ANNUAL REVIEWS

Annual Review of Vision Science

Two Sides of the Same Coin: Efficient and Predictive Neural Coding

Michael B. Manookin[1,3,4] and Fred Rieke[2,3]

[1]Department of Ophthalmology, University of Washington, Seattle, Washington, USA;
email: manookin@uw.edu

[2]Department of Physiology and Biophysics, University of Washington, Seattle, Washington, USA; email: rieke@uw.edu

[3]Vision Science Center, University of Washington, Seattle, Washington, USA

[4]Karalis Johnson Retina Center, University of Washington, Seattle, Washington, USA

Keywords

vision, efficient coding, prediction, information bottleneck

Abstract

Some visual properties are consistent across a wide range of environments, while other properties are more labile. The efficient coding hypothesis states that many of these regularities in the environment can be discarded from neural representations, thus allocating more of the brain's dynamic range to properties that are likely to vary. This paradigm is less clear about how the visual system prioritizes different pieces of information that vary across visual environments. One solution is to prioritize information that can be used to predict future events, particularly those that guide behavior. The relationship between the efficient coding and future prediction paradigms is an area of active investigation. In this review, we argue that these paradigms are complementary and often act on distinct components of the visual input. We also discuss how normative approaches to efficient coding and future prediction can be integrated.

INTRODUCTION

Successful behavior relies on the ability to quickly and efficiently extract salient information from the environment. This is a challenging task—the environment contains an almost unlimited amount of information, most of which is irrelevant to survival and behavior. In addition, the relevant pieces of information are often highly correlated with one another, such that some pieces of information can be predicted from others. An ideal coding system would utilize knowledge of these correlations to encode only those bits of information that are necessary for reconstructing the relevant properties of the environment (Attneave 1954, Barlow 1961).

Several issues make achieving such an ideal coding challenging. First, the dynamics of the environment are complex, and the bits of information that are relevant to behavior can change depending on the context. This intertwines the issue of learning the correlation structure of the current environment with that of identifying a good coding strategy (Młynarski & Hermundstad 2018, 2021). Second, time is of the essence for behavior. The brain must collect information from the environment and modify behavior within only a few hundred milliseconds, but neurons are severely limited in the amount of information that they can transmit on such brief timescales (Borst & Theunissen 1999, Koch et al. 2006). Thus, the brain must rapidly sift through these incoming sensory inputs and separate out the relevant information. A major goal of both sensory and computational neuroscience is to understand how the brain copes with these challenges and how these strategies develop over evolutionary, developmental, and behavioral timescales.

The efficient coding hypothesis is an influential paradigm for understanding how neural systems should operate to transmit dynamic sensory information (Attneave 1954, Barlow 1961). Efficient coding indicates that a group of neurons should represent sensory information as concisely as possible to utilize their limited metabolic and signaling resources (Laughlin 1981, Olshausen & Field 1996, van Hateren 1992). Neurons can accomplish this by tuning their representations to the statistical properties of their sensory inputs. For example, efficient coding predicts how the range of responses that a neuron can generate should be allocated—with a small fraction of the response range allocated to rare inputs and a large fraction to common inputs. These principles of efficient coding can explain many properties of neurons found in the early visual and auditory systems (Laughlin 1981, Lewicki 2002, Rieke et al. 1995, Sharpee et al. 2006, Smith & Lewicki 2006, Srinivasan et al. 1982).

The information theoretic concepts developed by Claude Shannon provide the foundation for understanding how efficient coding should be achieved (Shannon 1948). The key concept is that some aspects of the environment carry more information than others by virtue of how expected or unexpected they are (e.g., the statement that December days in Seattle are short is not particularly informative for someone who has lived in Seattle). To be efficient, the neural code should emphasize the informative aspects while ignoring or discarding the noninformative ones. Importantly, these discardable properties are identified based on the statistics of the environment (e.g., properties of a large collection of natural images) and not the properties that differentiate a specific environment from others (e.g., a specific natural image).

A simultaneous strength and weakness of information theory is that all bits of information are assigned equal value (Shannon 1948). This is clearly not the case for biological systems—different pieces of information can have different consequences for survival and reproductive success. Given the temporal delays associated with receiving sensory information about the environment, encoding that information, and then acting on it, information that cannot guide future actions is useless to an organism. Thus, at a minimum, the brain should prioritize those pieces of information that allow it to estimate what is most likely to occur in the future (Bialek et al. 2001, Rust & Palmer 2021, Salisbury & Palmer 2016). This estimate must use contextual information to separate predictions based on the current environment from predictions based entirely on the average statistics

of all environments. For example, while watching a horror movie, the building tension can predict coming events in ways that the average statistics of all movies cannot. This context-specific prediction is essential for successful behavior. This has led to a revised picture of efficient coding stating that all bits of information are not created equal and the assertion that neural codes should prioritize information about the future.

Herein lies a conundrum for those of us trying to understand these ideas about efficient neural processing and their implications for the neural systems that we study. Classical efficient coding and this updated picture both liberally use the term prediction, but they use it in different ways. In classical efficient coding, prediction is the idea that the brain has a (statistical) model of the state of the world that is updated on evolutionary, developmental, and behavioral timescales. This model captures correlations in the world that can be used to weight different aspects of the incoming sensory information by their likely information content. Removing correlations emphasizes those aspects of the incoming sensory information that distinguish a specific environment from the set of environments consistent with the world model. This view of prediction provides a useful paradigm for understanding several aspects of neural processing. One example is the notion that neural responses should be allocated based on the likelihood of particular stimuli (Laughlin 1981). Another example is that many features of the world change slowly, and thus the past state of the environment can be subtracted from the incoming sensory information to highlight those features of the input that are changing (Brenner et al. 2000, Fairhall et al. 2001, Hosoya et al. 2005). Related work from Rao & Ballard (1999) emphasizes the importance of predictability to neural response properties, i.e., of weighting inputs based on how much they vary from an internal model of the environment.

The other use of prediction specifically considers the importance of inferring what is likely to happen in the future. In the paradigm proposed by Bialek, Palmer, and colleagues, this is precisely what the brain should do—weight incoming sensory information based on its utility in predicting the future (Bialek et al. 2001, Palmer et al. 2015). This sense of prediction is similar to the idea of extrapolation or anticipation (Berry et al. 1999, Schwartz et al. 2007a). In some cases, coding strategies consistent with efficient coding and those consistent with predicting future world states are similar—e.g., from both perspectives, a stationary object in a scene need not consume coding resources. As we discuss below, however, in other cases, these distinct definitions of prediction can have distinct and even opposing implications for how neurons should behave (Salisbury & Palmer 2016).

THE ENCODING PROBLEM: AN EXAMPLE FROM LANGUAGE

Efficient coding depends strongly, both conceptually and quantitatively, on information theory. The central hypothesis of efficient coding is that neurons communicate information efficiently with one another; in the same way, natural language is considered an efficient communication system. In this section, we use language as an example of the relationship between efficient coding and information theory, with the aim of demonstrating how similar ideas can be applied to neural encoding.

