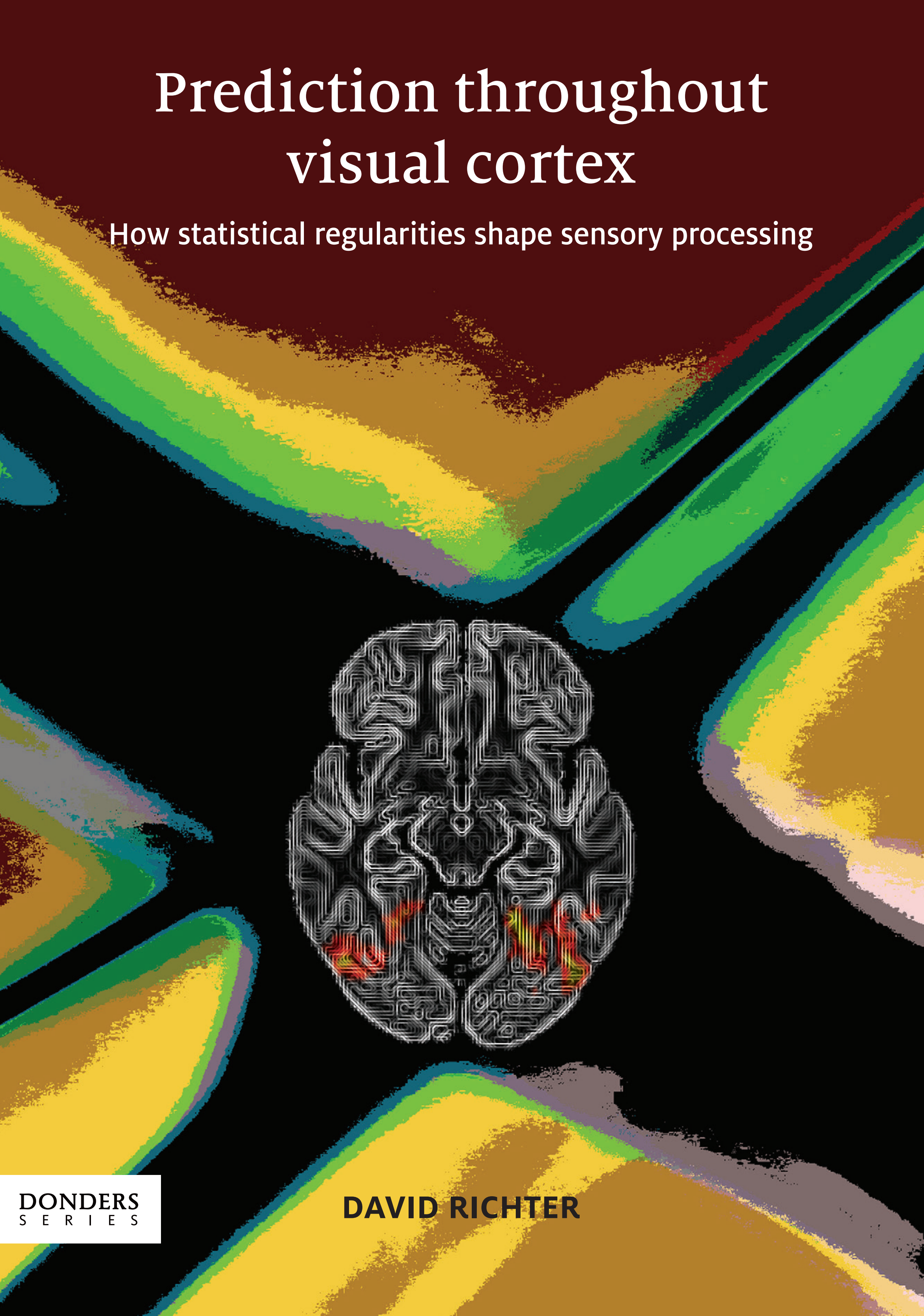
# Prediction throughout visual cortex

## How statistical regularities shape sensory processing



**DAVID RICHTER**

DONDERS
SERIES

# Prediction throughout visual cortex

## How statistical regularities shape sensory processing

DAVID RICHTER

# Prediction throughout visual cortex

## How statistical regularities shape sensory processing

*Coverdesign*
David Richter

*Lay-out*
Proefschriftenbalie, Nijmegen

*Print*
Ipskamp Printing, Enschede

ISBN 978-94-6421-219-8

Donders Institute for Brain, Cognition and Behaviour
Donders Centre for Cognitive Neuroimaging
Radboud University Nijmegen, The Netherlands

© David Richter, 2021

# Prediction throughout visual cortex

## How statistical regularities shape sensory processing

**Proefschrift**

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van decanen
in het openbaar te verdedigen op donderdag 11 maart 2021
om 16.30 uur precies

door

**David Richter**

geboren op 29 mei 1987
te Düsseldorf, Duitsland

**Promotor**
Prof. dr. F.P. de Lange

**Manuscriptcommissie**
Prof. dr. R.J. van Lier
Prof. dr. N.B. Turk-Browne *(Yale University, Verenigde Staten)*
Dr. C. Press *(Birkbeck, University of London, Verenigd Koninkrijk)*

# Prediction throughout visual cortex
## How statistical regularities shape sensory processing

**Doctoral Thesis**

to obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the Rector Magnificus prof. dr. J.H.J.M. van Krieken,
according to the decision of the Council of Deans
to be defended in public on Thursday, March 11, 2021
at 16.30 hours

by

**David Richter**

born on May 29, 1987
in Düsseldorf, Germany

**Supervisor**
Prof. dr. F.P. de Lange

**Doctoral Thesis Committee**
Prof. dr. R.J. van Lier
Prof. dr. N.B. Turk-Browne *(Yale University, United States of America)*
Dr. C. Press *(Birkbeck, University of London, United Kingdom)*

# Table of Contents

Chapter 1 Introduction 9

Chapter 2 Suppressed sensory response to predictable object stimuli throughout the ventral visual stream 25

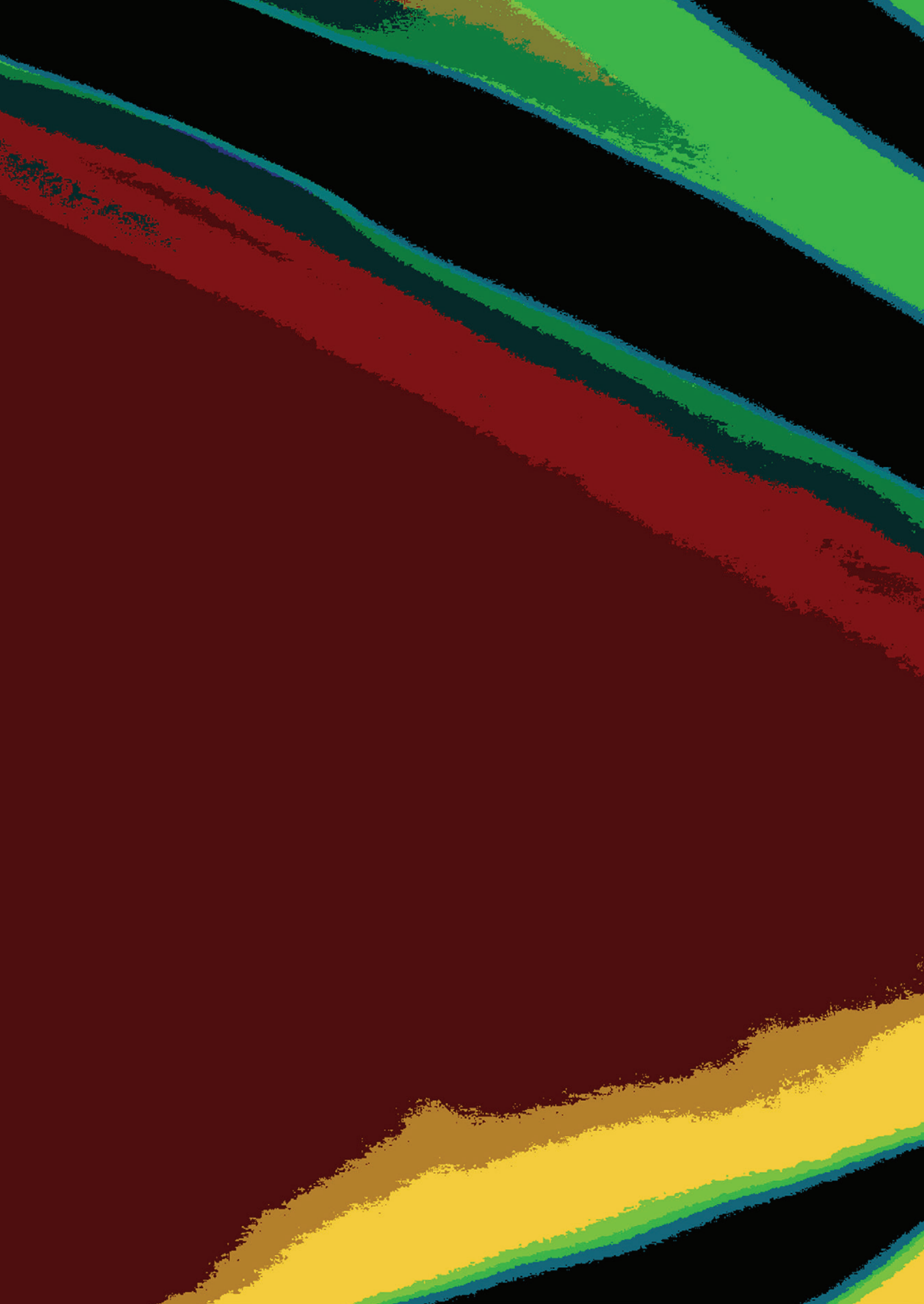
Chapter 3 Statistical learning attenuates visual activity only for attended stimuli 51

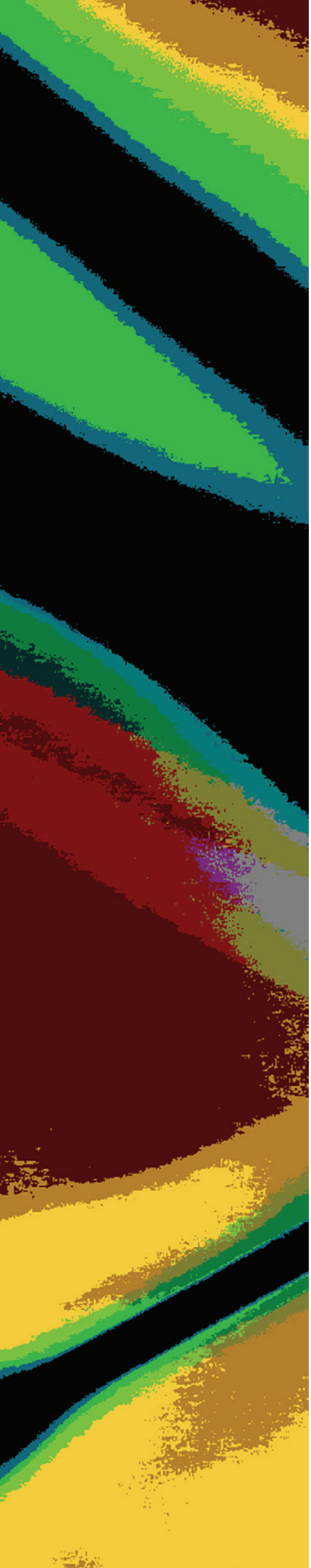
Chapter 4 Dampened sensory representations for expected input across the ventral visual stream 91

Chapter 5 Incidental statistical learning of unimodal but not cross-modal statistical regularities 147

Chapter 6 Discussion 189

References 207
Nederlandse samenvatting 221
Curriculum Vitae 225
List of Publications 226
Acknowledgments 227
Research Data Management 229
Donders graduate school for cognitive neuroscience 231





## CHAPTER 1

# Introduction

What do you see in the left image in Figure 1.1A? Initially, you probably perceive an arbitrary arrangement of black lines and shapes. However, if I tell you that this a degraded version of the picture of the cat on the right, your percept is likely to change to seeing the cat. Thus, only by virtue of the expectation to find a cat your conscious percept dramatically changed from an initially meaningless array of shapes to a coherent picture of a cat. This example demonstrates in a powerful and intuitive manner that perception appears to be fundamentally influenced by expectations. How expectations modulate perception, and in particular neural processing throughout the sensory brain, will be the focus of my thesis. But before establishing more specific questions, which I will address throughout this thesis, I will take a step back and assess how we can approach the study of perception and the influence of expectations.

Usually, conscious perception may appear to be definite and unambiguous, creating the intuition that the brain may simply register a definite sensory world. Indeed, traditionally perception has been construed as a feedforward process, moving from light sensitive cells in the retina, over simple contrast and edge detectors in early visual areas, to increasingly complex shape representations in higher visual areas [1,2]. Successive integration of sensory information, by means of feedforward and lateral excitatory and inhibitory connections between assemblies of neurons, allows for remarkably complex response properties to be constructed. However, accounting for the dramatic influence of expectation on perception, as you experienced in Figure 1.1A, is more challenging in a purely feedforward model.

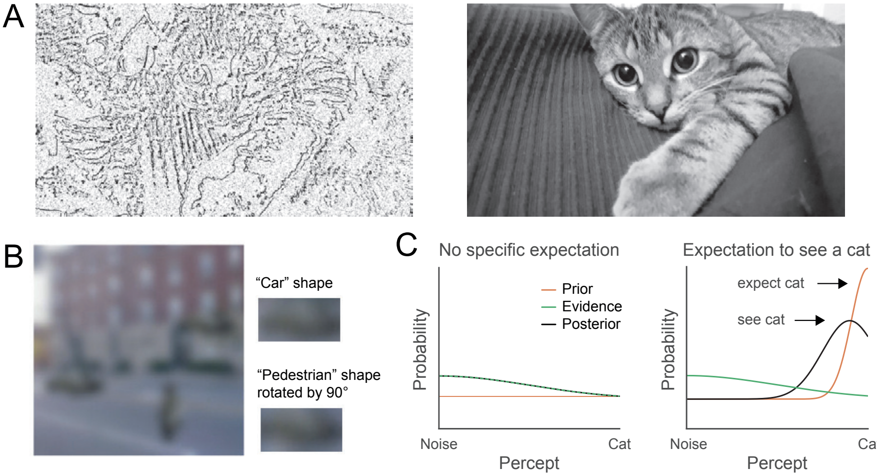


**FIGURE 1.1 Expectations modulate perception.**

(A) Low detail image demonstrating the effect of expectation on perception. Initially, the picture on the left may appear to contain only random shapes. However, upon reading (or noticing) that the image depicts the same cat as on the right, your percept may dramatically change into that of a cat, even though the bottom-up stimulus information remains identical. (B) Shown is a blurry city scene, with a car on the road on the left, and a pedestrian in the foreground on the right. Yet, in fact the blurry shapes of the "car" and "pedestrian" are identical, except for a 90 degree rotation, as demonstrated on the right. This image illustrates how expectations due to context can affect perception of individual objects. Image reproduced with permission from [3]. (C) Illustration of Bayesian inference. The abscissa denotes the possible percepts, ranging from visual noise (initial percept) to a cat (final percept), while the ordinate shows the associated probability. An enormous number of other percepts are also possible, but this depiction is simplified to two possibilities for illustration purposes. The left panel depicts your initial viewing of Figure 1.1A, that is, a situation without a specific prior (orange distribution) and ambiguous sensory input (evidence; green distribution). Thus, the posterior (black distribution) does match the evidence, a percept of visual noise, but also reflecting the large uncertainty about what the stimulus might depict (i.e., a very broad distribution). The right panel illustrates the situation after receiving an informative perceptual prior (expectation to see a cat), which is combined with the ambiguous sensory evidence, yielding a new posterior interpretation (seeing a cat), strongly influenced by the prior.

## Vision as perceptual inference

The intuition that perception is unambiguous, and simply reflects a definite sensory world, belies the true ambiguity that is in fact present in sensory input. Consider for example Figure 1.1B, depicting a city scene, with a car on the left and a pedestrian walking on the right. However, there is very little detail in the image and in fact the car and pedestrian are identical shapes, rotated only by 90 degrees. Yet, we do effortlessly identify a car and a pedestrian. How can we explain that so little visual detail suffices to determine what we see, and that two almost identical sensations create completely different percepts? Again, the answer is expectations. The combination of features resemble city scenes you have experienced before, and given this context you expect pedestrians on the sidewalk and cars on the road. In other words, your prior experience fills in the lack of visual detail. Curiously, the examples in Figure 1.1A and 1.1B thus suggest that what we perceive is a combination of prior knowledge, such as our expectations, and sensory input. Throughout this thesis I will approach perception as an inferential process, during which prior knowledge and current inputs are integrated in order to infer what the most likely causes for our sensations are, thereby creating our conscious percepts. This approach will allow me to investigate how expectations shape perception.

### BOX 1.1 Perception as Bayesian inference

The concept of perceptual inference dates back many years, with similar ideas already expressed by Hermann von Helmholtz in the middle of the 19th century [4]. Yet, it has only been in the past two decades that the idea has gained substantial traction and support in neuroscience, with prominent proposals establishing computational and implementational mechanisms underlying perceptual inference. Bayesian accounts of perception [5–7] constitute a particularly prominent approach to understanding the computational principles underlying perception. Bayesian inference is a method for updating believes by combining prior believes with new data (evidence). To illustrate what this means, let us return to the example of the cat in Figure 1.1A. In Bayesian terms your expectation to see a cat constitutes the prior. Initially, the prior probability of seeing a cat was very low, similar to the probability of a dog, a banana, or many other things. Thus, combined with the uncertain sensory evidence, you were unsure what the image might depict. This situation is illustrated in the left panel of Figure 1.1C – notice the low probability (i.e., high uncertainty) of the prior, evidence and posterior. However, by reading that this is the picture of the cat on the right, your prior believe of seeing a cat increased (right panel of Figure 1.1C; orange distribution). Ultimately, this prior 'moved' your interpretation of the ambiguous sensory evidence (green distribution) to the posterior interpretation of perceiving a cat (black distribution). In other words, on this account prior expectations and sensory data are continuously combined to infer the probable causes underlying sensation. Moreover, we can also appreciate that the influence of the prior will not only depend on the prior itself, but also on the reliability of the sensory evidence. That is, if sensory input is particularly noisy or ambiguous (e.g., Figure 1.1A), perception will be influenced more by prior knowledge [8] than if input is low in noise and unambiguous [9,10]. A fundamental strength of this account is that top-down influences on perception, such as expectations, can be readily integrated.

## Predictive coding

A particularly successful approach casting perception as a process of unconscious inference is predictive coding (e.g., [11–13]; reviews: [14,15]). Predictive coding implements computational principles of Bayesian inference (Box 1.1) by prediction error minimization. The idea is that at each step of the sensory hierarchy a mismatch between top-down prediction (prior) and bottom-up input (evidence) is computed. The resulting prediction error is weighted by its precision (i.e., sensory reliability; [16]), and in turn constitutes the bottom-up input for the next layer in the hierarchy. Thus, at each stage only the unpredicted signal has to be relayed as bottom-up input, hence reducing redundancy in sensory cortex [11]. Predictions are generated by internal models, given hidden variables, predicting the input specific to the respective layer in the cortical hierarchy. Thus, predictions and prediction errors are thought to be feature-specific [15]. Through recurrent information passing, that is the recursive relaying of predictions and prediction errors, errors can be minimized by adjusting predictions. This iterative processes aims to ultimately derive the interpretation yielding the smallest prediction error, hence reflecting the most likely cause of the sensory input. Thus, predictive coding is, in a sense, the inversion of the classical feedforward approach to perception, because top-down signals represent the (inferred) world, while bottom-up information is used as feedback to update the top-down representations (note, details of different predictive coding models can vary; for a review see: [17]).

# Expectation suppression

Let us reconsider the example of the cat in Figure 1.1A in the context of predictive coding. Initially, you did not know what to expect and the percept was incoherent. In other words, your uninformative predictions did not explain the bottom-up input well, thus resulting in a large prediction error. However, once your generative model was updated by reading that there is the cat in the stimulus, your prediction better matched the bottom-up input, resulting in smaller prediction errors and, through recursive error minimization, you converged on the coherent interpretation of a cat. Thus, we would expect that prediction errors to a predicted stimulus (expect cat) are smaller than the errors in response to an unpredicted stimulus (no expectation). In other words, predictions ‘explain away’ bottom-up activity. This hypothesis provides us an intriguing approach to assess the initially raised question: how do expectations modulate neural processing in the sensory brain?

Previous work has investigated prediction errors by inducing expectations through statistical regularity. For example, Kok et al. [18] presented participants with an auditory cue, which with 75% reliability predicted the orientation of a grating stimulus. The authors showed that neural responses in primary visual cortex were suppressed for expected compared to unexpected orientations. This phenomenon, also known as expectation suppression (reviews: [19,20]) is illustrated in Figure 1.2. In the figure, notice the relative suppression of neural activity to the stimulus when it is expected. According to predictive coding accounts, even a well-predicted stimulus evokes prediction errors, however these errors are resolved more quickly and are smaller in magnitude. Thus, expectation suppression appears to mirror the properties of prediction errors, that is, reduced responses to predicted stimuli. Crucially, expectation suppression is distinct from repetition suppression [21,22], and is found even when stimulus base rates and repetition frequencies are controlled for [18,23–28]. Moreover, expectation suppression has been demonstrated beyond primary visual cortex, in object selective visual areas [23,27,28] and in audition [22,29]. Thus, expectation suppression may be a key signature of how expectations modulate sensory processing, and therefore will be of particular interest in this thesis. Moreover, given that expectation suppression has been reported in different paradigms and sensory areas, we may hypothesize that prediction constitutes a general processing principle across cortex [12,19,30]. From this hypothesis we can derive specific characteristics which expectation suppression should have, if it does reflect a modulation by predictions, and if predictions do constitute a fundamental neural processing principle.

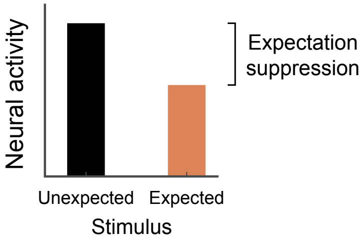


**FIGURE 1.2 Expectation suppression.**

Illustration of the suppression of neural activity by expectations. Sensory responses are suppressed if a stimulus is expected (e.g. expected a cat, and saw a cat), compared to the response to an unexpected stimulus (e.g. expected a banana, and saw a cat). I will refer to this relative suppression of neural responses as expectation suppression.

## Expectation suppression: a fundamental phenomenon of sensory processing?

First, sensory modulations by expectations, and in particular expectation suppression, should be evident throughout the sensory hierarchy. That is, if prediction is a core principle of sensory processing, we should find its effects throughout sensory cortex, given adequate predictions and stimuli. Second, predictive patterns common in our sensory world, such as associations between naturalistic object stimuli should be readily acquired and subsequently affect sensory processing. This contention asserts that, if prediction is a principle of sensory processing, prediction confirmations and violations of common stimuli, such as everyday objects, should result in expectation suppression. The rationale is that objects constitute a behaviorally relevant category of stimuli and that the association between such objects are prevalent in our sensory environment (e.g., a sidewalk predicting a pedestrian in Figure 1.1B), and hence should affect sensory processing. Third, learning predictions and subsequently utilizing them should occur automatically, without intention to learn or use the underlying predictions. This final assertion suggests that if perception is a form of unconscious inference, predictions should affect perception largely automatically and without any intention to learn or use these predictions.

While previous studies support some of these hypotheses, several discrepancies and gaps exist in the literature. On the one hand, some previous studies reported *enhanced* neural responses to predictable compared to random sequences of stimuli [31], and enhance responses to attended expected stimuli [32], thus suggesting that expectations may not always suppress neural responses. Moreover, previous studies in humans, providing evidence for expectation suppression, frequently investigated expectations instantiated by simple transitional probabilities. Given the paradigms used in these studies, predictions may have been learned and utilized explicitly by participants to predict upcoming input (e.g., [18,24,25,28]). Similarly, in studies with non-human primates it is unclear whether monkeys actively predicted the upcoming stimuli [23,27]. Hence, whether expectations suppression arises automatically and for task-irrelevant predictions, particularly following incidental learning of complex associations, remains unknown. If we can show that expectations arise automatically, for task-irrelevant predictions of object stimuli, and subsequently modulate perception, these results would provide additional support for the hypothesis that prediction is a fundamental neural process underlying perception. **Chapters 2 and 3** will address these questions and contribute to charting how expectations influence perceptual processing. Next, let us consider the properties of expectation suppression across visual cortex in more detail.

## Feature-specific predictions across the ventral visual stream?

Expectation suppression has been reported in response to prediction violations of object and face images in higher visual areas, such as inferior temporal cortex [23,27], and in humans in object and face selective areas [24,25,28]. However, from these results it remains unknown whether and how complex predictions of object and face stimuli modulate sensory processing across the entire ventral visual stream, particularly in early visual cortex. This question is non-trivial, because sensory neurons are tuned for different features across the visual hierarchy. V1 neurons are, for example, tuned to the orientation of contrasts [33], and neurons in LOC are selective for the shape of a stimulus [34]. Thus, when you expected a cat in Figure 1.1A, does this expectation of an entire object (a cat) translate into feature-specific shape predictions and even low level expectations of local oriented contrasts, relevant for response properties of neurons in V1 [33]? Hierarchical predictive coding accounts suggest that this is the case, as predictions and prediction errors are thought to be feature-specific [15]. Yet, as also noted by Walsh et al. [14] in their review of predictive processing, there is little work directly assessing this claim. Exploring the feature-specificity of predictions and prediction errors across the hierarchy is crucial, as it tests an important hypothesis of predictive coding and further informs us how expectations modulate perception. I will address this question using fMRI in **chapters 2-3** and forward modelling in **chapter 4**.

## Expectation suppression: an effect of prediction or attention?

If expectation suppression is a wide-spread and feature-specific neural phenomenon, as hypothesized above, this would however not necessarily mean that it does in fact reflect prediction error signals. That is, while I introduced expectation suppression here in terms of predictive processing accounts, the phenomenon itself is not uniquely accounted for by prediction based theories. For example, one may propose that surprise attracts attention [35–37], and attention in turn modulates the gain of sensory neurons [38,39]. Expectation suppression could therefore reflect increased attention towards unexpected (surprising) compared to expected stimuli [40]. Exploring how expectations modulate sensory processing is a key question in this thesis, hence evaluating this alternative account in more detail is crucial, as it fundamentally questions error coding in sensory cortex. **Chapters 2-3** contain experiments which provide valuable insight for addressing this alternative account, and I will revisit and discuss the evidence for and against it in the general discussion in **chapter 6**.

## Do expectations sharpen or dampen neural representations?

Finally, even if we can establish that expectation suppression reflects prediction errors, this would still not address *how* expectations modulate neural responses, or which neural populations are suppressed by prediction. Two accounts are commonly discussed in the expectation literature [19], which at the onset of the projects reported in this thesis only received limited empirical support. Sharpening accounts suggest that neural populations tuned *away* from the expected stimulus features are particularly suppressed by expectations [18,41]. This modulation results in an overall suppressed response, but sharpened population representation of the expected stimulus. In other words, on this account representations are enhanced by suppressing features not in line with the top-down predictions (Figure 1.3A, left panel). The middle image in Figure 1.3B illustrates the population representation modulated according to the sharpening account – note the increased (sharpened) contrast compared to baseline (left image in Figure 1.3B). In opposition to sharpening, dampening (or cancellation; [42]) accounts posit that neural populations tuned *towards* expected features are suppressed by expectations (Figure 1.3A, right panel). Thus, dampening results in a suppressed response, and a dampened population representation of the expected stimulus [23,43]. Predictable input is thus effectively filtered out on this account. In Figure 1.3B, notice the reduced (dampened) contrast.

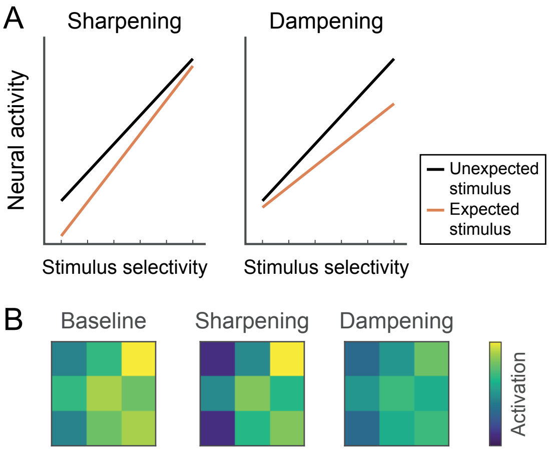


**FIGURE 1.3 Sharpening and dampening accounts of expectation suppression.**

(A) Illustration of the suppression of neural activity according to sharpening and dampening accounts of expectations. According to sharpening account (left), the suppression magnitude is larger for neural populations tuned away from the expected stimulus features, hence resulting in less expectation suppression the more selective a neuron is for a given stimulus. In contrast, dampening (right) proposes that the neurons tuned towards the expected stimulus features are most suppressed, thus expectation suppression positively scales with stimulus selectivity. (B) Depicted is the population response to an unexpected stimulus (i.e., baseline; left) and to an expected stimulus according to sharpening (middle) and dampening (right) respectively. Sharpening results in an increased contrast of the pattern present in the population response; i.e., a sharper representation of the expected stimulus. Dampening reduces the contrast of the pattern, thus resulting in a dampened representation of the expected stimulus. Note that the averaged response is equal for the sharpened and dampened response, only the specific population which experiences most suppression differs. Thus, both accounts result in expectation suppression, while making opposite prediction with respect to the neural population that is suppressed.

Comparing these two distinct accounts will yield insight into the neural modulations underlying expectation suppression, and suggest which functional role expectations may have in guiding perception, because both accounts are associated with different adaptive mechanisms. That is, sharpening may facilitate veridical representations, hence aiding in accurate and rapid perception, while dampening may reduce redundancy in sensory cortex and highlight novel information [19,44]. **Chapters 2 and 4** contrast predictions of sharpening and dampening accounts, using fMRI and forward models, in an effort to elucidate what type of neural modulation underlies perceptual expectations.

## Statistical learning and sources of perceptual priors

The questions outlined above primarily concern the consequences of expectations for sensory processing. Next, let us briefly consider how these expectations (priors) are formed in the first place. We have seen based on the example in Figure 1.1A that priors can be induced by explicitly receiving information – i.e., I showed you the original picture of the cat. However, in Figure 1.1B you did not require any instructions or additional information in order to make sense of the city scene. The relevant prior that cars usually drive on the road, you probably formed throughout your life, based on experiences in situations resembling the context present in the image. Yet, you may never have explicitly noticed that you have learned this prior. These simple examples suggest two things. Priors can be formed in different ways, likely involving distinct routes towards their acquisition. Second, one mechanism by which priors can be formed involves the extraction of statistical regularities from the sensory world; e.g., cars usually drive on roads. Statistical regularities do not only occur on a conceptual level (cars driving on roads), but also for low level sensory features (e.g., temporal regularity in visual features [45]). Thus, sensitivity to statistical regularities provides a tremendous source of information to predict future states of the world, from short-term sensory input to long-term, high level events unfolding in predictable ways. Indeed, predictive coding suggests that priors are formed by iteratively adjusting predictions by prediction error minimization, and thus priors come to represent the statistical regularities in the sensory world. Therefore, investigating how agents extract statistical regularities from the sensory environment over time, also known as statistical learning (reviews: [46–49]), is a particularly powerful approach to elucidate the neural mechanisms underlying perceptual inference, and will be a second focus in this thesis.

The acquisition of statistical regularities can occur incidentally [50], possibly even implicitly [31,51], across different modalities, including vision [36,52–54] and audition

[55–57]. Often statistical learning results in facilitated behavioral responses, such as faster and more accurate responses to expected stimuli [51,58,59]. And, arching back to expectation suppression, the sensory consequences of statistical learning are frequently a suppression of neural responses. In particular, sensory responses to sequentially presented stimuli have been shown to be suppressed for stimuli that were expected given the previous image [23,26]. Thus, statistical learning appears to provide a fundamental route towards learning sensory priors, and a different lens through which the sensory consequences of expectations can be investigated. While numerous studies demonstrated how and under which circumstances sensory priors may be acquired (reviews: [46,47,49]), several questions remain in the field of statistical learning. I will address some of these questions in this thesis, as an understanding of the mechanism for acquiring sensory priors can yield valuable insight into how expectations modulate perception.

## Are the sensory consequences of statistical learning automatic?

Statistical learning and its sensory consequences have been suggested to occur automatically, without intent or awareness [31,50,51,60]. However, it remains unclear whether expectation suppression, following incidental statistical learning, arises for unattended stimuli. That is, once acquired, do predictions necessarily impact sensory processing, or does this modulation hinge on actively attending the predictable stimuli? Previous studies have yielded mixed results, with some reporting expectation suppression for unattended stimuli [32,61], while others find no modulation of sensory responses by expectations without attention [62]. Additionally, while some previous studies manipulated task-relevance, their effectiveness of manipulating attention can be questioned, as will be elaborated on in **chapter 3**. Shedding light on the automaticity of the sensory consequences of statistical learning will yield new insight into the origins and consequences of expectation modulations in sensory cortex, and in particularly whether this modulation occurs pre-attentively. In **chapter 3** I will report an fMRI study we performed to probe the automaticity of expectation suppression and its dependence on attention.

## Does statistical learning depend on modality-specific and domain-general mechanisms?

Whether statistical learning depends on domain-general and modality-specific mechanisms remains debated [47,49,63]. That is, some mechanisms underlying statistical learning may be subject to modulations and constrains particular to the specific stimulus modality [63–67]. Moreover, a central neural mechanism has also been suggested to contribute to statistical learning, irrespective of the sensory modality [36,56,68–70]. But, whether domain-generality constitutes a crucial bottleneck for learning cross-modal associations has seen little investigation. The rationale is that comparing cross-modal to unimodal learning provides a window on modality-specific and domain-general contributions to statistical learning, as cross-modal learning requires additional integration of sensory information in a domain-general network, and hence cannot occur based on modality-specific mechanisms alone. I will explore the question of domain-generality and modality-specificity in **chapter 5**. Establishing whether modality-specific and domain-general constraints limit statistical learning may provide valuable insight into the underlying neural mechanisms supporting the extraction of statistical regularities over time, and in turn into the sensory consequences of statistical learning, expectation suppression. Moreover, besides cross-modal limitations to statistical learning, **chapter 5** will also investigate how the reliability of deterministic compared to probabilistic associations affects the acquisition of statistical regularities. Given that many perceptual priors may be associated with uncertain outcomes and concern association between sensory modalities, elucidating how people learn from non-deterministic and cross-modal associations can provide insight into how and to what extent incidental statistical learning may provide priors supporting perceptual inference.

## Overview of this thesis

In sum, at the core of this thesis is the question how expectations influence perception. A central question is whether prediction may constitute a fundamental operating principle of the sensory brain. I approach perception as a process of unconscious perceptual inference, in line with predictive processing accounts of perception [11–13], as this approach allows me to address these core questions guided by a well-established framework. In particular, I will focus on how expectations, derived from statistical regularities in the sensory world, modulate processing throughout the ventral visual stream. The key neural phenomenon will be expectation suppression, the attenuated sensory response to predicted stimuli, which I view through two lenses, as a signature of perceptual inference, possibly reflecting prediction errors, and a consequence of statistical learning.

