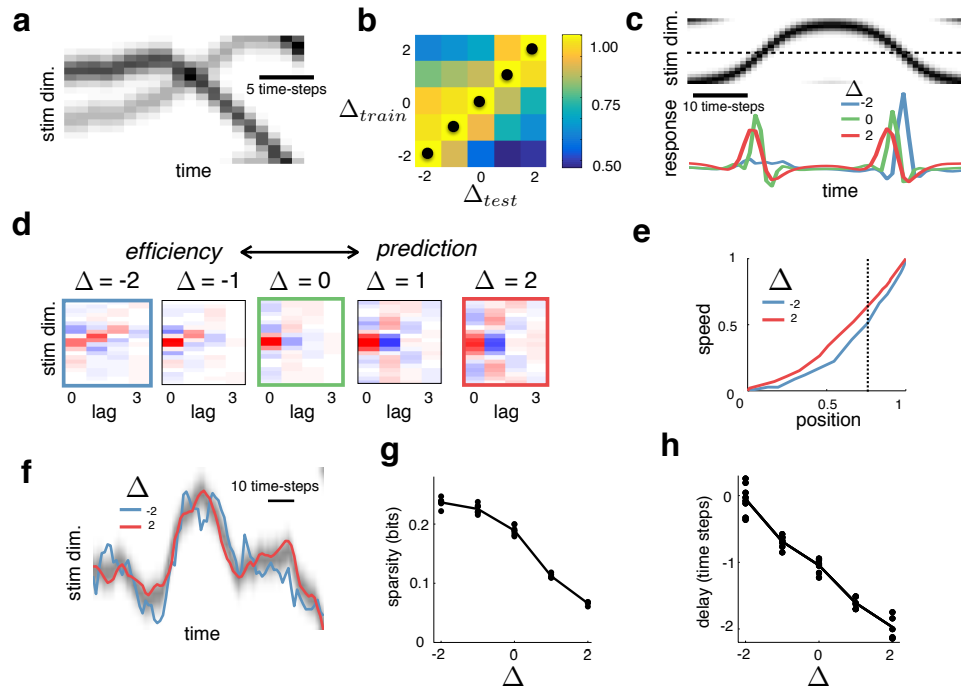
within their RF (which helps predictions). Intuitively, to be efficient and respond with greatest selectivity, the neuron first has to wait to process and recognize the “complete” stimulus feature; unavoidably, however, this entails a processing delay and leaves no information to be encoded predictively. This can be seen in Fig 4g-h, which shows how both the response sparsity and delay to stimuli within a cell’s RF decrease with  $\Delta$ . In SI section 3.4 we describe in detail why this trade-off between efficiency and prediction leads to direction selective filters when  $\Delta < 0$ , but not when  $\Delta > 0$  (SI fig. 5).

Beyond the effects on the optimal code of various factors explored in detail in the main paper, our framework further generalises previous efficient and sparse coding results to factors listed in Table 1 and discussed in SI Section 3.5. For example, decreasing the capacity,  $C$  (while holding  $\Delta$  constant at  $-2$ ) resulted in neurons being unselective to stimulus motion (SI fig 6a), with a similar result observed for increased input noise (SI fig 6b). Thus, far from being generic, traditional sparse temporal coding, in which neurons responded to local motion, was only observed in a specific regime (i.e.,  $\Delta < 0$ ,  $C \gg 0$  and low input noise,  $n \sim 0$ ).

## Discussion

Efficient coding has long been considered a central principle for understanding early sensory representations [4, 5], with well-understood implications and generalizations [9, 28]. It has been successful in predicting many aspects of neural responses in early sensory areas directly from the low-order statistics of natural stimuli [7, 29, 30, 31, 24], and has even been extended to higher-order statistics and central processing [32, 33]. However, a criticism of the theory is that it implicitly treats all sensory information as equal, despite empirical evidence that neural systems prioritise behaviourally relevant (and not just statistically likely) stimuli [34]. To overcome this limitation, Bialek and colleagues proposed an alternative theory, called predictive coding, which posits that neural systems encode maximal information about future inputs, given fixed information about the past [11, 12]. This theory is motivated by the fact that stimuli are only useful for performing actions when they are predictive about the future.