Imagine that we are given the task of sending a series of messages in English using a binary code (i.e., a series of 1s and 0s). The goal would be to create a one-to-one mapping between words in the English lexicon and different sequences of 0s and 1s such that we can encode any possible sentence. If we were given an unlimited amount of time to convey our messages, then this mapping between words and binary digits could be completely arbitrary as long as it was one to one. However, our task would be considerably more difficult if we were told that our code must be efficient, meaning that it must represent words with the fewest possible digits, allowing these messages to be sent as quickly as possible.
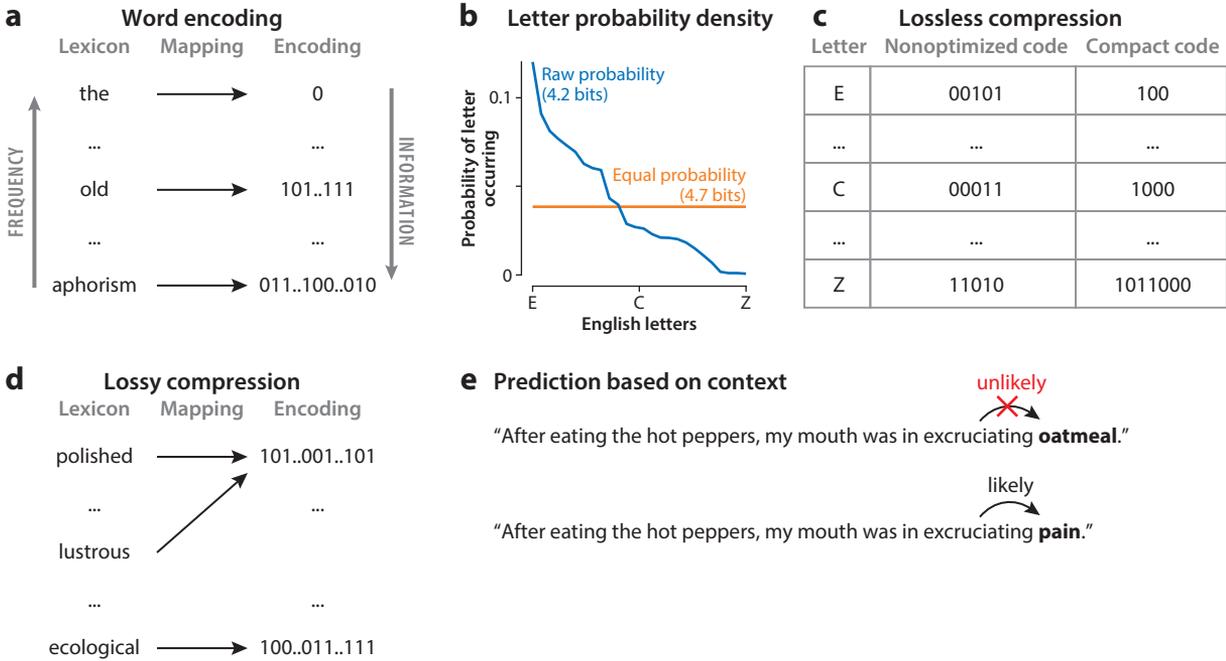
a   **Word encoding**

Lexicon	Mapping	Encoding
the	$\longrightarrow$	0
...		...
old	$\longrightarrow$	101..111
...		...
aphorism	$\longrightarrow$	011..100..010

FREQUENCY ↑   INFORMATION ↓

b   **Letter probability density**

Raw probability (4.2 bits)

Equal probability (4.7 bits)

Probability of letter occurring: 0.1 ... 0

English letters: E   C   Z

c   **Lossless compression**

Letter	Nonoptimized code	Compact code
E	00101	100
...	...	...
C	00011	1000
...	...	...
Z	11010	1011000

d   **Lossy compression**

Lexicon	Mapping	Encoding
polished	$\longrightarrow$	101..001..101
...		...
lustrous		
...		...
ecological	$\longrightarrow$	100..011..111

e   **Prediction based on context**

unlikely ✗

"After eating the hot peppers, my mouth was in excruciating **oatmeal**."

likely

"After eating the hot peppers, my mouth was in excruciating **pain**."

Figure 1

Essential efficient coding concepts. (*a*) Example word-encoding model in which words in the English lexicon are mapped to sequences of binary digits (i.e., bits). Mapping the most frequent words to the shortest binary sequences produces the most compact coding scheme. (*b*) Probability of letter occurrence for a corpus of written English. Computed entropy values for the letters are shown for the raw distribution (*blue*) compared to an equal distribution of letter occurrences (*orange*). (*c*) Two possible lossless compression schemes for the English letters. The nonoptimized code uses the same digit length for all of the letters regardless of frequency, whereas a more compact code could be generated by accounting for this frequency information. (*d*) Word-encoding scheme similar to that in panel *a* except that a certain amount of information loss is tolerated. This lossy compression scheme mitigates the impact of this information loss by using the same binary encoding to represent words with similar meanings. (*e*) Semantic context can be used to predict the words that are likely to come in the near future. In the example sentence, context indicates that "pain" is more likely to be the next word than "oatmeal."

According to Shannon, the motivation for constructing efficient codes is that any communication system is limited in its channel capacity—the capacity to encode information in a given time bin (Shannon 1948). This capacity is limited even with today's most advanced hardware, and it is severely limited for the wetware in the brain. Information theory provides the tools to use this limited coding capacity efficiently. To make our messages as short as possible, we need to represent the words that we transmit most frequently with the shortest binary sequences and words that are used infrequently with longer sequences (**Figure 1a**). The key is that the relative frequency of the words in the lexicon determines an efficient coding strategy.

To formalize these ideas, consider the process of encoding letters in English. Without knowledge of the frequency with which letters occur in English, we might assume that all letters occur with equal frequency. We can use information theory to determine the most efficient way to encode these letters as a sequence of binary digits. Entropy is the appropriate tool for this task. Entropy measures the randomness or uncertainty in a distribution (in our example, the distribution of possible letters). Our uncertainty is highest when we lack prior knowledge about the frequency of letters occurring and assume that all letters occur with equal probability. The minimal number of binary digits that we would need to uniquely encode all letters is determined by

computing the entropy of the probability distribution for these letters,

$$H(X) = -\sum_{x \in X} p(x)\log_2 p(x),$$

where $H(X)$ is the entropy of the distribution, and $p(x)$ is the probability of a particular letter occurring. If our 26 letters each occur with equal probability, then we would calculate the entropy in the following way:

$$H(X) = -\sum_{i=1}^{26} \frac{1}{26} \log_2 \frac{1}{26} = -\left(\frac{1}{26} \log_2 \frac{1}{26} + \frac{1}{26} \log_2 \frac{1}{26} + \ldots \right) = 4.7 \text{ bits.}$$

Each bit of information represents one binary character in our code—thus, in this case, we would need to use an average sequence length of 4.7 binary characters to represent the 26 English letters. In other words, the cost of transmitting a letter in this system is, on average, 4.7 bits. This value reflects the cost of transmitting the code when our uncertainty (entropy) is highest—when we lack prior knowledge about the alphabet.

We can construct a more compact and efficient code if we have prior knowledge about the relative frequencies of letter occurrence in written English. For example, the letter E occurs much more frequently than the letter Z (**Figure 1b**). The entropy from the distribution of letters in an actual English corpus gives a value of 4.2 bits. This tells us that we can compress the encoding of these 26 letters by using short sequences of binary digits for frequent letters and long sequences for infrequent letters. This is the goal of efficient coding: leveraging knowledge about the likelihood of inputs (letters or natural stimuli) to remap them into a new, more efficient format (sequences of binary digits or spikes). Consideration of correlations (e.g., between Q and U) could further increase efficiency, a point that we return to below.

A key feature of information theory is that the procedure depends only on the probability distribution of the inputs and not on the meaning of the letters or words. Shannon (1948) provided the tools that we need to encode the language, whether it is English or an alien language spoken on a distant planet that is completely indecipherable to us. This fact that meaning is not considered in classical information theory is a key point that becomes important in the discussion below.

Efficient coding, as described above, preserves information content. Another way that we could decrease the length of our encoding is by using a form of compression that relaxes that constraint (i.e., lossy compression; **Figure 1d**). For example, we might use the same sequence of bits to represent similar words in the lexicon—e.g., "polished" and "lustrous." In this example, subtle semantic differences between "polished" and "lustrous" would be lost, but the benefits of such compression may outweigh the costs, e.g., by allowing "polished" and "tarnished" to be coded distinctly. In the nervous system, such lossy compression is critical. The natural environment contains an almost unlimited amount of information, only a small fraction of which can be transmitted by neurons in the timescales relevant to behavior. The key concept in efficient compression is that two words that have similar meaning can be combined during encoding, and when resources are at a premium, they should be combined. This same concept applies to sensory processing—an efficient code would combine two stimuli that lead to the same inference (Bialek 2012, Hermundstad et al. 2014).

Predictive processing is another key aspect of efficient coding that applies to language. While reading or listening to a sentence, humans use contextual cues to actively predict the words and concepts that are likely to follow. In the examples above, we leverage knowledge about the relative frequencies of words or letters occurring in a large ensemble of texts to construct a more efficient code for the lexicon or alphabet. However, we need more than this knowledge about the average

frequency of words across the ensemble of texts to actively infer meaning from a particular sentence and to predict the words and concepts that are likely to follow. Consider the phrase, "After eating the hot peppers, my mouth was in excruciating…" Given the sequence of words in this phrase, we can surmise that the next word is almost certainly a synonym of "pain" and not related in meaning to "joy" or "oatmeal." In this case, we are using deeper contextual information about the words in the sentence to infer the overall meaning and predict what is likely to follow.