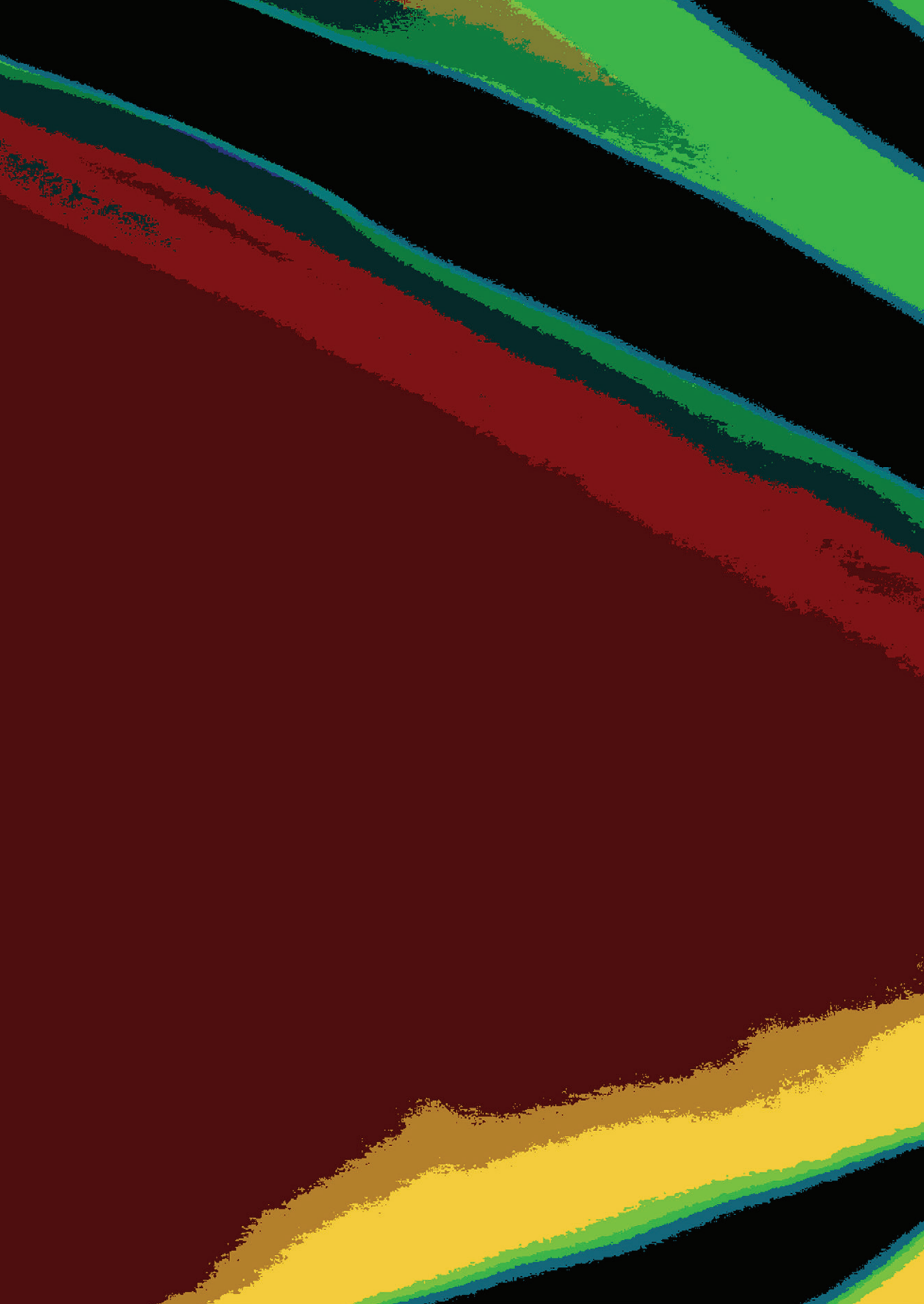
**Chapter 2** will start by exploring the extent to which complex object expectations, derived by incidental statistical learning, affect sensory processing throughout the ventral visual stream, as measured by fMRI. Moreover, in this chapter I will also investigate how population representations are modulated by expectations, contrasting the sharpening and dampening accounts.

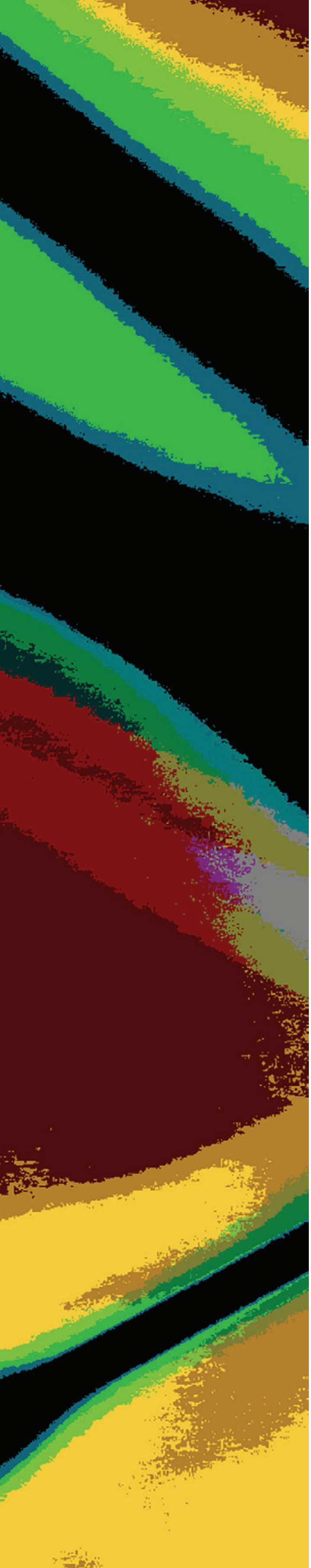
**Chapter 3** aims to assess whether the sensory consequence of statistical learning, expectation suppression, arises automatically or is dependent on attention, using fMRI. In addition, I further chart the characteristics of expectation suppression by investigating the stimulus-specificity of object level predictions across the ventral visual stream.

**Chapter 4** will assess what type of neural modulation best explains expectation suppression, using forward models and a combined analysis of the fMRI datasets of chapters 2 and 3. To this end I will contrast sharpening and dampening models of expectations, as well as models implementing a feature-unspecific suppression.

**Chapter 5** is devoted to exploring the limits of incidental statistical learning. A particular focus is on the modality-specificity of statistical learning, which is assessed by comparing learning of cross-modal and unimodal associations. Moreover, the capacity to learn from probabilistic compared to deterministic associations is explored.

**Chapter 6** summarizes and integrates the results presented in chapters 2-5. Moreover, an alternative account, casting expectation suppression as reflecting attention modulations instead of prediction error coding, will be discussed. I will additionally review potential explanations for the discrepancy in the expectation literature yielding sharpening and dampening respectively. Finally, I will highlight the core conclusions we can draw from the work presented in this thesis and the wider literature, focusing on how expectations influence perceptual processing.





## CHAPTER 2

# Suppressed sensory response to predictable object stimuli throughout the ventral visual stream

*This chapter has been published as:*
Richter, D., Ekman M., de Lange FP (2018) Suppressed Sensory Response to Predictable Object Stimuli throughout the Ventral Visual Stream. *J. Neurosci. 38*, 7452–7461. doi.org/10.1523/JNEUROSCI.3421-17.2018

# Abstract

Prediction plays a crucial role in perception, as prominently suggested by predictive coding theories. However, the exact form and mechanism of predictive modulations of sensory processing remain unclear, with some studies reporting a downregulation of the sensory response for predictable input, while others observed an enhanced response. In a similar vein, downregulation of the sensory response for predictable input has been linked to either sharpening or dampening of the sensory representation, which are opposite in nature. In the present study we set out to investigate the neural consequences of perceptual expectation of object stimuli throughout the visual hierarchy, using fMRI in human volunteers. Participants of both sexes were exposed to pairs of sequentially presented object images in a statistical learning paradigm, in which the first object predicted the identity of the second object. Image transitions were not task relevant; thus all learning of statistical regularities was incidental. We found strong suppression of neural responses to expected compared to unexpected stimuli throughout the ventral visual stream, including primary visual cortex (V1), lateral occipital complex (LOC), and anterior ventral visual areas. Expectation suppression in LOC scaled positively with image preference and voxel selectivity, lending support to the dampening account of expectation suppression in object perception.

# Introduction

Our environment is structured by statistical regularities. Making use of such regularities by anticipating upcoming stimuli is of great evolutionary value, as it enables the agent to predict future states of the world and prepare adequate responses, which in turn can be executed faster or more accurately [51,58,59]. Our brains are exquisitely sensitive to these statistical regularities [31,71–73]. In fact, it has been suggested that a core operational principle of the brain is prediction [74] and prediction error minimization [12]. Statistical learning is an automatic learning process by which statistical regularities are extracted from the environment [73], without explicit awareness or effort by the observer [50,75], even under concurrent cognitive load [76]. These statistical regularities can be used to form predictions about upcoming input, with effects of statistical learning being evident even 24 hours after exposure [51].

The neural consequences of perceptual predictions have been investigated extensively, but conflicting results have emerged. For example, Turk-Browne et al. [31] reported larger neural responses to predictable than random sequences of stimuli in human object-selective lateral occipital complex (LOC). However, contrary to this notion, neurons in monkey inferotemporal cortex (IT), the putative homologue of human LOC [77], showed reduced responses to expected compared to unexpected object stimuli [23,26]. This is in line with findings in human primary visual cortex (V1), which revealed that visual gratings of an expected orientation elicit a suppressed neural response compared to gratings of an unexpected orientation [18,78]. Even though there is superficial agreement between these studies, the exact form of expectation suppression, in terms of the underlying effect of expectations on the neural representations of stimuli, appeared to be opposite. Kok et al. [18] observed the strongest suppression in voxels that were tuned away from the expected stimulus, resulting in a sparse, sharpened population code. Electrophysiological studies in macaques on the other hand have reported a positive scaling of expectation suppression with image preference [23], suggesting that sensory representations are dampened for expected stimuli [79].

In sum, several discrepancies remain concerning the neural basis of perceptual expectation, which may be related to differences in species (macaque vs. human), measurement technique (spike rates vs. fMRI BOLD), and cortical hierarchy (early vs. late). In the current study, we set out to examine the existence and characteristics of expectation suppression throughout the visual hierarchy, using a paradigm that closely matches a set of previous studies on object prediction in macaque monkeys [23,27]. This allowed us to better compare and generalize between species, methods, and different levels of the cortical hierarchy. First we exposed participants to pairs of sequentially presented object images in a statistical learning paradigm. Next, we recorded neural responses, using whole-brain fMRI, to expected and unexpected object image pairs. By contrasting responses to expected and unexpected pairs we probed whether a suppression of expected object stimuli is evident throughout the ventral visual stream, and in particular in object-selective cortex. Moreover, by investigating expectation suppression as a function of image preference and voxel selectivity we contrasted sharpening with dampening (scaling) accounts of expectation suppression.

In brief, our results show that expectation suppression is ubiquitous throughout the human ventral visual stream, including object-selective LOC. Furthermore, we found that expectation suppression positively scales with object image preference and voxel selectivity within object-selective LOC. This suggests that object predictions dampen sensory representations in object-selective regions.

# Materials and Methods

## Participants

Twenty-four healthy, right-handed participants (17 female, aged 23.3 ± 2.4 years, mean ± SD) were recruited from the Radboud research participation system. The sample size was based on an a priori power calculation, computing the required sample size to achieve a power of 0.8 to detect an effect size of Cohen's d $\geq$ 0.6, at alpha = 0.05, for a two-tailed within subjects t-test. Participants were prescreened for MRI compatibility, had no history of epilepsy or cardiac problems, and normal or corrected-to-normal vision. Written informed consent was obtained before participation. The study followed institutional guidelines of the local ethics committee (CMO region Arnhem-Nijmegen, The Netherlands). Participants were compensated with 42 euro for study participation. Data from one subject was excluded due to excessive tiredness and poor fixation behavior. One additional subject was excluded from all ROI based analyses, since no reliable object-selective LOC mask could be established due to subpar fixation behavior during the functional localizer.

## Experimental Design and Statistical Analysis

### *Stimuli and experimental paradigm*

**Main task.** Participants were exposed to two object images in quick succession. Each image was presented for 500 ms without interstimulus interval, and an intertrial interval of 1500-2500 ms during behavioral training and 4110-6300 ms during fMRI scanning (see Figure 2.1A for a single trial). A fixation bullseye (0.5° visual angle in size) was presented throughout the run. For each participant 16 object images were randomly selected from a pool of 80 stimuli (also see: *Stimuli*). Eight images were assigned as leading images, i.e. appearing first on trials, while the other eight images served as trailing images, occurring second. Image pairs and the transitional probabilities between them were determined by the transitional probability matrix depicted in Figure 2.1B, based on the transition matrix used by Ramachandran et al. [27]. The expectation manipulation consisted of a repeated pairing of images in which the leading image predicted the identity of the trailing image, thus over time making the trailing image expected given the leading image. Importantly, the transitional probabilities governing the associations between images were task irrelevant, since participants were instructed to respond, by button press, to any upside-down versions of the images, the occurrence of which was not related to the transitional probability manipulation and could not be predicted. Upside-down images (target trials) occurred on ~9% of trials. Participants were not informed about the presence of any statistical regularities and instructed to maintain fixation on the central fixation bulls-eye. Trial order was fully randomized.

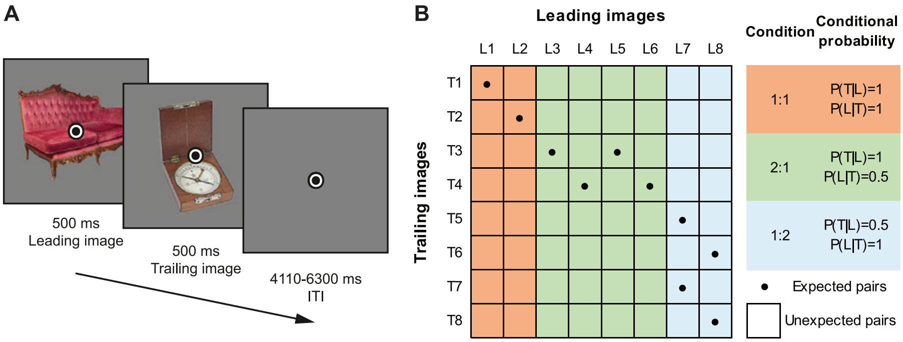


FIGURE 2.1 Paradigm overview.

(A) Depicts a single trial, with two example images and superimposed fixation bullseye. Leading images and trailing images were presented for 500 ms each, without interstimulus interval, followed by an intertrial interval of 4110-6300 ms (fMRI session; 1500-2500 ms during behavioral training). Participants responded to upside-down images by button press; the image at either position (leading or trailing) could be upside-down. (B) Shows the utilized image transition matrix determining image pairs. Eight leading images (L1 – L8) and eight trailing images (T1- T8) were used for each participant. Conditional probability conditions are highlighted and their respective conditional probabilities during training are listed on the right; condition 1:1 (orange), 2:1 condition (green), 1:2 condition (blue). Cells with dots indicate expected image pairs, while empty cells denote unexpected pairs.

During behavioral training only expected image pairs were presented on a total of 1792 trials, split into 8 blocks with short breaks in between blocks. Thus, during this session the occurrence of image L1 was perfectly predictive of image T1 (i.e. $P(T1|L1) = 1$; see Figure 2.1B). Apart from these trials, which constituted the 1:1 conditional probability condition, there were also trials with a 2:1 and 1:2 image pairing. In the 2:1 conditional probability condition the leading image was perfectly predictive of the trailing image (e.g. $P(T3|L3) = 1$), but two different leading images predicted the same trailing image, thereby reducing the conditional probability of the *leading* image given a particular trailing image (i.e. $P(L3|T3) = 0.5$). Lastly, the 1:2 condition consisted of a reduced predictive probability of the trailing image given the leading image, as such image L7 for instance was equally predictive of images T5 and T7 (i.e. $P(T5|L7) = 0.5$ and $P(T7|L7) = 0.5$).

On the next day participants performed one additional behavioral training block, consisting of 224 trials, and another 48 practice trials in the MRI during acquisition of the anatomical image. The task during the subsequent fMRI experiment was identical to the training session, except that also unexpected image pairs occurred. Nonetheless, the expected trailing image was still most likely to follow a given leading image, namely on 56.25% of trials compared to 6.25% for each unexpected trailing image (1:1 condition). It is important to note that each trailing image is only (un-)expected by virtue of its temporal context, i.e. which leading image it has been preceded by. Thus, each trailing images serves both as an expected and unexpected image depending on context. Additionally, trial order was fully randomized, thus rendering systematic effects of trial history unlikely. In sum, any difference between expected and unexpected occurrences cannot be explained in terms of different base rates of the trailing images, adaptation or trial history. Since intertrial intervals were longer in the fMRI session, and responses to upside-down images therefore occurred at a lower rate, potentially reducing participants' vigilance, the percentage of upside-down images was increased to ~11% of trials. As during the behavioral training session, in the main fMRI task participants were not informed about the presence of transitional probabilities, and there was no correlation between the image transitions and the occurrence of upside-down images. In total the MRI main task consisted of 512 trials, split into four equal runs, with an additional three resting blocks (each 12 sec) per run. Feedback on behavioral performance (percent correct and mean response time) was provided after each run. To ensure adequate fixation on the fixation bullseye, an infrared eye tracker (SensoMotoric Instruments, Berlin, Germany) was used to record and monitor eye positions.

**Functional localizer.** The main task was followed by a functional localizer, which was used for a functional definition of object-selective LOC for each participant, and to determine image preference for each voxel within visual cortex in an expectation neutral context. Finally, localizer data served as independent training data for the multi-voxel pattern analysis (see: *Data analysis, Multi-voxel pattern analysis*). In a block design each object image was presented four times, each time flashing at 2 Hz (300 ms on, 200 ms off) for 11 sec. The utilized stimuli were the same object images as shown during the fMRI main task. Additionally, a globally phase-scrambled version of each image [80] was shown twice, also flashing at 2 Hz for 11 sec. The order of objects images and scrambles was randomized. Participants were instructed to fixate the bullseye and respond by button press whenever the fixation bullseye dimmed in brightness.

**Questionnaire.** Following the fMRI session, participants filled in a brief questionnaire probing their explicit knowledge of the image transitions. Knowledge of each of the eight image pairs was tested by presenting participants with one leading image at a time, instructing them to select the most likely trailing image.

**Categorization task.** Finally, outside the scanner, participants performed a categorization task. During this task, participants indicated, by button press, whether the trailing image would fit into a shoebox (yes/no decision); similar to Dobbins et al. [81], and Horner and Henson [82]. This task was aimed at assessing any implicit reaction time or accuracy benefits due to incidental learning, since the statistical regularities, learned during the previous parts of the experiment, could be used to predict the correct response before the trailing image appeared. For each participant the same images and transitions were used as during their fMRI task. Furthermore, it was ensured that half of the trailing images in each conditional probability condition (1:1, 1:2, 2:1) fit into a shoebox, while the other half did not fit. A brief practice block was used to make sure that participants correctly classified the object images and understood the task. Participants were not informed about the intention behind this task, nor were they instructed to make use of the statistical regularities, in order to avoid influencing their behavior. A full debriefing took place after the categorization task.

**Stimuli.** Object stimuli were taken from Brady et al. [83], and consisted of a large collection of diverse full-color photographs of objects. Of this full set of images, a subset of 80 images was selected; 40 objects fitting into a shoebox, and 40 objects not fitting into a shoebox. Images spanned approximately 5° x 5° visual angle and were presented in full-color on a mid-grey background. During training stimuli were displayed on a LCD screen and back-projected during MRI scanning (EIKI LC-XL100 projector; 1024 x 768 pixel resolution, 60 Hz refresh rate), visible using an adjustable mirror. Since images were drawn at random per participant, each image could occur in any condition or position, thereby eliminating potential effects induced by individual image features.

### *fMRI data acquisition*

Functional and anatomical images were collected on a 3T Skyra MRI system (Siemens, Erlangen, Germany), using a 32-channel headcoil. Functional images were acquired using a whole-brain T2*-weighted multiband-8 sequence (time repetition [TR] / time echo [TE] = 730/37.8 ms, 64 slices, voxel size 2.4 mm isotropic, 50° flip angle, A/P phase encoding direction). Anatomical images were acquired with a T1-weighted magnetization prepared rapid gradient echo sequence (MP-RAGE; GRAPPA acceleration factor = 2, TR/TE = 2300/3.03 ms, voxel size 1 mm isotropic, 8° flip angle).

### *Data analysis*

**Behavioral data analysis.** Behavioral data from the categorization task was analyzed in terms of reaction time (RT) and accuracy. All RTs exceeding 3 SD above mean and below 200 ms were excluded as outliers (2.0% of trials). Since unexpected trailing image trials during the categorization task may require a change in the response, any differences in RT and accuracy between the expected and unexpected conditions may reflect a combination of surprise and response adjustment, thereby inflating possible RT and accuracy differences. Therefore, only unexpected trials requiring the same response as the expected image were analyzed, yielding an unbiased comparison of the effect of expectation. RTs for expected and unexpected trailing image trials were averaged separately per participant and subjected to a paired t-test. The error rate was also calculated separately for expected and unexpected trailing image trials per subject and analyzed with a paired t-test. Additionally, the effect size of both differences was calculated in terms of Cohen's $d_z$ [84]. All standard errors of the mean presented here were calculated as the within-subject normalized standard error [85] with Morey's [86] bias correction.

**fMRI data preprocessing.** fMRI data preprocessing was performed using FSL 5.0.9 (FMRIB Software Library; Oxford, UK; www.fmrib.ox.ac.uk/fsl; [87], RRID:SCR_002823). The preprocessing pipeline included brain extraction (BET), motion correction (MCFLIRT), temporal high-pass filtering (128 s), and spatial smoothing for univariate analyses (Gaussian kernel with full-width at half-maximum of 5 mm). No smoothing was applied for multivariate analyses, nor for the voxel-wise image preference analysis. Functional images were registered to the anatomical image using FLIRT (BBR) and to the MNI152 T1 2mm template brain (linear registration with 12 degrees of freedom). The first eight volumes of each run were discarded to allow for signal stabilization.

**Univariate data analysis.** To investigate expectation suppression across the ventral visual stream, voxel-wise general linear models (GLM) were fit to each subject's run data in an event-related approach using FSL FEAT. Separate regressors for expected and unexpected image pairs were modeled within the GLM. All trials were modeled with one second duration (corresponding to the duration of the leading and trailing image combined) and convolved with a double gamma haemodynamic response function. Additional nuisance regressors were added, including one for target trials (upside-down images), instruction and performance summary screens, first-order temporal derivatives for all modeled event types, and 24 motion regressors (six motion parameters, the derivatives of these motion parameters, the squares of the motion parameters, and the squares of the derivatives; comprising FSL's standard + extended set of motion parameters). The contrast of interest for the whole-brain analysis compared the average BOLD activity during unexpected minus expected trials, i.e. expectation suppression. Data was combined across runs using FSL's fixed effect analysis. For the across participants whole-brain analysis, FSL's mixed effect model FLAME 1 was utilized. Multiple comparison correction was performed using Gaussian random-field based cluster thresholding, as implemented in FSL, using a cluster-forming threshold of $z > 3.29$ (i.e. $p < 0.001$, two-sided) and a cluster significance threshold of $p < 0.05$. An identical analysis was performed to assess the influence of the different conditional probability conditions (see: *Main task*), except that the expected and unexpected event regressors were split into their respective conditional probability conditions (1:1, 1:2, 2:1), thus resulting in a GLM with six regressors of interest.

**Planned region of interest analyses.** Within each ROI (V1 and LOC; see: *Region of interest definition*), the parameter estimates for the expected and unexpected image pairs were extracted separately from the whole-brain maps. Per subject the mean parameter estimate within the ROIs was calculated and divided by 100 to yield an approximation of mean percent signal change compared to baseline [88]. These mean parameter estimates were in turn subjected to a paired t-test and the effect size of the difference calculated (Cohen's $d_z$). For the conditional probability manipulation, a similar ROI analysis was performed, except that the resulting mean parameter estimates were subjected to a 3x2 repeated measures ANOVA with conditional probability condition (1:1, 2:1, 1:2) and expectation (expected, unexpected) as factors. For this analysis we calculated eta-squared ($\eta^2$) as a measure of effect size.

**Multi-voxel pattern analysis.** Multi-voxel pattern analysis (MVPA) was performed per subject on mean parameter estimate maps per trailing image. These maps were obtained by fitting voxel-wise GLMs per trial for each subject, following the 'least squares separate' approach outlined in Mumford et al. [89]. In brief, a GLM is fit for each trial, with only that trial as regressor of interest and the remaining trials as one regressor of no interest. This was done for the functional localizer and main task data. The resulting parameter estimate maps of the functional localizer were used as training data for a multi-class SVM (classes being the eight trailing images), as implemented in Scikit-learn (SVC; [90], RRID:SCR_002577). Decoding performance was tested per subject on the mean parameter estimate maps from the main task data for each trailing image, split into expected and unexpected image pairs. The choice to decode mean parameter estimate maps, instead of single trial estimates, was made after observing that image decoding performance when decoding individual trials was close to chance, indicating a lack of sensitivity to detect potential differences between expected and unexpected image pairs. This decision was based on an independent MVPA collapsed over expected and unexpected image pairs, without inspection of the contrast of interest. Expected image pair trials are by definition more frequent, which may in turn yield a more accurate mean parameter estimate. Thus, stratification by random sampling was used to balance the number of expected and unexpected image pairs per trailing image, thereby removing potential bias. In short, for each iteration (n = 1,000) a subset of expected trials was randomly sampled to match the number of unexpected occurrences of that trailing image. Finally, decoding performance was analyzed in terms of mean decoding accuracy. To this end, the class with the highest probability for each test item was chosen as the predicted class and the proportion of correct predictions calculated. Mean decoding performances for expected and unexpected image pairs were subjected to a two-sided, one sample t-test against chance decoding performance (chance level = 12.5%). If decoding was above chance for the expected and unexpected image pairs, decoding performances between expected and unexpected pairs were compared by means of a paired t-test and the effect size was calculated. In short, the classifier was used to distinguish between the eight trailing images, after being trained on the single-trial parameter estimates from the functional localizer. The performance of the classifier was tested on the per-image parameter estimates from the main task split into the expected and unexpected condition.

**Image preference analysis.** For the voxel-wise image preference analysis the single trial GLM parameter estimate maps, as outlined in the *MVPA* section above, were utilized. Within each participant the parameter estimate maps of the functional localizer were averaged for each trailing image, thus yielding an average activation map induced by each trailing image in an expectation free, neutral context. The same was done for the main task data, but for expected and unexpected occurrence of each trailing image separately. Then, for each voxel, trailing images were ranked according to the response they elicited during the functional localizer. These rankings were applied to the main task data, resulting in a vector per voxel, consisting of the mean activation (parameter estimate) elicited by the trailing images during the main task, ranked from the least to most preferred image based on the context neutral, independent functional localizer data. This was done separately for expected and unexpected occurrence of each trailing image. Within each ROI the mean parameter estimates of expected and unexpected image pairs per preference rank was calculated. For each ROI linear regressions were fit to the ranked parameter estimates, one for expected and one for unexpected pairs. A positive regression slope would thus indicate that the ranking from the functional localizer generalized to the main task, which was considered a prerequisite for any further analysis. This was tested by subjecting the slope parameters across subjects to a two-tailed one sample t-test, comparing the obtained slopes against zero. Furthermore, this analysis assumes a linear relation between the response parameter estimates and preference rank. Of note, a strong non-linear relationship, in either of the expectation conditions, could pose a problem for the interpretation of the resulting slope parameter. Therefore, we tested for linearity, by comparing the model fit between the linear model and a second order polynomial model. The data was deemed sufficiently linear, if the fit of the linear model was superior to the fit of the non-linear model as index by a smaller Bayesian information criterion (BIC; [91]). If these requirements were met for the expected and unexpected conditions, the difference between slope parameters was compared by a two-tailed paired t-test. If the amount of expectation suppression (i.e. unexpected minus expected) indeed scales with image preference (i.e. dampening), then we should find the slope parameter for the unexpected condition regression line to be significantly larger than for the expected condition. The opposite prediction, a larger slope parameter for the expected condition, is made by the sharpening account. For this comparison the effect size was also calculated in terms of Cohen's $d_z$.

The rationale of this analysis is that a dampening mechanism suppresses responses in highly active neurons (i.e. those neurons which are tuned towards the expected feature) more than in less active neurons (those which are tuned away). Thus, when responses within a voxel are strong to a particular image more neurons can be suppressed by dampening than when a less preferred image is shown. Since a neural sharpening mechanism, opposite to dampening, would particularly suppress less active neurons compared to highly active ones, the reverse pattern would be evident under sharpening.

In addition to the ROI based approach, we also performed a whole brain version of the image preference analysis in order to provide an overview of where dampening or sharpening might be evident beyond our a priori defined ROIs. The analysis was identical to the ROI based approach, outlined above, except for that the amount of expectation suppression per voxel and preference rank was calculated in order to display results across the whole brain. Regressions were thus fit to expectation suppression as function of image preference rank for each voxel and subject. The fit was constrained to voxel in which the response to expected and unexpected stimuli showed a significant positive slope with preference rank, thereby indicating that the image preference ranking generalized from the localizer to the main task. Unlike in the ROI based approach, the data was spatially smoothed using a Gaussian kernel with full-width at half-maximum of 8 mm. The slope parameters across subjects were tested against zero in each voxel. Since in this analysis expectation suppression was expressed as a function of image preference rank, from least to most preferred, positive slopes indicate support for dampening, while negative slopes are evidence for sharpening.

**Bayesian analyses.** In order to assess whether any non-significant results constituted a likely absence of an effect, or rather indicated a lack of statistical power to detect possible differences, corresponding Bayesian tests were performed. All Bayesian analyses were carried out in JASP ([92], RRID:SCR_015823) using default settings; i.e. paired and one-sample t-tests used a Cauchy prior width of 0.707 and repeated measures ANOVAs used a prior with the following settings: *r scale fixed effects* = 0.5, *r scale random effects* = 1, *r scale covariates* = 0.354. The number of samples of the RM ANOVA was increased to 100,000 and Bayes Factors for the inclusion of the respective factors are reported ($BF_{inclusion}$), which yields the evidence for the inclusion of that factor averaged over all models in which the factor is included [93]. Interpretations of the resulting Bayes Factors are based on the classification by Lee and Wagenmakers [94].

**Region of interest definition.** The two a-priori regions of interest, object-selective LOC and V1, were defined per subject based on data that was independent from the main task. In order to obtain object-selective LOC, GLMs were fit to the functional localizer data of each subject, modelling object image and scrambled image events separately with a duration corresponding to their display duration. First-order temporal derivatives, instruction and performance summary screens, as well as motion regressors were added as nuisance regressors. The contrast, object images minus scrambles, thresholded at $z > 5$ (uncorrected; i.e. $p < 1e\text{-}5$), was utilized to select regions per subject selectively more activated by intact object images compared to scrambles [95,96]. The threshold was lowered on a per subject basis, if the LOC mask contained less than 300 voxels in native volume space. The individual functional masks were constrained to anatomical LOC using an anatomical LOC mask obtained from the Harvard-Oxford cortical atlas (RRID:SCR_001476), as distributed with FSL. Finally, a decoding analysis of object images (also see: *Multi-voxel pattern analysis*) was performed using a searchlight approach (6 mm radius) on the functional localizer data, using a k-fold cross-validation scheme with four folds. This MVPA yielded a whole brain map of object image decoding performance, based on which the 200 most informative LOC voxels (in native volume space) in terms of image identity information were selected from the previously established LOC masks. This was done to ensure that the final masks contain voxels which best discriminate between the different object images. Freesurfer 6.0 ('recon-all'; [97], RRID:SCR_001847) was utilized to extract V1 labels (left and right) per subject based on their anatomical image. Subsequently, the obtained labels were transformed back to native space using 'mri_label2vol' and combined into a bilateral V1 mask. The same searchlight approach mentioned above was used to constrain the anatomical V1 masks to the 200 most informative V1 voxels concerning object identity decoding. To verify that our results were not unique to the specific (but arbitrary) ROI size, we repeated all ROI analyses with ROI masks ranging from 50 to 300 voxels in steps of 50 voxels.

### Software

FSL 5.0.9 (FMRIB Software Library; Oxford, UK; www.fmrib.ox.ac.uk/fsl; [87], RRID:SCR_002823) was utilized for preprocessing and analysis of fMRI data. Additionally, custom Matlab (The MathWorks, Inc., Natick, Massachusetts, United States, RRID:SCR_001622) and Python (Python Software Foundation, RRID:SCR_008394) scripts were used for additional analyses, data extraction, statistical tests, and plotting of results. The following toolboxes were used: NumPy ([98], RRID:SCR_008633), SciPy ([99], RRID:SCR_008058), Matplotlib ([100], RRID:SCR_008624), PySurfer (https://pysurfer.github.io/, RRID:SCR_002524), Mayavi ([101], RRID:SCR_008335), and Scikit-learn ([90], RRID:SCR_002577). Whole-brain results are displayed using Slice Display [102] using a dual-coding data visualization approach [103], with color indicating the parameter estimates and opacity the associated z statistics. Additionally, PySurfer was used to display whole-brain results on an inflated cortex, with surface labels from the Desikan-Killiany atlas [104]. Bayesian analyses were performed using JASP 0.8.1.1 ([92], RRID:SCR_015823). Stimulus presentation was done using Presentation® software (version 18.3, Neurobehavioral Systems, Inc., Berkeley, CA, RRID:SCR_002521).

## Results

### Expectation suppression throughout the ventral visual stream

We first examined expectation suppression within our a priori defined ROIs, V1 and object-selective LOC. We observed a significantly larger BOLD response to unexpected compared to expected image pairs, both in V1 ($t_{(21)} = 3.20$, $p = 0.004$, Cohen's $d_z = 0.68$, Figure 2.2C) and object-selective LOC ($t_{(21)} = 5.03$, $p = 5.6e\text{-}5$, Cohen's $d_z = 1.07$, Figure

2.2C). To ensure that the results are not dependent on the (arbitrarily chosen) mask size of the ROIs, the analyses were repeated for ROIs of sizes between 50-300 voxels (691-4147mm$^3$); the direction and statistical significance of all effects was identical for all ROI sizes.

A whole-brain analysis, investigating effects of perceptual expectation across the brain, revealed an extended statistically significant cluster (Figure 2.2A, black contours) of expectation suppression across the ventral visual stream. As also evident in Figure 2.2B, cortical areas showing significant expectation suppression included large parts of bilateral object-selective LOC, bilateral fusiform gyrus, bilateral inferior parietal cortex and right posterior parahippocampal gyrus. Thus, there is substantial support for a wide-spread expectation suppression effect across the ventral visual stream.

Next, we assessed the neural effect of the conditional probability conditions within V1 and LOC. While this analysis confirmed a weaker response for expected items in V1 ($F_{(1,21)} = 6.39$, $p = 0.020$, $\eta^2 = 0.233$) and LOC ($F_{(1,21)} = 19.50$, $p = 2.4\text{e-}4$, $\eta^2 = 0.481$), there was no significant modulation by conditional probability, nor an interaction between conditional probability and expectation in either V1 (conditional probability: $F(2,42) = 2.02$, $p = 0.145$, $\eta^2 = 0.088$; interaction: $F_{(2,42)} = 1.19$, $p = 0.315$, $\eta^2 = 0.053$) or LOC (conditional probability: $F_{(2,42)} = 1.90$, $p = 0.162$, $\eta^2 = 0.083$; interaction: $F_{(2,42)} = 0.92$, $p = 0.407$, $\eta^2 = 0.042$). Bayesian analyses yielded very strong support for the effect of expectation in LOC ($\text{BF}_{\text{Incl.}} = 35.403$), but provided moderate evidence that conditional probability did not have an effect ($\text{BF}_{\text{Incl}}$. = 0.327), and neither did the interaction of expectation and conditional probability ($\text{BF}_{\text{Incl.}} = 0.290$). In V1 results remained inconclusive, since there was only anecdotal evidence against an effect of expectation ($\text{BF}_{\text{Incl.}} = 0.426$) and conditional probability ($\text{BF}_{\text{Incl.}} = 0.373$), but moderate evidence against an effect of the interaction ($\text{BF}_{\text{Incl.}} = 0.172$). Thus, since there is no evidence for an effect of the conditional probability manipulation, we collapse across the three different conditional probability conditions for all subsequent analyses.