Compared to efficient coding, predictive coding has remained relatively unexplored (though see later for alternative definitions of predictive coding, which have received more attention). Existing work only considered the highly restrictive scenario where neurons maximise information encoded in their *instantaneous* responses [12, 21, 17]. In this case (and sub-



**Fig. 4. Efficient versus predictive coding of a “gaussian-bump” stimulus.** (a) Stimuli consisted of gaussian-bumps that drifted stochastically along a single spatial dimension (with circular boundary conditions). (b) Information encoded by neural responses about the stimulus at varying lag,  $\Delta_{test}$ , after optimization with varying  $\Delta_{train}$ . Black dots indicate the maximum for each column. (c) Response of example neuron to a test stimulus (above), after optimisation with  $\Delta = -2$  (blue),  $\Delta = 0$  (green), and  $\Delta = 2$  (red). (d) Spatio-temporal encoding filters for an example neuron, after optimisation with different  $\Delta$ . (e) Circular correlation between the reconstructed speed of a moving gaussian blob and its true speed, versus the circular correlation between the reconstructed position and its true position, obtained from neural responses optimised with  $\Delta = \pm 2$  (red and blue curves). Curves were obtained by varying  $\gamma$  in Eq (3), to find codes with different channel capacities. (f) Linear reconstruction of the stimulus trajectory, obtained from neural responses optimised with  $\Delta = \pm 2$  (red and blue curves). The full stimulus is shown in grayscale. While coding capacity was chosen to equalize the mean reconstruction error for both models (vertical dashed line in panel e), the reconstructed trajectory was much smoother for the predictive (red) than for the efficient (blue) coding model. (g) Response sparsity (defined as the negentropy of neural responses), versus  $\Delta$  (dots = individual neurons; line = population average). (h) Delay between stimulus presented at a neuron’s preferred location and each neuron’s maximum response, versus  $\Delta$ .

ject to some additional assumptions, such as gaussian stimulus statistics and instantaneous encoding filters), predictive coding is formally equivalent to slow feature analysis [18]. This is the exact opposite of efficient coding, which (at low noise/high capacity) predicts that neurons should temporally decorrelate their inputs [14].

To clarify the relation between efficient and predictive coding, we developed a unified framework that can treat both theories [22, 11, 23]. We investigated what happens when the neural code is optimised to be both predictive and temporally efficient (Fig b). In this case, the optimal code depends critically on the channel capacity (i.e. signal-to-noise ratio), which describes how much information the neurons can encode about their input. At high capacity (i.e. low-noise), neurons always temporally decorrelate their input. At finite capacity (i.e. mid-to high-noise), however, the optimal neural code varies qualitatively depending on whether the goal is to reliably predict the future or efficiently reconstruct the past.

When we investigated predictive coding of natural stimuli, we found solutions that are qualitatively different from known sparse coding results, in which individual neurons are tuned to directional motion of local edge features [27]. In contrast, we found that neurons optimised for predictive coding are selective for motion speed but not direction (Fig 3 and SI Fig 4). Surprisingly, however, the neural population as a whole encodes motion even more accurately than before (Fig 4e). We show that these changes are due to an implicit trade-off between maintaining a sparse code (which is efficient) and re-

**Table 1.** List of factors determining the optimal neural code. The first three factors are explored in detail in the main text.

Factor	Control parameter	Effect
Coding capacity	$C$	Fig 2
Decoding window	$\tau$	Fig 2
Decoding lag	$\Delta$	Fig 2-4
Input noise magnitude	noise $n$ , added to stim. $y$	SI 3.5.1
Temp. corr. in spiking	2 <sup>nd</sup> term in Eq (1)	SI 3.5.2
Stim. prediction window	$Y_{(t+\Delta_1:t+\Delta_2)}$ in Eq (1)	SI 3.5.3
Encoding model	parametric form of $p(r x)$	SI 3.5.4

sponding quickly to stimuli within each cell’s RF (which aids predictions; Fig 4g-h).

It is notable that, in our simulations, strikingly different conclusions are reached by analysing single neuron responses versus, the responses of the entire neural population. Specifically, looking only at single neuron responses would lead one to conclude that when performing predictive coding, neurons *did not* encode motion direction; looking at the neural population responses reveals that the opposite is true. This illustrates the importance of population-level analyses of neural data, and how, in many cases, single neuron responses can give a false impression of which information is represented by the population.

A major challenge in sensory neuroscience is to derive the observed cell-type diversity in sensory areas from a normative theory. For example, in visual area V1, one observes a

range of different cell-types, some of which have spatio-temporally separable RFs, and others which do not [35, 36]. The question arises, therefore, whether the difference between cell-types emerges because different subnetworks fulfill qualitatively different functional goals. One hypothesis, suggested by our work, is that cells with separable RFs have evolved to encode predictive information, while cells with non-separable RFs evolved to optimise coding efficiency. More generally, the same hypothesis could explain the existence of multiple cell-types in the mammalian retina [37], with each cell-type mosaic implementing an optimal code for a particular choice of optimisation parameters, e.g., channel capacity or prediction lag.