Such predictive inference is an integral part of natural language processing. One study recorded the eye movements of children who listened to a sentence while presented with two different pictures. Children as young as two years old made anticipatory eye movements toward the picture that correctly completed a sentence prior to hearing the word (Mani & Huettig 2012). For example, while listening to the sentence, "The boy eats a big cake," children made anticipatory eye movements toward a picture of a cake prior to hearing the word "cake." This ability to use information from the recent past to estimate what is likely to happen in the future is a key aspect of successful behavior across sensory, motor, and cognitive modalities (Bialek et al. 2001, Rust & Palmer 2021, Salisbury & Palmer 2016).

This process of inference and prediction relies on constructing a model of language and of the world—cake is something that we eat, and hot peppers can cause pain. This relates closely to knowledge gained from the ensemble of our language experiences—in the course of our reading and hearing of natural language, we have created these semantic associations between cake and eating and between heat and pain. Thus, we need context-specific knowledge to perform prediction. This requires a set of tools beyond what is available from classical information theory. We demonstrate below that ideas developed by the Reverend Thomas Bayes will prove our salvation.

EFFICIENT ENCODING OF VISUAL SCENES

The language example highlights the importance of allocating coding resources based on the probability of observing specific letters or words. This same strategy is central to constructing efficient neural representations of sensory information. To understand this concept, it is helpful to first consider a situation in which this coding strategy would be inefficiently matched to the statistical properties of the input—the case of the linear neuron trying to encode a Gaussian input.

Consider the case in which a neuron receives some sensory input from the environment and then encodes that input to downstream neurons using a sequence of spikes (**Figure 2a**). The neuron could use a variety of encoding strategies. In this example, we constrain the dynamic range of this neuron to fire between 0 and 10 spikes in a given time bin and assume that all that matters is the spike count. The simplest coding strategy would be to choose a linear mapping in which the number of spikes used to represent a given input would be determined by multiplying the input by a scale factor and adding a constant offset (i.e., $y = mx + b$; **Figure 2b**).

This mapping is a good strategy if all possible inputs occur with equal probability. However, if the distribution of the inputs is not uniform, then the linear encoding strategy is inefficient because it weights inputs that occur infrequently (on the tails of the distribution) in the same way as those that occur much more frequently (near the mean of the distribution). The result is that the neuron uses a significant portion of its spike outputs to represent events that occur infrequently, leaving relatively fewer outputs for inputs that occur frequently (**Figure 2b**). This means that the precision with which the inputs are encoded does not take into account the frequency with which they occur. In efficient coding, the goal is to encode the entire distribution of inputs as accurately as possible given a limited output bandwidth, or equivalently, to minimize the error between input and the neural representation. In the vernacular of information theory, the goal is to maximize the mutual information between the input and the neuron's spike output. How can this goal be achieved?
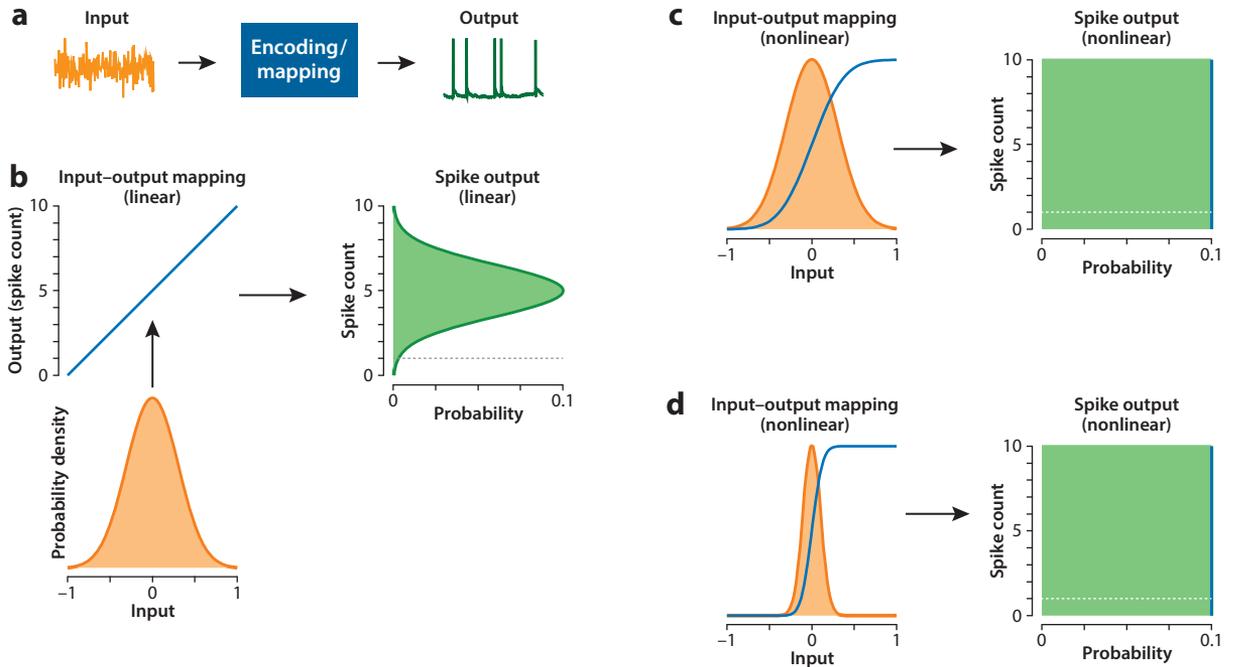
Figure 2

Comparison of encoding approaches for mapping a Gaussian input to the neural spike output. (*a*) Diagram illustrating the encoding process. The encoder receives a Gaussian input and maps it to spike output. (*b*) A linear encoding applies equal weighting to infrequent values and frequent values, dedicating a large proportion of the output range to values that occur infrequently. (*c*) Nonlinear mapping scheme in which a cumulative Gaussian function is used to encode the input distribution. This approach weights the inputs according to their likelihood of occurrence, resulting in optimal use of the dynamic range. (*d*) Nonlinear encoding scheme similar to that in panel *c* except that the slope of the nonlinearity is increased to account for the narrowing of the input distribution. This also results in ideal use of the output range.

Information theory provides the optimal solution to this problem. Instead of a linear mapping, the information-theoretic approach uses a nonlinear mapping between input and output when the input distribution is not uniform. In the case of a Gaussian input distribution with independent and additive neural noise, this optimal mapping function is the cumulative Gaussian function. This mapping is efficient because it redistributes the inputs so that they are represented in the output according to their probability of occurrence, and as a consequence, each output occurs with equal probability (**Figure 2c**). For the assumed noise properties, equalizing the probability of each possible output maximizes the encoded information. In a key study, Laughlin (1981) demonstrated that neurons in the fly visual system show the behavior predicted by efficient coding and information theory—the neuron's input–output relationship closely resembled a cumulative Gaussian predicted based on the near-Gaussian distribution of intensities in natural images.

Importantly, efficient coding and information theory not only tell us this optimal input–output mapping when the input statistics are static, but also tell us how this mapping should change when the input statistics vary. For example, information theory indicates that, when the distribution of inputs narrows, the slope of the input–output mapping should steepen to optimally utilize the output range (**Figure 2d**). Numerous empirical studies have shown that this is what occurs when the input statistics change in this way—when the input distribution changes, neurons change the slope of their input–output curve in the appropriate direction (Baccus & Meister 2002,

Fairhall et al. 2001, Kim & Rieke 2001, Rieke 2001, Smirnakis et al. 1997). More recent work has investigated how incomplete knowledge of the input statistics (e.g., due to dynamic changes in those statistics) impacts efficient coding strategies (Hermundstad et al. 2014; Młynarski & Hermundstad 2018, 2021).

The principles of efficient coding have been used to explain other fundamental properties of cells in the early visual system. One of the earliest successes related the receptive-field structures of early visual neurons to the structure of natural scenes. Barlow's (1961) formulation of efficient coding states that groups of neurons should limit the amount of common sensory information that they encode, thereby decreasing information redundancy and increasing the efficiency of the neural code. Achieving such independence in neural signaling is challenging because many different properties of natural environments are correlated, and these correlations are sufficiently complex that we do not even have good descriptions of them. Efficient coding has largely focused on so-called second-order correlations—e.g., correlations in light intensity between two spatial locations.