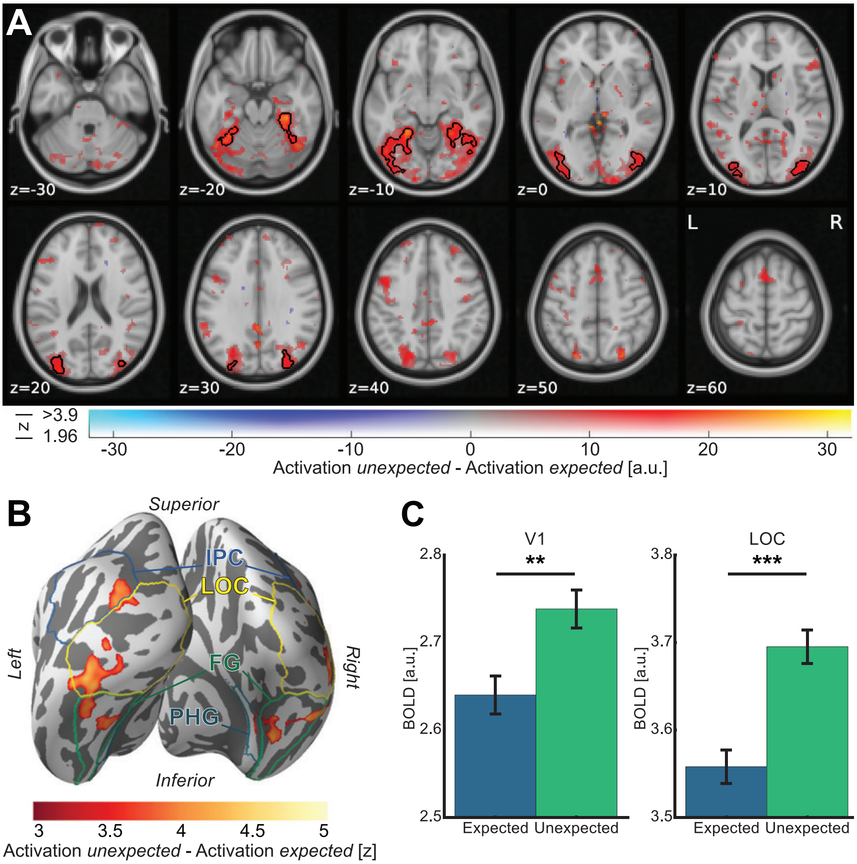


**FIGURE 2.2 Univariate fMRI results.**

(A) Expectation suppression throughout the ventral visual stream. Displayed are parameter estimates for unexpected image pairs minus expected pairs overlaid on the MNI152 2mm template. Color represents the parameter estimates, with red-yellow clusters indicating expectation suppression, and opacity depicting the associated z statistics. Black contours outline statistically significant clusters (GRF cluster corrected), which include significant expectation suppression in superior and inferior divisions of LOC, temporal occipital fusiform cortex, and posterior parahippocampal gyrus. (**B**) Expectation suppression displayed on an inflated cortex reconstruction. Z statistics of the expectation suppression contrast (cluster thresholded) are displayed. Visible are large clusters showing significant expectation suppression in LOC, fusiform gyrus (FG), inferior parietal cortex (IPC) and posterior parahippocampal gyrus (PHG). (**C**) Expectation suppression within V1 and object-selective LOC. Displayed are parameter estimates ± within-subject standard error for responses to expected and unexpected images pairs. In both ROIs, V1 (left bar plot) and LOC (right bar plot), BOLD responses to unexpected image pairs were significantly stronger than to expected image pairs. ** $p < .01$, *** $p < .001$.

## Perceptual expectations dampen sensory representation in LOC

To examine whether sharpening or dampening of sensory representations underlies the observed expectation suppression effect in V1 and LOC, an image preference analysis was conducted. In short, BOLD responses were regressed on image preference rank, with dampening predicting a steeper slope for unexpected compared expected images and sharpening predicting the opposite (see *Methods* for details). First, we tested whether the relation between voxel-level BOLD responses and image preference rank was better described by a linear model than a polynomial model. There was higher model evidence for linear compared to non-linear response profiles in both areas and conditions (V1, expected: $BIC_{linear} = 97.02 < BIC_{polynomial} = 97.25$; V1, unexpected: $BIC_{linear} = 95.85 < BIC_{polynomial} = 96.09$; LOC, expected: $BIC_{linear} = 94.82 < BIC_{polynomial} = 95.01$; LOC, unexpected: $BIC_{linear} = 94.99 < BIC_{polynomial} = 95.28$). Furthermore, results, depicted in Figure 2.3A, reveal positive slopes within V1 (expected: $t_{(21)} = 9.11$, $p = 9.6e\text{-}9$, Cohen's $d_z = 1.94$; V1 unexpected: $t_{(21)} = 9.90$, $p = 2.3e\text{-}9$, Cohen's $d_z = 2.11$), as well as in LOC (expected: $t_{(21)} = 3.39$, $p = 0.003$, Cohen's $d_z = 0.72$; LOC unexpected: $t_{(21)} = 7.14$, $p = 4.8e\text{-}7$, Cohen's $d_z = 1.52$), confirming that the image preference ranking from the functional localizer data generalized to the main task. This indicates a stable, reproducible sensory code and allows for an analysis of the difference in slopes between expected and unexpected image pairs. Crucially, image preference slopes were significantly steeper for unexpected than expected image pairs in LOC ($t_{(21)} = 2.18$, $p = 0.041$, Cohen's $d_z= 0.47$). This means that the amount of expectation suppression (i.e. the difference in the two regression lines in Figure 2.3A) increased with the image preference rank in object-selective LOC. A control analysis confirmed that the results were independent of the number of voxels in the ROI mask (mask sizes 50-300 voxels). There was no statistically significant difference in slopes between the expectation conditions in V1 ($t_{(21)} = 1.20$, $p = 0.242$, Cohen's $d_z = 0.26$), regardless of the number of voxels in the ROI mask (50-300 voxels). In order to explore whether there was evidence for the absence of dampening in V1 a Bayesian t-test was performed on the difference of the image preference slopes (unexpected vs. expected) in the V1 ROI. This analyses yielded a $BF_{10} < 1/3$ for all V1 ROI sizes, except for the 200 voxel mask ($BF_{10} = 0.423$). Together this suggests that there is moderate evidence for the absence of dampening in V1.

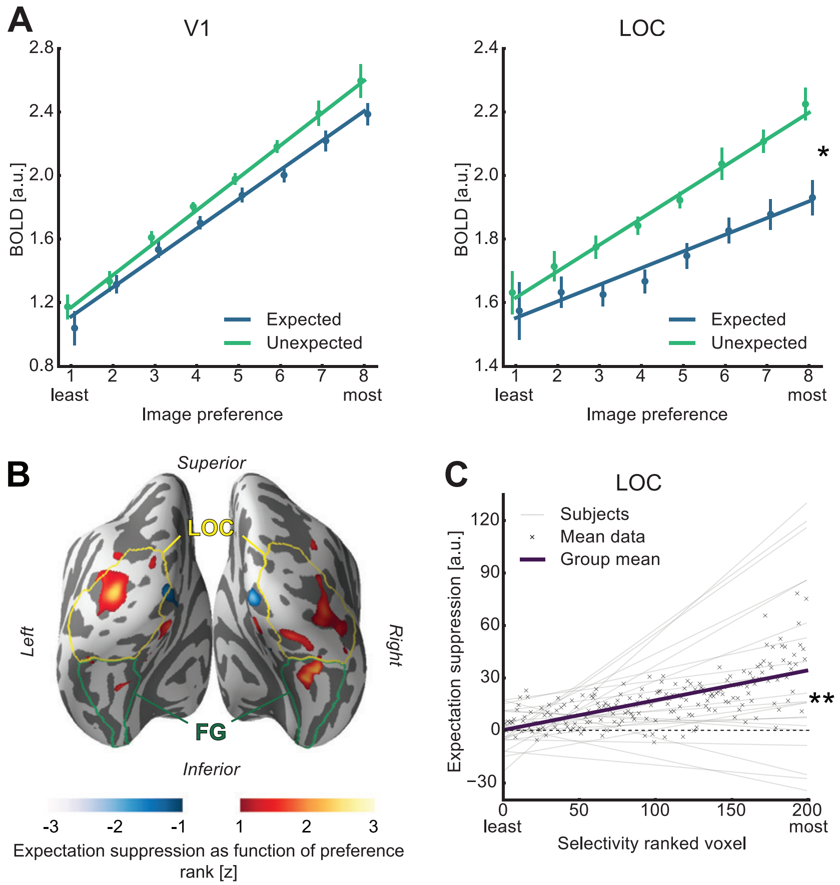


**FIGURE 2.3** **Results of image preference analysis.**

(**A**) Image preference analysis results in V1 and object-selective LOC. Parameter estimates ± within-subject standard error are displayed as a function of voxel-wise image preference, ranked from the least to the most preferred image rank based on the functional localizer. Superimposed is the mean regression line fit of the subject-wise regressions for expected and unexpected image pairs separately (see *Methods*). The left line plot shows responses to expected and unexpected image pairs within the V1 ROI. The fitted regression lines for expected and unexpected are parallel; i.e. no difference in slopes. The right plot displays image preference results for object-selective LOC, showing a steeper slope for the unexpected image pair regression line compared to the corresponding expected image pair regression line. (**B**) Image preference analysis results displayed on an inflated cortex reconstruction. Z statistic (uncorrected) of expectation suppression as function of image preference rank is shown in color, with red indicating more suppression for preferred stimuli (dampening) and blue indicating less suppression for preferred stimuli (sharpening). Visible are clusters showing a dampening effect largely in bilateral LOC and to a lesser extend in fusiform gyrus (FG). (**C**) Expectation suppression (unexpected – expected) as function of voxel selectivity, ranked from the least to the most selective voxels, in object-selective LOC. Displayed are the linear models per subject, the mean linear model (group mean), and the mean data for each selectivity ranked voxel. The amount of expectation suppression increases as a function of voxel selectivity. * $p < .05$, ** $p < .01$.

In order to provide an additional overview of the localization of the dampening effect beyond our a priori ROIs we performed a whole brain analysis of the image preference analysis. Results depicted in Figure 2.3B, using a liberal threshold, suggest clusters of dampening to be primarily located in LOC and to a lesser degree in fusiform gyrus.

After showing a dampening of representations in object-selective LOC, we further explored whether this dampening at the voxel level is likely to reflect neural dampening, as also evident in Meyer and Olson [23] and Kumar et al. [79]. A key problem is that under certain conditions a neural *sharpening* mechanism can produce voxel level dampening, as also suggested by Alink et al. [105] in the case of repetition suppression. Thus, we performed an additional analysis in which we analyzed expectation suppression (i.e. unexpected minus expected) as a function of voxel selectivity (i.e. slope of the response amplitude to preference ranked images). We reasoned that under a dampening account, selective voxels, showing strong responses to some, but weak responses other stimuli, are on average more likely to yield strong expectation suppression than low selectivity voxels. Sharpening on the other hand predicts the opposite pattern, since highly selective voxels should be less suppressed by sharpening, or even enhanced in their response, because more activated neurons are on average tuned towards the expected stimulus, compared to voxels with lower selectivity. For this analysis we first established a voxel selectivity ranking. The rank was based on the slope of activity regressed onto image preference during the localizer for each voxel. The rationale is that voxels which are more selective in their response yield a larger slope parameter, since the activity elicited by different images shows a larger difference than in voxels with low selectivity (i.e. those that respond similarly to different images). After obtaining the slope parameter of image preference per voxel, we ranked voxels by this slope coefficient, reflecting voxel selectivity during the localizer. Next we regressed expectation suppression during the main task onto voxel selectivity rank. As explained above, we reasoned that dampening predicts a positive slope for this regression, while sharpening would predict a negative slope. Results from LOC, depicted in Figure 2.3C, showed a significant positive slope of expectation suppression with voxel selectivity ($t_{(21)}$ = 3.00, $p = 0.007$, Cohen's $d_z = 0.64$), demonstrating that highly selective voxels are more suppressed by expectation than less selective ones. These effects were present for all LOC ROI mask sizes from 50-300 voxel. Thus, also the selectivity analysis provides evidence that neural responses are dampened by expectations in LOC. Results in V1 were inconclusive with no significant effect of voxel selectivity on expectation suppression ($t_{(21)} = 1.80$, $p = 0.086$, Cohen's $d_z = 0.38$) and only weak anecdotal evidence for the absence of an effect in the corresponding Bayesian t-test ($BF_{10} = 0.887$).

In another complementary analysis, we reasoned that if the reduced activity for expected items is associated with a reduction of noise (sharpening), it is expected to be associated with an increase in classification accuracy in a MVPA [18]. Conversely, a dampening of the representation is predicted to be associated with a decrease in classification accuracy for expected image pairs [79]. Generally, image identity could be classified well above chance (12.5%) in V1 (expected: 27.9%, $t_{(21)} = 10.89$, $p = 4.3e\text{-}10$, Cohen's $d_z = 2.32$; unexpected: 30.2%, $t_{(21)} = 15.70$, $p = 4.5e\text{-}13$, Cohen's $d_z = 3.35$), and LOC (expected: 18.5%, $t_{(21)} = 5.69$, $p = 1.2e\text{-}5$, Cohen's $d_z = 1.21$; unexpected: 19.5%, $t_{(21)} = 6.76$, $p = 1.1e\text{-}6$, Cohen's $d_z = 1.44$). While a trend towards better decoding performance for unexpected images was visible in both ROIs, in line with dampening of the sensory response, this difference was not statistically significant (V1: $t_{(21)} = 1.93$, $p = 0.067$, Cohen's $d_z = 0.41$; LOC: $t_{(21)} = 1.16$, $p = 0.260$, Cohen's $d_z = 0.25$). Bayesian t-tests of this difference also remained inconclusive in both ROIs (V1: $BF_{10} = 1.073$; LOC: $BF_{10} = 0.403$).

## Expectation facilitates image categorization

In order to assess whether concurrent to the described neural effects also behavioral benefits of expectation are evident, data from the categorization task was analyzed. Results demonstrate that participants categorized expected trailing images faster ($M$ = 524.4 ms, $SEM$ = 3.8 ms) than unexpected items ($M$ = 537.4 ms, $SEM$ = 3.8; $t_{(21)} = 2.40$, $p = 0.026$, Cohen's $d_z = 0.51$; Figure 2.4A). A similar, albeit not statistically significant trend ($t_{(21)} = 1.19$, $p = 0.247$, Cohen's $d_z = 0.25$) was visible in terms of error rates (Figure 2.4B). Analysis of the questionnaires showed that on average participants correctly identified 4.0 ± 2.3 (± SD) of the eight image pairs.

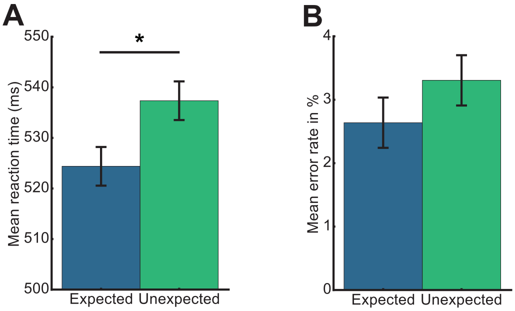


**FIGURE 2.4 Facilitation of behavioral responses by expectations.**

Behavioral data analysis from the categorization task indicates incidental learning of image transitions. Mean values ± within-subject standard error are shown. (A) Shows mean RT to expected and unexpected trailing images. RTs were significantly faster to expected trailing images compared to unexpected images. (B) Shows the corresponding mean error rates. * $p < .05$.

### Spatial extent of expectation suppression

In a post hoc analysis we investigated whether the expectation suppression effect in V1 and LOC was spatially unspecific, or constrained to regions activated by the object stimuli. The reasoning was that a spatially unspecific effect indicates that at least part of the observed expectation suppression may be due to arousal changes in response to unexpected compared to expected trailing images, while a constrained effect may point towards a spatially specific top-down modulation. To investigate this, the amount of expectation suppression was compared between voxels significantly activated by object stimuli and those that were not. The split into activated and not activated voxels was performed using data from the functional localizer, with activated voxels being defined as all voxels within anatomically defined V1 and LOC respectively, which exhibited a significant activation by object images ($z > 1.96$; i.e. $p < 0.05$, two-sided), while non-activated voxels were defined as voxels displaying no significant activation, nor deactivation ($-1.96 < z < 1.96$). ROI masks were constrained to gray matter voxels. In both ROIs, activated and non-activated voxels showed evidence of expectation suppression (V1, activated voxels: $t_{(21)} = 3.01$, $p = 0.007$, Cohen's $d_z = 0.64$; V1, non-activated voxels: $t_{(21)} = 2.17$, $p = 0.041$, Cohen's $d_z = 0.46$; LOC, activated voxels: $t_{(21)} = 4.11$, $p = 0.0005$, Cohen's $d_z = 0.88$; LOC, non-activated voxels: $t_{(21)} = 2.51$, $p = 0.021$, Cohen's $d_z = 0.53$). In LOC expectation suppression was significantly stronger in voxels that were activated by the stimuli than in non-activated voxels ($t_{(21)} = 2.20$, $p = 0.039$, Cohen's $d_z = 0.47$). However, in V1 this difference was not statistically significant ($t_{(21)} = 1.09$, $p = 0.286$, Cohen's $d_z = 0.23$). A Bayesian analysis of V1 data remained inconclusive, yielding only anecdotal evidence for the absence of a difference between activated and non-activated voxels ($BF_{10} = 0.379$).

## Discussion

We set out to investigate the neural effects of perceptual expectation and demonstrated that, after incidental learning of transitional probabilities of object images, expectation suppression is evident throughout the human ventral visual stream. Importantly, the amount of expectation suppression scaled positively with image preference and voxel selectivity in LOC, suggesting that dampened sensory representations underlie expectation suppression in object-selective areas, in line with results from monkey IT [23,79].

## Dampening of sensory representation in object-selective cortex

The suppression of expected stimuli, evident throughout the ventral visual stream in the present study, extends and supports previous research showing expectation suppression in early visual areas [18,78,106] and monkey IT [23,26]. The observed suppression may constitute an efficient and adaptive processing strategy, which filters out predictable, irrelevant objects from the environment. Conversely, the stronger response to unexpected objects may serve to render unexpected stimuli more salient. This surprise response to unexpected stimuli may draw attention towards these stimuli, as also reasoned by Meyer and Olson [23]. Such capture of attention is adaptive since unexpected events may provide particularly relevant information. It is important to note that the utilized paradigm did not manipulate attention towards expected or unexpected stimuli in a top-down fashion. In fact, unexpected and expected stimuli were only distinguishable by the context in which they occurred. Therefore, if unexpected stimuli do indeed automatically capture attention [37,107], then any attentional modulation must temporally follow the expectation effect, and not vice versa.

Given the absence of a neutral condition, we cannot differentiate whether the observed expectation suppression effect constitutes a suppressed response for expected stimuli, or an enhanced response to unexpected ones, or both. While there is evidence for both, expectation suppression and surprise enhancement [26,108], the present data cannot speak to this issue, but only concerns the relative difference between expected and unexpected stimuli.

We showed that the amount of expectation suppression scales with image preference in object-selective LOC, as also demonstrated in monkey IT [23]. Scaling indicates that expectation suppression in object-selective areas does *not* merely signal an unspecific surprise response, but rather that sensory representations are dampened by expectations, since the neural population most responsive to the expected stimulus is also most suppressed. Accordingly, we also demonstrated that expectation suppression scales positively with voxel selectivity. This result further supports the dampening account of expectation, since selective voxels contain more highly responsive neurons, tuned towards the expected stimulus features, which are also most suppressed by dampening. While there are some scenarios in which neural sharpening could account for some of the results presented here in isolation, the joint set of observations can only be accounted for by a dampening process at the neural level. Thus, our results lend support to the notion that neural responses are dampened by expectations in object selective LOC. Functionally, a dampening of sensory representations is in line with an adaptive mechanism, which filters out behaviorally irrelevant, predictable objects from the environment.

If expectation suppression, and the underlying representational dampening in LOC, represents an adaptive neural strategy, one might expect behavioral benefits to correlate with the neural effects. Although we observed behavioral benefits for expected stimuli during the categorization task, the present study cannot answer whether expectation suppression is associated with behavioral benefits, since during the fMRI task, and central to the interpretation above, expectations were task irrelevant. Task relevant predictions, necessary in order to investigate this question, may in turn change the underlying neural dynamics. In fact, it has been suggested that, at least in early visual areas, attention can reverse the suppressive effect of expectation [32].

While we did observe expectation suppression in V1, we did not find conclusive evidence for, or against, dampening or sharpening. These results cannot be explained by the absence of image preference in V1 for the utilized stimuli, as the preference ranking itself was reliable. Since a stimulus unspecific suppression was evident in V1, it is possible that object specific expectations were resolved at a higher level in the cortical hierarchy and only the results of the prediction (expected or unexpected) was relayed to V1 as feedback. Alternatively, a dampening effect may exist in V1, albeit of a smaller magnitude than in LOC, yielding an effect below detection threshold for the present study. Suppression in V1 may also have arisen due to spatially unspecific effects across V1, such as arousal changes, after the resolution of expectations in higher cortical areas. This interpretation is supported by the fact that expectation suppression was not significantly larger in stimulus-driven than non-stimulus-driven voxels.

Finally, the present results are at odds with a previous study that observed a sharpening of the sensory population response in V1 by expectation [18]. While we did not find evidence for a sharpening of responses in V1, we did observe dampening in LOC, in line with studies of monkey electrophysiology [23,79]. Thus, our data shows that the disagreement in previous studies, suggesting sharpening in human V1 [18] and dampening in monkey IT [23,79] are unlikely to be caused by differences between species or recording methods. We briefly discuss three factors that may account for the opposite results. First, Kok et al. [18] and the present study employed different stimuli (grating vs. object stimuli), tailored to investigate the population response in different areas of the visual hierarchy (V1 vs. LOC). Given that we did not find evidence for sharpening in V1, the opposite results cannot be explained by a general difference between the sensory areas, but rather an interaction between stimulus type and sensory area. Second, we induced expectations by prolonged exposure prior to scanning, while in Kok et al. [18] expectations were learnt and updated during the experiment. Interestingly, while expectation suppression has been shown in monkey

IT when expectations were induced by long-term exposure [23,26], this effect was not found when expectations were induced during the experiment [109,110]. Finally, there are differences between the studies in task demands. In the current study, we examined neural activity elicited by expected and unexpected *non-target* stimuli, i.e. stimuli that did not require a response by the observer. On the other hand, all stimuli in Kok et al. [18] were *target* stimuli, requiring a discrimination judgment by the observers. Given that attentional selection is known to sharpen stimulus representations [111], this difference in task setup could explain the opposite results.

## Prediction errors and predictive coding

Within a hierarchical predictive coding framework, prior expectations about an upcoming stimulus act as top-down signals predicting the bottom-up input based on generative models of the agent [12]. These predictions are then compared to the actual bottom-up input resulting in a mismatch signal, the prediction error (PE). Expectation suppression, as evident in the present data, and previously observed by others (e.g. [18,43,112]), matches the properties of a PE signal. That is, the ensuing PE is smaller for expected compared to unexpected trailing images, since the mismatch between prediction and input is smaller, thus resulting in expectation suppression, as evident here throughout the ventral visual stream. Furthermore, a dampening of object representations in LOC can well be explained within predictive coding as a result of the stronger and prolonged resolution of prediction errors elicited by unexpected images.

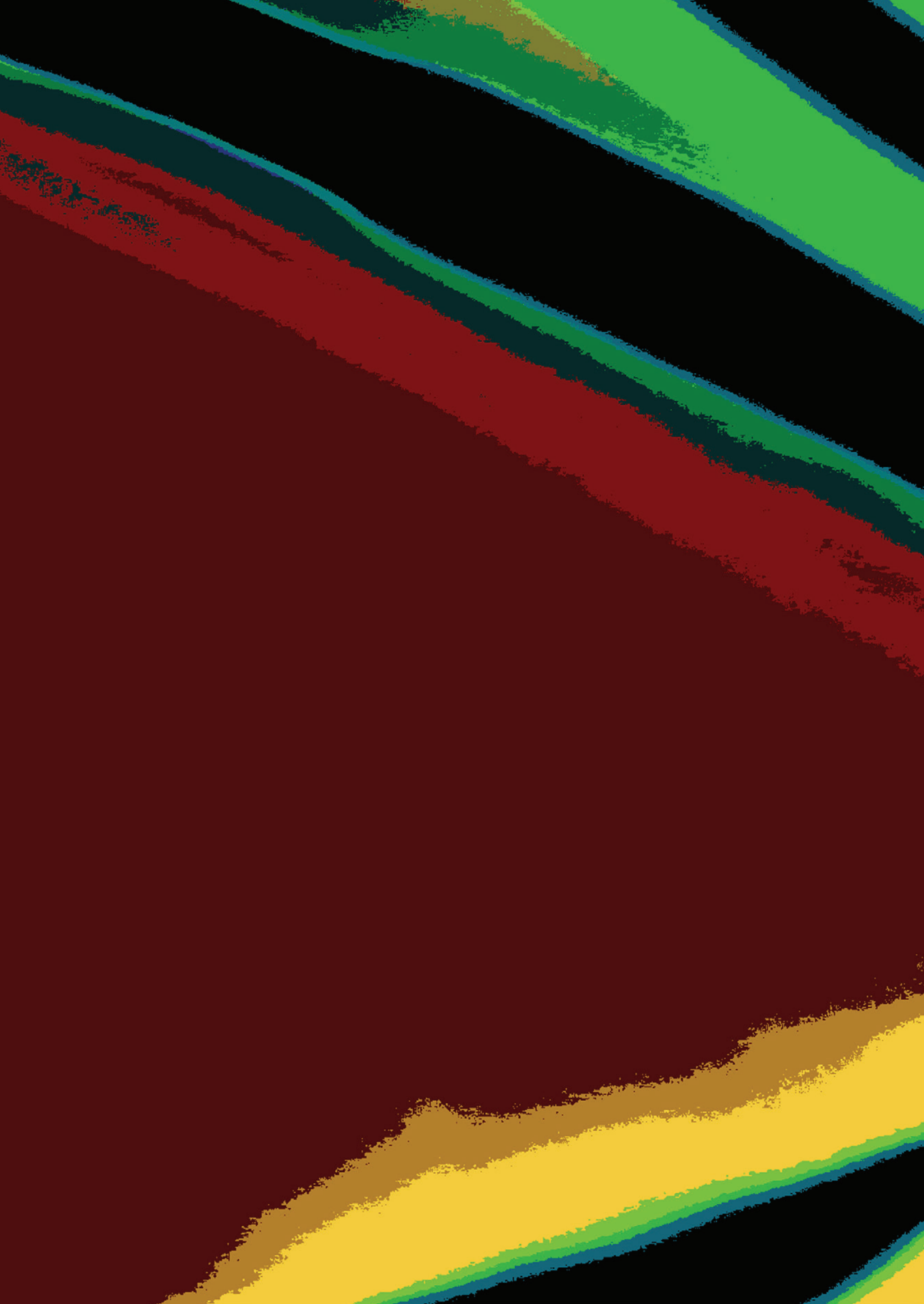
Alternatively, our results could partially be explained by changes in arousal, potentially reflecting globally enhanced responses following surprising stimuli. This explains why expectation suppression in V1 was spatially unspecific and some suppression was evident in non-stimulus driven voxels in LOC. However, such unspecific upregulation of activity cannot readily account for the stimulus- and spatially specific response modulations in LOC, while predictive coding explains these effects well.

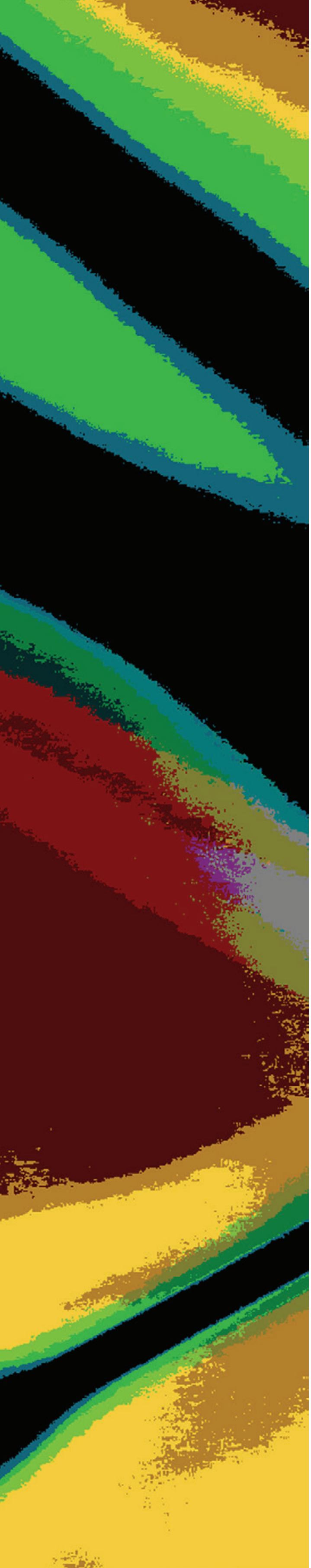
## No systematic modulation of expectation suppression by conditional probability

The present results do not provide evidence for a systematic modulation of expectation suppression by conditional probabilities. This is somewhat surprising given that a modulation has been demonstrated in monkey TE [27]. Furthermore, it is only by virtue of the difference in conditional probability that a trailing image can be considered expected or unexpected. Thus, by its nature expectation suppression should be sensitivity to conditional probability. We believe that this null result may be due to a lack of sensitivity of the associated analysis. The complexity of the transition matrix and the relatively small difference in conditional probability between the conditions, as well as the split of the available data into the three conditions may have all led to a reduction in sensitivity. Thus, to further elucidate the nature of expectation suppression future research in humans is required, possibly utilizing simplified paradigms or extended exposure to the image transitions.

## Conclusion

Taken together, our results demonstrate that expectation suppression is a widespread neural mechanism of perceptual expectation in the ventral visual stream, which increases with image preference and voxel selectivity. Perceptual expectations thus lead to a dampening of sensory representations in object-selective cortex, possibly supporting our ability to filter out irrelevant, predictable objects.





CHAPTER 3

# Statistical learning attenuates visual activity only for attended stimuli

*This chapter has been published as:*
Richter, D., de Lange, F.P. (2019) Statistical learning attenuates visual activity only for attended stimuli. *eLife.* 1–27. doi.org/10.7554/eLife.47869

# Abstract

Perception and behavior can be guided by predictions, which are often based on learned statistical regularities. Neural responses to expected stimuli are frequently found to be attenuated after statistical learning. However, whether this sensory attenuation following statistical learning occurs automatically or depends on attention remains unknown. In the present fMRI study, we exposed human volunteers to sequentially presented object stimuli, in which the first object predicted the identity of the second object. We observed a reliable attenuation of neural activity for expected compared to unexpected stimuli in the ventral visual stream. Crucially, this sensory attenuation was only apparent when stimuli were attended, and vanished when attention was directed away from the predictable objects. These results put important constraints on neurocomputational theories that cast perception as a process of probabilistic integration of prior knowledge and sensory information.

# Introduction

Previous experience constitutes a valuable source of information to guide perception and behavior. Extracting statistical regularities from past input in the environment to form expectations about the future has been shown to improve behavior in myriad ways [51,58,59]. Indeed, the acquisition of statistical regularities is thought to occur automatically [31] and affects behavior even in the absence of an intention to learn, or an awareness of, the regularities [50,75]. Given the significant behavioral and perceptual relevance of expectations, it is perhaps not surprising that the brain shows a remarkable sensitivity to statistical regularities. Many studies documented attenuated neural responses for expected compared to unexpected object stimuli in ventral visual regions subserving object recognition, both in terms of single unit spiking activity in monkeys [23,26] and in terms of non-invasively measured BOLD activity in humans ([24,25,113]; for a review see: [19]). This reduced response to expected stimuli has frequently been interpreted, within a predictive processing framework [11,12,114], as signifying a reduction of prediction errors elicited by the stimulus when sensory input matches prior expectations. However, it remains largely unknown whether this sensory attenuation process to predicted visual stimuli is automatic, as its relation to statistical learning may suggest, or only apparent when the predictable stimuli are attended.

Indeed, research on visual statistical learning in monkeys has typically not manipulated attention, but only required monkeys to passively fixate in order to obtain reward [23,26], thereby precluding conclusions pertaining to the dependence of these predictive processes on attention. Many studies in humans, providing evidence for suppressed responses to expected stimuli, did require participants to attend the predictable stimuli (e.g., [18,24,25,113]). On the other hand, den Ouden et al. [61] demonstrated attenuated responses to task-irrelevant expected stimuli, suggesting the possibility that the sensory consequences of statistical learning may not depend on attention. Similarly, Kok et al. [18] showed that the sensory attenuation for grating stimuli with an expected orientation was independent of whether the orientation feature was attended or not. Importantly however, in both these studies the expected or unexpected stimulus was the only stimulus presented on the screen, so even though the stimuli were not relevant, attention was not effectively disengaged by other stimuli. Without competition, it is likely that even a task-irrelevant stimulus will receive some attention.

Thus, at present it remains unclear whether statistical learning automatically results in altered neural responses to expected compared to unexpected visual stimuli, or whether this process hinges on the stimuli being attended. In order to answer this question, we exposed participants to sequentially presented pairs of object images. The first image predicted the identity of the second image, thereby making an image expected depending on temporal context. We recorded responses to expected and unexpected object images using whole-brain fMRI while participants performed one of two tasks. Either participants categorized the predictable, second object image as (non-)electronic (rendering the object images attended), or they classified a concurrently shown character (letter or symbol), presented within the fixation dot, as (non-)letter (rendering the object images unattended).

In brief, our results demonstrate strong sensory attenuation for expected object images within the ventral visual stream. Crucially however, expectation suppression was only evident when objects were attended and vanished when participants attended the concurrently presented alphanumeric characters at fixation. This suggests that sensory attenuation induced by statistical learning is not the result of an automatic integration of prior knowledge with incoming information, but hinges on attention, thus constraining neurocomputational theories of perceptual inference.

## Results

We exposed participants to statistical regularities by presenting object image pairs in which the leading image predicted the identity of the trailing image. During a learning session, participants performed a detection task of unpredictable upside-down images. On the next day, in the MRI scanner, participants were shown the same object image pairs, however unexpected trailing images were also presented; i.e., images which were predicted by a different leading image. Crucially, participants either classified the trailing object as (non-)electronic, thus actively attending the predictable object, or classified a concurrently presented, but unpredictable, trailing character as (non-)letter, thus not attending the predictable object.

### Attention is a prerequisite for perceptual expectations

First, we investigated whether the sensory attenuation for expected object stimuli was equally present when participants attended the objects or not, focusing on our a priori defined ROIs (see Figure 3.1A): primary visual cortex (V1), object-selective lateral occipital complex (LOC), and temporal occipital fusiform cortex (TOFC). In all three regions, expectation suppression was robustly present when participants attended the objects (V1: $t_{(33)} = 3.573$, $p = 0.001$, $d_z = 0.613$; LOC: $t_{(33)} = 3.860$, $p = 5.0\text{e-}4$, $d_z = 0.662$; TOFC: $t_{(33)} = 5.133$, $p = 1.2\text{e-}5$, $d_z = 0.880$), but absent when participants attended the characters at fixation; i.e., when the predictable objects were unattended

(V1: $t_{(33)} = -0.216$, $p = 0.830$, $d_z = -0.037$; LOC: $t_{(33)} = -0.831$, $p = 0.412$, $d_z = -0.143$; TOFC: $t_{(33)} = 0.072$, $p = 0.943$, $d_z = 0.012$). Indeed, Bayesian analyses showed moderate support for the null hypothesis ($BF_{10} < 1/3$) of no expectation suppression in all three regions during the character categorization task (V1: $BF_{10} = 0.188$; LOC: $BF_{10} = 0.253$; TOFC: $BF_{10} = 0.184$). The robustness of this distinct pattern of expectation suppression for the two conditions was statistically confirmed by an interaction analysis (expectation by attention interaction, V1:, $F_{(1,33)} = 7.706$, $p = 0.009$, $\eta^2 = 0.189$; LOC: $F_{(1,33)} = 12.580$, p = 0.001, $\eta^2 = 0.276$; TOFC: $F_{(1,33)} = 16.955$, $p = 2.4e\text{-}4$, $\eta^2 = 0.339$).