Testing such hypotheses rigorously against quantitative data would require us to generalise our work to nonlinear encoding and decoding models (Table 1, final row). Here we focused on a linear decoder to lay a solid theoretical foundation and permit direct comparison with previous sparse coding models, which also assumed a linear decoder [27, 8, 26]. In addition, a linear decoder forces our algorithm to find a neural code for which information can be easily extracted by downstream neurons performing biologically plausible operations. While the linearity assumptions simplify our analysis, the framework can easily accommodate non-linear encoding and decoding. For example, we previously used a “kernel” encoding model, where neural responses are described by a non-parametric & non-linear function of the input [23]. Others have similarly used a deep convolutional neural network as an encoder [38].

In the future it would be interesting to investigate how our ideas relate to sensory processing in a hierarchy. Hierarchical processing has long been discussed in the context of efficient coding [24], where neurons at each layer are assumed to remove residual statistical dependencies in their inputs [39, 19]. In contrast, hierarchical predictive coding, in which neurons

at each layer encode maximal information about their future inputs, has not yet been explored.

“Predictive coding” has been used to describe different approaches. Here, we understood the term in information-theoretic context, implying that neurons preferentially encode stimuli that carry information about the future [11]. However, predictive coding has also been used to imply that neurons encode “surprising” stimuli, i.e., those not predictable from past inputs [4, 40, 41]. Elsewhere, predictive coding describes a particular type of hierarchical processing, in which feed-forward projections encode an error signal, equal to the difference between bottom-up sensory inputs and top-down predictions from higher sensory areas [42, 43]. These alternative definitions of predictive coding are not equivalent. For example, sensory stimuli can be surprising based on past inputs, but not predictive about the future [44]. Likewise, previous theories of hierarchical predictive coding do not address which sensory information should be preferentially encoded or alternatively, discarded. Clarifying the relationship between these inequivalent definitions of predictive coding and linking them mathematically to coding efficiency provided one of the initial motivations for our work. In past work, alternative coding theories are often expressed using very different mathematical frameworks, impeding comparison between them, and sometimes leading to confusion. In contrast, by using a single mathematical framework to compare different theories—efficient, sparse and predictive coding—we were able to see exactly how they relate to each other, the circumstances under which they make opposing or similar predictions, and what happens when they are combined.