Second-order correlations occur because objects occupy space (Dong & Atick 1995, Field 1987, Ruderman & Bialek 1994). Larger objects cause correlations across larger distances—for example, a large, white beach ball will produce more correlations at the level of the retinal photoreceptors than a baseball viewed at the same distance. The spatial extent of correlations can be described by the spatial frequency content of images, terminology that comes from Fourier analysis (Dong & Atick 1995, Field 1987, Ruderman & Bialek 1994). Fourier analysis conveniently and completely describes images as a combination of cosine waves that vary in their amplitudes, frequencies, and phases. Two-point correlations are captured by the dependence of the amplitude on spatial frequency. The beach ball in our example is composed of lower spatial frequencies than the baseball because we can fit fewer beach balls in a given region of space. Thus, an image with more two-point correlations occurring over short spatial distances contains more power at high spatial frequencies.

The description above is far from complete because correlations between three or more points in space, described by the phase structure, also contribute strongly to structure in natural images. This concept is illustrated in **Figure 3**. We swapped the phase information between two natural images while maintaining the two-point correlations. This manipulation fundamentally changes the original images, making it appear that the images themselves have been swapped (Simoncelli & Olshausen 2001, Turner et al. 2019). Most natural images are similarly distinguished by differences in the phase information, rather than by differences in two-point correlations (Oppenheim & Lim 1981, Piotrowski & Campbell 1982).

Information about the spatial frequency content of natural images typically comes from averaging across a large ensemble of images (Dong & Atick 1995, Field 1987, Ruderman & Bialek 1994). Just as in our language example, since this distribution of spatial frequencies is not flat (i.e., some spatial frequencies are overrepresented relative to others), we can leverage information about these ensemble statistics to construct a more efficient code. Specifically, if we assume that the brain's internal model of the environment has this information about the average ensemble statistics, then we can remove these correlations from the raw light inputs early in visual processing and thus improve the efficiency of visual representations (Attneave 1954, Barlow 1961). Thus, one strategy for encoding natural images more efficiently is to reduce redundancies by suppressing low spatial frequencies without disturbing the phase information in the images (Atick & Redlich 1992, Field 1987). If the noise is distributed equally across spatial frequencies, then this strategy can be achieved by whitening the incoming visual inputs to remove the bias at low spatial frequencies (Laughlin 1981).

As mentioned above, the spatial frequency and phase information are distinguished by the types of image correlations that they represent. Spatial frequency information represents
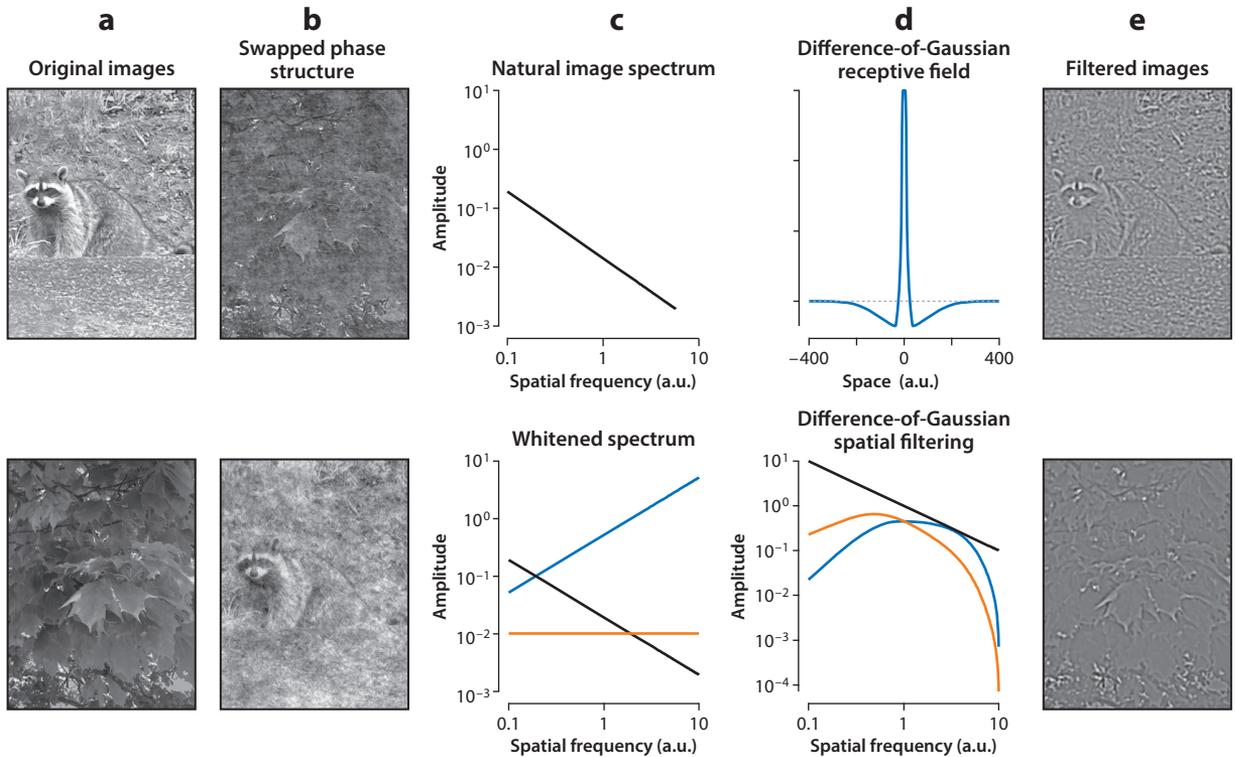
a Original images
b Swapped phase structure
c Natural image spectrum / Whitened spectrum
d Difference-of-Gaussian receptive field / Difference-of-Gaussian spatial filtering
e Filtered images

Figure 3

Correlation structure of natural images. (*a*) Images of a racoon (*top*) and leaves (*bottom*). (*b*) The images in panel *a* with the original two-point correlation (spatial-frequency) structure but with the higher-order correlation (phase) structure swapped between the images. (*c*) Probability distribution of spatial frequencies (two-point correlations) across a large ensemble of natural images (Field 1987). Both axes are shown on a logarithmic scale. Applying the high-pass filter (*blue*) whitens the natural image spectrum so that all spatial frequencies are represented with equal probability (*orange*). (*d*) Example difference-of-Gaussian filter previously thought to describe the spatial filtering properties of retinal ganglion cells. In the frequency domain, this structure forms a bandpass filter (*bottom*, *blue*) that attenuates both high and low frequencies (*orange*). (*e*) The images in panel *a* filtered with a difference-of-Gaussian filter. The filter significantly enhances the boundaries between changes in contrast.

correlations between two points in space, while phase information comprises correlations among three or more points in space. This distinction means that we can apply a spatial filter that attenuates low-spatial-frequency pairwise correlations in the image relative to high-spatial-frequency pairwise correlations, while leaving higher-order correlations (i.e., phase information) intact. Specifically, this spatial filter would remove the bias at low frequencies, essentially flattening the spatial frequency distribution, as in the case of the optimized input–output mappings (**Figure 2**).

The average distribution of amplitudes of different spatial frequencies in natural images is shown on a logarithmic scale in **Figure 3c** (*top*). Applying a high-pass spatial filter flattens this distribution so that all spatial frequencies are represented with equal amplitude (**Figure 3c**, *bottom*). In reality, the retina contains several noise sources, which can limit the efficacy of high-pass filtering visual inputs. The optimal solution under efficient coding will vary depending on our assumptions about these noise sources. For example, noise in the photoreceptors is largely uncorrelated between neighboring photoreceptors, and thus, it occurs at higher spatial frequencies. This form of noise limits the utility of whitening at high spatial frequencies; because of this, the optimal efficient coding solution becomes a bandpass filter that reduces redundancy in the image

by attenuating low frequencies and decreases the effects of photoreceptor noise by also attenuating higher frequencies (Atick & Redlich 1992, Field 1987).

This ideal spatial filter resembles the center-surround receptive field structure of many retinal ganglion cells (Atick & Redlich 1990). The canonical center-surround receptive-field structure computes the difference in light intensities over the receptive-field center and surrounding regions, which removes spatial correlations in the image. This model is appealing for its simplicity and its ability to explain the receptive-field structures of early visual neurons.