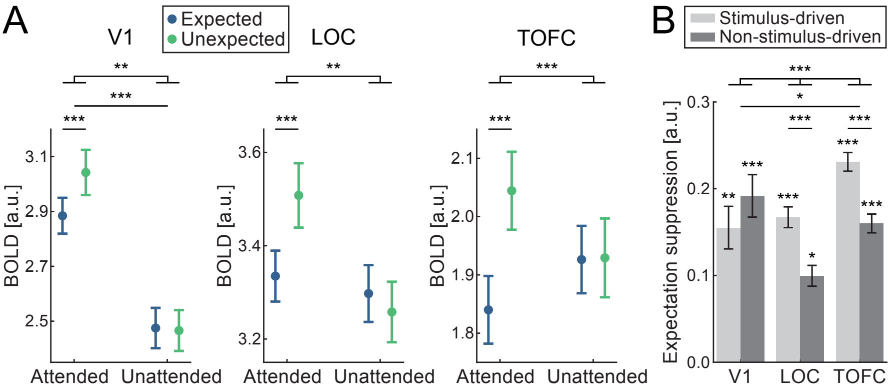


**FIGURE 3.1 Expectation suppression within the ventral visual stream depends on attention.**

(**A**) Displayed are parameter estimates +/- within-subject SE for responses to expected (blue) and unexpected (green) object stimuli during the objects attended task (attended) and objects unattended task (unattended). In all three ROIs, V1 (left), LOC (middle), and TOFC (right) BOLD responses were significantly suppressed in response to expected stimuli during the objects attended task. No difference was found between BOLD responses to expected and unexpected stimuli during the objects unattended task. The interaction effect between expectation and attention condition was significant in all three ROIs. (**B**) Expectation suppression in primary visual cortex is stimulus unspecific, and specific only in higher visual areas. Displayed is the average expectation suppression effect (BOLD responses, unexpected minus expected) split into stimulus-driven (light gray) and non-stimulus-driven (dark gray) gray matter voxels. Data are shown for the three ROIs, V1 (left bars), LOC (middle bars), and TOFC (right bars). Expectation suppression in LOC and TOFC was significantly larger for stimulus-driven than non-stimulus-driven voxels, while no such difference was evident in V1, indicating that expectation suppression in V1 was stimulus unspecific. Error bars indicate within-subject SE. Note, that the ROI masks in panel A and B differ, for details see: *ROI definition* and *Stimulus specificity analysis* in the *Materials and Methods* section. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Thus, in V1, LOC, and TOFC, there was a significant suppression of BOLD responses for expected compared to unexpected object stimuli exclusively during the object categorization task. No such modulation of BOLD responses by expectation was observed in the objects unattended condition in any of the three a priori ROIs, and in fact, there was moderate evidence for the absence of such a modulation when objects were unattended. We repeated all ROI analyses within the same ROIs but with different ROI sizes in order to ensure that our results were not dependent on the a priori but arbitrarily defined ROI mask size. Results were highly similar (i.e., the same effects showing statistically significant results) to those mentioned above within all three ROIs (V1, LOC, TOFC) for all tested ROI sizes, ranging from 100 to 400 voxels (800 $mm^3$ - 3200 $mm^3$) in steps of 100 voxels. Thus, our results do not depend on the exact ROI size but represent responses within the respective areas well.

We also examined how expectation modulated neural activity outside our predefined ROIs by performing a whole-brain analysis. Results of this whole brain analysis are illustrated in Figure 3.2A. The upper row in Figure 3.2A shows extensive clusters of expectation suppression throughout the ventral visual stream when objects were attended, but no difference when the objects were unattended (middle row), leading to a significant interaction (bottom row). These results complement our ROI-based analysis by showing that the observed expectation suppression effect is not unique to the a priori defined ROIs but evident throughout the ventral visual stream.

Outside the ventral visual stream, additional clusters of expectation suppression are evident in anterior insula and the frontal operculum, the precentral and inferior frontal gyrus, superior frontal gyrus and supplementary motor cortex, superior parietal lobule, as well as parts of the cerebellum. All significant clusters are summarized in a table in Supplementary File 3.1. Again, all these non-sensory clusters showed reduced activity for expected objects only when the object stimuli were attended and categorized. There was no significant modulation of activity by expectation anywhere in the whole brain analysis when the objects were unattended.

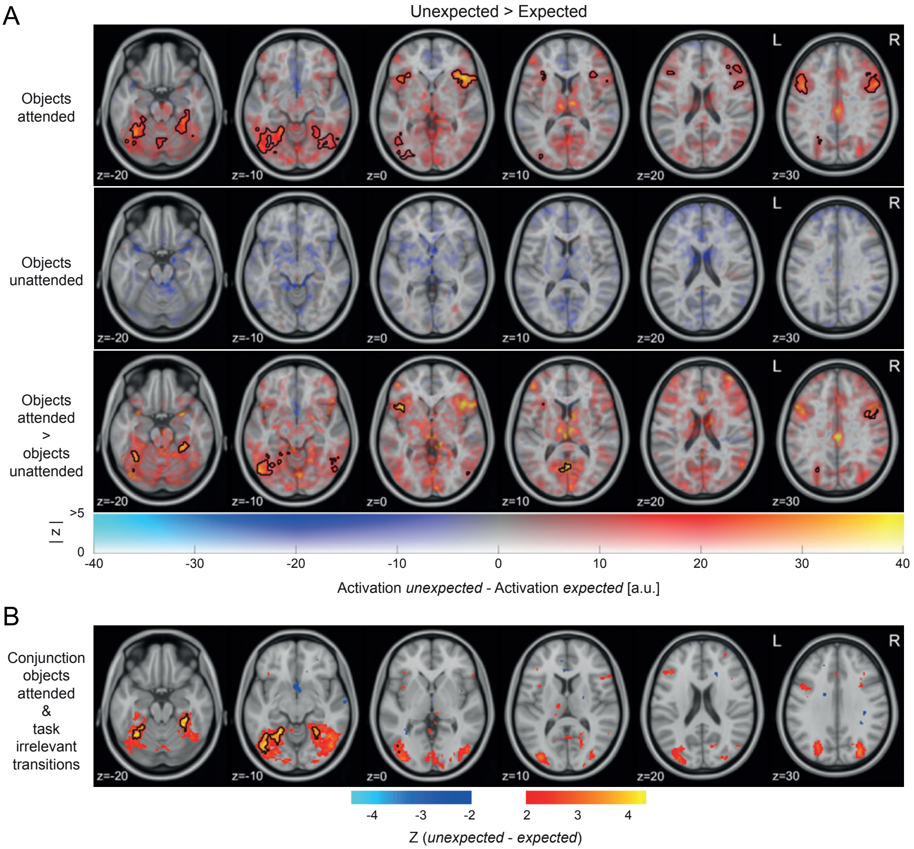

A
Unexpected > Expected
Objects attended
Objects unattended
Objects attended > objects unattended
L
R
z=-20
z=-10
z=0
z=10
z=20
z=30
|z|
>5
0
-40
-30
-20
-10
0
10
20
30
40
Activation *unexpected* - Activation *expected* [a.u.]
B
Conjunction objects attended & task irrelevant transitions
-4
-3
-2
2
3
4
Z (*unexpected* - *expected*)

**FIGURE 3.2 Expectation suppression across cortex for attended object stimuli only.**

(A) Widespread expectation suppression across cortex in the objects attended condition. Displayed are parameter estimates for unexpected minus expected image pairs overlaid onto the MNI152 2mm template. Color indicates unthresholded parameter estimates: red-yellow clusters represent expectation suppression. Opacity represents the z statistics of the contrasts. Black contours outline statistically significant clusters (GRF cluster corrected). Significant clusters included major parts of the ventral visual stream (early visual cortex, LOC, TOFC), anterior insula, and inferior frontal gyrus during the objects attended condition (upper row). No significant clusters were evident in the objects unattended condition (middle row). The interaction (attended > unattended; bottom row) showed significant clusters similar to those of the attended condition, albeit less extensive. (B) Expectation suppression across the ventral visual stream for attended objects, but with task-irrelevant predictions. Displayed are z statistics of the contrast unexpected minus expected of the conjunction: attended *task-relevant predictions* ∪ *task-irrelevant predictions*; data of task-irrelevant predictions from: [113]. Exclusively the ventral visual stream clusters showed significant expectation suppression in this conjunction, while all non-sensory area clusters were no longer significant. Thus, only the ventral visual stream clusters displayed a sensitivity to conditional probabilities, irrespective of whether predictions were task-relevant or task-irrelevant, as long as the predictable stimuli were attended.

## Expectation suppression requires attention to the stimuli, but not their predictable relationship

During the object categorization task, the ability to form expectations about the trailing object stimulus was helpful for the participants, and indeed expected object stimuli were categorized more quickly and accurately (see Figure 3.5A and *Expectations facilitate object classification*). This begs the question whether the expectation suppression effect that we observed throughout multiple brain areas during the object categorization task reflects differences in task engagement. Participants had an incentive to (implicitly or explicitly) use their knowledge of the predictable relationship between the leading and trailing image to prepare their object categorization response. In order to examine which brain regions exhibited expectation suppression irrespective of the relevance of the predictable relationship between stimuli, we performed a conjunction analysis that highlighted regions that showed significant expectation suppression both in the current study (during the object categorization task) and in a similar study that we published previously [113]. During this latter study, participants also attended the object stimuli, but were asked to press a button whenever an object appeared that was flipped upside-down. Upside-down images occurred rarely, and importantly, were not related to the (implicitly learned) statistical regularities. Figure 3.2B shows the whole-brain results of this conjunction analysis. Significant, bilateral clusters of expectation suppression were evident throughout most of the ventral visual stream. However, none of the non-sensory clusters showed significant expectation suppression during both experiments. Thus, only in the ventral visual stream we found strong and robust evidence for expectation suppression, regardless of whether the predictable relationship was task-relevant or task-irrelevant, as long as the predictable object pairs were attended.

## Stimulus specificity of the neural modulation by expectation

Next, we investigated the stimulus specificity of expectation suppression. Stimulus specificity concerns the question whether only stimulus-driven voxels or also voxels that were not (strongly) driven by the object stimuli displayed expectation suppression. The rationale was that an unspecific suppression effect (i.e., expectation suppression that is also evident in not stimulus-driven voxels) may result from global non-sensory effects, such as changes in general arousal or global surprise signals. On the other hand, stimulus specific suppression effects, being limited to stimulus-driven voxels, are rather suggestive of a more specific suppression mechanism that selectively operates on the neural populations that represent the expected stimulus; e.g., the dampening of stimulus-specific prediction errors as a result of a match between prediction and input.

All three ROIs were split into two populations of gray matter voxels, according to their stimulus responsiveness (stimulus-driven: responding to the object images; not stimulus-driven: not significantly responding to the object images), using independent data from the localizer run. There were strong differences between the ROIs in terms of the stimulus specificity of expectation suppression (Figure 3.1B; ROI x drive interaction: $F_{(1.245,41.080)} = 7.651$, $p = 0.005$, $\eta^2 = 0.188$). Whereas there was clear evidence for a larger expectation suppression effect in stimulus-driven than not stimulus-driven voxels in higher visual areas (LOC: $t_{(33)} = 3.991$, $p = 3.4e\text{-}4$, $d_z = 0.684$; TOFC: $t_{(33)} = 4.654$, $p = 5.1e\text{-}5$, $d_z = 0.798$), suppression was not significantly different between stimulus-driven and not stimulus-driven voxels in V1 ($t_{(33)} = -1.057$, $p = 0.298$, $d_z = -0.181$). Indeed, a Bayesian analysis indicated moderate support for the absence of a difference between stimulus-driven and not stimulus-driven voxels in V1 ($BF_{10} = 0.307$). Of note, all sub-populations in all three ROIs showed significant expectation suppression (all $p < 0.05$), suggesting that there is a general suppression of activity for expected stimuli in visual cortex, irrespective of whether the visual cortical area is driven by the stimuli. However, in later visual cortical areas (LOC and TOFC) there was significantly more expectation suppression in neuronal subpopulations that were driven by the stimulus, implying a more selective suppression mechanism in these areas.

## Surprising stimuli elicit a larger pupil dilation

In view of the suggestion that a global, stimulus unspecific response modulation may partially account for expectation suppression, we performed an exploratory analysis to examine whether surprising stimuli were associated with a stronger pupil dilation in our task. Pupil responses have been with linked with changes in arousal [115,116], which in turn may account for the stimulus unspecific suppression component. Moreover, pupil dilation scales with surprise [117–119]. Thus, this account would predict enhanced pupil dilation to unexpected compared to expected stimuli when objects were attended.

There was indeed a larger pupil diameter for unexpected compared to expected trailing images during the objects attended task (Figure 3.3, left). This difference emerged gradually starting ~600 ms after the onset of the trailing object image, and was significant between 1.5-2.8 seconds, as assessed with a cluster permutation test ($p_{\text{cluster}} = 0.017$). When objects were unattended, no significant difference in pupil diameter was found between the expectation conditions, and in fact, no timepoint surpassed the cluster formation threshold (i.e., all timepoints $p > 0.05$ uncorrected; Figure 3.3, right). However, the expectation induced difference in pupil diameter was not reliably different between attended and unattended stimuli ($p_{\text{cluster}} = 0.393$). Thus, the data showed that the pupil was significantly more dilated for unexpected than expected objects when the images were attended, mirroring the results of the neural data – albeit, without a reliable difference between attended and unattended stimuli. This tentatively suggests that the enhanced BOLD responses to unexpected stimuli might be partially accounted for by a global mechanism, such as increased arousal in response to surprising stimuli.

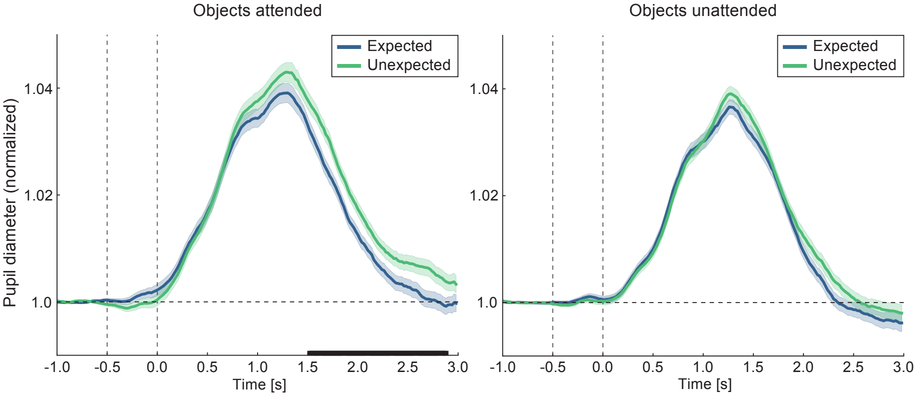


**FIGURE 3.3 Larger pupil dilations in response to unexpected compared to expected stimuli during the objects attended task.**

Displayed are pupil diameter traces over time, relative to trailing image onset. Pupil diameter data for expected (blue) and unexpected (green) image pairs are shown for the objects attended task (left) and objects unattended task (right). The black line on the abscissa denotes statistically significant differences in pupil dilations between expected and unexpected images (cluster permutation test, $p < 0.05$). In the objects attended condition significantly larger pupil dilations in response to unexpected images are evident between 1.52 to 2.88 seconds after trailing image onset (left). No significant difference is found in the objects unattended condition (right), nor in the interaction between conditions. The first vertical dashed line indicates leading image onset, the second vertical line trailing image onset. Shaded areas denote within-subject SE. Timepoints from -1.0 to -0.5 seconds served as baseline period.

## Expectation suppression and pupil dilations to surprising stimuli are associated

We explored whether expectation suppression and pupil dilation differences between unexpected and expected objects were associated. In other words, we sought for evidence of an association between the effect of expectations on pupil dilation and the expectation induced neural response attenuation. For this analysis we rank correlated expectation suppression magnitudes with pupil dilation differences for each participant. Results, displayed in Figure 3.4A, suggest that, when objects were attended, expectation suppression in V1 was more pronounced for trailing images that also resulted in larger pupil dilation differences ($t_{(31)} = 2.464$, $p = 0.019$, $d_z = 0.436$). This association was not reliable in LOC ($t_{(31)} = 1.413$, $p = 0.167$, $d_z = 0.250$; $BF_{10} = 0.466$) or TOFC ($t_{(31)} = 1.401$, $p = 0.171$, $d_z = 0.248$; $BF_{10} = 0.458$). There was no correlation of pupil dilation differences and expectation suppression when stimuli were unattended in any of the ROIs (V1: $t_{(31)} = -0.159$, $p = 0.875$, $d_z = -0.028$; $BF_{10} = 0.191$; LOC: $t_{(31)} = -0.125$, $p = 0.901$, $d_z = -0.022$; $BF_{10} = 0.190$; TOFC: $t_{(31)} = 0.177$, $p = 0.861$, $d_z = 0.031$; $BF_{10} = 0.192$). There was no significant overall difference in the correlation strength between attended and unattended stimuli ($F_{(1,31)} = 1.892$, $p = 0.179$, $\eta^2 = 0.058$), nor between ROIs ($F_{(1.558,48.293)} = 0.134$, $p = 0.823$, $\eta^2 = 0.004$), nor their interaction ($F_{(2,62)} = 0.482$, $p = 0.603$, $\eta^2 = 0.015$). Thus, when stimuli were attended there was evidence for an association of pupil dilation and expectation suppression in V1.

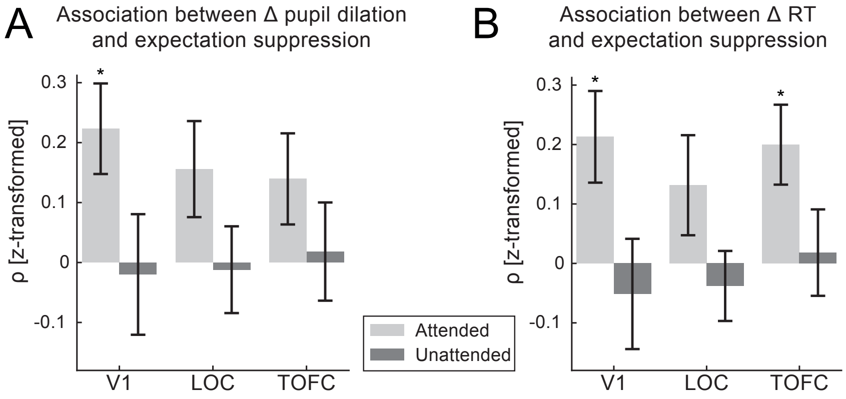


**FIGURE 3.4 Expectation suppression is associated with pupil dilation differences and behavioral benefits of expectations.**

(**A**) Correlation of expectation suppression magnitude and pupil dilation differences due to expectation. When predictable objects are attended, trailing images that induce larger pupil dilation differences are also showing larger expectation suppression magnitudes in V1. No such association is evident when objects are unattended. (**B**) Correlation of expectation suppression magnitude and RT benefits due to expectation. When predictable objects are attended, larger RT benefits are associated with larger expectation suppression effects in V1 and TOFC. This association is absent when objects are unattended. Error bars indicate within-subject SEM. * $p < 0.05$.

## Expectations facilitate object classification

In order to assess whether, concurrent with the neural effects of expectations, behavioral benefits of expectations were evident, we analyzed behavioral responses during MRI scanning in terms of reaction times (RTs) and response accuracy. Overall, the objects attended (classify electronic items) and objects unattended task (classify characters at fixation) showed very similar response accuracies (attended: 94.3% ± 5.4% vs. unattended: 94.0% ± 6.6%, mean ± SD) and only minor differences in RTs (attended: 574 ± 150 ms vs. unattended: 602 ± 131 ms, mean ± SD). This supports the notion that both tasks were of approximately equal difficulty.

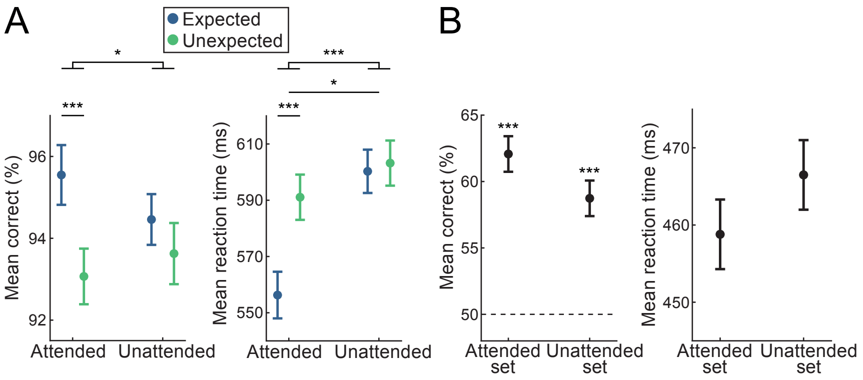


**FIGURE 3.5 Behavioral results demonstrate statistical learning.**

(**A**) Behavioral benefits of expectations demonstrate statistical learning. Displayed are mean accuracy (left) and mean reaction time (right) +/- within-subject SE. Responses to expected stimuli are significantly more accurate and faster, an effect exclusively observed during the objects attended condition. Thus, object identity expectations benefit behavioral performance during object classification and do not impact letter classification. (**B**) Pairs of both the objects attended image set and the objects unattended image set were classified significantly above chance, indicating a learning of the pairs for both conditions. Displayed are mean accuracy (left) and mean reaction time (right) during the post-scanning pair recognition task, +/- within-subject SE. The dashed line indicates chance level. During the pair recognition task, no differences in either classification accuracy (left) or response speed (right) were observed between pairs previously belonging to the objects attended task compared to the objects unattended task. * $p < 0.05$, *** $p < 0.001$.

During the object categorization task, participants could benefit from the foreknowledge of the identity of the trailing object image, as they were asked to categorize the trailing image. Such a benefit would however not be expected during the character categorization task, as the participants could fully ignore the object stimuli during this task. This is precisely what we observed, both in terms of accuracy and RTs (Figure 3.5A). During the object categorization task, participants were more accurate ($W = 457$, $p = 3.2e\text{-}4$, $r_B = 0.536$) and faster ($W = 9$, $p = 3.8e\text{-}9$, $r_B = -0.970$) for expected compared to unexpected trailing object stimuli. Conversely, during the character categorization task, no such benefit was observed in terms of accuracy ($t_{(33)} = 1.600$, $p = 0.119$, $d_z = 0.274$; $BF_{10} = 0.582$) or RT ($W = 252$, $p = 0.447$, $r_B = -0.153$; $BF_{10} = 0.273$). The robustness of this distinct pattern of behavioral advantage for expected stimuli for the two conditions was statistically confirmed by an interaction analysis (accuracy: $F_{(1,33)} = 5.203$, $p = 0.029$, $\eta^2 = 0.136$; RT: $F(1,33) = 37.543$, $p = 6.6e\text{-}7$, $\eta^2 = 0.532$).

## Neural and behavioral effects of expectations are associated

In order to explore whether the observed expectation suppression is associated with the behavioral benefits due to expectations, we correlated the magnitude of expectation suppression and the expectation induced RT benefits. Results, illustrated in Figure 3.4B, show that when the predictable objects were attended, behaviorally observed expectation RT benefits and neurally observed expectation suppression were associated in both, V1 ($t_{(33)} = 2.442$, $p = 0.020$, $d_z = 0.419$) and TOFC ($t_{(33)} = 2.236$, $p = 0.032$, $d_z = 0.384$), but no reliable correlation was found in LOC ($t_{(33)} = 1.384$, $p = 0.176$, $d_z = 0.237$, $BF_{10} = 0.439$). There was no association in any ROI when objects were unattended (V1: $t_{(33)} = -0.418$, $p = 0.679$, $d_z = -0.072$, $BF_{10} = 0.199$; LOC: $t_{(33)} = -0.374$, $p = 0.711$, $d_z = -0.064$, $BF_{10} = 0.196$; TOFC: $t_{(33)} = 0.179$, $p = 0.859$, $d_z = 0.031$, $BF_{10} = 0.186$). On average correlations were not reliably larger when objects were attended than when they were unattended (attention: $F_{(1,33)} = 2.920$, $p = 0.097$, $\eta^2 = 0.081$). The pattern of results was similar in all ROIs ($F_{(1.636,53.988)} = 0.615$, $p = 0.513$, $\eta^2 = 0.018$; interaction: $\mathrm{F}_{(1.461,48.203)} = 0.381$, $\mathrm{p} = 0.619$, $\eta^2 = 0.011$). Thus, there is some evidence that when the objects were attended, participants showed larger benefits (faster RTs) for expected trailing images for which they also showed larger magnitudes of expectation suppression in V1 and TOFC. These results suggest that the neural and behavioral effects of expectations are associated.

## No differences in association strength between attended and unattended object pairs

An alternative explanation for the absence of sensory attenuation for expected object stimuli during the character categorization task is that statistical regularities for the objects that are presented during this condition have simply not been learned. This explanation may be unlikely, because the vast majority of exposure to the expected pairs takes places in the learning session, during which the same task (upside-down image detection) was used for all image pairs. However, it is nonetheless important to ensure that statistical regularities were learned for the image pair sets of the object and the character categorization task. To empirically address this, we tested whether participants had explicit knowledge of the statistical regularities for all object pairs. During this post-scanning pair recognition task, participants were asked to indicate which one of two trailing images was more likely given the leading image. Participants indicated the correct trailing image with above chance accuracy for both, the set of object pairs that was previously attended (Figure 3.5B;

performance = 62.1% ± 1.8%, mean ± SE; $t_{(33)} = 6.803$, $p = 4.6e\text{-}8$, $d_z = 1.167$) and the set that was previously unattended (performance = 58.7% ± 2.2%; $t_{(33)} = 3.905$, $p = 2.2e\text{-}4$, $d_z = 0.670$). There was no statistically significant difference in accuracy on the pair recognition task between these sets of objects ($W = 365$, $p = 0.256$, $r_B = 0.227$; $BF_{10} = 0.737$). Reaction times were also similar for both sets of objects (objects previously attended: RT = 458.8 ± 25.4 ms; objects previously unattended: RT = 466.5 ± 25.9 ms; $t_{(33)} = -1.208$, $p = 0.236$, $d_z = -0.207$; $BF_{10} = 0.358$). Thus, the image pairs belonging to both task conditions (objects attended and unattended tasks) were reliably learned, most likely during the extensive behavioral training session, and there was no evidence for a significant difference in the learning of associations for the two sets of object pairs. This strongly suggests that the differences in sensory attenuation between the two attention conditions are unlikely to be explained by differences in the strength of the association between the object pairs.

## Visual processing continues in the absence of attention

Finally, one may wonder whether the lack of expectation suppression when objects were unattended is due to the fact that object stimuli simply did not elicit strong activity in the ventral visual stream, as they were not in the focus of attention. Although all three ROIs showed reliable above-baseline activity also when objects were unattended (Figure 3.1A), and activity in LOC and TOFC was of similar amplitude during both conditions, the overall activity level may partly represent stimulus-unrelated activity. Therefore, in an explorative analysis, we assessed the strength of stimulus-specific activity in our three ROIs, by means of a decoding analysis of the trailing images. In brief, a multi-class decoder was trained to differentiate between the six trailing images per attention condition. The classifier was trained on data obtained in an independent localizer run, during which participants performed a separate task (detection of dimming of fixation dot). Performance of this decoder was tested on the mean parameter estimates per trailing image for each of the two attention conditions of the main MRI task data. Because each task was comprised of six trailing images, chance performance was 16.7%. One-sample t-tests or Wilcoxon signed rank test (as applicable) showed that in each of the three ROIs (V1, LOC, TOFC) and tasks (objects attended, objects unattended) object identity could be decoded above chance (V1 attended: 81.1%; $W = 595$, $p = 3.3e\text{-}7$, $r_B = 1$; V1 unattended: 84.8%; $W = 595$, $p = 3.2e\text{-}7$, $r_B = 1$; LOC attended: 37.3%; $t_{(33)} = 6.303$, $p = 4.0e\text{-}7$, $d_z = 1.08$; LOC unattended: 38.0%; $W = 583$, $p = 9.7e\text{-}7$, $r_B = 0.96$; TOFC unattended: 25.0%; $W = 476$, $p = 0.002$, $r_B = 0.60$), except in TOFC in the attended condition (TOFC attended: 19.6%; $W = 383$, $p = 0.143$, $r_B = 0.287$; $BF_{10} = 0.388$).

Moreover, decoding accuracy was not different between the objects attended and unattended conditions in any of the ROIs (V1: $t_{(33)} = -1.197$, $p = 0.240$, $d_z = -0.205$, $BF_{10} = 0.354$; LOC: $t_{(33)} = -0.214$, $p = 0.832$, $d_z = -0.037$, $BF_{10} = 0.188$; TOFC: $t_{(33)} = -1.726$, $p = 0.094$, $d_z = -0.296$, $BF_{10} = 0.697$). This suggests that the object stimuli evoked a reliable stimulus-specific activity pattern in all three sensory regions, which was not significantly different in strength between the two tasks (object categorization and character categorization). Note, the participants' task during the localizer run, which we used to train the classifier, was to detect a dimming of the fixation dot. As such, object stimuli were unattended during the localizer run, which may render the training data more similar in terms of attention allocation to the objects unattended task than the objects attended task. This may explain why decoding accuracy is similar, or even higher, for unattended compared to attended objects. More importantly, overall visual processing of the object stimuli was clearly present even when the objects stimuli were not attended, as the identity of the objects could be reliably decoded from neural activity patterns throughout the ventral visual stream when objects were unattended.

## Discussion

In the present study, we set out to investigate how sensory attenuation following visual statistical learning is modulated by attention. In line with previous studies [18,25,106,113,120] we found a significant and wide-spread attenuation of neural responses to expected compared to unexpected stimuli. Crucially, we showed that attending to the predictable stimuli is a prerequisite for this expectation suppression effect to arise. While unattended objects led to reliable and stimulus-specific increases in neural activity, and object pairs were equally learned for these stimuli, there was no differential activity depending on whether the trailing object was expected or unexpected. Additionally, we found that higher visual areas exhibited stimulus specific expectation suppression, whereas early visual cortex showed a global, stimulus unspecific suppression, possibly arising from a general increase in arousal in response to surprising stimuli.

### Attention is a prerequisite for expectation suppression

Our results show that a core neural signature of perceptual expectations, expectation suppression [18,25,106,113], is only evident when attention is directed to the predictable object stimuli. Specifically, when participants engaged in an object categorization task, we found a wide-spread reduction of neural activity for expected compared to unexpected stimuli throughout the ventral visual stream (V1, LOC, TOFC), as well as several non-sensory areas (anterior insula, inferior frontal gyrus, precentral gyrus, and superior parietal lobule). Strikingly, no modulation of neural activity by expectation was found when attention was drawn away from the object stimuli.

Interestingly, by directly comparing our present data with a previous dataset, in which we used a similar design (reported in [113]), we established that expectation suppression is present throughout the ventral visual stream irrespective of whether predictions are task-irrelevant, as long as the object stimuli are attended. In contrast, the larger activity for surprising stimuli in non-sensory areas (insular, frontal and parietal cortex) was only observed in the context of task-relevant expectations. This suggests that neural activity in the ventral visual stream is modulated by conditional probabilities, as long as the stimuli are attended, while the modulations in non-sensory regions are probably reflecting differences in task demands, given that unexpected stimuli were more difficult to categorize (reflected by a cost in speed and accuracy). During the object classification task, unexpected objects may require response inhibition, reevaluation of the category, and thus a new response decision. Given that the anterior insula has been associated with task control, action evaluation [121], as well as general attentional processes [122], and inferior frontal gyrus with response inhibition [123,124], the interpretation that the expectation modulation in non-sensory clusters may reflect task related aspects, but not conditional probabilities per se, appears well-supported by previous research.

Finally, our results also demonstrate that larger expectation suppression effects in V1 and TOFC are associated with increased reaction time benefits afforded by expectations when people are judging the predictable objects. This suggests that the observed expectation suppression effect may not merely constitute an epiphenomenon of more resource efficient neural processing. Instead, given the present data, it is plausible that the behavioral advantage of predicting stimuli may partially be rooted in improved and more effective sensory processing already at the early stages of visual processing. Predictions may thus help in converging more rapidly on an interpretation of the current sensory input, thereby contributing to faster reactions to expected than unexpected stimuli.

## No perceptual predictions without attention

Our results corroborate and extend earlier work by Larsson and Smith [62], who observed that stimulus expectation only affected repetition suppression when the stimuli were attended. However, they appear at odds with several previous studies that have reported expectation suppression in the visual system for stimuli that were not task-relevant and thus appeared unattended [18,32,61]. However, in all these studies, while the predictable stimuli were task-irrelevant, attention was not effectively drawn away by a competing stimulus that required attention. While our attention manipulation is also based on task-relevance, we do engage attention elsewhere using a competing task. This is a crucial difference between the present and previous studies, because it is likely that any supraliminal stimulus, in the absence of competition, will be attended to some degree, even if it is not task-relevant, especially if the stimulus is surprising [125]. Indeed, synthesizing earlier and current findings, we can conclude that expectation suppression in the visual system occurs irrespective of exact task goals and relevance of the predictable objects and their predictable relationship, but it is abolished by drawing attention away from the stimuli. This suggests that the integration of prior knowledge and sensory input is gated by attention – i.e., prior knowledge only exerts an influence on stimuli that are in the current focus of attention, instead of automatically and pre-attentively modulating sensory input as an obligatory component of perceptual processing.

It is however possible that other, more 'stubborn' prior expectations [126] that are derived over longer (ontogenetic or phylogenetic) time scales may persist even when attention is drawn away, such as perceptual fill-in during the Kanizsa illusion [127]. Therefore, it is crucial to discriminate between different types of predictions, as expectations of different sources may rely on different neural mechanisms and therefore have distinct properties. Similarly, for simple stimuli, such as oriented gratings [18,32] or simple sequences [128], the resolution of expectations may depend less on recurrent processing throughout the visual hierarchy than for complex objects. Thus, it is conceivable that the automaticity of predictive processing partially depends on the complexity of the predictable stimuli and their association, with increasing complexity requiring increasing processing across the hierarchy, and in turn a focus of attention on the predictable stimuli.

## Specific vs. unspecific surprise responses

In LOC and TOFC expectation suppression was largest in neural populations that were driven by the stimuli. Surprisingly, this was not the case in V1, where the suppression was uniformly present in the population that was driven by the stimuli and the population that was not. This replicates the results of our previous study [113] and suggests that the expectation suppression we observe in V1 is not the result of a stimulus-specific reduction in prediction error responses of neurons processing the stimulus. Rather, they suggest that the observed expectation suppression effect in V1 may be accounted for by a more general response modulation. Widespread nonperceptual modulations of visual cortical activity have been documented in response to unexpected events [129,130] and have been suggested to be linked to the cholinergic or noradrenergic system [131,132]. Interestingly, both the cholinergic and noradrenergic systems have also been associated with fluctuations in pupil dilation [133]. In line with this, we found a significantly enhanced pupil dilation in response to unexpected stimuli when the objects were attended. This suggests two possible global mechanisms which may partially account for the observed unspecific expectation suppression effect. Given that both pupil dilation [115,116] and the noradrenergic system [134] are associated with arousal changes, it is possible that expectation suppression is partially accounted for by an increased arousal in response to surprising stimuli. A related explanation is that enhanced pupil dilation to surprising stimuli [117–119] results in enhanced retinal illumination, which in turn leads to stronger responses in early visual areas [135], which could potentially also contribute to stimulus unspecific expectation suppression in V1. These interpretations are further supported by the fact that expectation suppression and pupil dilation differences between unexpected and expected attended stimuli were associated, with trailing images that elicit larger pupil dilation differences also showing more pronounced expectation suppression in V1.

It is unlikely however that these explanations can fully account for the observed expectation suppression effect across the visual hierarchy, given the stimulus-specificity of suppression in LOC and TOFC. Also, it is important to bear in mind that earlier studies, using different stimuli and paradigms, did observe stimulus-specific expectation effects in V1 [18,136]. Combined, the evidence suggests that the resolution of prediction errors crucially depends on the visual areas that are specifically coding the feature that is diagnostic of an expectation confirmation or violation, while areas below this level may only witness an unspecific, global modulation in their response, signifying the binary expectation confirmation or violation.

## Attention and prediction errors

Within the predictive coding framework, it has been suggested that attention modulates the gain of prediction error units [16]. On first glance, our results may not appear compatible with the suggestion that attention modulates the gain of prediction errors, because we observe a stimulus specific bottom-up signal (prediction error) when stimuli are unattended, but no difference in the size of this prediction error between expected and unexpected stimuli. However, it is conceivable that the gain modulation of activity in prediction error units only occurs after the initial feedforward activity sweep, once the object predictions are strongly activated and start exerting an effect on the resolution of the prediction error. In particular, the response to unexpected attended stimuli may be upregulated by attention, while prediction errors for expected attended stimuli are rapidly resolved, thus resulting in the difference in activity for attended objects. On the other hand, when attention is drawn away from the object stimuli, a reduced gain on prediction error units results in the observed attenuation of overall BOLD responses, and an absence of a reliable difference between expected and unexpected stimuli. A closely related, but conceptually distinct, interpretation is that attention constitutes a (modulation of the) prior itself [114,137]. On this account, attention boosts relevant predictions, as during the object classification task, thus leading to wide-spread expectation suppression, due to larger prediction errors for unexpected compared to expected stimuli. However, when attention is disengaged from the object stimuli, object predictions are not generated, and thus do not exert an effect on sensory processing.

## Interpretational limitations

One may wonder whether the character categorization task at fixation may have drawn attention away from the objects so forcefully that the object stimuli were no longer processed by sensory cortex. It is important to note here that, although attention was engaged at fixation by the character categorization task, this task was of trivial difficulty. Thus, it seems unlikely that attentional resources were exhaustively engaged by the task, preventing any processing of the surrounding object stimuli, thereby causing the absence of predictive processing. Indeed, behavioral performance was at ceiling during both tasks. Furthermore, even when objects were unattended reliable visual processing took place, as evident by strong responses and object-specific neural patterns in the visual ventral stream. This suggests that in-depth visual processing of object stimuli did occur in the absence of attention, but predictive processes in particular ceased.