**Acknowledgements.** This work was supported in part by the Austrian Science Fund grant FWF P25651.

1. M. Carandini, J. B. Demb, V. Mante, D. J. Tolhurst, Y. Dan, B. A. Olshausen, J. L. Gallant, & N. C. Rust (2005) *J. Neurosci.* 25(46):10577–10597.
2. C. M. Ziemba, J. Freeman, J. A. Movshon, & E. P. Simoncelli (2016) *Proc Nat'l Acad Sci USA*, 113(22):3140–3149.
3. R. Q. Quiroga, L. Reddy, G. Kreiman, C. Koch, & I. Fried (2005) *Nature*, 435(7045):1102–1107.
4. F. Attneave (1954) *Psychol. Rev.* 61(3):183–93.
5. H. B. Barlow (2961) “Possible principles underlying the transformation of sensory messages.” In *Sensory Communication*, ed. WA Rosenblith, pp. 217–34, Cambridge, MA: MIT Press.
6. R. Linsker (1988) *IEEE Computer*, 21, 105–117.
7. J. J. Atick & A. N. Redlich (1992) *Neural Computation*, 4(2):196–210.
8. B. A. Olshausen & D. J. Field (1996) *Nature*, 381(6583):607–609.
9. E. C. Smith & M. S. Lewicki (2006) *Nature*, 439(7079):978–982.
10. A. C. Lin, A. M. Bygrave, A. de Calignon, T. Lee, & G. Miesenbock (2014) *Nat. Neurosci.* 17(4):559–568.
11. W. Bialek, R. De Ruyter Van Steveninck, & N. Tishby “Efficient representation as a design principle for neural coding and computation,” presented at the 2006 IEEE International Symposium on Information Theory, pp. 659–663.
12. S. E. Palmer, O. Marre, M. J. Berry II, & W. Bialek (2015) *Proc Natl Acad Sci USA*, 112(22):6908–6913.
13. Y. Dan, J. J. Atick, & R. C. Reid (1996) *J. Neurosci.*, 16(10):3351–3362.
14. D. W. Dong, & J. J. Atick (1995) *Network: Computation in Neural Systems* 6(2):159–178.
15. N. Brenner, W. Bialek & R. de Ruyter van Steveninck (2000) *Neuron*, 26, 695–702.
16. A. L. Fairhall, G. D. Lewen, W. Bialek & R. de Ruyter van Steveninck (2001) *Nature*, 412, 787–92.
17. F. Creutzig & H. Sprekeler (2008) *Neural Comp.*, 20:1026–1041.
18. P. Berkes & L. Wiskott (2005) *J. Vis.* 5(6).
19. E. P. Simoncelli & B. A. Olshausen (2001) *Ann. Rev. of Neurosci.* 24(1):1193–1216.
20. B. A. Olshausen & D. J. Field (2004) *Curr. Op. in Neurobio.* 14(4):481–487.
21. L. Buesing & W. Maass (2010) *Neural Comp.*, 22(8):1961–1992.
22. N. Tishby, F. C. Pereira, & W. Bialek (1999) “The information bottleneck method,” in *Proc. of the 37th Annual Allerton Conference on Communication, Control and Computing*, pp. 368–377.
23. M. Chalk, O. Marre, & G. Tkačik (2016) “Relevant sparse codes with variational information bottleneck,” in *Advances in Neural Information Processing Systems*, 1957–1965
24. A. Hyvärinen, J. Hurri & P. O. Hoyer (2009) *Natural Image Statistics*, Springer-Verlag.
25. A. V. Oppenheim & J. S. Lim (1981) *Proceedings IEEE* 69: 529–541.
26. A. J. Bell, & T. J. Sejnowski (1997) *Vision research*, 37(23), 3327–3338.
27. J. H. van Hateren, A. van der Schaaf (1998) *Proc. R. Soc. London Ser. B* 265:359–66.
28. G. Tkačik, J. S. Prentice, V. Balasubramanian & E. Schneidman (2010) *Proc Nat'l Acad Sci USA*, 107, 14419–24.
29. Y. Karklin & E. P. Simoncelli (2011) “Efficient coding of natural images with a population of noisy Linear-Nonlinear neurons,” in *Advances in Neural Information Processing Systems* 24, 999–1007.
30. E. Doi, J. L. Gauthier, G. D. Field, J. Shlens, A. Sher, M. Greschner, T. A. Machado, L. H. Jepson, K. Mathieson, D. E. Gunning, A. M. Litke, L. Paninski, E. J. Chichilnisky & E. P. Simoncelli (2012) *J. of Neurosci.* 32(46):16256–16264.
31. V. Balasubramanian & P. Sterling (2009) *J Physiol*, 587, 2753–2767.
32. G. Tkačik, J. S. Prentice, J. D. Victor, V. Balasubramanian (2010) *Proc Nat'l Acad Sci USA*, 107, 18149–54.
33. A. M. Hermundstad, J. J. Briguglio, M. M. Conte, J. D. Victor, V. Balasubramanian, G. Tkačik (2014) *eLife*, 10.7554.
34. C. K. Machens, T. Gollisch, O. Kolesnikova, & A. V. M. Herz (2005) *Neuron*, 47(3):447–456.
35. G. C. DeAngelis, I. Ohzawa & R. Freeman (1995), *Trends. Neurosci.* 17:451–458
36. N. J. Priebe, S. G. Lisberger, J. A. Movshon (2006) *J. Neurosci.* 26(11):2941–2950.
37. T. Baden, P. Berens, K. Franke, M. R. Roson, M. Bethge & T. Euler (2016) *Nature*, 529(7586):345–350.
38. A. Alemi, I. Fischer, J. V. Dillon, & K. Murphy (2016) *arXiv:1612.00410*.
39. G. Chechik, A. Globerson, N. Tishby, & Y. Weiss (2005) *J. of Machine Learning Res.*, 6(165–188).
40. M. Boerlin, & S. Deneve (2011) *PLoS Comput Biol*, 7(2), e1001080.
41. T. Hosoya, S. A. Baccus & M. Meister (2005) *Nature*, 436(7047), 71–77.
42. R. P. Rao, & D. H. Ballard (1999) *Nature neuroscience*, 2(1):79–87.
43. W. Schultz & A. Dickinson (2000) *Ann. Rev. of Neurosci.* 23(1):473–500.
44. J. Salisbury & S. Palmer (2016) *Journal of Statistical Physics*, 162(5):1309–1323.