However, these early models suggesting that the retina provided relatively simple front-end filtering for vision missed the mark in certain ways (Atick & Redlich 1990, 1992; Campbell & Robson 1968). For example, the classical center-surround receptive-field model poorly predicts retinal responses to natural scenes—the very stimuli that it evolved to encode (Heitman et al. 2016, Turner & Rieke 2016). Linear mechanisms do not mix different spatial frequencies and thus can perform only a limited set of computations. Indeed, the retina contains many different types of nonlinear processing beginning in the cone photoreceptors, and these nonlinearities are key components of complex and varied computations, including orientation and direction selectivity (Angueyra et al. 2022, Gollisch & Meister 2010). Moreover, empirical work demonstrated that decorrelating the retinal output relies more on these circuit nonlinearities than on the canonical receptive-field structure (Pitkow & Meister 2012). More generally, nonlinear mechanisms can perform computations that depend on multiple spatial frequencies.

As mentioned above, mechanisms that reduce redundancies by attenuating pairwise image correlations do not significantly disturb the phase information. The phase information in a natural image contains critical information needed to detect lines, contours, and boundaries between objects. Thus, this information can be used to predict the likely intensity values in a region of space when those intensities are not known—a perceptual phenomenon known as filling in. This is one example of how this phase information can be used; the key idea is that decorrelation mechanisms operating on pairwise correlations will produce distinct effects on the spatial frequency and phase information in an image.

IN THE CHAPEL OF THE REVEREND BAYES

Prediction, in the efficient coding sense, is the idea that the brain contains a model of the environment and that it updates this model based on incoming sensory evidence (Rao & Ballard 1999). This idea is very old—in the eighteenth and nineteenth centuries, Kant and von Helmholtz both proposed similar ideas about the process of perception involving an updating of prior beliefs based on new evidence [see Kant 1998 (1781), von Helmholtz 2000 (1867)]. The mathematical formulation of this idea has also been around for a long time, originally conceived by the English minister Thomas Bayes and formalized by the French mathematician Pierre Simon Laplace in the eighteenth century. In recent years, the field of neuroscience has gained several critical insights by reframing efficient coding in the context of Bayesian statistics (Dayan et al. 1995; Friston 2010; Młynarski & Hermundstad 2018, 2021). In this framework, for example, the correlation structure of natural images is captured by a statistical Bayesian model.

Bayes and Laplace provided a simple mathematical formula for determining the likelihood of something occurring given a prior knowledge or belief about certain properties of the world (i.e., the prior). In the face of new evidence (i.e., the likelihood), this prior is updated to form a revised model (i.e., the posterior; **Figure 4a**). This formula has been applied to a wide range of problems inside and outside of neuroscience. In the context of vision, the idea is that the brain has a model of the external environment that forms a belief about the current state of the world. The brain then updates this model as it accumulates new sensory information from the environment. In the
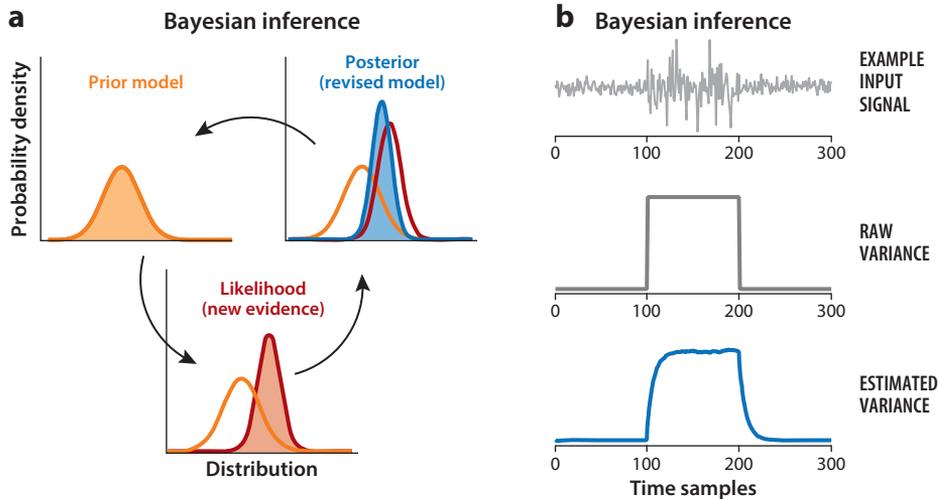
Figure 4

Bayesian paradigm for updating beliefs based on new evidence. (*a*) General schematic for Bayesian inference. A belief state or internal model of an environmental variable is represented by the prior probability distribution (*orange*). New evidence (the likelihood; *red*) influences and updates the prior (according to Bayes' theorem) to create a new belief state or revised model (the posterior probability; *blue*). (*b*) Application of Bayesian inference to estimate the variance of an input signal. (*Top*) An example of a Gaussian input signal that switches between low and high variance. (*Middle*) The raw variance of the input signal as a function of time. (*Bottom*) A Bayesian estimation of the raw variance.

context of efficient coding, this internal model represents the correlation structure contained in the ensemble of natural images, which allows the early visual system to remove these predictable correlations (i.e., decorrelation or whitening).

The Bayesian approach makes specific and testable predictions about how neurons should behave following changes in the input statistics (DeWeese & Zador 1998; Młynarski & Hermundstad 2018, 2021; Wark et al. 2009). For example, when the variance of incoming signals changes over time, a Bayesian system responds by updating its belief or model about the external environment (DeWeese & Zador 1998). This updating process produces stereotypical changes in output when the variance increases or decreases (**Figure 4***b*). When the variance of incoming signals changes, the new distribution of input values (i.e., the likelihood) shifts the posterior distribution away from the prior. Several time steps are required before the prior reflects the new input distribution, resulting in a smoothed estimate of the variance (**Figure 4***b*). This response pattern, which is characteristic of a Bayesian-optimal system encoding an estimate of variance, has been found in a subset of sensory neurons; these sensitizing or facilitating cells have been described in several vertebrates, including primates (Appleby & Manookin 2019; Kastner & Baccus 2011, 2013; Nikolaev et al. 2013).

Młynarski & Hermundstad (2018, 2021) further demonstrated that encoding via a Bayesian updating of the prior minimizes the time required to adjust to the new environmental context or input statistics. In addition to the Bayesian encoding of variance, they identify another strategy that minimizes the error between the current model and the new context (Młynarski & Hermundstad 2018, 2021). Minimizing this error predicts than an encoder adapts to the changing input statistics in a manner similar to adapting neurons described experimentally (Baccus & Meister 2002, Fairhall et al. 2001, Kim & Rieke 2001, Rieke 2001, Smirnakis et al. 1997). These cells show transient changes in firing following changes in the variance or mean of the input distribution.

Młynarski & Hermundstad (2018, 2021) also predicted that these two distinct encoding strategies will have differences in how they respond when the mean of the input distribution changes. The first strategy, which resembles neural sensitization, shows a much more sustained change in firing following increases or decreases in the mean. The second strategy, which resembles adapting cells and seeks to minimize the overall error in the estimate, shows fast and transient changes following either an increase or decrease in the mean. Similar patterns of activity have been observed in mouse ganglion cells following changes in mean luminance (Tikidji-Hamburyan et al. 2015). However, we currently do not know whether this behavior during mean adaptation maps cleanly to the adapting and sensitizing cell behavior observed during changes in the stimulus variance.

Both the adaptation and sensitization strategies share the common goal of learning the rules of the environment with incomplete information. In the Bayesian framework, these rules are contained in a statistical model of the environment—the prior. Changes in the environmental context produce shifts in the distribution of sensory inputs (i.e., the likelihood), which update the prior to produce the posterior distribution—a revised model of the environment. These strategies diverge in their coding objectives when the input statistics change, and these differing objectives result in contrasting response properties (Młynarski & Hermundstad 2018, 2021).

PREDICTIVE ENCODING OF VISUAL INPUTS

For the sake of simplicity and tractability, Shannon's (1948) original formulation of information theory implicitly assigned equal value to all bits of information. However, we know that this cannot reflect reality—some bits of information carry more intrinsic value to an organism than others (Bialek et al. 2001). For example, in observing a lion stalking through the grass nearby, the information needed to identify the lion's location and heading carry more behavioral value than subtleties in the hue and size of its mane. These latter properties can only be appreciated from a safe distance. To complicate matters, the relevance of different bits of information can depend on the context. For example, when accounting for all the members of a pride from a safe distance, the lion's mane may matter more than the direction that the lion is moving—the absence of a mane on any of the lions may indicate that one or more of the male members has not been accounted for and may be a legitimate reason for alarm.