Another alternative explanation of the present results could be that predictive relationships were not learned for the set of objects that were used during the character categorization task, thereby accounting for the absence of a prediction effect. The pair recognition task at the end of the experiment however showed that associations were learned for both image pair sets. Thus, a lack of visual processing or absence of learning cannot account for the observed results. Also, it is worth noting that initially the used probabilistic associations ($P(\text{expected}|\text{cue}) = 0.5$) may appear less strong than in some previous studies, e.g.: Egner et al. [24], Kok et al. [18], and Summerfield et al. [120] used $P(\text{expected}|\text{cue}) = 0.75$. However, the likelihood ratio of expected / unexpected stimuli ($0.5 / 0.1 = 5$) used here is actually larger (i.e., each unexpected image is more surprising) than in the cited studies ($0.75 / 0.25 = 3$). Moreover, similar probabilistic associations have been successfully employed in studies investigating neural effects of statistical learning in both non-human primates [23] and humans [113]. In short, the utilized conditional probabilities are comparable to previous studies investigating statistical learning. Finally, it is worth emphasizing that neither adaptation nor familiarity effects can account for the observed results, because all trailing objects served both as expected and unexpected images, depending only on temporal context (i.e., the leading image).

### Conclusion

In sum, our results suggest that visual statistical learning results in attenuated sensory processing for predicted input, but only when this input is attentively processed. Thus, attention seems to gate the integration of prior knowledge and sensory input. This places important constraints on neurocomputational theories that cast perceptual inference as a process of automatic integration of prior and sensory information.

## Materials and Methods

### Preregistration and Data Availability

The present study was preregistered at Open Science Framework (OSF) before any data were acquired. The preregistration document is available at DOI: 10.17605/OSF.IO/36TE7. All procedures and criteria outlined in the preregistration document were followed, unless explicitly specified in the Method section below. In this manuscript, only research question 1 of the preregistration document is addressed. All data analyzed in the present paper are available here: http://hdl.handle.net/11633/aacg3rkw

### Participants and Data Exclusion

Our target sample size was n = 34. This sample size was chosen to ensure 80% power for detecting at least a medium effect size (Cohen's d ≥ 0.5) with a two-sided paired t-test at an alpha level of 0.05. In total, 38 healthy, right-handed participants were recruited from the Radboud University research participation system. The study followed institutional guidelines of the local ethics committee (CMO region Arnhem-Nijmegen, The Netherlands). We excluded four participants, following our exclusion criteria (see preregistration document and *Data Exclusion*) resulting in the desired sample size of n = 34 participants (25 female, age 24.9 ± 4.8 years, mean ± SD) for data analysis. Of these four exclusions, three exhibited excessive motion during scanning, and one was caused by the participant falling asleep, thus resulting in an incomplete data set.

### *Data Exclusion*

The following preregistered criteria were utilized for the rejection of data. If any of the following criteria applied, data from that participant were excluded from all analyses. (1) Subpar fixation behavior during scanning, indicative by a total duration of closed eyes exceeding 3 SD above the group mean – only trials with stimuli were considered in this analysis; i.e., null events and instruction or performance screens were not included. (2) Excessive relative motion larger than ½ voxel size (i.e., 1mm) during MRI scanning, as indexed by the total number of these motion events exceeding 2 SD above the group mean. (3) Task performance during MRI scanning indicating frequent attentional lapses, as indicated by a mean error rate 3 SD above the group mean.

A fourth rejection criterion, outlined in the preregistration document, based on chance level performance during the post-scan pair recognition task (see: *Pair recognition task* and *2AFC task* in the preregistration document), was not enforced. This decision was based on feedback by participants, indicating that the short ITI during this task made it very challenging, even for participants who reported to have learned most of the associations. Thus, the preregistered pair recognition task based exclusion criterion would not fulfill the desired function of reliably indicating which participants did not explicitly learn the associations, as participants struggled with the task due to its fast pace. Indeed, the enforcement of the criterion would have resulted in the rejection of an additional nine participants (~26% of participants) from data analysis, which was deemed too stringent.

## Stimuli and experimental paradigm

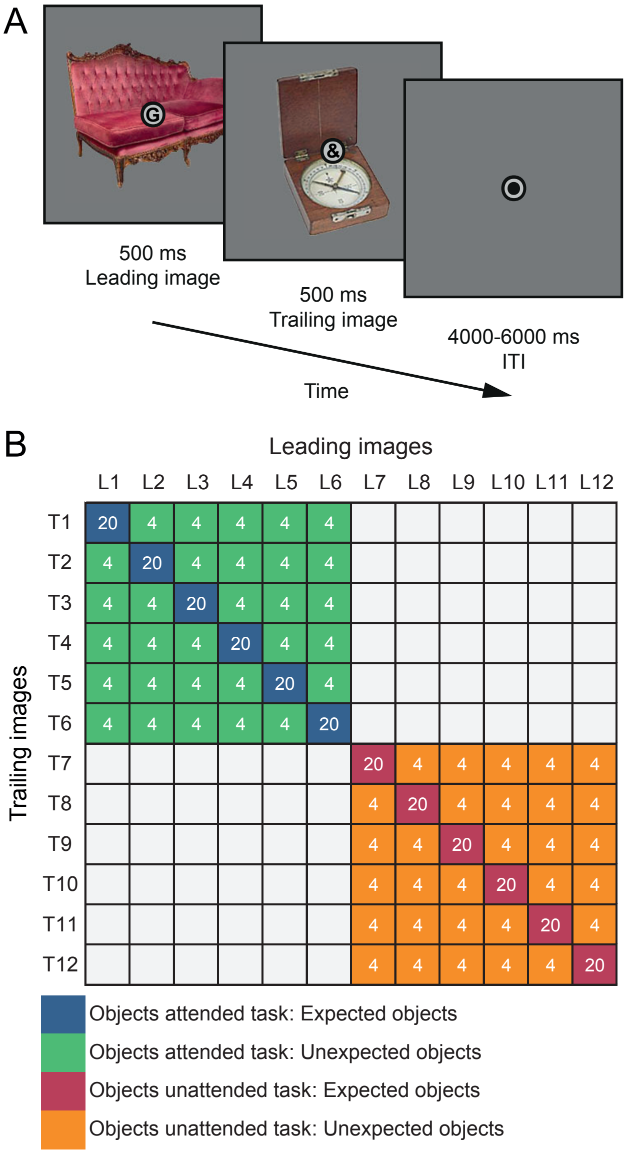
### *Experimental paradigm*

The experiment consisted of two sessions on two consecutive days. On each day the same stimuli were used for each participant, but different tasks were employed.

**Learning session – day one.** On each trial participants were exposed to two images of objects in quick succession (see Figure 3.6A for a single trial). Each stimulus was presented for 500 ms without an interstimulus interval and an intertrial interval between 1000-2000 ms. Each participant saw 24 different object images, 12 of which only occurred as leading images (i.e., as the first image on a trial), while the remaining 12 occurred only as trailing images (i.e., as the second image on a trial). Importantly, during the learning session the leading image was perfectly predictive of the identity of the trailing image [P(trailing|leading) = 1]. In other words, there were 12 image pairs during learning. While participants were made aware of the existence of such regularities, the regularities were not task-relevant. On 20% of trials, one of the two object images was presented upside-down – either the leading or the trailing image could be flipped upside-down. Crucially, whether an image was upside-down could not be predicted and was completely randomized. Participants were instructed to press a button as soon as an upside-down image occurred. Both speed and accuracy were emphasized. On trials without an upside-down image, no response was required. Throughout the entire trial, a fixation bull's-eye (outer circle 0.7° visual angle) was superimposed at the center of the screen. Within the inner circle of the fixation bull's-eye (0.6° visual angle) alphanumeric characters (letters or symbols) were presented (~0.4° visual angle). The characters were presented at the same time and for the same duration as the object stimuli – i.e., two characters per trial, each for 500 ms. As with the object images, there were 12 leading characters and 12 trailing characters. However, unlike the objects, the identity of the characters, including whether a letter or symbol occurred, was randomized and thus unpredictable. Participants were instructed that they could ignore these characters, but to maintain fixation on the fixation bulls-eye. In total each participant performed 960 trials during the learning session split into four runs, with a brief break in between runs. Thus, each of the image pairs occurred 80 times during the learning session. The learning session took approximately 60 minutes.

**FIGURE 3.6 Experimental paradigm.**

(**A**) A single trial is displayed, starting with a 500 ms presentation of the leading object and the leading character superimposed at fixation. Next, without ISI, the trailing object and trailing character are shown for 500 ms. Each trial ends with a 4000-6000 ms ITI (MRI session; 1000-2000 ms ITI learning session), showing only a fixation dot. (**B**) Statistical regularities depicted as image transition matrix with object pairs and trial numbers during MRI scanning. L1 to L12 represent leading objects, while T1 to T12 represent the trailing objects. Leading and trailing objects were randomly selected per participant from a larger pool of images - i.e., leading images of one participant may occur as trailing objects of another participant, in a different task, or not at all. Blue cells denote expected object pairs of the objects attended (object categorization) task, while green indicates unexpected object pairs of the objects attended task. Red denotes expected objects of the objects unattended (character categorization) task, and orange indicates unexpected objects of the objects unattended task. Each participant was also assigned 12 leading and 12 trailing characters (6 letters, 6 symbols each). Unlike the object images, there was no association between leading and trailing characters – i.e., the identity of the leading and trailing character was unpredictable. White numbers represent the total number of trials of that cell during MRI scanning. In total 120 trials of each of the four conditions were shown during MRI scanning per participant. In the behavioral learning session, participants performed an orthogonal oddball detection task, during which only expected pairs were shown (i.e., only the diagonal of the matrix), for a total of 80 trials per expected pair (960 trials total).



| | L1 | L2 | L3 | L4 | L5 | L6 | L7 | L8 | L9 | L10 | L11 | L12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| T1 | 20 | 4 | 4 | 4 | 4 | 4 | | | | | | |
| T2 | 4 | 20 | 4 | 4 | 4 | 4 | | | | | | |
| T3 | 4 | 4 | 20 | 4 | 4 | 4 | | | | | | |
| T4 | 4 | 4 | 4 | 20 | 4 | 4 | | | | | | |
| T5 | 4 | 4 | 4 | 4 | 20 | 4 | | | | | | |
| T6 | 4 | 4 | 4 | 4 | 4 | 20 | | | | | | |
| T7 | | | | | | | 20 | 4 | 4 | 4 | 4 | 4 |
| T8 | | | | | | | 4 | 20 | 4 | 4 | 4 | 4 |
| T9 | | | | | | | 4 | 4 | 20 | 4 | 4 | 4 |
| T10 | | | | | | | 4 | 4 | 4 | 20 | 4 | 4 |
| T11 | | | | | | | 4 | 4 | 4 | 4 | 20 | 4 |
| T12 | | | | | | | 4 | 4 | 4 | 4 | 4 | 20 |

🟦 Objects attended task: Expected objects
🟩 Objects attended task: Unexpected objects
🟥 Objects unattended task: Expected objects
🟧 Objects unattended task: Unexpected objects

**fMRI session – day two.** Day two of the experiment took place one day after the learning session. First, participants performed an additional 240 trials of the same upside-down task as during the learning session in order to refresh the learned associations. Then participants performed two new tasks in the MRI scanner. During MRI scanning, trials were similar to the learning session, using the same stimulus presentation durations, except for longer intertrial intervals (4000-6000 ms, randomly sampled from a uniform distribution). Another change to the paradigm during MRI scanning was a reduction of the probability of the trailing image given the leading image; $P(trailing\_expected|leading) = 0.5$. Thus, now only in 50% of trials a leading image was followed by its expected trailing image. In the remaining 50% of trials, one of the other five trailing images would occur, making these images unexpected given that particular leading image (i.e., each unexpected trailing image had $P(trailing\_unexpected|leading) = 0.1$). This was achieved by splitting the original 12x12 transition matrix from day one into two 6x6 matrices (see Figure 3.6B). One 6x6 matrix was used for each of the two tasks participants performed in the MRI (object categorization and character categorization tasks; see below). Thus, each expected trailing image was five times more likely given its leading image than any of the unexpected trailing images. Furthermore, each trailing image was only (un-) expected by virtue of the leading image it followed, which in turn also ensured that all images occurred equally often throughout the experiment, excluding confounds due to stimulus frequency or familiarity. During MRI scanning, an infrared eye tracker (SensoMotoric Instruments, Berlin, Germany) was used to monitor and record the position and pupil size of the left eye, at 50 Hz. Finally, after MRI scanning, a brief pair recognition task was performed – for details see *Pair recognition task* below.

**Object categorization task.** During the object categorization task participants were required to categorize, as quickly and accurately as possible, the trailing object on each trial as either electronic or non-electronic. Thus, during this task it was beneficial to be able to predict the identity of the trailing object using the learned associations. Failing to respond, or responding later than 1500 ms after trailing image onset, was considered a miss. Because the 12x12 transition matrix was split into two 6x6 matrices, one for this task, one for the character categorization task, it was ensured that both 6x6 matrices contained three electronic and three non-electronic objects as trailing and leading images, ensuring an equal base rate of both categories. Before performing this task, it was explained that 'electronic' would be any object that contains any electronic components or requires electricity to be used. Furthermore, it was ensured that participants could correctly classify each object by displaying all objects on screen and requesting participants to verbally categorize and name each object before entering the MRI.

**Character categorization task.** Trials of the character categorization task were identical to the object categorization task, except that participants were instructed to categorize the trailing character on each trial as a letter (of the standard Latin alphabet: A, B, D, E, G, H, J, K, M, N, R, S) or non-letter (i.e., a symbol or letter of a non-Latin alphabet: €, $, =, +, ϕ, Ͽ, £, ‡, Ӵ, ל, !, ?). While the presentation onset and duration of the characters coincided with the presentation of the object images, the identity of the trailing character was not predictable. As with the object images, six characters (three letters, three non-letters) were assigned as leading characters and six were assigned as trailing characters (three letters, three non-letters) for each of the two tasks (object and character categorization task). This was done to ensure that the character categorization task was as similar as possible to the object categorization task, and that exposure to the individual characters was as frequent as to the objects. Thus, in short, the rationale of the character categorization task was to draw attention away from the object stimuli and towards the characters, without imposing a heavy load on attentional or cognitive resources. Indeed, both tasks were designed to yield task performance at ceiling level. For both the object and character categorization tasks, feedback on behavioral performance was provided at the end of each run.

**Procedure, MRI session.** First, participants performed a brief practice run consisting of 50 trials (~5 minutes in duration) of either the object or character categorization task in the MRI. However, during the practice run, no unexpected trailing images occurred in order to retain the strong expectations built up during the learning session. Additionally, during the practice run, an anatomical image was acquired. After the practice run, two runs of the object or character categorization task were performed. Each run (~14 minutes) consisted of 120 trials and 7 null events of 12 seconds. Next, a practice run of the other task followed – i.e., if the object categorization task was performed first, the character categorization task would now follow, or vice versa. The task order was counter-balanced across participants. The practice run was again followed by two runs of the second task. After this, participants performed one functional localizer run (see: *localizer*). Finally, participants did a pair recognition task (see: *Pair recognition task*), assessing the learning of the object pairs. Once finished, participants were fully debriefed, and any remaining questions were addressed.

**Localizer.** We included a localizer session to define object-selective LOC for each participant and to constrain region of interest (ROI) masks to the most informative voxels using data from an independent, context-neutral run (i.e., without expectations). The functional localizer consisted of a repeated presentation of the previously seen trailing images and their phase-scrambled version. Images were presented for 12 seconds at a time, flashing at 2Hz (300 ms on, 200 ms off). At some point during stimulus presentation, the middle circle of the fixation dot would dim. Participants were instructed to press a button, as fast as possible, once they detected the dimming of the fixation dot. Each trailing image was presented 6 times. Additionally, a phase-scrambled version of each trailing image was presented 3 times. Furthermore, 12 null events, each with a duration of 12 seconds were presented. The presentation order was fully randomized, except for excluding direct repetitions of the same image and ensuring that each trailing image once preceded and once followed a null event in order to optimize the design.

**Pair recognition task.** The rationale of this task was to assess the learning of the object pairs (i.e, statistical regularities) and to compare whether participants learned the regularities during the objects attended task better than during the character categorization task. The pair recognition task followed the MRI session and consisted of the presentation of a leading image followed by two trailing images, one on the left and one on the right of the fixation dot. Participants were instructed to indicate, by button press, which of the two trailing images was more likely given the leading image. In order to prevent extensive learning during this task, a few trials with only unexpected trailing images were shown. Furthermore, participants were instructed that a response was required on each trial, even when they were unsure. Stimulus durations and intertrial intervals were identical to the learning session, i.e., 500 ms leading image, 500 ms trailing images, and a variable intertrial interval (1000-2000 ms randomly sampled from a uniform distribution). A response had to be provided within 1500 ms after trailing image onset, or otherwise the trial was counted as a miss. Participants performed one block of this task, consisting of 240 trials.

### *Stimuli*

Sixty-four full color object stimuli were used during the experiment. The object images were a selection of stimuli from Brady et al. [83], comprising typical object stimuli which were clearly electronic or non-electronic in nature (stimuli can be found here, DOI: 10.17605/OSF.IO/36TE7). Of these 64 object stimuli, 24 were randomly selected per participant, of which 12 were randomly assigned as leading images, while the other 12 served as trailing images. Thus, each specific image could occur as leading image for one participant, as trailing image for another participant, and not at all for a third participant, thereby minimizing the impact of any particular image's features. Images spanned approximately 5° x 5° visual angle on a mid-gray background, both during the learning session and MRI scanning. During the learning session stimuli were presented on an LCD screen (BenQ XL2420T, 1920 x 1080 pixel resolution, 60 Hz refresh rate). During MRI scanning, stimuli were back-projected (EIKI LC-XL100 projector, 1024 x 768 pixel resolution, 60 Hz refresh rate) on an MRI-compatible screen, visible using an adjustable mirror mounted on the head coil.

We calculated the average relative luminance of the object stimuli by converting the stimulus images from sRGB to linear RGB, then calculated the relative luminance for all pixels (where relative luminance $Y = 0.2126*R + 0.7152*G + 0.0722*B$; [138]), and finally averaged the obtained luminance values, thereby obtaining the mean relative luminance per image. On this relative luminance scale, 0 would be a completely black image, while 1 would be a white image. The average relative luminance of the stimulus set was 0.225, while the relative luminance of the mid gray background, presented during the ITI, was 0.216.

### *fMRI data acquisition*

Anatomical and functional images were acquired on a 3T Prisma scanner (Siemens, Erlangen, Germany), using a 32-channel head coil. Anatomical images were acquired using a T1-weighted magnetization prepared rapid gradient echo sequence (MP-RAGE; GRAPPA acceleration factor = 2, TR/TE = 2300/3.03 ms, voxel size 1 mm isotropic, 8° flip angle). Functional images were acquired using a whole-brain T2*-weighted multiband-6 sequence (time repetition [TR] / time echo [TE] = 1000/34.0 ms, 66 slices, voxel size 2 mm isotropic, 75° flip angle, A/P phase encoding direction, FOV = 210 mm, BW = 2090 Hz/Px). To allow for signal stabilization, the first five volumes of each run were discarded.

## Data analysis

### *Behavioral data analysis*

Behavioral data from the main task MRI runs were analyzed in terms of reaction time (RT) and response accuracy. Trials with RT < 200 ms, RT > 1500 ms, or no response were rejected as outliers from RT analysis (1.56% of trials). The two factors of interest were expectation status (expected vs. unexpected) and attention (objects attended vs. objects unattended task). Thus, a 2x2 repeated measures analysis of variance (RM ANOVA) was used to analyze behavioral data, with the additional planned simple main effects analyses of expected vs. unexpected within each task condition using two-sided paired t-tests. For these tests, RT and accuracy data per participant were averaged across trials and subjected to the analyses. For all paired t-tests, the effect size was calculated in terms of Cohen's $d_z$ [84], while partial eta-squared (η2), as implemented in JASP [139], was used as a measure of effect size for the RM ANOVA. Standard errors of the mean were calculated as the within-subject normalized standard error of the mean [85] with bias correction [86]. Data from the pair recognition task were analyzed by means of two-sided paired t-tests or Wilcoxon signed rank test, if the normality assumption was violated, comparing RTs and response accuracies between image pairs belonging to the attended vs. unattended conditions. Effect size for Wilcoxon signed rank test was calculated as the matched rank biserial correlation ($r_b$; [139]).

### *fMRI data preprocessing*

fMRI data were preprocessed using FSL 5.0.11 (FMRIB Software Library; Oxford, UK; www.fmrib.ox.ac.uk/fsl; [87], RRID:SCR_002823). The preprocessing pipeline consisted of the following steps: brain extraction (BET), motion correction (MCFLIRT), grand mean scaling, temporal high-pass filtering (128 seconds). For univariate analyses, data were spatially smoothed (Gaussian kernel with full-width at half-maximum of 5 mm), while for multivariate analyses no spatial smoothing was applied. FSL FLIRT was used to register functional images to the anatomical image (BBR) and the anatomical image to the MNI152 T1 2mm template brain using linear registration (12 degrees of freedom). Registration to the MNI152 standard brain was only applied for whole-brain analyses, while all ROI analyses were performed in each participant's native space in order to minimize data interpolation.

### *fMRI data analysis*

FSL FEAT was used to fit voxel-wise general linear models (GLM) to each participant's run data in an event-related approach. In these first-level GLMs, expected and unexpected image pair events were modeled as two separate regressors with a duration of one second (the combined duration of leading and trailing image) and convolved with a double gamma haemodynamic response function. An additional regressor of no interest was added to the GLM, modeling the instruction and performance summary screens. Moreover, the first temporal derivatives of these three regressors were added to the GLM. Finally, 24 motion regressors (FSL's standard + extended set of motion parameters) were added to account for head motion, comprised of the six standard motion parameters, the squares of the six motion parameters, the derivatives of the standard motion parameters and the squares of the derivatives. The contrast of interest, expectation suppression, was defined as the BOLD response to unexpected minus expected images. FSL's fixed effects analysis was used to combine data across runs. Because each run either used the objects attended or objects unattended (character categorization) task, two separate regressors were used in the fixed effects analysis, one for the objects attended task, one for the objects unattended task. Finally, across participants, data were combined using FSL's mixed effects analysis (FLAME 1). Gaussian random-field cluster thresholding was used to correct for multiple comparisons, using the updated default settings of FSL 5.0.11, with a cluster formation threshold of $p < 0.001$ (one-sided; i.e., $z \geq 3.1$) and cluster significance threshold of $p < 0.05$.

### *Region of interest (ROI) analysis*

ROI analyses were conducted in each participant's native space. The three a priori defined and preregistered ROIs were V1, object-selective LOC and TOFC. The choice of these ROIs was based on our previous study [113], in which we found significant expectation suppression in these cortical areas. For each ROI the mean parameter estimate was extracted from the participant's parameter estimate maps, representing the expected and unexpected images. This was done separately for the objects attended and objects unattended tasks, thus resulting in four parameter of interest. The parameter estimates were divided by 100 to yield percent signal change relative to baseline [88]. For each ROI, these data were submitted to a 2x2 RM ANOVA with expectation (expected, unexpected) and attention (objects attended, objects unattended) as factors. Simple main effects were calculated for the expectation effect in each of the attention conditions using two-sided paired t-tests. As applicable, Cohen's $d_z$ or partial eta-squared ($\eta 2$) were calculated as measures of effect size. Again, the within-subject normalized standard error of the mean [85] with bias correction [86] was calculated as an indicator of the standard error.

**ROI definition.** All ROIs were preregistered and defined a priori, based on previous results, and refined using independent data. The three ROIs were V1, object selective LOC, and TOFC. V1 was defined based on each participant's anatomical image, using Freesurfer 6.0 for cortex segmentation (recon-all; [97], RRID:SCR_001847). The resulting V1 labels were transformed into native volume space using 'mri_label2vol' and merged into one bilateral mask. LOC masks were created in each participant's native space using data from the functional localizer. Object selective LOC was defined as bilateral clusters, within anatomical LOC, showing a significant preference for intact compared to scrambled object stimuli [95,96]. To this end, one regressor modeling intact objects and one regressor modeling scrambled objects were fit to each participant's localizer data. Additional regressors of no interest were added to the model, with one regressor modeling instruction and performance screens, the temporal derivatives of all regressors, and the 24 motion regressor as also described above (see: *fMRI data analysis*). The contrast of interest, objects minus scrambles, was constrained to anatomical LOC, and the largest contiguous clusters in each hemisphere were extracted per participant. By default, the contrast was thresholded at $z >= 5$ (uncorrected; i.e., $p < 1e-6$). The threshold was lowered on a per participant basis if the resulting LOC clusters were too small; i.e., bilateral mask with less than 400 voxels in native volume space. The TOFC ROI mask was created using an anatomical temporal-occipital fusiform cortex mask from the Harvard-Oxford cortical atlas (RRID:SCR_001476), as distributed with FSL. This mask was further constrained to voxels showing a significant expectation suppression effect on the group level in our previous study, using an independent data set (Figure 2A in [113]). The resulting mask was transformed from MNI space to each participant's native space using FSL FLIRT.

Finally, each of the three ROI masks were constrained to the 300 voxels forming the most informative neighborhoods concerning object identity decoding. This was done by performing a multi-voxel pattern analysis (see: *Multi-voxel pattern analysis (MVPA)*) on the localizer data set per participant, decoding object identity. This ensured that the final masks contained the voxels that were from the most informative neighborhoods in each respective mask. It was not required that the final mask formed one contiguous cluster. In order to verify that our results did not depend on the a priori defined but arbitrary number of voxels in the ROI masks, we repeated all ROI analyses with masks ranging from 100-400 voxels (i.e., 800 $mm^3$ to 3200 $mm^3$) in steps of 100 voxels.

### *Multi-voxel pattern analysis (MVPA)*

A decoding analysis was performed on each participant's localizer data. For this analysis, not spatially smoothed mean parameter estimate maps were obtained per localizer trial by fitting a GLM with only one trial as regressor of interest and all remaining trials as one regressor of no interest [89]. Subsequently, these parameter estimate maps were used in a multi-class, linear SVM-based decoding analysis (SVC function, Scikit-learn; [90], RRID:SCR_002577), with the 12 trailing images as classes. The analysis was performed on the localizer data across the whole brain using a searchlight approach (6 mm radius) and stratified 4-fold cross-validation. Finally, the resulting decoding accuracy maps were used to constrain the ROI masks (see *ROI definition*).

We employed a similar decoding analysis to determine whether object-specific neural activity in the visual ventral stream was equally present during both the objects attended and unattended tasks. As above, a multi-class decoder with linear SVMs was used to decode object images. The per trial parameter estimates of the localizer run served as training data. For each main task run voxel-wise GLMs were fit with a regressor for each trailing image per expectation condition. As in the other fMRI analyses, the 24 motion regressors and temporal derivatives were added to the model (see *fMRI data analysis*). Finally, the decoder was tested on the obtained trailing image parameter estimates per run. As each attention condition consisted of six trailing images, chance performance of this decoder was at 16.7% (1/6).

### *Stimulus specificity analysis*

In an effort to further explore the nature of expectation suppression throughout the ventral visual stream, we investigated the stimulus specificity of the suppression effect. The key question here was if expectation suppression was primarily present in stimulus-driven voxels within a given area, or whether most voxels in an area showed the effect, regardless of whether or not they were stimulus-driven.

In order to investigate specificity, we obtained anatomically defined masks of our three ROIs (V1, LOC, TOFC). For V1 the unconstrained, anatomically defined Freesurfer V1 mask was used (see *ROI definition*). Anatomical LOC and TOFC were defined using the Harvard-Oxford cortical atlas. FSL FAST was used to obtain a gray matter mask for each participant based on their anatomical scan. Masks were transformed to the participant's native EPI space. Next, the three ROI masks were constrained to the participant's gray matter voxels. Within the resulting ROI masks, using the contrast object stimuli compared to baseline from the functional localizer run, voxels were split into two categories, stimulus-driven ($z > 1.96$; i.e., $p < 0.05$, two-sided), and not stimulus-driven, but also not deactivated, voxels ($-1.96 < z < 1.96$). Average expectation suppression was compared between ROIs split into stimulus-driven vs. not stimulus-driven voxels. Thus, a 3x2 RM ANOVA with ROI (V1, LOC, TOFC) and stimulus-driven (stimulus-driven vs not stimulus-driven) as factors was used for analysis. Greenhouse-Geisser correction was applied, if Mauchly's sphericity test indicated a violation of the sphericity assumption. Furthermore, the simple main effect of stimulus-driven vs. not stimulus-driven was assessed within each ROI. Additionally, to test for the presence of any expectation suppression, the amount of suppression was compared against zero using one sample t-tests.

### *Pupillometry*

In order to investigate whether pupil dilation effects accompany expectation suppression, we analyzed the pupil diameter data recorded during MRI scanning. A priori, two participants were rejected from this analysis, as the experiment log book indicated that pupil diameter data was unreliable for these two participants, leaving 32 participants for pupillometry. First, blinks were detected using a velocity based method, following the procedure outlined by Mathôt [140]. A blink was defined as a negative velocity peak (eyes closing), followed by a positive velocity peak (eyes opening) within a time period of 500 ms. The velocity threshold was set to 5 (arbitrary units). An additional 100 ms were added as padding before and after the detected blink onset and offset. If padding resulted in overlapping blink windows, consecutive blinks were considered as one long blink. Linear interpolation was used to replace missing data during blinks (18.05% of data). Note, this number includes the padding, and all time periods of no interest, such as null events, instruction and performance screens, as well as recording periods before and after MRI run onset; i.e., periods during which participants were free to close their eyes. Remaining missing data, not following a typical blink profile, were excluded from analysis, again adding a padding of 100 ms (3.07% of data). Similarly, outlier data with implausible velocity profiles were also rejected from the analysis, using the same velocity-based threshold as for blink detection but without the criterion of a negative peak followed by a positive peak (5.30% of data). Thus, data interpolation was only applied for short time intervals, which represent a clear blink, in order to avoid interpolation based on artifacts or over exceedingly long time periods. Finally, pupil data were smoothed using a Hanning window of 200 ms, and epoched into trials from 1 second before trailing image onset to 4 seconds after trailing image onset. The data of each trial were baseline corrected by diving the pupil diameter estimates by the mean diameter during the baseline period, 0.5 to 0 seconds before leading image onset. As a final data quality check, all trials exceeding pupil diameter values 7 SDs above the mean pupil diameter were rejected (3.01% trials). Trials with expected trailing images and unexpected trailing images were averaged separately for each participant. The difference between unexpected minus expected was subjected to a cluster-based permutation test (100,000 permutations; two-sided $p < 0.05$; cluster formation threshold $p < 0.05$) in order to assess statistical significance. Data from the objects attended and the objects unattended tasks were analyzed separately.

### *Linking pupil and neural measures*

In an exploratory analysis we sought to provide additional evidence for an association between pupil dilation and expectation suppression. To this end, we correlated expectation suppression with pupil dilation differences between expected and unexpected objects per trailing image. First, we obtained per trailing image parameter estimates by fitting a voxel-wise GLM to the fMRI data for each run, following the same procedure as for the main fMRI data analysis, outlined in *fMRI data analysis* and *Region of interest (ROI) analysis*. The only difference was that a separate regressor per trailing image and expectation condition was fit, thus resulting in a model with 12 regressors of interest (6 trailing images * 2 expectation conditions). As before, data was combined across runs using FSL's fixed effect analysis. The resulting parameter estimate maps were extracted for each ROI (V1, LOC, TOFC) and converted to percent signal change. Finally, for each participant we calculated expectation suppression for each trailing image (expectation suppression = $BOLD_{unexpected} - BOLD_{expected}$). Similarly, we calculated the difference in pupil dilation between unexpected and expected occurrences of each trailing image. For this we extracted the preprocessed (see:

*Pupillometry*) pupil size estimates for each trial and calculated the mean pupil size within the time window that showed a significant difference in pupil dilation between unexpected compared to expected attended stimuli on the group level (Figure 3.3, left panel); i.e., 1.52 to 2.88 seconds after trailing image onset. Next, we calculated the average difference in pupil size for each trailing image for unexpected compared to expected occurrences, thus yielding six pupil size difference scores (unexpected – expected) for both attention tasks per participant. Spearman's rank correlation was then used to estimate the correlation between the pupil dilation differences and expectation suppression magnitudes for each participant. Therefore, this correlation expresses the correlation in ranks of pupil dilation differences and expectation suppression magnitude for the trailing images, with positive correlations indicating that trailing images with large expectation suppression effects are also associated with larger pupil dilation differences. The obtained correlation coefficients were Fisher z-transformed and compared against zero (no correlation) using one-sample t-tests for each ROI and attention condition. We also submitted the coefficients to a repeated measures ANOVA with ROI and attention as factors.

### *Linking behavioral and neural measures*

In another exploratory analysis we investigated the relationship between behavioral and neural benefits of expectations by correlating expectation suppression with the behavioral RT benefit for expected stimuli observed during MRI scanning. First, we calculated the RT benefit for each trailing image during the main fMRI task (RTbenefit = $RT_{unexpected} - RT_{expected}$, per trailing image). Within each ROI we then correlated expectation suppression per trailing image (see: *Linking pupil and neural measures*) with RT benefit per trailing image using Spearman's rank correlation. Thus, this correlation coefficient indicates the rank correlation of expectation induced RT benefits and expectation suppression magnitude for the different trailing images. For statistical inference across participants, we Fisher z-transformed the correlation coefficients, and tested whether the observed correlation coefficients differ from zero (no correlation) in each condition by performing one-sample t-tests for each ROI and attention task separately. Finally, we also compared the magnitude of the correlations between ROIs and attention tasks using a 3x2 repeated measures ANOVA with ROI and attention condition (task) as factors.

### *Bayesian analyses*

In order to further evaluate any non-significant tests, in particular simple main effects, we performed the Bayesian equivalents of the above outlined analyses. JASP 0.9.0.1 ([139]; RRID:SCR_015823) was used to perform all Bayesian analyses, using default settings. Thus, for Bayesian t-tests a Cauchy prior width of 0.707 was chosen. Qualitative interpretations of Bayes Factors are based on criteria by Lee and Wagenmakers [94].

## Software

MRI data preprocessing and analysis was performed using FSL 5.0.11 (FMRIB Software Library; Oxford, UK; www.fmrib.ox.ac.uk/fsl; [87], RRID:SCR_002823). Custom Python 2.7.13 (Python Software Foundation, RRID:SCR_008394) scripts were used for additional analyses, data handling, statistical tests and data visualization. The following Python libraries and toolboxes were used: NumPy 1.12.1 ([98], RRID:SCR_008633), SciPy 0.19.0 ([99], RRID:SCR_008058), Matplotlib 1.5.1 ([100], RRID:SCR_008624), Statsmodels 0.8.0 (www.statsmodels.org) and Scikit-learn 0.18.1 ([90], RRID:SCR_002577). Additionally, Slice Display [102], a Matlab 2017a (The MathWorks, Inc., Natick, Massachusetts, United States, RRID:SCR_001622) data visualization toolbox, was used for displaying whole-brain results. JASP 0.9.0.1 ([139], RRID:SCR_015823) was used for Bayesian analyses and RM ANOVAs. Stimuli were presented using Presentation® software (version 18.3, Neurobehavioral Systems, Inc., Berkeley, CA, RRID:SCR_002521).

## Supplemental analyses

### *Pupil dilation is associated with larger BOLD responses*

In order to provide additional support for the hypothesis that pupil dilation differences may partially underlie expectation suppression in V1, we examined the relationship between pupil dilation and the BOLD response. First, we extracted per trial pupil size data and parameter estimate maps from the fMRI main task data for V1. Pupil size data was preprocessed as described in *Pupillometry*, and extracted from a three-second time window, starting with trailing image onset and ending 2.5 seconds after trailing image offset; thus, the time window covered the full duration shown in Figure 3.3 after trailing image onset. fMRI data was preprocessed as outlined in *fMRI data preprocessing*. Next, for each trial we fitted a GLM with only one trial as regressor of interest and all remaining trials as regressors of no interest [89]. Per participant, we extracted the per trial parameter estimate maps, averaged within the V1 ROI, and z scored the mean parameter estimates per condition separately in order to remove potential effects of mean differences between the conditions. We also z scored the pupil size estimates per condition for the same reason. Next, we fitted per participant a GLM with the mean BOLD parameter estimates (one per trial) as predicted variable and a regressor with pupil size for each expectation and attention condition combination (i.e., four regressors of interest) as predictors. Statistical inference across subjects was performed by subjecting the thus obtained parameter estimates of the four regressors of interest to a 2x2 repeated measures ANOVA, as with our main ROI analysis; i.e., with attention and expectation as factors. Furthermore, in order to assess whether the BOLD response was influenced by pupil dilation at all we performed one-sample t-tests comparing the obtained parameter estimates against zero for each condition separately. Additionally, we performed a similar analysis, but split the fMRI data into stimulus-driven vs. non-stimulus-driven V1 gray matter voxels (see *Stimulus specificity analysis* for details on the ROI mask creation). This analysis thus results in a 2x2x2 repeated measures ANOVA with expectation, attention and stimulus-responsiveness as factors.