Bialek and colleagues (2001; see also Rust & Palmer 2021, Salisbury & Palmer 2016) formulated a revised picture of efficient coding that incorporates behavioral importance, with a focus on accurately estimating future states of the environment. The central idea is that information about the past that is not useful in predicting what will occur in the future cannot be used to guide behavior and, thus, should be discarded from neural representations. Current instantiations of this idea focus on predictions of future events rather than an explicit consideration of behavioral significance, since the latter is much harder to quantify. Returning to our lion-in-the-grass analogy, as the lion moves through the grass, its past positions can help us estimate its trajectory and where it is likely to be in the near future. Thus, this past positional information contains significant, behaviorally relevant predictive information.

A central consideration in the prediction problem is the recognition that small groups of neurons are limited in the amount of information that they can transmit on the timescales relevant to behavior. In early vision, a key timescale is the 200–400 ms period of fixation between ballistic eye movements. Many ganglion cells will have a dynamic range of approximately 0–20 spikes in this time period, but the visual information impinging on their receptive fields contains light intensity, contrast, motion, orientation, and chromatic information with a much larger dynamic range (Koch et al. 2006). While this example is overly simplistic (e.g., timing of spikes can contribute to information encoding), it captures the essence of the problem: The information content of the
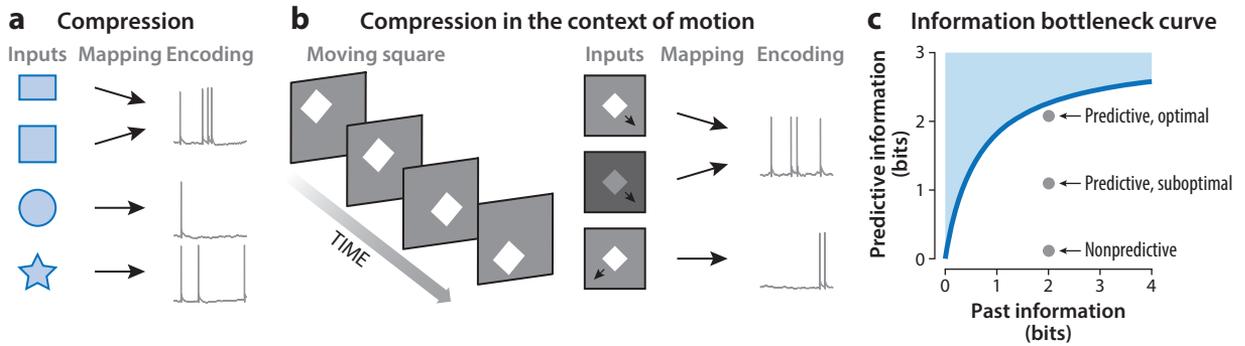
a Compression

Inputs Mapping Encoding

b Compression in the context of motion

Moving square

Inputs Mapping Encoding

TIME

c Information bottleneck curve

Predictive information (bits)

Past information (bits)

● ← Predictive, optimal

● ← Predictive, suboptimal

● ← Nonpredictive

Figure 5

Compression and predictive coding. (*a*) Example lossy compression scheme in which similar objects (four-sided objects) are mapped to the same spike output. (*b*) Separate frames of a movie in which a square moves from the upper left to the lower right corner of the screen (*left*). An efficient encoding of motion direction would map the squares moving in the same direction with the same spike output regardless of the mean luminance (*right top* versus *middle*); squares moving in different directions would be mapped to different spike outputs even if the mean luminance was identical (*right top* versus *bottom*). (*c*) Example information bottleneck curve showing the relationship between predictive information (*y* axis) and past information (*x* axis). The solid curve demarcates the maximal predictive information that can be encoded given the number of bits of past information encoded. The gray circles represent three possible outcomes for an encoder, such as a neuron. Values higher on the *y* axis encode more predictive information, but only the value near the boundary encodes that information in a nearly ideal way.

natural environment, and that encoded by the photoreceptors, far exceeds the coding capacity of sensory neurons (Gollisch & Meister 2008).

A limited coding capacity means that retinal circuits must compress the raw light inputs from the photoreceptors, and this process necessarily throws away certain pieces of information. This is termed lossy compression. As mentioned above, a key principle in this type of compression is that, if two or more input patterns carry similar information for the computation in question, they should be represented with the same neural encoding pattern (Bialek et al. 2001). Consider, for example, the direction-selective ganglion cells that are found in many vertebrate retinas. The task of these cells is to detect the direction of a locally moving object, and because information about the hue of the object (e.g., redness versus greenness) may be irrelevant for this computation, red and green moving objects with identical motion trajectories can have the same representation in the spike outputs of these cells (Barlow & Levick 1965, Sabbah et al. 2017, Taylor et al. 2000).

These concepts also apply to predictive encoding—in creating a compact representation of the visual inputs, past information that does not provide information about the future should be discarded. Furthermore, different pieces of past information that provide the same information about the future should be represented with the same neural representation or spike pattern (Bialek 2012, Hermundstad et al. 2014). Tishby and colleagues (Bialek et al. 2001, Tishby et al. 1999) developed the information bottleneck paradigm to determine how compressing information about the past affects the information that a neuron can signal about the future. This approach identifies a set of optimal mappings that determines the largest amount of future information that can be encoded for a given amount of past information (**Figure 5c**). This relationship allows us to assess the predictive encoding of different neurons and populations of neurons with varying dynamic ranges (Liu et al. 2021, Palmer et al. 2015).

A challenge in studying predictive encoding is finding a good experimental paradigm, as the type of information that should be predicted can be difficult to ascertain in some contexts (Bialek et al. 2001). Motion is an excellent paradigm for studying prediction because the type of information that should be predicted is clear—a group of neurons must use information about the past

positions of an object to determine where it is likely to be in the future. Consider, for example, the motion of a baseball—if the ball's trajectory is completely deterministic, then knowing the past positions of the ball allows prediction of the position of the ball at any arbitrary time in the future. However, if the ball's trajectory has a random component, then this component limits the ability to estimate future positions.

How efficiently is this predictive motion information encoded in ganglion cells when object motion is not completely deterministic? To answer this question, Palmer and colleagues (2015) measured the ability of salamander ganglion cells to predict the future positions of a bar moving with a noisy trajectory. They then leveraged the information bottleneck paradigm to determine the efficiency of predictive encoding. They found that this encoding approached the information bottleneck boundary for many of the cells, indicating that these cells encoded nearly all of the predictive information possible given the dynamic range of their spike outputs. These studies focused on the behavior of retinal neurons in response to translational motion, but other types of behaviorally salient motion occur during natural vision, such as when objects approach or move away from an observer (Fitzgerald et al. 2011, Hu & Victor 2010, Nitzany & Victor 2014). The encoding of these types of motion and translational motion by several primate ganglion cell types also approaches the information bottleneck boundary (Liu et al. 2021).

Recall that the bottleneck boundary reflects the behavior of an ideal encoder. In other words, a piece of hardware or software that is specifically designed to extract predictive motion information but has limited coding capacity cannot perform better than the bottleneck boundary. The fact that the retina can perform this task in a nearly optimal way suggests that the ethological importance of motion prediction has placed strong selective pressures on this computation in salamanders, primates, and likely many other vertebrate species. It also highlights the importance of considering behavioral relevance, in this case, predictions of the future, in approaches to neural coding.

Bayesian statistics also provides us with an excellent framework for understanding predictive processing. For example, having an internal model of how lions move through the grass can help a prey animal predict their future positions and determine appropriate strategies to avoid being in those positions. In the same way, an internal model of other environmental statistics can help an organism more accurately predict what will happen in the future. For example, having an internal model or belief about the typical range of speeds at which objects move can be helpful when trying to estimate motion in noisy environments (Darlington et al. 2017, Yang et al. 2012).

Consistent with this idea, psychophysical studies have shown that human motion perception is biased toward slow speeds when the signal-to-noise ratio is low—the same object moving at the same speed is perceived as moving more slowly when the object's contrast is decreased (Stone & Thompson 1992, Thompson 1982, Weiss et al. 2002). Stocker & Simoncelli (2006) demonstrated that this speed bias in motion perception is consistent with a Bayesian model based on an internal model of the likely speed distribution in nature, which is also biased toward slower speeds. This paradigm has also been applied to eye movements that occur while tracking moving objects (Lisberger 2010). Neurons in the frontal eye fields that initiate anticipatory eye movements show a bias for both target speed and direction that reflects the optimal Bayesian updating of a prior model based on changes in sensory inputs (Darlington et al. 2018).

SUMMARY AND FUTURE DIRECTIONS

Despite being treated separately in much of the literature, the concepts of efficient coding and future prediction are intimately related and complement each other at both a conceptual and a mechanistic level. Classical efficient coding mechanisms use knowledge of a large ensemble of natural inputs to prevent expected correlations from reaching the neural output. These mechanisms

primarily operate on correlations between two points in space or time to suppress long-range (i.e., low-frequency) correlations when the signal-to-noise ratio is favorable (**Figure 3e**). These processes that remove long-range correlations can benefit the goal of future prediction. Spatial decorrelation, for example, can improve the ability to locate a moving target—a key asset to motion prediction processes (Bialek et al. 2001, Palmer et al. 2015, Salisbury & Palmer 2016).

Prediction benefits from knowing the two-point correlation structure, such as the location of an object in space over time, but these correlations do not indicate the object's velocity. This information is contained in the higher-order spatiotemporal correlations (i.e., phase information). This principle also applies to the time domain. For example, in measuring temporal changes in the light intensity, the correlation times can be computed from the two-point correlations in the intensity values, but the higher-order correlations are needed to determine whether the intensity is increasing or decreasing.

Normative studies using a Bayesian approach to understanding neural processing have produced key insights into strategies that the visual system can use to cope with the challenges of encoding in dynamic environments (Młynarski & Hermundstad 2018, 2021). These studies can explain, from first principles, many properties of cells that either adapt or sensitize to changes in stimulus variance and to changes in mean light intensity (Appleby & Manookin 2019; Kastner & Baccus 2011, 2013; Nikolaev et al. 2013; Tikidji-Hamburyan et al. 2015). A similar approach has been applied to efficient coding in the visual cortex (Rao & Ballard 1999). In this paradigm, descending inputs from the cortex suppress ascending early sensory inputs that are predicted by the internal model of the visual environment.

A major open question is how to apply this important work to understanding predictive coding of future inputs. Hermundstad's work (Hermundstad et al. 2014; Młynarski & Hermundstad 2018, 2021) extends the efficient coding paradigm to understand the process of inference when an organism has only partial knowledge about the world. This idea should also be applicable to the information bottleneck. For example, we currently do not have a tractable way of calculating the information bottleneck when it is unclear which pieces of information are maximally informative (Bialek et al. 2001, Salisbury & Palmer 2016). Thus, one possible approach is to utilize a Bayesian approach that weights different pieces of information by their potential behavioral relevance to determine which information should be predicted and thus provide a tractable way of computing the bottleneck (Hermundstad et al. 2014; Młynarski & Hermundstad 2018, 2021). Such an approach would directly link predictive coding with another important and poorly understood issue in sensory neuroscience—how to quantify the behavioral relevance of different inputs.

SUMMARY POINTS

1. Efficient coding approaches to visual processing leverage ensemble statistics that reflect the average properties of natural environments (Dong & Atick 1995, Field 1987, Ruderman & Bialek 1994). The efficient coding hypothesis indicates that these average statistical properties should be removed from neural representations of sensory input to reduce information redundancy and improve efficiency (Barlow 1961, Rao & Ballard 1999).

2. An extension of this efficient coding paradigm asserts that the brain should leverage information from past sensory inputs to actively predict what is likely to occur in the future (Bialek et al. 2001, Friston 2010, Salisbury & Palmer 2016).

3. Efficient coding and prediction mechanisms both perform key and complementary functions, and they can act on distinct components of sensory inputs. Classical efficient coding largely acts on correlations between two points in space and time to suppress the long-range correlations that are overrepresented in the natural environment. Predicting future sensory inputs requires access to these two-point correlations and also to higher-order correlations among three or more spatiotemporal points.

FUTURE ISSUES

1. Previous studies have applied the information bottleneck approach under conditions in which the relevant stimulus statistics were completely known (Liu et al. 2021, Palmer et al. 2015). However, in the natural environment, organisms must operate in many contexts in which it is unclear what information should be prioritized because the relevant stimulus statistics are dynamic. Normative approaches have leveraged Bayesian statistics to understand the trade-offs inherent in neural coding during conditions of uncertainty (Młynarski & Hermundstad 2018, 2021). Can we also leverage the power of this Bayesian approach to determine the appropriate inputs to the information bottleneck when the relevant statistics are uncertain?

2. Work aimed at understanding how neural coding helps to predict the future has focused on visual motion (Berry et al. 1999, Johnston & Lagnado 2015, Liu et al. 2021, Palmer et al. 2015, Schwartz et al. 2007b). However, prediction should also occur across sensory contexts (Bialek et al. 2001). Understanding how widespread predictive processing is in the brain and the coding trade-offs associated with prediction across sensory contexts should contribute to a deeper understanding of the principles of neural processing.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

LITERATURE CITED

Angueyra JM, Baudin J, Schwartz GW, Rieke F. 2022. Predicting and manipulating cone responses to naturalistic inputs. *J. Neurosci.* 42(7):1254–74

Appleby TR, Manookin MB. 2019. Neural sensitization improves encoding fidelity in the primate retina. *Nat. Commun.* 10:4017

Atick JJ, Redlich AN. 1990. Towards a theory of early visual processing. *Neural Comput.* 2(3):308–20

Atick JJ, Redlich AN. 1992. What does the retina know about natural scenes? *Neural Comput.* 4(2):196–210

Attneave F. 1954. Some informational aspects of visual perception. *Psychol. Rev.* 61(3):183–93

Baccus SA, Meister M. 2002. Fast and slow contrast adaptation in retinal circuitry. *Neuron* 36(5):909–19

Barlow HB. 1961. The coding of sensory messages. In *Current Problems in Animal Behaviour*, ed. WH Thorpe, OL Zangwill, pp. 331–60. Cambridge, UK: Cambridge Univ. Press

Barlow HB, Levick WR. 1965. The mechanism of directionally selective units in rabbit's retina. *J. Physiol.* 178(3):477–504

Berry MJ II, Brivanlou IH, Jordan TA, Meister M. 1999. Anticipation of moving stimuli by the retina. *Nature* 398(6725):334–38

Bialek W. 2012. *Biophysics: Searching for Principles*. Princeton, NJ: Princeton Univ. Press

Bialek W, Nemenman I, Tishby N. 2001. Predictability, complexity, and learning. *Neural Comput.* 13(11):2409–63

Borst A, Theunissen FE. 1999. Information theory and neural coding. *Nat. Neurosci.* 2(11):947–57

Brenner N, Bialek W, de Ruyter van Steveninck R. 2000. Adaptive rescaling maximizes information transmission. *Neuron* 26(3):695–702

Campbell FW, Robson JG. 1968. Application of Fourier analysis to the visibility of gratings. *J. Physiol.* 197(3):551–66

Darlington TR, Beck JM, Lisberger SG. 2018. Neural implementation of Bayesian inference in a sensorimotor behavior. *Nat. Neurosci.* 21(10):1442–51

Darlington TR, Tokiyama S, Lisberger SG. 2017. Control of the strength of visual-motor transmission as the mechanism of rapid adaptation of priors for Bayesian inference in smooth pursuit eye movements. *J. Neurophysiol.* 118(2):1173–89

Dayan P, Hinton GE, Neal RM, Zemel RS. 1995. The Helmholtz machine. *Neural Comput.* 7(5):889–904

DeWeese M, Zador A. 1998. Asymmetric dynamics in optimal variance adaptation. *Neural Comput.* 10(5):1179–202

Dong DW, Atick JJ. 1995. Statistics of natural time-varying images. *Network* 6(3):345–58

Fairhall AL, Lewen GD, Bialek W, de Ruyter van Steveninck RR. 2001. Efficiency and ambiguity in an adaptive neural code. *Nature* 412(6849):787–92

Field DJ. 1987. Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A* 4(12):2379–94

Fitzgerald JE, Katsov AY, Clandinin TR, Schnitzer MJ. 2011. Symmetries in stimulus statistics shape the form of visual motion estimators. *PNAS* 108(31):12909–14