Increased pupil dilations were associated with larger BOLD responses regardless of whether stimuli were attended and expected (attended expected: $t_{(31)} = 3.006$, $p = 0.005$, $d_z = 0.531$; attended unexpected: $t_{(31)} = 4.392$, $p = 1.2e\text{-}4$, $d_z = 0.776$; unattended expected: $t_{(31)} = 5.228$, $p = 1.1e\text{-}5$, $d_z = 0.924$; unattended unexpected: $W = 452$, $p = 2.1e\text{-}4$, $r_B = 0.712$). Results are shown in Figure 3.3–figure supplement 1. Pupil dilation led to slightly stronger BOLD increases when objects were unattended than attended ($F_{(1,31)} = 5.563$, $p = 0.025$, $\eta^2 = 0.152$), but independent of whether stimuli were expected or unexpected ($F_{(1,31)} = 0.054$, $p = 0.817$, $\eta^2 = 0.002$; interaction: $F_{(1,31)} = 2.261$, $p = 0.143$, $\eta^2 = 0.068$). Thus, pupil dilation had a positive effect on overall BOLD responses in V1.

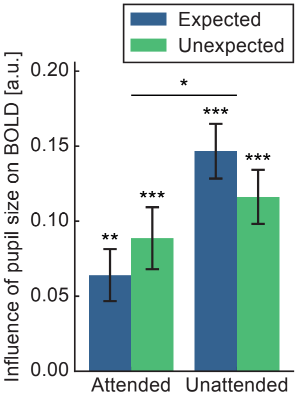


**FIGURE 3.3–FIGURE SUPPLEMENT 1 Pupil dilation influences BOLD responses in V1.** Displayed are the parameter estimates of the influence of pupil size on BOLD responses in V1. BOLD responses increase with larger pupil dilations regardless of whether stimuli were attended and expected. However, pupil dilation influenced BOLD responses more when objects were unattended than attended. Whether stimuli were expected or unexpected did not change the association between BOLD and pupil dilation. Error bars indicate within-subject SEM. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

### *Pupil dilation influences BOLD responses more in non-stimulus-driven V1 voxels*

Next, we assessed whether the same association would hold in stimulus-driven and non-stimulus driven V1 voxels. Figure 3.3–figure supplement 2 shows that there was indeed a reliable, positive association between BOLD responses and pupil dilation within both stimulus-driven and non-stimulus-driven voxels. Again, larger pupil dilations were predictive of enhanced BOLD responses when object stimuli were attended and unattended, as well as for expected and unexpected objects (all t-tests $p < 0.05$). Interestingly, this association was somewhat larger in non-stimulus-driven than stimulus-driven voxels ($F_{(1,31)} = 9.267$, $p = 0.005$, $\eta^2 = 0.230$), suggesting that the association between BOLD and pupil dilation is particularly strong for those neural populations that are not driven by our object stimuli. This is in line with earlier observations that non-stimulus-driven activations (possibly reflecting neuromodulation) are greater in regions that represent more peripheral parts of the visual field [129]. There was also a stronger association of pupil dilation and BOLD responses when objects were unattended ($F_{(1,31)} = 5.042$, $p = 0.032$, $\eta^2 = 0.140$), but the magnitude of the association was not affected by whether a stimulus was expected or not ($F_{(1,31)} = 0.008$, $p = 0.928$, $\eta^2 = 2.6e\text{-}4$). Moreover, no interaction effect was observed (all interactions $p > 0.1$).

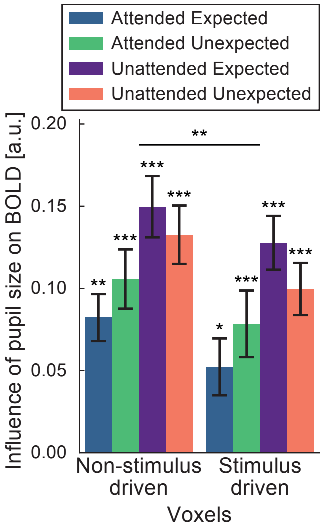


**FIGURE 3.3–FIGURE SUPPLEMENT 2 Pupil dilation influences BOLD responses more in non-stimulus-driven than stimulus-driven V1 voxels.**

Displayed are the parameter estimates of the influence of pupil size on BOLD responses in V1. BOLD responses increase with larger pupil dilation. This association was stronger in non-stimulus-driven (left) than stimulus-driven (right) V1 gray matter voxels. Error bars indicate within-subject SEM. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Thus, to summarize, our results show that pupil dilation has a substantial, positive association with V1 BOLD responses, regardless of whether stimuli were attended and expected, for both stimulus-driven and non-stimulus-driven neural populations. This result is expected, given that pupil dilation has been related to other processes known to correlate with BOLD responses such as mental effort, arousal and attention (for a review see: [141]). Moreover, increases in retinal illumination due to larger pupil dilation can also result in increased BOLD activity [135]. These results support our suggestion that larger pupil dilations in response to unexpected stimuli, possibly reflecting general arousal mechanisms, may partially account for expectation suppression in V1. However, it should also be noted that the association between pupil dilation and the BOLD response is not solely observed when objects were attended, as pupil dilation is likely a general reflection of vigilance and arousal [115,116], which is expected to fluctuate also when the objects are not attended. That this association is more pronounced in non-stimulus-driven voxels, further supports the possibility that expectation suppression in V1, including the suppression observed in non-stimulus-driven voxels, may partially reflect non-perceptual effects such as arousal changes, which are reflected by larger pupil dilations in response to surprising stimuli.

### *No differences in pupil dilation during baseline*

We assessed pupil size during baseline to ensure that differences in pupil dilation between expectation conditions or attention tasks do not simply reflect difference in baseline (e.g., pre-stimulus arousal differences). Pupil data was preprocessed using the same pipeline as described in *Pupillometry*, except for that pupil size was extracted in raw units during the baseline period. Per participant, pupil size was then averaged for each attention and expectation condition separately. Mean pupil estimates were then compared between conditions using a 2x2 repeated measures ANOVA, with expectation and attention as factors. Additionally, a Bayesian repeated measure ANOVA was conducted to quantify the evidence for the absence of a difference in pupil size during baseline.

Results showed that there was no difference in baseline pupil size before attended compared to unattended stimuli ($F_{(1,31)} = 5.226$, $p = 0.484$, $\eta^2 = 0.016$, $BF_{inclusion} = 0.254$), nor before expected compared to unexpected stimuli ($F(1,31) = 0.001$, $p = 0.926$, $\eta^2 = 2.8e\text{-}4$, $BF_{inclusion} = 0.136$; interaction: $F_{(1,31)} = 6.2e\text{-}4$, $p = 0.955$, $\eta^2 = 1.0e\text{-}4$, $BF_{inclusion} = 0.042$). Figure 3.3–figure supplement 3 shows the pupil size in raw units during the baseline period. Thus, data suggest that pupil size, and thereby likely arousal, during baseline was of a similar magnitude during both attention tasks and expectation conditions, thereby rendering an explanation of the observed phasic differences in pupil size based on differences in baseline pupil size unlikely.

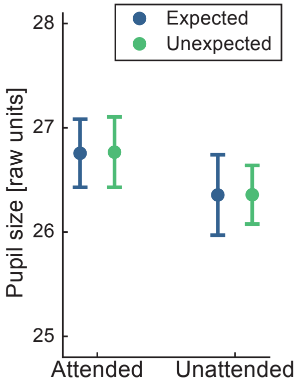


**FIGURE 3.3–FIGURE SUPPLEMENT 3** **No difference in baseline pupil size between attention tasks, nor expectation conditions.**

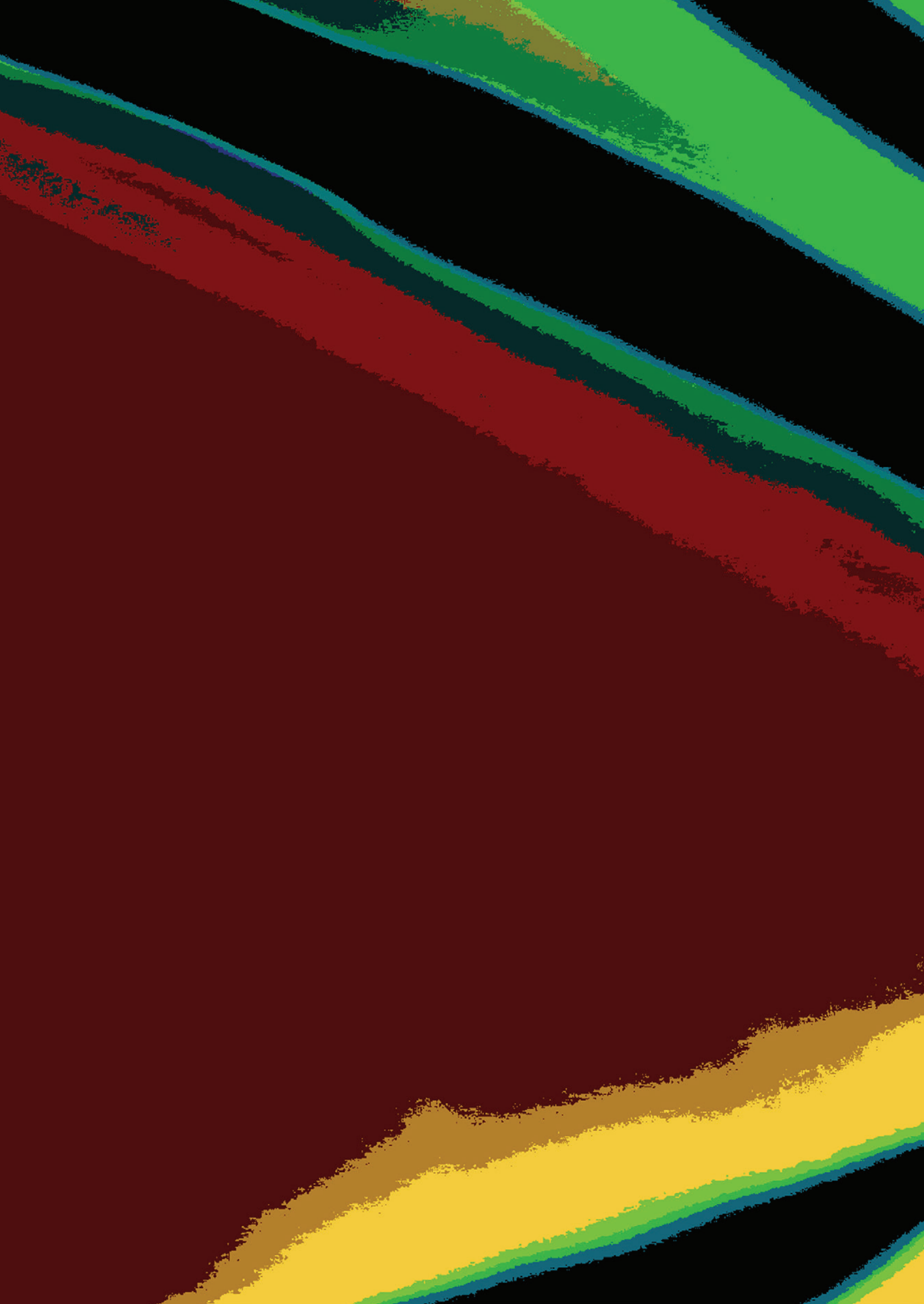
Displayed are mean pupil sizes during the baseline period in raw units for expected and unexpected trials during the objects attended and unattended task. Pupil sizes during baseline were similar for trials with expected and unexpected object stimuli, as well as during both tasks. Error bars indicate within-subject SEM.

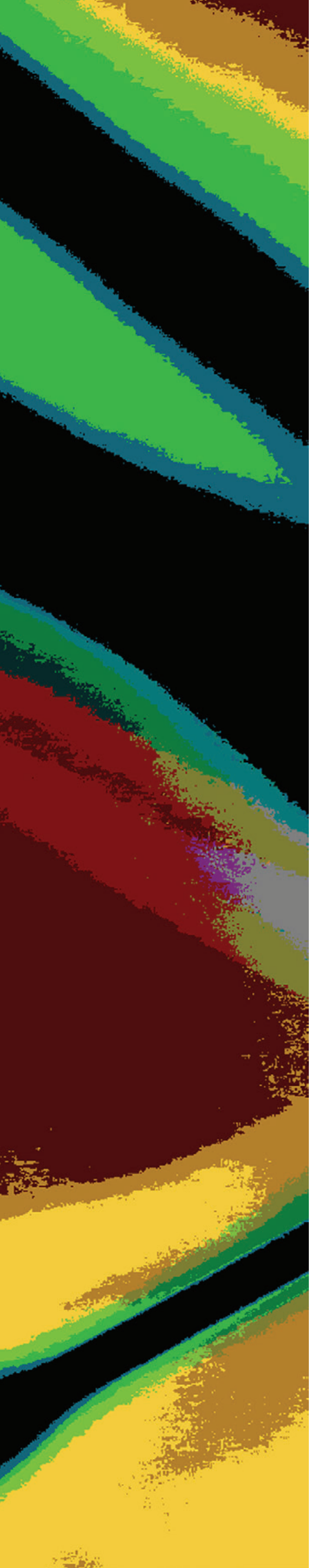
# Supplemental Information

**SUPPLEMENTARY FILE 3.1 Overview of expectation suppression across cortex.**

Brain areas showing significant expectation suppression (GRF cluster corrected). Listed are significant clusters with their respective area label, MNI coordinate of the peak z value, the number of voxels in the cluster, as well as the p value of the cluster and its max z statistic. For large clusters (n voxels > 700) additional local z maxima (*z* > 3.72; i.e., *p* < 0.0001, one-sided) are also shown with area label, MNI coordinates and max z statistic. Unexp. = unexpected image pairs; Exp. = expected image pairs; Att. = objects attended task; Unatt. = objects unattended (characters attended) task.

| Contrast | Area label | MNI coordinates | | | n voxels | *p* cluster | max *z* |
|---|---|---|---|---|---|---|---|
| | | x | y | z | | | |
| Unexp. > Exp. | Lateral Occipital Cortex, inferior division | -48 | -72 | -14 | 1536 | 4.1e-15 | 4.61 |
| (Att.) | Temporal Occipital Fusiform Cortex | -34 | -52 | -18 | | | 4.58 |
| | Lingual Gyrus | -22 | -52 | -10 | | | 4.51 |
| | Temporal Fusiform Cortex, posterior division | 30 | -38 | -20 | 535 | 3.6e-7 | 4.96 |
| | Lateral Occipital Cortex, inferior division | 48 | -68 | -12 | 373 | 1.6e-5 | 4.24 |
| | Precentral Gyrus | 48 | 4 | 34 | 1456 | 1.5e-14 | 4.68 |
| | Frontal Operculum Cortex | 44 | 20 | -2 | | | 4.63 |
| | Inferior Frontal Gyrus, pars opercularis | 50 | 14 | 28 | | | 4.56 |
| | Frontal Orbital Cortex | 34 | 26 | -4 | | | 4.36 |
| | Precentral Gyrus | -42 | -2 | 36 | 471 | 1.5e-6 | 4.36 |
| | Frontal Operculum Cortex | -40 | 18 | 0 | 156 | 0.0079 | 3.94 |
| | Superior Frontal Gyrus | 4 | 18 | 56 | 626 | 6.0e-8 | 4.39 |
| | Superior Parietal Lobule | -26 | -56 | 48 | 329 | 5.0e-5 | 4.31 |
| | Superior Parietal Lobule | 30 | -48 | 46 | 173 | 0.0046 | 4.36 |
| | Cerebellum, Vermis VI | -4 | -64 | -18 | 128 | 0.0210 | 4.61 |
| | Cerebellum, Left Crus I | -10 | -76 | -30 | 126 | 0.0226 | 4.32 |
| Unexp. > Exp. | - | - | - | - | - | | - |
| (Unatt.) | | | | | | | |
| [Unexp. > Exp. | Lateral Occipital Cortex, inferior division | -46 | -70 | -12 | 745 | 6.4e-9 | 4.48 |
| (Att.)] | Temporal Occipital Fusiform Cortex | -42 | -62 | -14 | | | 4.28 |
| > | Inferior Temporal Gyrus, temporooccipital part | -46 | -50 | -16 | | | 4.12 |
| [Unexp. > Exp. | Temporal Fusiform Cortex, posterior division | 30 | -38 | -24 | 173 | 0.0053 | 4.77 |
| (Unatt.)] | Lateral Occipital Cortex, inferior division | 50 | -66 | -14 | 139 | 0.0161 | 3.87 |
| | Precentral Gyrus | 38 | 8 | 26 | 222 | 0.0012 | 4.32 |
| | Frontal Operculum Cortex | -40 | 16 | 2 | 119 | 0.0322 | 3.81 |
| | Lateral Occipital Cortex, superior division | -22 | -62 | 36 | 117 | 0.0345 | 3.75 |
| | Cerebellum, Left Crus II | -10 | -76 | -34 | 125 | 0.0260 | 3.83 |
| | Precuneous Cortex | 0 | -62 | 12 | 116 | 0.0358 | 3.84 |





## CHAPTER 4

# Dampened sensory representations for expected input across the ventral visual stream

*This chapter is based on:*
Richter, D., Heilbron, M., and de Lange, F.P. (in preparation). Dampened sensory representations for expected input across the ventral visual stream.

# Abstract

Expectations, derived from previous experience, can help in making perception faster, more reliable and informative. A key neural signature of perceptual expectations is expectation suppression, an attenuated neural response to expected compared to unexpected stimuli. While expectation suppression has been reported using a variety of paradigms and recording methods, it remains unclear what neural modulation underlies this response attenuation. Sharpening models propose that neural populations tuned away from an expected stimulus are particularly suppressed by expectations, thereby resulting in an attenuated, but sharper population response. In contrast, dampening models suggest that neural populations tuned *towards* the expected stimulus are most suppressed, thus resulting in a dampened, less redundant population response. Empirical support is divided, with some studies favoring sharpening, while others support dampening. A key limitation of previous studies is the ability to draw inferences about neural-level modulations based on population (e.g., voxel) level signals, which integrate over millions of neurons (e.g., the BOLD response). Indeed, recent simulations of repetition suppression showed that opposite neural modulations can lead to comparable population-level modulations. Forward models provide one possible solution to this inference limitation. We used forward models to implement both sharpening and dampening models, mapping individual neural modulations to voxel-level data. By comparing simulated neural responses to a combined analysis of two previously published fMRI studies, we show that a feature-unspecific gain modulation underlies expectation suppression in early visual cortex. However, in higher-order object selective visual areas feature-specific gain modulations, suppressing neurons particularly tuned towards the expected stimulus, best explain the empirical fMRI data. Thus, our results are in line with the dampening account, suggesting that expectations may reduce redundancy in sensory cortex, and promote updating of internal models on the basis of surprising information.

# Introduction

The perceptual system faces at least two challenges: to represent the world as quickly and accurately as possible, and to promote processing of novel information. Relying on previous experience to guide perception may help to meet both challenges [44], and is advantageous to an agent acting in an information rich environment. Indeed, deriving expectations from previous experience aids performance, enabling faster and more accurate responses to expected events [9,51,58,59]. Within cortex, the consequences of prior expectations are evident during sensory processing in both early and higher visual areas [19]. One well-established neural consequence of prediction in perception is expectation suppression: the attenuation of sensory responses to expected compared to unexpected stimuli. Expectation suppression has been reported in several sensory modalities and species, using different recording methods, in a wide range of paradigms (for reviews see: [19,20]).

However, it remains unclear what neural mechanism underlies this phenomenon. On the one hand, population sharpening models propose that expectations preferentially suppress neurons tuned away from the expected stimulus [18,41,142]. By inhibiting information that is inconsistent with top-down expectations, such a sharpening process would bias perception in line with our expectations, echoing Bayesian models of perception [5,6]. The net result is a response that is reduced in amplitude, but carries a sharper, more reliable representation of the stimulus. On the other hand, dampening (or cancellation [42]) models argue that expectations preferentially suppress neurons tuned *towards* expected stimuli [23,43,79,113,143]. By cancelling information in line with prior expectations, the brain would reduce redundancy in the sensory stream, while at the same time favoring processing of novel or surprising information. Thus, on this account responses are reduced and neural representations of expected stimuli dampened.

To date, studies that tried to arbitrate between these two accounts have yielded mixed results. In line with population sharpening, some studies [18,142] found that expected stimuli evoked weaker BOLD responses, but that stimulus identity was more accurately decoded from those same BOLD responses – suggesting a sharper representation. Moreover, it was found that expectation suppression was weaker in voxels that responded more strongly to the expected stimulus, in agreement with the hypothesized suppression of inconsistent information [18]. However, other studies using similar techniques found the opposite pattern of effects: reduced classification accuracies [143] or pattern similarities [43] for expected stimuli, and larger suppression magnitudes for preferred compared to non-preferred stimuli [113], in line with dampening accounts. One possible explanation for these inconsistencies is that the observed BOLD or MEG signal integrates over millions of neurons, making it difficult to infer neural-level mechanisms from population-level measurements. Indeed, in the domain of sensory adaptation, Alink et al. [105], building on work by Weiner et al. [144], recently showed that the relation between neural-level mechanisms and voxel-level results can be rather counter intuitive. Their simulations suggest, for instance, that a dampening-like mechanism at the neural-level can, in principle, manifest as a sharpening-like result at the voxel-level – and vice versa. To overcome these interpretational difficulties, Alink et al. [105] proposed a forward modelling approach to explicitly model which underlying neural-level mechanism could best explain the observed voxel-level adaptation results. While adaptation and expectation are distinct phenomena [22,62,110], they share some key characteristics. This make an analogous approach suitable to investigate expectation suppression.

Here, we build on and extend the approach of Alink et al. [105] and Weiner et al. [144], by using forward models to elucidate the neural mechanism underlying expectation suppression in the ventral visual stream. First, we analyzed and integrated data of two previously published studies [113,145], which manipulated perceptual expectations by presenting human volunteers (n = 56) with expected and unexpected objects images. For both studies, the effects of expectation were characterized in terms of eight fMRI outcome metrics, both univariate and multivariate. These metrics were based on previous studies, where they were interpreted as evidence for either sharpening [18,142] or dampening accounts [43,113,143]. This resulted in a specific pattern of effects of expectation within three regions of interest (ROIs) across the ventral visual stream: primary visual cortex (V1), object selective lateral occipital complex (LOC), and temporal occipital fusiform cortex (TOFC). Next, we used forward models to explicitly model which neural mechanism best explained the observed effects in each ROI. We implemented a set of six distinct models, which all predict a suppression of neural responses to expected stimuli, but differ in terms of the underlying mechanism of that suppression. In particular, we defined dampening as a local feature-specific gain modulation [113,143], in which the gain of neural populations tuned *towards* the expected stimulus features is reduced. Conversely, we defined population sharpening as a remote feature-specific gain modulation [18,142], in which the gain of neural populations tuned *away* from the expected stimulus features is reduced. Moreover, we modeled previously suggested feature-unspecific effects as a global gain modulation [113,145]. As an additional competitor, we also implemented response tuning models, that narrow the width of the response function. These models have been suggested in the wider literature, as underlying response modulations for related phenomena like attention and adaptation [146–148].

To foreshadow the results, we show that perceptual expectations in the ventral visual stream are best modeled by a feature-specific local gain modulation of neural responses. Thus, our results, particularly in higher visual areas LOC and TOFC, are in line with dampening accounts of expectations, which advocate a suppression of neural responses particularly for neural populations tuned *towards* the expected stimulus features. This dampening of neural responses suggests that perceptual expectations, derived from statistical regularities, may reduce information redundancy and bias information processing towards surprising, novel information.

## Results

In a first step, we analyzed the empirical fMRI data using eight outcome metrics used in previous studies investigating population sharpening and dampening [18,43,105,113,142,143]. Next, using independent fMRI data, we validated the implemented stimulus feature spaces, which were used to model neural responses to each object stimulus in an ROI specific fashion. We then performed the simulation and analyzed the fit of the different models to the empirical results. Mimicking the interpretation in most empirical studies, we first assessed the qualitative fit in terms of the sign of the slopes per outcome metric. Additionally, we analyzed the fit in a more detailed, quantitative fashion by calculating the mean squared error (MSE) between the simulated and empirical results, and assessed which model type best explained the empirical data. Finally, we explored which parameter values resulted in the optimal fit, thereby exploring the circumstances under which the models best explain the observed data.

### Empirical fMRI data

First, we analyzed the empirical fMRI data. In brief, we utilized eight different outcome metrics, based on analyses used in previous studies (e.g. [18,43,105,113,142,143]). (1) Mean amplitude modulation (MAM) due to a stimulus being expected vs unexpected; i.e., expectation suppression. (2) Within-class correlation (WC) and (3) between-class correlation (BC) between stimuli, for expected and unexpected occurrences of the stimuli. (4) Classification performance (CP) defined as the difference between BC-WC for expected and unexpected stimuli respectively, as well as (5) classification performance using linear support vector machines (SVM). (6) Amplitude modulation by amplitude (AMA), the amplitude modulation by expectation as a function of voxel mean amplitude. (7) Amplitude modulation by selectivity (AMS), the amplitude modulation by expectation as function of voxel mean selectivity. (8) Image preference analysis (IP), assessing the mean amplitude modulation within a voxel as a function of image preference. Details are described in: *Materials and Methods, MRI outcome metrics*.

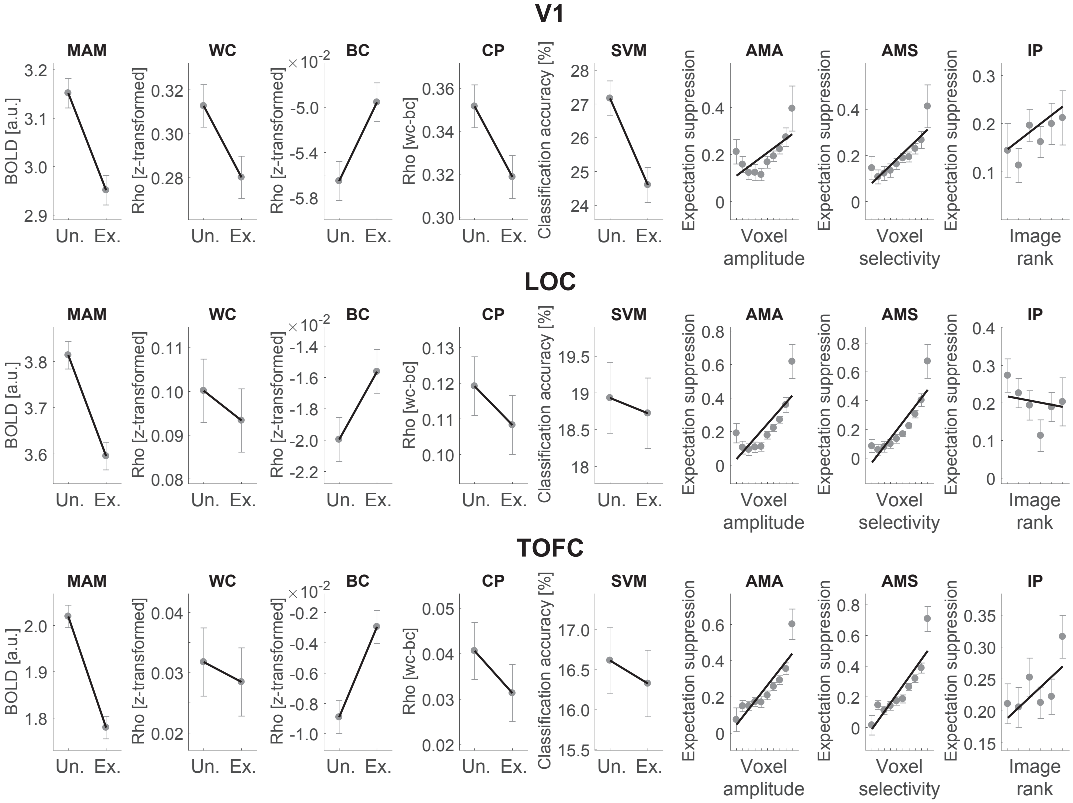


FIGURE 4.1 **Empirical fMRI results.**

Shown are empirical fMRI data analysis results from V1, LOC, and TOFC from a combined analysis of data from Richter and de Lange ([145]; n = 34) and Richter et al. ([113]; n = 22). These results constitute the data features the subsequent simulations are compared against. *Un.* = Unexpected trials, i.e., when the trailing image was unexpected given the leading image. *Ex.* = expected trials, i.e., when the trailing image was expected given the leading image. Outcome metrics: *MAM* = mean amplitude modulation (expectation suppression), WC = within-class correlation, *BC* = between-class correlation, *CP* = classification performance (BC-WC), *SVM* = support vector machine based classification (chance level is ~15%), *AMA* = amplitude modulation by amplitude, *AMS* = amplitude modulation by selectivity, *IP* = image preference analysis (amplitude modulation by image preference). For display purposes, only IP data of images 2-7 from Richter et al. [113] are displayed, but all image data were analyzed.

Figure 4.1 depicts the fMRI results from the three ROIs, V1, LOC, and TOFC. Data from Richter and de Lange [145], and Richter et al. [113] was combined by pooling participants. In all ROIs a substantial modulation of MAM is evident, with expected stimuli being suppressed relative to unexpected ones; i.e., constituting expectation suppression, the key phenomenon of interest. Furthermore, clear difference between WC and BC emerge between expected and unexpected stimuli, resulting in improved classification accuracies (CP, SVM) for unexpected stimuli in V1, while LOC and TOFC show a similar albeit less reliable pattern. Moreover, voxel with larger mean amplitude (AMA) and selectivity (AMS) show more expectation suppression ($\text{BOLD}_{\text{unexpected}} - \text{BOLD}_{\text{expected}}$) in all ROIs. A similar, albeit less clear trend, is also evident within voxels, with larger suppression for more preferred stimuli (IP) in V1 and TOFC. Since the goal of the analysis is only to estimate data features, which are subsequently used to compare the simulation against, we do not report inferential statistics here. However, for completeness, a full set of statistics, corresponding to the results displayed in Figure 4.1, are summarized in supporting tables S4.1 (V1), S4.2 (LOC), and S4.3 (TOFC). Overall, empirical results are comparable between the three ROIs, with differences mainly emerging in terms of variability and effect sizes, while the sign of the effects (slopes) are generally the same.

## Feature space models explain neural variance in target ROIs

Because we modeled neural responses to different stimuli we had to establish for each ROI a feature space model, which reliably describes the object stimuli in a manner relevant to the neural responses in the target ROIs (V1, LOC, TOFC). V1 feature space was defined by the predominant orientation of the object stimuli, as V1 neurons are tuned to stimulus orientation [33]. LOC responses were modeled by shape complexity, based on Vernon et al. [34]. TOFC feature space was derived from human-rated semantic similarity, which is thought to correlate with complex visual features [149]. Additional details are described in *Materials and Methods, Feature space*.

In order to verify that each feature space captured significant variance in neural response, we performed a representation similarity analysis (RSA), using independent localizer data. In brief, for each participant we correlated the feature space (model) representational dissimilarity matrix (RDM) with the neural RDM. Subsequently, we compared the obtained correlation coefficients against zero (i.e., no correlation between feature space and neural RDM). RSA results, depicted in Figure 4.2, show that the feature spaces explained a significant amount of neural variance in their target ROIs (V1, orientation feature space: $t_{(55)} = 6.23$, $p = 6.9\text{e-}8$, $d_z = 0.83$; LOC, shape feature space: $t_{(55)} = 5.21$, $p = 2.9\text{e-}6$, $d_z = 0.70$; TOFC, semantic feature space: $t_{(55)} = 3.03$, $p = 0.004$, $d_z = 0.40$). Detailed results of all associated tests are summarized in Table S4.4 and S4.5. In sum, the designed feature spaces reliably capture neural variance in their target ROIs, thereby validating the usefulness of these ROI specific feature space models.

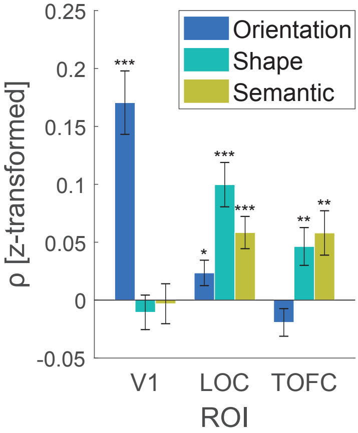


**FIGURE 4.2 Stimulus feature space models explain neural variance in target ROIs.**

Shown are RSA results, in terms of Fisher z-transformed Spearman's Rho (ρ), per ROI (V1, LOC, TOFC) and feature space model (blue = orientation feature space, green = shape complexity feature space, yellow = semantic similarity feature space). In V1 only the orientation feature space (blue) explains significant neural variance. In LOC all three models explain some neural variance, however numerically the shape complexity model (green) outperforms both other feature space models. In TOFC the semantic similarity and shape complexity feature space models explain significant neural variance, however the semantic similarity model explains numerically the most variance of neural responses. Thus, the designed feature spaces reliably capture neural variance in their target ROIs, validating the usefulness of the feature space models. Error bars indicate the SEM. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

## Simulation procedure

An overview of the simulation procedure is depicted in Figure 4.3A, and details are described in *Materials and Methods, Simulation*. In brief, we model neural responses to object stimuli using neural response functions, which model neural responses depending on the neuron's response tuning and the properties of the stimulus in feature space.

Biased sampling [150] and macroscale maps [151] are two leading account of how stimulus selectivity arises in fMRI voxel data. We used a simple implementation in line with these accounts by random sampling a limited number of neurons with different feature tunings to form voxels [105]. As a consequence of the limited random sampling, simulated voxels showed distinct response preferences for different stimuli akin to the responses seen in empirical fMRI data. For more details see: *Materials and Methods, Simulation, Simulating voxels*. We then presented to these simulated voxels the same stimuli, on the same number of trials, as to the human volunteers during the fMRI experiments.

Neural responses for expected stimuli were modulated according to six distinct models, depicted in Figure 4.3B. Two classes of modulations were employed. Gain modulations linearly scaled the responsiveness of neurons, without modulating the shape of the response function. Tuning modulations narrowed the shape of the response function in feature space, but did not affect the peak amplitude. Additionally, three distance functions were implemented, determining where in feature space, relative to the expected stimulus feature, the modulation was applied. For global models modulations were applied equally across neural populations in feature space. In local models neural populations tuned towards the expected feature value were modulated, while in remote models neural populations tuned away from the expected feature value were modulated. Thus, the local gain modulation model represented dampening (cancellation) accounts, and the remote gain modulation model population sharpening accounts.

Simulated data was analyzed using the analysis pipeline as designed for the empirical data, relying on the eight outcome metrics described above. The entire procedure was repeated for each model type across a large parameter grid (n = 7,820; for details see: *Materials and Methods, Simulation, Parameter grid*), extensively exploring the three free model parameters: $a$ (suppression magnitude), $b$ (effect of distance in feature space), and $\sigma$ (width of neural response function).