Friston K. 2010. The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11(2):127–38

Gollisch T, Meister M. 2008. Rapid neural coding in the retina with relative spike latencies. *Science* 319(5866):1108–11

Gollisch T, Meister M. 2010. Eye smarter than scientists believed: neural computations in circuits of the retina. *Neuron* 65(2):150–64

Heitman A, Brackbill N, Greschner M, Sher A, Litke AM, Chichilnisky EJ. 2016. Testing pseudo-linear models of responses to natural scenes in primate retina. bioRxiv 045336. **https://doi.org/10.1101/045336**

Hermundstad AM, Briguglio JJ, Conte MM, Victor JD, Balasubramanian V, Tkačik G. 2014. Variance predicts salience in central sensory processing. *eLife* 3:e03722

Hosoya T, Baccus SA, Meister M. 2005. Dynamic predictive coding by the retina. *Nature* 436(7047):71–77

Hu Q, Victor JD. 2010. A set of high-order spatiotemporal stimuli that elicit motion and reverse-phi percepts. *J. Vis.* 10(3):9

Johnston J, Lagnado L. 2015. General features of the retinal connectome determine the computation of motion anticipation. *eLife* 4:e06250

Kant I. 1998 (1781). *Critique of Pure Reason*, transl. P Guyer, AW Wood. Cambridge, UK: Cambridge Univ. Press

Kastner DB, Baccus SA. 2011. Coordinated dynamic encoding in the retina using opposing forms of plasticity. *Nat. Neurosci.* 14(10):1317–22

Kastner DB, Baccus SA. 2013. Spatial segregation of adaptation and predictive sensitization in retinal ganglion cells. *Neuron* 79(3):541–54

Kim KJ, Rieke F. 2001. Temporal contrast adaptation in the input and output signals of salamander retinal ganglion cells. *J. Neurosci.* 21(1):287–99

Koch K, McLean J, Segev R, Freed MA, Berry MJ II, et al. 2006. How much the eye tells the brain. *Curr. Biol.* 16(14):1428–34

Laughlin S. 1981. A simple coding procedure enhances a neuron's information capacity. *Z. Naturforsch. C* 36(9–10):910–12

Lewicki MS. 2002. Efficient coding of natural sounds. *Nat. Neurosci.* 5(4):356–63

Lisberger SG. 2010. Visual guidance of smooth-pursuit eye movements: sensation, action, and what happens in between. *Neuron* 66(4):477–91

Liu B, Hong A, Rieke F, Manookin MB. 2021. Predictive encoding of motion begins in the primate retina. *Nat. Neurosci.* 24(9):1280–91

Mani N, Huettig F. 2012. Prediction during language processing is a piece of cake—but only for skilled producers. *J. Exp. Psychol. Hum. Percept. Perform.* 38(4):843–47

Młynarski WF, Hermundstad AM. 2018. Adaptive coding for dynamic sensory inference. *eLife* 7:e32055

Młynarski WF, Hermundstad AM. 2021. Efficient and adaptive sensory codes. *Nat. Neurosci.* 24(7):998–1009

Nikolaev A, Leung K-M, Odermatt B, Lagnado L. 2013. Synaptic mechanisms of adaptation and sensitization in the retina. *Nat. Neurosci.* 16(7):934–41

Nitzany EI, Victor JD. 2014. The statistics of local motion signals in naturalistic movies. *J. Vis.* 14(4):10

Olshausen BA, Field DJ. 1996. Natural image statistics and efficient coding. *Network* 7(2):333–39

Oppenheim AV, Lim JS. 1981. The importance of phase in signals. *Proc. IEEE* 69(5):529–41

Palmer SE, Marre O, Berry MJ II, Bialek W. 2015. Predictive information in a sensory population. *PNAS* 112(22):6908–13

Piotrowski LN, Campbell FW. 1982. A demonstration of the visual importance and flexibility of spatial-frequency amplitude and phase. *Perception* 11(3):337–46

Pitkow X, Meister M. 2012. Decorrelation and efficient coding by retinal ganglion cells. *Nat. Neurosci.* 15(4):628–35

Rao RP, Ballard DH. 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* 2(1):79–87

Rieke F. 2001. Temporal contrast adaptation in salamander bipolar cells. *J. Neurosci.* 21(23):9445–54

Rieke F, Bodnar DA, Bialek W. 1995. Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proc. Biol. Sci.* 262(1365):259–65

Ruderman DL, Bialek W. 1994. Statistics of natural images: scaling in the woods. *Phys. Rev. Lett.* 73(6):814–17

Rust NC, Palmer SE. 2021. Remembering the past to see the future. *Annu. Rev. Vis. Sci.* 7:349–65

Sabbah S, Gemmer JA, Bhatia-Lin A, Manoff G, Castro G, et al. 2017. A retinal code for motion along the gravitational and body axes. *Nature* 546(7659):492–97

Salisbury JM, Palmer SE. 2016. Optimal prediction in the retina and natural motion statistics. *J. Stat. Phys.* 162(5):1309–23

Schwartz G, Harris R, Shrom D, Berry MJ II. 2007a. Detection and prediction of periodic patterns by the retina. *Nat. Neurosci.* 10(5):552–54

Schwartz G, Taylor S, Fisher C, Harris R, Berry MJ II. 2007b. Synchronized firing among retinal ganglion cells signals motion reversal. *Neuron* 55(6):958–69

Shannon CE. 1948. A mathematical theory of communication. *Bell Syst. Tech. J.* 27(3):379–423

Sharpee TO, Sugihara H, Kurgansky AV, Rebrik SP, Stryker MP, Miller KD. 2006. Adaptive filtering enhances information transmission in visual cortex. *Nature* 439(7079):936–42

Simoncelli EP, Olshausen BA. 2001. Natural image statistics and neural representation. *Annu. Rev. Neurosci.* 24:1193–216

Smirnakis SM, Berry MJ, Warland DK, Bialek W, Meister M. 1997. Adaptation of retinal processing to image contrast and spatial scale. *Nature* 386(6620):69–73

Smith EC, Lewicki MS. 2006. Efficient auditory coding. *Nature* 439(7079):978–82

Srinivasan MV, Laughlin SB, Dubs A. 1982. Predictive coding: a fresh view of inhibition in the retina. *Proc. R. Soc. Lond. B* 216(1205):427–59

Stocker AA, Simoncelli EP. 2006. Noise characteristics and prior expectations in human visual speed perception. *Nat. Neurosci.* 9(4):578–85

Stone LS, Thompson P. 1992. Human speed perception is contrast dependent. *Vis. Res.* 32(8):1535–49

Taylor WR, He S, Levick WR, Vaney DI. 2000. Dendritic computation of direction selectivity by retinal ganglion cells. *Science* 289(5488):2347–50

Thompson P. 1982. Perceived rate of movement depends on contrast. *Vis. Res.* 22(3):377–80

Tikidji-Hamburyan A, Reinhard K, Seitter H, Hovhannisyan A, Procyk CA, et al. 2015. Retinal output changes qualitatively with every change in ambient illuminance. *Nat. Neurosci.* 18(1):66–74

Tishby N, Pereira FC, Bialek W. 1999. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, pp. 368–77. Champaign, IL: Univ. Ill.

Turner MH, Rieke F. 2016. Synaptic rectification controls nonlinear spatial integration of natural visual inputs. *Neuron* 90(6):1257–71

Turner MH, Sanchez Giraldo LG, Schwartz O, Rieke F. 2019. Stimulus- and goal-oriented frameworks for understanding natural vision. *Nat. Neurosci.* 22(1):15–24

van Hateren JH. 1992. A theory of maximizing sensory information. *Biol. Cybern.* 68(1):23–29

von Helmholtz H. 2000 (1867). *Concerning the Perceptions in General*, Vol. 3, transl. JPC Southall. Bristol, UK: Thoemmes Press

Wark B, Fairhall A, Rieke F. 2009. Timescales of inference in visual adaptation. *Neuron* 61(5):750–61

Weiss Y, Simoncelli EP, Adelson EH. 2002. Motion illusions as optimal percepts. *Nat. Neurosci.* 5(6):598–604

Yang J, Lee J, Lisberger SG. 2012. The interaction of Bayesian priors and sensory data and its neural circuit implementation in visually guided movement. *J. Neurosci.* 32(49):17632–45