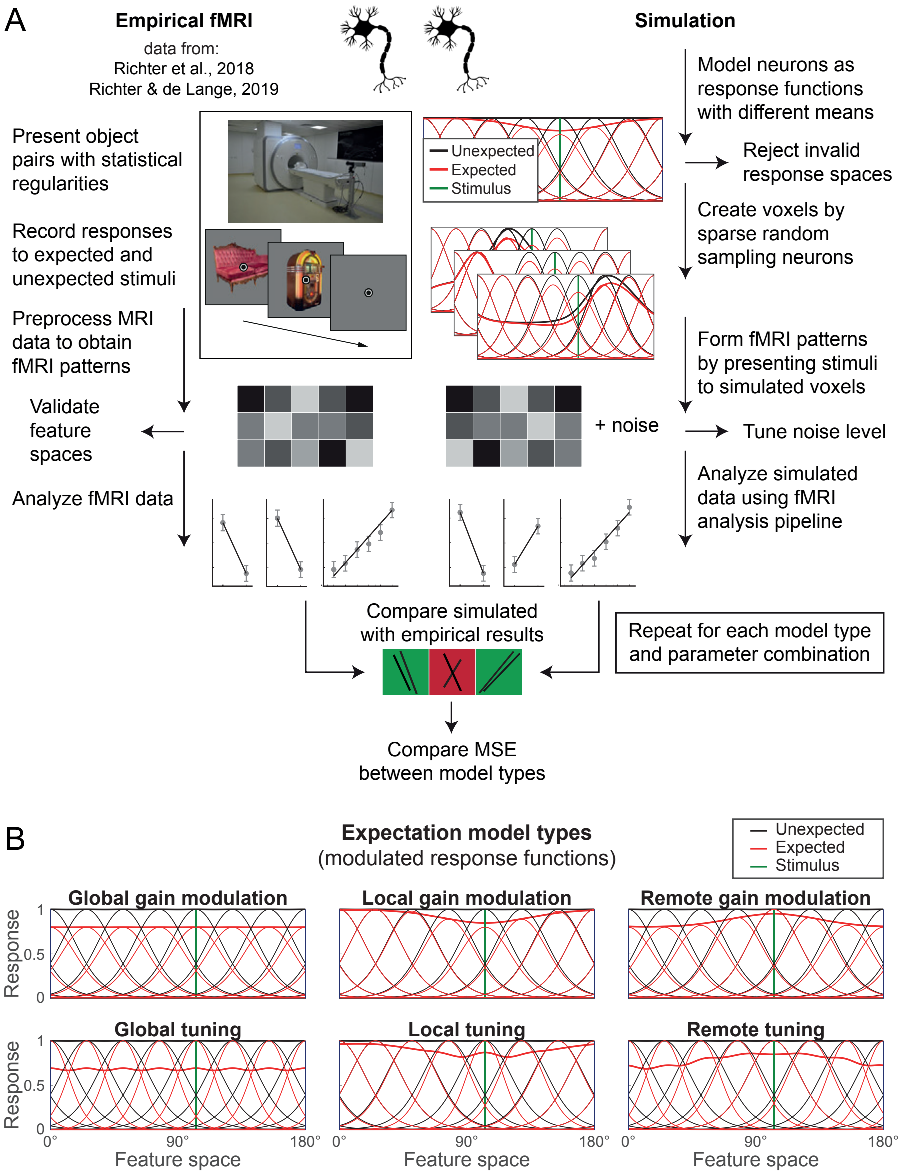

A
Empirical fMRI
data from:
Richter et al., 2018
Richter & de Lange, 2019
Simulation
Present object pairs with statistical regularities
Record responses to expected and unexpected stimuli
Preprocess MRI data to obtain fMRI patterns
Validate feature spaces
Analyze fMRI data
Model neurons as response functions with different means
Unexpected
Expected
Stimulus
Reject invalid response spaces
Create voxels by sparse random sampling neurons
Form fMRI patterns by presenting stimuli to simulated voxels
+ noise
Tune noise level
Analyze simulated data using fMRI analysis pipeline
Compare simulated with empirical results
Repeat for each model type and parameter combination
Compare MSE between model types
B
Expectation model types
(modulated response functions)
Unexpected
Expected
Stimulus
Global gain modulation
Local gain modulation
Remote gain modulation
Global tuning
Local tuning
Remote tuning
Response
1
0.5
0
0°
90°
180°
Feature space

**FIGURE 4.3 Simulation procedure and expectation models.**

(A) Overview of the empirical data acquisition and analysis, as well as the simulation procedure. On the left, the procedure for empirical fMRI data is depicted. Human volunteers were presented with object images, the identity of which was governed by statistical regularities, thereby making the objects expected or unexpected by virtue of the preceding image. The resulting fMRI data was analyzed using eight different outcome metrics (for details see: *Materials and Methods, Empirical fMRI data and MRI outcome metrics*). The right side illustrates the simulation procedure. Starting from the top, neural responses to different object stimuli were simulated in a ROI specific fashion, using empirically validated custom feature spaces (Figure 4.2). Responses to expected stimuli were modulated using six different models. A fine-tuned amount of noise was added to the simulated response patterns, based on estimates of noise in the empirical data. Next, simulated data was analyzed using the analysis metrics also used for the empirical fMRI data analysis. Finally, simulated and empirical results were compared in terms of the sign of the slopes per outcome metric, as well as in a more fine-grained fashion by calculating the mean squared error (MSE). For details see: *Materials and Methods, Simulation*. (**B**) Neural response functions, and their modulation by expectation. Depicted are illustrations of the six neural response modulation models. Thin black lines denote unmodulated response functions across feature space. Red lines are modulated responses. Thick lines indicate the normalized summed response. The depicted example is from V1, thus representing a circular feature space. Green shows the position of an example stimulus in feature space. Starting from the top left: Global gain modulation reduces the amplitude of the modulated response by a multiplicative factor (*a*) evenly across feature space. Local gain modulation (dampening model) reduces the amplitude by a multiplicative factor, however the magnitude of the response modulation depends on the distance between the expected stimulus and the response function (effect of distance is modulated by the *b* parameter). Remote gain modulation (sharpening model) is identical to local gain modulation, except that neural populations tuned away from the expected stimulus are modulated. The tuning models (bottom row) reduce the width of the response function, with the magnitude of the reduction controlled by parameter a, thereby resulting in a more selective response. As with gain modulation models, the three distance functions apply. For details see: *Materials and Methods, Simulation, Modulation by expectations*. Note that the modulation by expectation is conditional on an expected stimulus being presented, thus modelling a top-down modulation. As such, the red curves should not be seen as full tuning curves, but rather as an illustration how expectation differentially modulates different neural populations as a function of their tuning. See supporting Text S4.1 and Figure S4.1 for a discussion and illustration of alternative implementations.

## Simulation results

### *Voxel-level results can be accounted for by opposite neural models*

With the empirical fMRI results established, we performed the simulation. First, we analyzed results by comparing the sign of slopes per outcome metric between the empirical and simulated results, following the procedure from Alink et al. [105]. The rationale for this approach is that such qualitative interpretations of analysis results are used in most empirical fMRI and MEG studies. For example, improved classification accuracies of expected stimuli have been used as evidence for population sharpening [18,142], and decreased accuracies for dampening [143]; indeed, similar qualitative interpretations apply for the other metrics as well, e.g. [43,113].

In V1, all six model types could, at least under one parameter combination, match the sign of the slopes of all eight outcome metrics found in the empirical data. Similarly in LOC, all model types succeeded in fitting the sign of seven of the eight empirical outcome metrics. Finally, in TOFC, three model types fit the sign of all eight empirical results (remote gain modulation, local and remote tuning), while three model types matched seven outcome slopes (global and local gain modulation, global tuning). These results show that no single voxel-level outcome, nor the combination of all eight outcome metrics, was uniquely characteristic of sharpening or dampening, nor any of the other implemented neural mechanisms. Thus, relying on a qualitative interpretation of only the sign of the slopes of the voxel-level results had only limited utility for the inference about underlying neural modulations, because all six models could replicate the observed slopes of (almost) all outcome metrics across the three ROIs. This conclusion is surprising and contrasts with the results reported by Alink et al. [105], who observed that only one model could qualitatively account for all outcome metrics with a single set of parameters. Therefore they concluded that this model best explained the results overall. In our case, such qualitative reasoning alone does not suffice, because all models had at least one set of parameters that qualitatively matched (almost) all outcome metrics, potentially because we searched a finer and broader grid of parameters, and simulated stimuli throughout the modeled feature-space (see *Discussion*).

While the present data showed that a similar qualitative analysis does not suffice to uniquely identify the best model, results depicted Figure 4.4, also demonstrated that the proportion of parameter combinations that fit (almost) all outcome metrics differed substantially between the six model types. In all three ROIs, global and local gain modulations showed the most robust fit across different parameter value combinations, fitting a maximal number of eight outcome metrics in V1 with 93% and 86% parameterizations respectively, compared to 13% for the next best model type, local tuning. Similar results were evident in LOC, with the sign of seven metrics successfully fit by global gain modulations in 82% of parameterizations and 77% of local gain modulations, while local tuning did so only in 42% of parameter combinations. In TOFC, global gain modulations fit seven outcome metrics in 93% and local gain modulations in 81% of parameterizations, compared to local tuning with 51%. In short, these results suggest that global and local gain modulations are less sensitive to the exact parameter values in producing the observed fMRI results. This robustness in turn increases the probability of these two model types reliably explaining the observed empirical results. That said, there is a substantial number of parameter combinations under which competing models do reliably produce the observed fMRI results as well. Moreover, a slope of the same sign does not necessarily accurately describe how well the simulated results fit the empirical results, as slope coefficients can differ drastically. Therefore, a more fine grained, quantitative approach is necessary to evaluate the model fit.

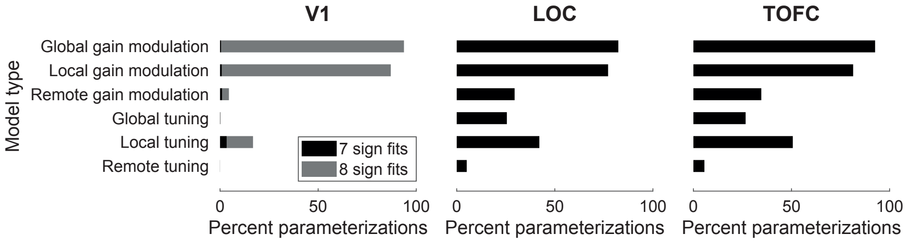


**FIGURE 4.4 Qualitative assessment of model fits.**

Displayed are the percentages of parameterizations for each model type that fit all (8 sign fits), or all but one (7 sign fits) of the empirical fMRI outcomes. In V1 (left), all model types fit all eight outcome metrics under at least one parameterization. Note: due to a very small percentage of fits, the 8 and 7 sign fits are barely visible for some models (e.g., remote tuning in V1). The percentage of good fits is noticeably larger for global and local gain modulations compared to all other model types. A similar, albeit smaller difference is also evident in LOC (middle) and TOFC (right), with a larger proportion of parameter combinations showing a fit to the sign of slopes for global and local gain modulation models. Worse fits, that is model parameterizations with less than seven outcome metric fits, are not displayed.

### *Perceptual expectations are best explained by a local gain modulation*

Next, we quantitatively analyzed the fit of the simulated to the empirical results by calculating the mean squared error (MSE) for each model type and parameter combination. In brief, we compared the relative slope of the simulated and empirical results for each outcomes metric. This difference in slopes was squared and the average per model type and parameter combination calculated. Thus, this MSE reflects how well each parameterization of each model types fits the empirical fMRI results. Results, depicted in Figure 4.5, show that in all three ROIs, V1, LOC and TOFC, the best fitting model type (lowest MSE) was local gain modulation (solid orange line). In fact, in LOC and TOFC, several parameterizations of the local gain modulation outperformed all other model types by a substantial margin. This suggests that the superior performance of local gain modulation is stable (i.e., not driven by noise) and robust to changes in the exact parameter values. Moreover, the mean MSE of the best 100 parameterizations (squares in Figure 4.5) of the local gain modulation model was lower than the mean MSE of all other model types. In sum, in intermediate and higher visual areas, LOC and TOFC, local gain modulations best explained the empirical results. In V1 results were less clear, with the global gain modulation model performing similar to local gain modulation. On average the 100 best global gain modulation models even outperformed local gain modulations, suggesting a broader, less specific gain suppression of neural responses in V1 compared to the local suppression in LOC and TOFC.

Remote gain modulation models, by contrast, performed poorly in all three ROIs. The optimal parameterization of the remote gain modulation model performed worse than any of the 100 best parameterizations of the local gain modulation in V1 and LOC, and worse than the ~20 best parameterizations in TOFC, again demonstrating a robustness of the results to noise and changes in the exact parameter values. Generally speaking, local modulations outperformed global and remote modulations in all ROIs, for both model classes (gain modulation and tuning), with the exception of global gain modulations in V1. Similarly, gain modulations outperformed tuning models in all ROIs, across the three distance domains (local, global, remote). In sum, our results favor local gain modulations, in line with dampening accounts, as underlying perceptual expectation suppression across intermediate (LOC) and higher visual areas (TOFC), and global or local gain modulations in early visual cortex (V1).

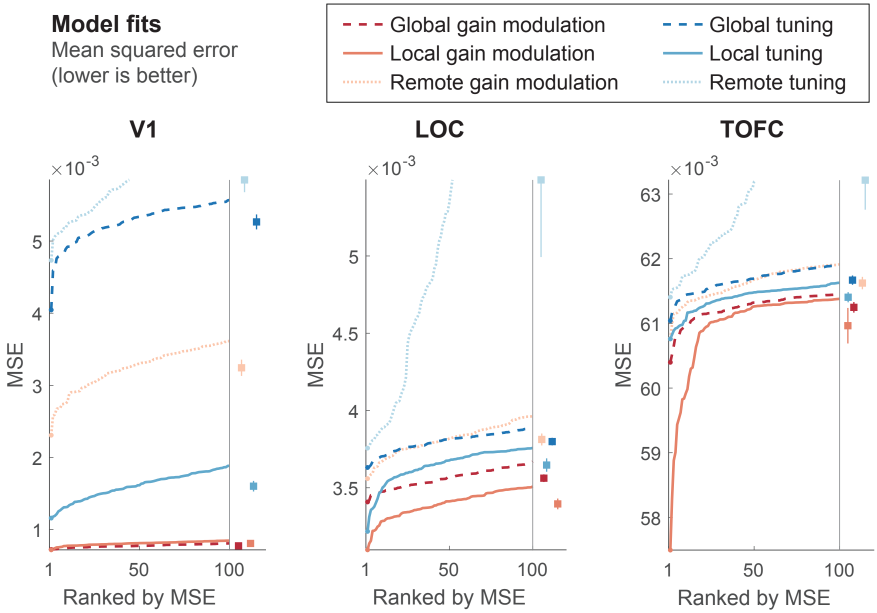


**FIGURE 4.5 Feature-specific local gain modulations best explain expectation suppression.** Displayed are model fits in terms of mean squared error (MSE) in the three ROIs (V1, LOC, TOFC) for the six model types. Modeled were two model classes: gain modulations (red-orange lines) and tuning modulations (blue lines), across three distance domains: global (dashed lines), local (solid lines), remote (dotted lines). Local gain modulations, representing dampening accounts, reduce the gain of neural populations tuned towards the expected stimulus features. In both, LOC (middle panel) and TOFC (right panel), local gain modulations outperformed all competing models in terms of the lowest MSE, thus providing the best model fit. In fact, several parameterizations of the local gain modulation models outperformed the next best model, indicating a robust superior fit. Moreover, in LOC and TOFC, the mean performance of the best 100 local gain modulation model parameterizations outperformed the mean of the best 100 parameterizations of any other model type (squares on the right in each panel). Error bars indicate 95% confidence intervals. In V1, both local and global gain modulations performed similarly, with local gain modulations constituting the best model type in terms of

the lowest MSE, while the mean of the best 100 global gain modulation parameterizations was lower than for local gain modulations. In sum, local gain modulations performed better than all other model types in LOC and TOFC, while both global and local gain modulations perform well in V1, with evidence for more robust fits for global gain modulations models in V1.

### *Suppression is local in higher visual areas, and global in early visual cortex*

In order to further explore the nature of local gain modulations, and the similar performance of local and global gain modulations in V1, we further investigated which parameter values resulted in the best model fits. Of particular interest were the $a$ (suppression magnitude) and $b$ (distance) parameter values for local gain modulations. To this end Figure 4.6, shows histograms of the parameter value distribution for the 100 best fitting global and local gain modulation models. The $a$ parameter represents suppression magnitude; i.e., the multiplicative gain modulation that expectations induce. As can be seen in Figure 4.6A, the median $a$ parameter of the best global gain modulation models was 0.8 in LOC and TOFC, and 0.85 in V1. Similarly, for the best local gain modulation models the median $a$ value was 0.75 in LOC, 0.67 in TOFC, and 0.85 in V1. Thus, on average (mean over ROIs and the two gain modulation models) the optimal $a$ parameter value was approximately 0.79, corresponding to a suppression of neural responses to approximately 79% of its unsuppressed response, if a stimulus was expected.

The $b$ parameter determines, in non-global models, the influence of distance between stimulus and neural populations on the suppression magnitude. Small $b$ parameter values are associated with localized feature-specific suppression, affecting only neural populations with similar response preferences. Large $b$ values on the other hand, correspond to global suppression, with local and remote models being identical to global models when $b = \infty$, implying that all neural populations are affected equally, irrespective of their tuning. In order to distinguish global from local/remote models, we limited the $b$ parameter grid to $b <= 2.3$ (i.e., approximately ¾ π; feature space spanned 0 to π. Thus, $b = 2.3$ constitutes a fairly broad, but not global, suppression profile; for an illustration and more details see: *Materials and Methods, Simulation, Modulation by expectations and Parameter grid*). As evident in Figure 4.6B, in LOC and TOFC, the median $b$ parameter of the 100 best fitting local gain modulation models was $b = 1.0$ and 0.9 respectively, suggesting a localized gain modulation of neural responses. In contrast, in V1 the median b parameter was 1.65, in a distribution of well-fitting $b$ values noticeably skewed towards large values. Statistical tests confirmed that the distribution of $b$ values was indeed significantly higher in V1 compared to LOC ($W = 3667$, $p = 9.6e\text{-}7$) and TOFC ($W = 4066$, $p = 3.2e\text{-}11$). In other words, local gain modulation models that did perform well in V1, were predominantly showing a wide-spread (i.e., fairly global) suppression of neural populations, compared to the more localized suppression in LOC and TOFC. These results support a similar conclusion as those depicted in Figure 4.5, which showed that global and local gain modulation models performed similarly well in V1, while global gain modulations were clearly inferior to the feature-specific local gain modulations in LOC and TOFC.

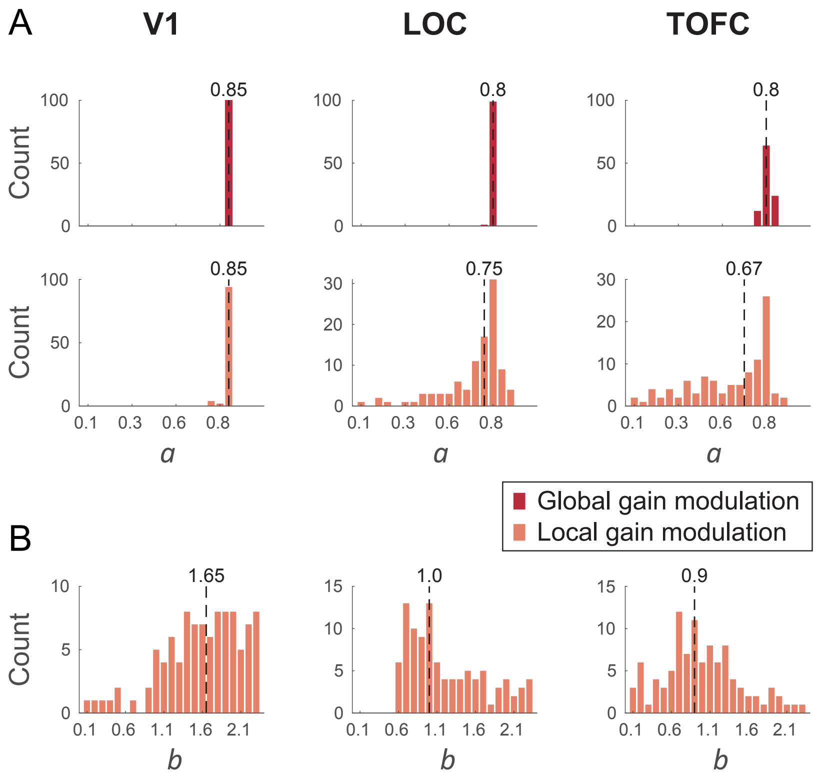


**FIGURE 4.6 Gain modulations are local in LOC and TOFC, and global in V1.**

Histograms of parameter values associated with the 100 best models for global (red) and local (orange) gain modulations in V1 (left), LOC (middle), and TOFC (right). Black dashed vertical lines indicate the median of the distribution. (**A**) Shows the count of specific *a* parameter values among the 100 best models. The *a* parameter reflects the magnitude of suppression. On average, an a parameter value of approximately 0.79 best fit results across both gain modulation models and all three ROIs, suggesting that response rates of neurons are reduced to ~79% of their unsuppressed response by perceptual expectations – albeit values differ somewhat between ROIs for local gain modulations. (**B**) Shows histograms of *b* parameter values of the 100 best local gain modulation models. Amongst the best models are larger *b* parameter values in V1 (right skewed distribution; median = 1.65), and smaller *b* values in LOC and TOFC (left skewed distributions with a smaller median LOC = 1.0 and TOFC = 0.9). The *b* parameter controls the effect of distance, with small *b* parameters being associated with localized suppression effects, while large *b* parameters reflect more wide-spread, and less feature-specific suppression. Thus, successful gain modulation models in LOC and TOFC tend to modulate neural responses in a localized feature-specific fashion, compared to a rather global modulation in V1. See *Materials and Methods, Simulation, Modulation by expectations and Parameter grid* for an illustration of the b parameter value range.

In sum, expectation suppression in higher visual areas appears to be a feature-specific local gain modulation, in support of dampening accounts. In contrast, a more global, feature-unspecific gain modulation seems to underlie expectation suppression in early visual cortex. Initially it may appear surprising that a simple suppression model, such as a global gain modulation, should account for the observed results in V1, as it is not in line with either sharpening or dampening. However, these results do in fact supplement earlier observations that reported stimulus-unspecific suppression in early visual cortex [145].

## Discussion

Predictions, based on statistical regularities in the sensory input, can be useful in guiding perception. In particular, predictions may aid to represent the world in a veridical fashion, as well as promote processing of novel and surprising information [44]. In the present study we investigated the neural mechanism underlying a widely reported neural signature of perceptual predictions, expectation suppression: the attenuation of responses to expected compared to unexpected stimuli (for a review see: [19]). On the one hand, population sharpening suggest that expectations sharpen sensory representations in line with expectations, by suppressing neurons tuned *away* from the expected stimulus [18,41,142]; modeled here as a remote gain modulation. On the other hand, the dampening (or cancellation) account proposes that expectations dampen sensory representations, by suppressing neurons tuned *towards* the expected stimulus [23,43,79,113,143]; modeled as a local gain modulation.

We tried to arbitrate between population sharpening and dampening accounts of perceptual expectations by employing forward models and a large range of fMRI outcome metrics previously used by empirical studies investigating expectation suppression [18,43,113,142,143]. Our approach comprised two steps: first, we established the effects of expectations in terms of these outcome metrics for three ROIs, throughout the ventral visual stream, based on a large (n = 56) combined analysis of fMRI data from two prior studies [113,145]. Next, we used forward models to quantitatively assess which underlying neural mechanism could best explain the observed effects of expectations. As neural mechanism we modeled the two accounts of interest, sharpening and dampening, as well as alternative models based on the wider literature. In brief, we show that perceptual expectations in the ventral visual stream are best explained by a feature-specific local gain modulation, in line with dampening. These results suggest that expectations, as investigated here, selectively suppress neurons tuned towards expected stimulus features, and may thereby serve to reduce information redundancy in sensory areas and highlight surprising, and novel information.

## No qualitative pattern of fMRI results is necessarily unique to sharpening or dampening

First, we demonstrated that all model types, both gain modulation and tuning models, across all three distance domains, could fit the empirical fMRI results on (almost) all fMRI outcome metrics. The utilized fMRI outcome metrics have been employed by previous studies to arbitrate between different accounts underlying expectation suppression [18,43,113,142,143]. In V1, all model types could, under at least one parameter combination, qualitatively fit all fMRI outcomes; i.e., the sign of the slope of the outcome metrics. Similar results were evident in higher visual areas, LOC and TOFC, with all model types matching the sign of at least seven of the eight metrics. Dampening and sharpening accounts make opposite predictions in terms of the neuronal population that is most suppressed by expectations. Thus, it is counter-intuitive that opposite neural-level modulations can qualitatively fit the same voxel-level results on a broad range of analyses, even under biologically inspired constraints (see: *Materials and Methods, Simulation, Response requirements*). These results highlight a crucial limitation of relying on heuristics in the interpretation of fMRI results, particularly if a limited number of fMRI analyses are utilized, echoing conclusions drawn by Alink et al. [105].

However, in contrast to Alink et al. [105], our results also show that relying on a purely qualitative interpretation of the results is not sufficient to reliably distinguish between population sharpening and dampening accounts, even if a combination of several fMRI outcome metrics is used. Indeed, all models could qualitatively fit the empirical results in V1, and a similar pattern was evident in LOC and TOFC. The additional flexibility of the models in our simulation is likely a consequence of modelling stimuli across the entire feature space, instead of only two specific feature values [105]. Moreover, a substantially finer resolution and broader scope of the explored parameter grid contributed to additional model flexibility, even though we also enforced additional response constraints based on plausible neural responses (see: *Materials and Methods, Simulation, Response requirements*). This versatility necessitates a more fine-grained, quantitative analysis of the model fits.

## Perceptual expectations dampen sensory representations

In higher visual areas, LOC and TOFC, local gain modulation models outperformed all other implemented models. In fact, not just the best performing model (smallest MSE) was local gain modulation in both ROIs, but this superior fit to the empirical data was stable and robust to variations in the precise parameter values of the model; i.e., several local gain modulation models fit the empirical data better than any other model. Moreover, the average MSE of the 100 best local gain modulation models was lower than the average MSE of the 100 best models of any other type in both LOC and TOFC, further supporting that local gain modulations best explain expectation suppression. Thus, our results, across intermediate and higher visual areas in humans, converge on a similar neural mechanism as underlying expectation suppression as previously proposed based on electrophysiological recordings in non-human primates [23,79]. Interestingly, previous work also shows that local gain modulations underlie stimulus adaptation as well [105], thus suggesting comparable neural modulation accounting for both phenomena, adaptation and expectation suppression. Similar neural modulation should however not be mistaken as an identity of the two phenomena [22,62,110].

Having established that local gain modulations underlie expectation suppression in the ventral visual stream, it is worth considering what functional role expectations may have in perception according to the dampening account. The hallmark of dampening (here modeled as local gain modulation) is a suppression of responses in neural populations tuned *towards* the expected stimulus features [23,43,113,143]. By suppressing neurons tuned towards the expected stimuli, a dampening of neural responses reduces redundancy in the sensory system. That is, if a stimulus was well predicted by internal models, there is no need to vigorously respond to that stimulus, as it presents little new information. Indeed, information is particularly relevant to an agent, in so far as it is novel information, because such information is valuable for updating internal models of the world, which in turn can promote adaptive behavior. Moreover, suppressing uninformative, well predicted input, may additionally preserve processing and attentional resources. Therefore, it seems adaptive that expectations can guide perception by suppressing expected input and highlighting unexpected stimuli; i.e., events that are informative.

## Reconciling sharpening and dampening accounts

As discussed above, the present results support the dampening account of expectation suppression, with expectations highlighting novel information, and deemphasizing expected, predictable input. These results may appear incompatible with the competing population sharpening account, which, in line with Bayesian views of perception [5,6], suggests that predictions sharpen representations in line with expectations. Indeed, our implementation of sharpening, remote gain modulation, performed poorly in all three ROIs. However, there are attempts to reconcile these seemingly incompatible accounts. For instance, hierarchical predictive coding theory proposes that sharpening and dampening occur in parallel but in different neural populations – namely, prediction and error neurons, which would reside in superficial and deep layers of cortex, respectively [12]. Note, however, that so far there has been no direct evidence for the existence of these two neuron types, and that this proposal does not explain why the present results would only reflect the dampening process occurring in the error neurons.

Alternatively, Press et al. [44] recently proposed that both processes, population sharpening and dampening, operate during different processing stages. The rationale is that initial processing relies on prior knowledge to sharpen sensory representations, followed by a late processing stage, dampening neural representations of the expected stimulus. Thereby, this account promises to unify results in the literature that initially appear incompatible (e.g., [18,41,142] vs. [23,43,79,113,143]). At the same time this proposal also acknowledges the adaptive value of predictions in fulfilling both challenges facing perception, veridical representations aided by prior knowledge, as well as using prior knowledge to reduce redundancy and help in information seeking and updating of internal models.

That said, the present results only provide evidence in line with the hypothesized late dampening stage. However, given that in the analyzed fMRI datasets (for details see: [113] and [145]) object stimuli were presented at full contrast, without visual noise, and for a fairly long duration (500 ms), it is conceivable that in this context a representational sharpening stage is of little relevance for veridical perception and good task performance. Consequently, the sharpening processing stage may have had little impact on the overall BOLD signal, which represents (indirectly) the integrated neural response over an extended time period. Thus, in the present data, the later representational dampening stage may have dominated the observed BOLD signal. On the other hand, in Kok et al. [18], participants performed a more perceptually demanding discrimination task, which may have placed emphasis on the sharpening stage, thereby resulting in the observed suppression of neural populations tuned away from the expected stimulus. Future work is required to directly assess the two stage processing account, and whether the dominance of one or the other process can be tipped by task demands and stimulus characteristics; e.g., perceptually challenging paradigms resulting in a sharpening of representations, overruling the dampening effect in the BOLD signal.

## Feature-unspecific suppression

While our results in higher visual areas support the dampening account, results in V1 do not so readily fit this interpretation. In fact, results in V1 suggest that perceptual expectation may suppress neural responses in a largely feature-unspecific fashion, with both global and local gain modulation models performing well. Amongst the best performing local gain modulation models in V1, broader suppression profiles (larger *b* parameter values) were evident than in LOC and TOFC. In other words, well performing local gain modulations models in V1 were more global in their suppression, affecting neural populations across different feature tunings. Moreover, in contrast to higher visual areas, on average the best global gain modulation models outperformed all other model types, including local gain modulations, in V1. These results may initially seem surprising, given that neither of the outlined accounts, dampening or sharpening, predicted this pattern of results. However, we previously suggested that expectation suppression in V1 may indeed be feature-unspecific [145]. In particular, we showed that voxels in V1, but not in LOC and TOFC, which were *not* significantly activated by the stimuli showed comparable amounts of expectation suppression to stimulus-driven voxels [145]. Combined with the present results, it appears that perceptual expectations, following statistical learning of object pairs, may result in feature-unspecific suppression of neural responses in early visual cortex, affecting neurons irrespective of their feature tuning.

Given that the here investigated expectations concern object identity predictions, it is intriguing to note that only object selective visual areas (LOC and TOFC) yielded feature-specific suppression, while early visual areas did not. Thus, it is plausible that only cortical areas whose response properties are particularly diagnostic of an expectation confirmation/violation show feature-specific expectation suppression, while lower visual areas, in this case V1, inherit an unspecific feedback signal from higher visual areas. This interpretation of the feature-specificity of perceptual expectations depending on stimulus characteristics and neural tuning properties, is further supported by noting that Kok et al. [18] presented oriented grating stimuli and accordingly observed representational sharpening in early visual cortex; i.e., the visual area particularly selective for stimulus orientation predictions. Thus, feature-specific perceptual expectation may manifest in the cortical areas tuned for features particularly diagnostic of an expectation violation or confirmation, while lower sensory areas in turn may only inherit unspecific surprise signals during subsequent feedback.

## Limitations

The present results have to be interpreted with some limitations in mind. First, while we synthesized empirical fMRI data from two separate studies, resulting in a large dataset (n = 56) including separate tasks, it is crucial to consider the type of perceptual expectations investigated in these datasets. In particular, both studies probed visual expectations extracted incidentally from statistical regularities. Thus, it remains unclear whether the perceptual expectations studied here, following incidental statistical learning, involve similar neural mechanisms and consequences as explicitly learned expectations. It is possible that different routes towards the acquisition of statistical regularities exist, relying on different neural mechanisms [47], thus raising the question whether the resulting sensory consequences may also differ.

The implemented forward models necessarily involve an oversimplification of the neural mechanisms and responses in visual cortex. For example, we relied on limited random sampling of neurons to form voxels with different response profiles, mirroring the large scale response preferences evident in empirical fMRI data. While our approach is in line with biased sampling [150] and macroscale maps [151], two leading accounts of voxel selectivity, there certainly are more refined and biologically plausible implementations. Additionally, we chose to bypass the complex dynamics involved in the hemodynamics underlying the BOLD signal, which constitutes a significant oversimplification. Moreover, alternative theories of stimulus selectivity may not be in agreement with our model (e.g., stimulus vignetting [152]), and thus our results cannot speak for such mechanisms. That said, the primary assumptions of our implementation appear fairly robust; i.e., a monotonic relationship between neural activity and voxel responses, and that voxel-level response preferences indirectly reflect neural tunings. Therefore, even though our model constitutes an oversimplification of the associated neural and hemodynamic processes, if these two core assumptions hold, our results are nonetheless likely to be informative about how expectations modulate neural responses.

Another limitation worth considering are the utilized feature space models. There is ample room to improve the feature space definitions with more complex implementations, and thereby increase the amount of explained variance. This may in turn help in yet more clearly distinguishing between the different predictions of the neural mechanism underlying expectation suppression. That said, we did show that even our simple custom feature spaces reliably capture neural response variance in their target ROIs, and are sufficient to yield distinguishable characteristics between population sharpening and dampening accounts.

By casting a broad parameter grid, and assessing results according to a variety of parameter combinations, instead of only the best fitting parameterization per model, we demonstrate the stability and robustness of our results to noise and changes in parameter tuning. This in turn also increases the robustness of the drawn conclusions. By inspecting well-fitting parameter values, such as the magnitude of suppression ($a$ parameter) among the best performing local and global gain modulation models, we noted that the average suppressed response is approximately

79% of the unmodulated response. Interestingly, this number well matches the magnitude of expectation suppression reported by Kaposvari et al. (Figure 6A/B in [26]) in terms of multi-unit firing rates recorded in monkey IT. While no formal link can be established, this convergence in suppression magnitudes between simulation and electrophysiological recordings provides additional confidence in the validity of the present simulation results.

Finally, the present results should not be seen in isolation. Using a novel approach in the investigation of expectation suppression, moving from neural-level models to voxel-level data, we converge on a similar conclusion as previous studies using a variety of other analyses and recording methods [23,43,79,113,143]. Combined these results provide robust evidence in favor of the dampening account.

### Conclusion

In sum, we show that, in intermediate and higher visual areas, perceptual expectations, following statistical learning of associations between object images, result in a feature-specific suppression of neural responses. This feature-specific suppression is particularly affecting neurons tuned towards the expected stimulus features. As a result this suppression dampens neural representations of expected stimuli, thereby potentially reducing redundancy in sensory cortex and emphasizing processing of surprising, novel information. Additionally, feature-unspecific suppression occurs in lower visual areas, such as V1, possibly as a consequence of unspecific feedback from higher visual areas. Whether feature-specificity depends on the type of predicted stimuli remains to be investigated. Moreover, whether the here supported dampening can operate in concert with a sharpening of representations during different stages of visual processing [44] poses an intriguing avenue for future research.

## Materials and Methods

### Empirical fMRI data

This section briefly describes the experimental protocol of the two fMRI datasets; for a full description see: Richter and de Lange [145] and Richter et al. [113].

#### *Experimental paradigm*

In both experiments, participants (n = 34 and n = 22 after data exclusion) were presented with two full-color object images in quick succession. Each image was presented for 500ms, without interstimulus interval, and an intertrial interval of approximately five seconds; Figure 4.7A depicts a single trial. Crucially, the identity of the trailing (second) image was predictable given the identity of the leading (first) image. Thus, each trailing image could either be expected or unexpected given the leading image. In Richter and de Lange [145], the transition matrix during a learning session consisted of 12 leading and 12 trailing images with deterministic associations (i.e., only expected pairs) on a total of 960 trials. During fMRI scanning, a subset of six by six images was shown using probabilistic associations (50% reliability; i.e., the expected image was five times more likely than any unexpected image; Figure 4.7B, left panel) on a total of 240 trials. Transitions were task-irrelevant during the learning session (unpredictable oddball detection), but could aid task performance during fMRI scanning (classification of trailing images). In Richter et al. [113], eight leading and eight trailing images were shown (Figure 4.7B, right panel). During the learning session (2,012 trials; 100% reliability) as well as during fMRI scanning (512 trials; ~56% reliability) statistical learning was incidental; i.e., the statistical regularities were not related to, or helpful in performing the task (unpredictable oddball detection). Only non-oddball trials (456 trials), without any behavioral responses, were analyzed. Additionally, both studies had one localizer run during which stimuli were presented in an expectation neural context for ~12 seconds, one at a time, flashing at 2 Hz.

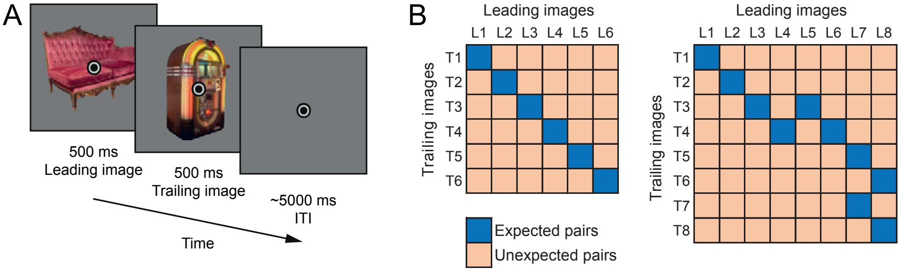


**FIGURE 4.7 Experimental paradigm and image transition matrix.**

(**A**) Depicts a single trial. A leading image (500ms) is followed by a trailing image (500ms), the identity of which is (un-)expected given the leading image. The trials end with an ITI of ~5000ms. (**B**) Shows the image transition matrices determining the association between images. Expected image pairs are denoted in blue. Each trailing image occurs as expected and unexpected image, with expectation status depending only on the leading image on a given trial. The transition matrix on the left is from the experiment reported in Richter and de Lange [145], while the matrix on the right is from Richter et al. [113].

### *fMRI data acquisition*

Data were acquired on a 3T Prisma [145] and 3T Skyra [113] scanner, using 32-channel head coils. In both cases, a whole-brain T2*-weighted multiband sequence was used to acquire functional MRI data. Data from Richter and de Lange [145] was acquired using a multiband 6 sequence with 2mm isotropic voxel size, and data from Richter et al. [113] with a multiband 8 sequence with 2.4mm isotropic voxel size. T1-weighted images were acquired in both studies using a magnetization prepared rapid gradient echo sequence (MP-RAGE) sequence with 1mm voxel size.

### *fMRI data preprocessing*

Preprocessing of the empirical fMRI data was performed using FSL 6.0 (FMRIB Software Library; Oxford, UK; www.fmrib.ox.ac.uk/fsl; [87]; RRID:SCR_002823). The preprocessing pipeline included: brain extraction (BET), motion correction (MCFLIRT), and temporal high-pass filtering (128 s). No spatial smoothing was applied, as voxel patterns were of primary interest. Functional images were aligned to the middle volume of the localizer run. All analyses were performed in native space in order to avoid unnecessary data interpolation.

### *fMRI data preparation*

The preprocessed fMRI data was further analyzed using the least squares separate approach by Mumford et al. [89] and Turner et al. [153]. A separate GLM was fit for each trial, consisting of one regressor of interest, modelling the response to the stimuli on the current trial. Additionally, regressors of no interest were added, consisting of a regressor per trailing image type (excluding the current trial), one regressor modelling events of no interest (instruction events), and 24 motion regressors (FSL's standard + extended set of motion parameters). Regressors were convolved with a standard double-gamma HRF. Finally, the parameter estimates for each trial and ROI were extracted separately, which constitute the pattern of responses to the stimuli presented on each trial.

### *Region of interest masks*

The same region of interest (ROI) masks were used as described in Richter and de Lange [145] and Richter et al. [113]. In brief, three ROIs were defined a priori: primary visual cortex (V1), object selective lateral occipital complex (LOC), and temporal occipital fusiform cortex (TOFC; akin to inferior temporal cortex). All three ROIs were defined anatomically and functionally. Moreover, ROI masks were further constraint to the 200 (data from [113]) or 300 (data from [145]) most informative voxels for decoding object identity using independent localizer data. Thus, the ROIs represent stimulus responsive voxels, across three different levels of the ventral visual stream. All three levels of the hierarchy were considered interesting, as it was not clear whether expectations may modulate responses in similar or distinct ways across the visual hierarchy.

## fMRI outcome metrics

The empirical and the simulated fMRI data were analyzed using the same analysis pipeline. In total, eight different outcome metrics were assessed. The reasoning for relying on this large number of diverse outcome metrics is based on Alink et al. [105], showing that, in the context of stimulus adaptation, neural models can show great flexibility in fitting empirical fMRI results, and that only by combining a range of outcome metrics one can successfully distinguish between the best performing models. We modified and extended the set of outcome metrics, resulting in the analyses summarized below. The utilized outcome metrics are based on previous studies investigating the neural mechanism underlying expectation suppression, building on a diverse set of studies supporting sharpening and dampening accounts.

### *Mean amplitude modulation (MAM)*

MAM probes the univariate differences in response amplitude between expected and unexpected stimuli. Thus, this metric indexes the commonly reported expectation suppression effect [18,25,43,106,113,142,145].

### *Within-class correlation (WC)*

WC assesses the correlation of neural patterns between different presentation instances of the same object stimulus, and potential difference in the size of this correlation between expected and unexpected occurrences of the stimuli. Thus, a large WC coefficient indicates that the same stimulus, presented on different trials, is represented in a similar fashion.

### *Between-class correlation (BC)*

Similar to WC, BC measures the correlation in neural responses. BC concerns the correlation in representations between different object stimuli. A low BC thus indicates that different stimuli elicit dissimilar response patterns. Parameter estimates are z-scored before WC and BC calculation. The two correlational metrics (WC, BC) are similar to representational analyses used by Blank and Davis [43].

### Classification performance (CP)

CP is defined as the difference between WC-BC, similar to the classification approach originally outlined in Haxby et al. [154]. A higher CP thus indicates that object representations are more distinct.

### Support vector machine classification (SVM)

SVM is defined as the decoding accuracy using linear SVMs. Object identity was decoded after SVMs were trained on independent localizer data. While different in the underlying method, the interpretation of SVM is similar to CP, in that a higher classification accuracy indicates more distinct neural representations. As above, data was z-scored. The classification metrics (CP, SMV) are akin to analyses used by Kok et al. [18], Yon et al. [142] and Han et al. [143].

### Amplitude modulation by amplitude (AMA)

AMA concerns the magnitude of the amplitude modulation (i.e., expectation suppression = $\text{response}_{\text{unexpected}}$ - $\text{response}_{\text{expected}}$) as a function of mean voxel amplitude. In other words, it expresses whether the amount of expectation suppression increases (or decreases) as a function of the average responsiveness of a voxel within an ROI. This metric thereby indexes whether expectation suppression scales with general responsiveness. Data was binned into 10 equally sized bins. The responsiveness ranking was established on independent localizer data. This analysis is based on Alink et al. [105].

### Amplitude modulation by selectivity (AMS)

AMS, similar to AMA, also expresses the magnitude of expectation suppression, but as a function of voxel selectivity. Selectivity is established based on independent localizer data, by fitting a GLM to each voxel's response regressed onto the response amplitude ranked images. Thus, a highly selective voxel responds strongly to some images and weakly to others, while a low selectivity voxel responds similarly to all images. AMS thereby assesses whether the response selectivity of a voxel correlates with the amount of expectation suppression experienced by that voxel. Note: the ROI masks contain the most informative voxels concerning object identity decoding (see *Region of interest masks*). Thus, all voxels in the ROI are (likely to be) stimulus driven. The AMS metric is based on Richter et al. [113] and Alink et al. [105].

### *Image preference (IP)*

IP indexes the amount of expectation suppression within each voxel as a function of image preference. Image preference is established based on localizer data, and ranks the response of a voxel to each trailing image by amplitude (i.e., image preference rank). Thus, IP expresses whether the magnitude of expectation suppression differs within a voxel depending on whether the displayed image is a preferred or non-preferred stimulus for this voxel. The IP outcome metric is based on Richter et al. [113] and Kok et al. [18].

## Feature space

For each ROI a feature space was defined on a neural response theoretical basis. V1 neurons are thought to be orientation selective [33], and thus V1 feature space was defined by the predominant orientation of the stimulus. LOC has been shown to represent shape complexity [34], which formed the basis of the LOC feature space. TOFC represents complex visual features, which appear related along semantic categories [149], thereby suggesting a feature space representing semantic similarity. Thus, for each ROI a one dimensional feature space was constructed, along which each object stimulus could be expressed as a point in feature space. Finally, each feature space was validated using representational similarity analysis (RSA), performed on independent localizer fMRI data. In order for a feature space to be considered usable, it should account for a statistically significant amount of neural response variance in its designated ROI (see *fMRI data analysis for feature space validation*).

### *V1 feature space*

For primary visual cortex, feature space was defined by the predominant orientation of each object stimulus. To this end Gabor filters of different frequencies (from 4/√2 to the hypotenuse of the length of the input image; [155]) and orientations (in steps of 20 degrees) were constructed, and the Gabor energy for each orientation extracted. Energy was averaged over the different frequencies and the orientation with maximal energy was used. This orientation thus represents the maximal orientation energy present in the object stimulus and thereby determined the positon of a stimulus in V1 feature space. Figure S4.2 shows the arrangement of the object stimuli in orientation feature space. Given that orientation is circular, this feature space was modeled as circular (i.e., feature space ranged from 0 to π). The resolution of feature space was set to 180 * π.

### *LOC feature space*

For LOC, feature space was defined by shape complexity. Following Vernon et al. [34], we calculated several metrics describing the complexity of each object's shape. These metrics included (1) the number of concavities, as well as (2) the area of these concavities, (3) the area and (4) perimeter of the smallest convex hull encompassing the object, and (5) the area of the smallest circle enclosing the object, as well as (6) the ratio of the silhouette of the object and the smallest circle. In brief, the number and area of the concavities describe shape complexity, because the number of concavities is directly related to the number of protrusions, which in turn is considered a metric for complexity (i.e., more complex shapes have more protrusions). Similarly, the area and perimeter of the smallest convex hull, encompassing the object, will be larger for complex objects with more protrusions. The area and ratio of the smallest circle encompassing the object relative to the silhouette indexes how compact an object is, with less compact (complex) objects yielding larger values. Because we aimed to describe feature space along one dimension for each ROI, the shape complexity metrics were subjected to a principle component analysis. Finally, the first principle component was extracted and used as the primary shape complexity descriptor. The arrangement of the stimuli in shape complexity feature space is depicted in Figure S4.3. As the shape complexity metric is arbitrary in its range relative to LOC responses (only relative positions and distances are interpretable), the same feature space resolution was used as for V1. However, LOC feature space was not circular, as the shape complexity dimension arranges objects from the most to the least complex (i.e., non-circular).

### *TOFC feature space*

For TOFC, feature space was defined by semantic similarity. The central idea is that TOFC responses are related to complex visual feature, which are correlated along semantic categories [149]. We used a multiple-arrangement task [156], in which a separate sample of participants (n = 32; 16 females; age 26.9 ± 8.5 years) arranged object stimuli in an arena by their semantic similarity. This behavioral study, like the MRI studies, followed institutional guidelines of the local ethics committee (CMO region Arnhem-Nijmegen, The Netherlands; METC no. 2014-288), including informed consent. Participants were instructed to arrange each object display by the similarity of the objects. It was emphasized that the arrangement should be made along semantic/categorical similarities, and not by low level features. Distances between objects thus represent the semantic dissimilarity between each object pair. Sixteen participants arranged the object arrays that the fMRI participants in Richter et al. [113] were exposed to, while the other 16 participants arranged the objects shown to participants in Richter et al. [145]. Each arrangement trial was the object matrix shown to one participant in the fMRI studies, thus constituting the context in which participants were exposed to the object stimuli, as semantic similarities are sensitive to context (e.g., cats and dogs are more similar to one another in a set of images also containing inanimate objects, but more dissimilar in a set comprised only of mammals). In addition, randomly sampled arrangements, as well as one arrangement of all object images in the database for each study, were appended until 75 minutes of experiment time passed. Finally, the individual representational dissimilarity matrices (RDMs) were collapsed by averaging, thus resulting in one RDM for the stimulus set of each fMRI experiment, which describe the human-rated, semantic dissimilarity (distance) of each object to each other object. Finally, multidimensional-scaling was used to extract a single dimension to describe the position of each object along the semantic similarity feature space. Figure S4.4 depicts the object stimuli arranged along the subjective similarity feature space. The precise values of this TOFC feature space are arbitrary, and only relative positions are meaningful. As with LOC, the feature space was non-circular.

### *fMRI data analysis for feature space validation*

Each participant's localizer run was analyzed in an event-related approach using FSL FEAT, modeling each object stimulus as a regressor of interest. Twenty-four motion regressors (FSL's standard + extended set of motion parameters), as well as other regressors of no interest (instruction screen) were added to the model. The contrasts of interest were the parameter estimate maps representing the responses to each object image compared to baseline (no visual stimulation). These contrast parameter estimates were extracted for each ROI and used to validate the three above described feature spaces.

In order to validate the feature space for each ROI, we performed RSA. In brief, for each participant of the MRI experiments, the positions in feature space of the objects shown to that participants were determined using the above outlined feature space models. From these positions in feature space, a RDM was calculated describing the dissimilarity between each object pair in the relevant feature space; i.e., for V1 distance in predominant orientation, for LOC distance in shape complexity, and for TOFC distance in semantic category. Next, the neural RDM was constructed by extracting the parameter estimates from the localizer run for each ROI separately (also see: *Region of interest masks*). The parameter estimates were z-scored and pairwise correlations between object representations were calculated. Thus, this RDM (1 – correlation) indexed the neural representational dissimilarity during the localizer run. Finally, we correlated the neural and feature space RDMs for each participant and

ROI using Spearman's rank correlation. Thus, this correlation coefficient describes the correlation between the feature space RDM and the neural RDM for each participant and ROI. After Fisher z-transforming the obtained correlation coefficients, data was combined across participants by subjecting the obtained correlation coefficients to a two-sided, one-sample t-test, comparing the correlation coefficients against zero (i.e., no correlation). A significant, positive correlation would thus indicate that the constructed feature space does account for variance in the neural responses, which was considered a requirement for each feature space in its designated ROI.

## Simulation

The following section first describes the simulation of neural responses and their modulation by expectations. Next, we outline the sampling of neurons to voxels, as well as the estimation and addition of noise to the voxel responses. Finally, requirements imposed on simulated neural responses are described, as well as the analysis of model fits.

### *Neural responses*

We modeled neural responses using neural response functions in ROI specific feature spaces. In the simplest case, a neural response was described as a Gaussian distribution with a mean in feature space and a standard deviation (σ). The full feature space was covered with eight neural response functions, each with a different mean, following the implementation of Alink et al. [105]. The maximal response of each neural population was normalized to one. Given that orientation feature space is circular, the circular normal (von Mieses) distribution was used for modelling V1 responses. Figure 4.8 (left panel) shows an example for V1. As the veridical standard deviation (σ) is unknown, σ was a free parameter in the model.

Because non-circular feature spaces yield significantly lower summed responses at their extremes, due to a lack of overlap of response functions beyond the boundary of the feature space (see Figure 4.8, right panel), each non-circular feature space was clipped to the central region of feature space in which the summed response exceeded at least 95% of the maximal response. In other words, this criterion avoids that some stimuli elicit a significantly reduced response simply due to its position at the extremes of feature space, as there is no theoretical or bilogical reason why a stimulus at the end of, for example the semantic category feature space, should elicit lower responses in TOFC than a stimulus in the center of that feature space.

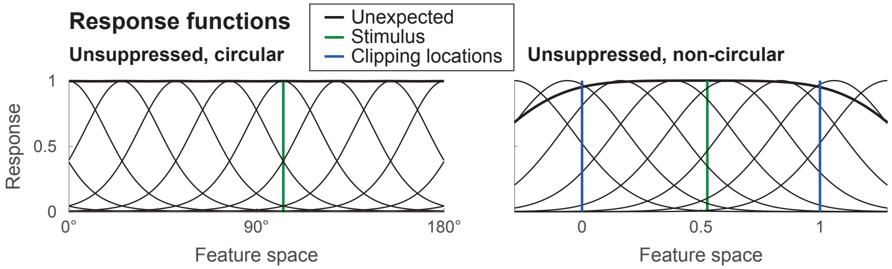


**FIGURE 4.8 Unsuppressed neural response functions.**

Depicted is an example set of unmodulated neural response functions (i.e., response to an unexpected stimulus). The left panel shows response functions covering a circular feature space (V1), while the right panel shows a non-circular feature space (LOC or TOFC). The thin black lines are individual response functions with different means along the feature space. Summed responses (normalized to one) are shown as the thick black line. The green line denotes an example stimulus. Blue lines show the clipping boundaries of the non-circular feature space – i.e., the region of feature space in which all stimuli are placed. The boundary is determined for each non-circular feature space by the minimal response criterion of 95% (see: Neural responses).

### *Modulation by expectations*

Next, neural responses were modulated by expectations, if a trial contained an expected trailing image, thus implementing a modulation of the neural response following a top-down modulation. This modulation can for instance result after recurrent message passing, in line with predictive coding accounts, representing a more extensive resolution of prediction errors for expected input. We did consider, and ruled out, alternative model implementation in which expectations are instantiated irrespective of an expected stimulus being shown. However, only local (dampening) models could result in expectation suppression, the core phenomenon of interest, in these alternative implementation. Thus, these alternative implementations rule out remote (sharpening) and global models by design. For a more detailed discussion and illustration see supplemental Text S4.1 and Figure S4.1. In the present stimulation, the suppression of neural responses by expectations was modeled using two primary classes of modulations. Gain modulation models reduced the response of a neuron in a linear fashion. The magnitude of suppression was determined by a free parameter ($a$). In the simplest case, $a$ was an unspecific multiplicative reduction; i.e., if $a$ = 0.7, the maximal response was reduced to 70% and all other responses were scaled proportionally in the global gain modulation model. On the other hand, tuning models modulated responses by reducing the width of the response functions. Thus, this modulation did not affect the maximal response of a neuron, but sharpened its response by making the responses more selective, with the $a$ parameter determining the extend of this tuning modulation.

Besides the two main model types (gain modulation and tuning), we also modeled three different distance functions. Distance functions determined where the effect of the response modulation occurred, relative to the expected stimulus in feature space. Global models affected the response functions equally across parameter space; i.e., global models do not have the parameter influencing the distance function. Local models exerted the modulation for neural populations close to the expected stimulus in feature space, and reduced in their modulation strength the further a neural population was tuned away from the expected stimulus in feature space. Remote models were opposite to local models, as they exerted the influence at the opposite side of feature space from the expected stimulus. A free parameter (*b*) influenced the distance over which the response modulation changed; i.e., large *b* parameter resulted in a broad influence across feature space. In fact, remote and local models are equivalent to global models if $b = \infty$. Figure 4.9 illustrates the effect different *b* parameter values have on suppression magnitudes across feature space.

The combination of two model types and three distance functions resulted in six models, which were used to describe response modulations by expectations. Thus, the response to an expected stimulus was the sum of the modulated response functions at the point in feature space of that expected stimulus. Figure 4.3B shows the six models in response to an example stimulus.

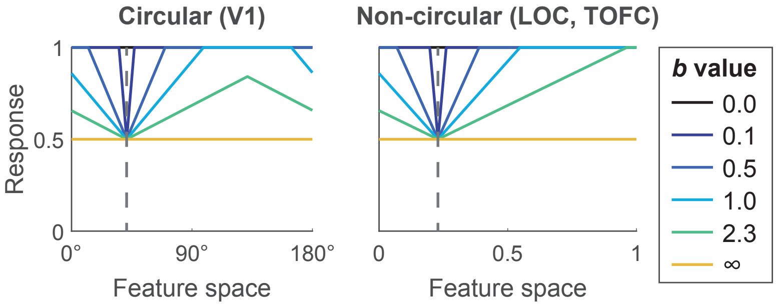


**FIGURE 4.9 Influence of the b parameter on suppression magnitude.**

Illustration of response suppression as a function of distance from an expected stimulus in feature space for different values of *b*. The dashed vertical lines indicate the position of an example stimulus in feature space. The example model is local gain modulation with a fixed *a* parameter value of $a = 0.5$ for illustration purposes. As the value for *b* increases (brighter colors), the broader the suppression profile becomes. For $b = \infty$ local/remote gain modulations are equal to global gain modulation. Grid search for b was limited to $0.1 >= b <= 2.3$. Note the extreme locality of suppression for small *b* values. For example for $b = 0.1$ neuronal populations tuned ~5 degrees away from the expected stimulus orientation would not experience any suppression. On the other extreme, for large b values, such as $b = 2.3$, local gain modulations would still substantially (suppression below 80% of responsiveness) affect neural populations tuned to exactly the opposite orientation of the expected stimulus (i.e., 90 degrees away in V1 feature space).

### *Model formulation*

Model types:
Gain modulation: $f_i(j)=c(i,j)\times g(x_j;\mu_i,\sigma)$
Tuning: $f_i(j)=g(x_j;\mu_i,c(i,j)\times\sigma)$

Distance functions:
Global: $c(i,j)=a$
Local: $c(i,j)=\min(1, a+|\frac{d(i,j)}{b}|(1-a))$
Remote: $c(i,j)=max(a,1-|\frac{d(i,j)}{b}|(1-a))$

Where,
$f_i(j)$ = activity of neural population i given stimulus j.
$c(i,j)$ = suppression given stimulus $j$, neural population $i$, and applicable distance function.
$x_j$ = position of stimulus $j$ in feature space.
$\mu_i$ = mean of neural population $i$ in feature space.
$\sigma$ = width of response function (free parameter).
$a$ = suppression magnitude parameter (free parameter).
$b$ = distance parameter (free parameter).
$d(i,j)$ = distance of stimulus $j$ to the mean of the neural population $i$ in feature space.

### *Parameter grid*

Given the above formulation, the present simulation has one free parameter determining the unmodulated response: the width of the neural response functions (σ). Furthermore, the modulated response functions contain two additional free parameters: the amount of suppression (*a*) and the effect of distance in feature space (*b*). A wide grid search was utilized to cover plausible parameter combinations (also see: *Response requirements*). The σ parameter value ranged from 0.1 to 6. Given the feature space width of π, these values represent a large range of σ values. Step size for σ was 0.1 from 0.1 to 1, 0.5 from 1.5 to 3, and 1 from 4 to 6. The *a* parameter was explored from *a* = 0.05 to 1, in steps of 0.05. An *a* parameter value of 0.05 corresponds to maximal suppression (e.g. for global gain modulation models this would result in a reduction of neural responses by 95% due to expectations), while *a* = 1 corresponds to no modulation by expectations. The final parameter, *b*, spanned values from *b* = 0.1 to 2.3 (i.e., approximately ¾ π), in steps of 0.1. Small *b* values represent maximal locality of modulations, while for *b* = ∞ local/remote would be identical to global models. We limited *b* to 2.3 in order to properly distinguish global from non-global models. However it should also be noted that $b = 2.3$ already represents a fairly global modulation, given that feature space ranges from 0 to $\pi$ (see Figure 4.9). Combined, the three parameters resulted in a total grid size of 7820 parameter combinations, which were explored for each of the six model types. As a grid search was utilized, two concerns need to be addressed. One, the parameter grid must cover the whole parameter space of interest, that is, the explored grid needs to be broad enough and/or bounded by theoretical or mathematical reasons. Two, the step size of explored values must be sufficiently small to accurately sample the error landscape. We addressed both concerns by showing that the error landscape was smooth and contained the minima well within the explored bounds, which additionally were limited by theoretical considerations. Results and an additional discussion of the parameter ranges are presented in Figures S4.5, S4.6, S4.7 and S4.8.

### *Simulating voxels*

Biased sampling [150] is arguably a leading account of how stimulus selectivity arises in fMRI voxel data. Simplified, the idea is that voxels pool over millions of neurons in a biased fashion, with different neural tunings being overrepresented in different voxels. Alternatively, global biases (maps) have been suggested to underlie the large scale response preferences evident in voxel-level data [151]. As in Alink et al. [105], we used a simple implementation in line with both macroscale maps and biased sampling, by random sampling a limited number of neurons with different feature tunings to form simulated voxels (eight per voxel). Spatial information was not modeled in the present simulation, as none of the ROI based analyses methods utilize spatial information beyond classical multivariate pattern analysis. The consequence of this limited random sampling procedure was simulated voxel-level data, which showed response preference for different stimuli. While this approach certainly constitutes a crude method for sampling neurons to form voxels, bypassing the complexities involved in the mapping of neural responses to BOLD signals (hemodynamics, etc.), it does succeed in creating voxels, which mirror the response profile seen in empirical fMRI data. In fact, the central assumptions of this approach are only that voxel-level selectivity reflects in an indirect manner neural tunings, and that there is some monotonic relationship between neural activity and voxel responses. In order to further improve the similarity of the simulated and empirical data, we created the same number of voxels as were analyzed in the empirical fMRI data, and added a customized amount of noise to the simulated voxels.

### *Noise estimation*

For each ROI and σ parameter combination, we added a custom amount of Gaussian noise to the voxel responses. The appropriate noise magnitude was determined by performing a separate SVM-based decoding analysis on the empirical localizer data and simulated localizer data. In brief, unmodulated neural responses to example stimuli were simulated and decoding was performed using linear SVMs. Iteratively more noise (noise parameter = 1 to 100 in steps of 1) was added to the simulation until the decoding performance of the simulated data was less than the decoding performance of the empirical fMRI data. Subsequently, the noise value yielding results closest to the empirical results was chosen as the noise level for the ROI and σ combination. This method ensures that no σ value is biased due to significant signal-to-noise ratio (SNR) advantages. Moreover, estimating noise levels further increases the comparability of the simulated and empirical data, in particular in terms of SNR, and accounts for potential effects of ROI specific SNR levels on the observed results.

### *Response requirements*

Given that the precise parameter values resulting in the most biologically plausible responses are unknown, we performed a broad grid search across parameter space. However, a broad grid search will inevitably result in some parameter combinations yielding implausible neural responses. The aim of the current study is to elucidate what type of neural modulation may underlie expectation suppression, and not to show the theoretical flexibility of unconstraint computational models. Thus, we enforced three biological plausibility criteria to the constructed neural responses spaces (NRS) that any parameter combination had to fulfill in order to be considered for the main simulation.

First, any unmodulated NRS had to cover the feature space reasonably well. This criterion rejects NRSs with too much variability or even ‘holes’ in its responses. An example of an excessively variable NRS is depicted in Figure 4.10A. Note that the unmodulated summed neural response (thick black line in Figure 4.10A) to some stimuli in feature space are drastically reduced compared to the response to other stimuli. In the case of V1, this would mean that neural responses to an oriented bar of e.g., 20 degrees would be more than twice as large as the maximum possible response to an oriented bar of ~30 degrees. Indeed, more extreme cases would even result in a de-facto blindness to certain feature values. While response biases on the population level certainly exist (e.g., in V1 to cardinal orientations; [157,158]), the variable response criterion enforced here does not concern the population (voxel) response, but the maximal response of any possible neuron to a particular feature value. Given that neural responses, at least to the unmodulated stimuli, should be fairly uniform (i.e., there are at least some neurons that respond to a given stimulus in the relevant feature space), we rejected any unmodulated NRS that had a point in feature space to which the unmodulated summed response was less than 75% of the maximal response. This ensured a reasonably uniform responsiveness to unmodulated stimuli.

Second, a similar minimum response criterion was also enforced for modulated NRSs. However, this criterion was set to a more liberal threshold to allow for a larger modulation range. In particular, the threshold was set such that the modulated summed response to any point in feature space was at least 10% of the maximal summed response. This criterion avoids outright blindness (i.e., zero response) to expected features. To put the threshold of 10% in perspective, one can consider that the ratio of the maximal spiking rates of expectation-sensitive IT neurons (population average) compared to baseline (no visual stimulation) is approximately 24%, as reported by Meyer and Olson [23]. In other words, we reject NRSs for which the expectation modulation was more than twice as strong as the response to visual stimulation compared to baseline in visually driven IT neurons.

Third, neurons of any unmodulated NRS had to be sufficiently selective. This criterion rejects NRSs in which neural tuning is implausibly uniform, as depicted in Figure 4.10B. The rationale for this criterion is based on the notion that feature selective neurons were simulated, and that feature selective neurons ought not to respond (almost) equally to all features. In particular, we rejected any unmodulated NRS in which the response of a neuron is >75% of its maximal response ½ π away from its mean. In other words, this criterion precluded neurons which would respond maximally to an orientation of 0 degrees and still respond with >75% of its maximum to orientations of 90 degrees.

In total, these response requirement criteria resulted in the rejection of 3.2% of NRS in V1 and 42.2% in LOC and TOFC. Rejection numbers were noticeably larger in LOC and TOFC, because V1 feature space was circular, utilizing von Mieses distributions instead of Gaussian distributions in LOC and TOFC. Thereby, the number of NRS with too low responses for any feature space position for very small values, or too unspecific responses for large values in the unmodulated NRS, was significantly lower in V1 than LOC and TOFC. However, more important than differences in rejection percentages between ROIs were possible differences between model types within the same ROI. Differences between model types were relevant to consider, as radically different rejection percentages would result in some model types having more valid parameterization, and thereby a higher chance of fitting the empirical results simply due to a larger number of valid simulations. Reassuringly, differences in rejection percentages between different model types were less than 10% between almost all model types in all three ROIs. Details are summarized in Table 4.1. Thus, each model type was sampled approximately equally often, with minor differences due to poor feature space coverage being slightly more prevalent in some model types (see rejection criteria above). In sum, the similar rejection rates imply that any potential differences in the performance of different model types are unlikely to be explained by differences in the number of sampled parameter combinations. Moreover, NRS rejections ensured that the obtained results are more likely to be meaningful, as the implemented rejection criteria support biological plausibility of the considered models.

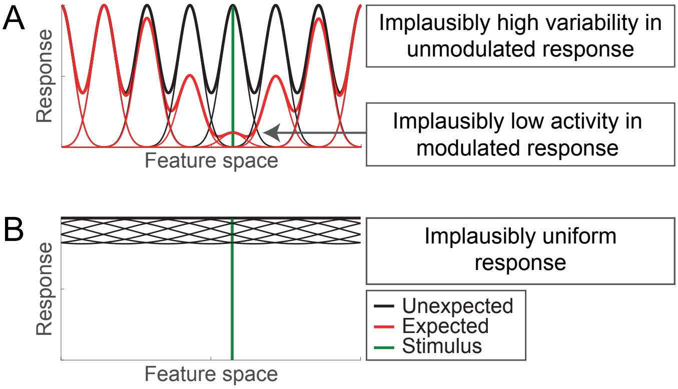


**FIGURE 4.10 Implausible responses.**

Shown are neural responses, which were considered biologically implausible, and thus rejected from the simulation. (**A**) Shows a neural response space for which the parameterization resulted in an implausibly high variability in the unmodulated neural response (thick black line). On this parameterization, the maximal unmodulated response to feature values would vary with more than 50%; e.g., the maximum possible responses of any neuron to a stimulus of 100 degrees orientation would be more than 50% larger than the response to a stimulus of 120 degrees. Any difference exceeding 75% in the unmodulated responses was deemed implausible. Note: this does not affect population difference due to an overrepresentation of neurons preferentially responding to specific features, but concerns the maximal possible response of any neurons type. The thick red line shows an implausibly low modulated response; i.e., the maximal response to this expected stimulus would be ~5% of the unmodulated response, which was considered biologically implausible, given that the approximate difference between maximal responsiveness and baseline (no visual stimulation) for expectation-sensitive, visually driven IT neurons is ~24% (see: [23]). (**B**) Depicts an example of an implausibly uniform neural response space. Thin lines indicate individual neural response functions, which are implausibly broadly tuned; i.e., the neurons are hardly selective for any feature, as their response exceeds 75% of their maximal response irrespective of the feature, which was deemed implausible.

**TABLE 4.1 Percentage of neural response space rejections.**

Shown are the percentages of neural response space rejections for each model type and ROI. While there are large differences in rejection rates between circular (V1) and non-circular (LOC and TOFC) feature space ROIs, importantly within each ROI rejection rates between models are fairly similar (differences of usually <10%). This suggests that differences in the number of sampled parameter combinations are unlikely to account for differences in the observed results, as all model types are sampled approximately equally often.

| | ROI | |
|---|---|---|
| **Model types** | **V1** | **LOC & TOFC** |
| Global gain modulation | 10.0% | 41.8% |
| Local gain modulation | 0.2% | 35.5% |
| Remote gain modulation | 2.0% | 39.0% |
| Global tuning | 4.1% | 46.8% |
| Local tuning | 0.7% | 38.5% |
| Remote tuning | 2.4% | 45.9% |

### *Simulation procedure*

After establishing the model formulation and creating voxels per simulated participant, we proceeded to present object stimuli to the simulated voxels. In fact, we simulated the presentation of the same stimuli, on the same number of trials as was used in the empirical data collection. In other words, each participant of the empirical fMRI data was simulated with the original trial matrix. We simulated each participant 50 times to ensure that results were not driven by noise, resulting in a total of n = 2800 simulated participants. In total 909,600 trials were simulated for each model type and parameter combination. Additionally, we also simulated localizer data by adding ½ of the estimated noise level, thereby mimicking the higher SNR afforded by the localizer run's design, again using the same number of simulated localizer trials as in the empirical localizer data. Finally, the simulated fMRI responses were analyzed using the same analyses as for the empirical fMRI data (see *fMRI outcome metrics*).

### *Model comparison*

The central question in the present study concerns which type of neural modulation best explains the empirical results. Frequently in fMRI studies statistically significant outcome metrics are interpreted in a binary fashion; i.e., either the effect reduces or increases the response. Mimicking this binary approach, we first analyzed our data by comparing the sign of simulated outcome metrics with the empirical results. Thus, in this analysis a model either can or cannot account for a particular outcome metric. By summing the number of matching slopes we can determine how many outcome metrics each model can account for. Additionally, we assessed how robust the results are to changes in the precise parameterization; i.e., how many different parameterizations of each model could explain a large number of outcome metrics. To this end, we calculated for each model type the percentage of valid model parameterization (i.e., parameter combinations that fulfilled the criteria listed in: *Response requirements*) that resulted in the maximal number of matching signs of slopes. This metric thus gives an indication not only how well each model type can account for the sign of the slopes under one ideal parameterization, but how robust the result is to changes in the parameter values, with more robust models being preferable.

Besides this binary approach, we also performed a more fine-grained quantitative analysis of the model fit. In particular, we calculated the mean squared error (MSE) for each model type and parameter combination. We first calculated the normalized slope for each outcome metric and model parameterization. A normalization step was necessary, because the empirical and simulated data, depending on the outcome metric, encompass different numerical ranges. Normalization was done by rescaling to the unexpected condition's value per outcome metric (for MAM, WC, BC, CP, SVM) or to the maximal suppression value (for AMA, AMS, IP). The slope for each outcome metric was calculated on these normalized values, thus resulting in relative, and thereby comparable, expectation effect slopes for each outcome metric. Thus, for example for MAM the resulting slope coefficient would represent the expectation suppression magnitude relative to unmodulated (unexpected) responses. We then compared these slopes for each simulated result (i.e., model type and grid point combination) to the empirical fMRI results by calculating the mean squared error (MSE). The MSE was defined as the averaged, squared difference between simulated and empirical results. Thus, each model and parameter combination resulted in one MSE value, with the smallest MSE constituting the model that most closely fit the empirical fMRI results. MSEs were calculated for each ROI separately, and averaged across datasets, weighted by the number of participants in the empirical dataset.

We investigated the resulting MSEs by displaying the MSE of the 100 best fitting parameter combinations for each model type. Particular attention was devoted to the best fitting model, as it most likely describes the mechanism underlying the empirical data. However, also the robustness to changes in the exact parameterization was of interest, thus we chose to display a range of best fitting model parameterizations. Additionally, we also investigated the performance of each model type across the best model parameterizations by calculating the average MSE, and associated 95% confidence interval, of the best 100 models.

Finally, we inspected the parameter values under which the models performed well. For this purpose, we were interested in the *a* and *b* parameters, as the best parameter values may provide additional information in characterizing the neural mechanism underlying perceptual expectations. In particular, the a parameter value gives an indication of the suppression magnitude, while the b parameter indexes how localized, in feature space, expectation suppression affects responses of neural populations.

## Software and data availability

MRI data was preprocessed and analyzed using FSL 6.0 (FMRIB Software Library; Oxford, UK; www.fmrib.ox.ac.uk/fsl; [87]; RRID:SCR_002823). Additional fMRI data analysis was performed using custom Python (Python Software Foundation, RRID:SCR_008394), with NumPy ([98]; RRID:SCR_008633), and Matlab 2018b (The MathWorks, Inc., Natick, Massachusetts, United States, RRID:SCR_001622) scripts. The simulations were performed using Matlab 2018b. Data and code will be available upon publication here: http://hdl.handle.net/11633/aaddrwao

# Supplemental Information

**TABLE S4.1 Statistics from empirical fMRI data analysis in V1.**

Shown are the results of the analysis of fMRI data from V1 per outcome metric (MAM, WC, BC, CP, SVM, AMA, AMS, IP; see *fMRI outcome metrics* for details). For each outcome metric the intercept and slope are reported, as well as the associated t-statistic and p value.

| | Statistic | | | |
|---|---|---|---|---|
| **Outcome metric** | **Intercept** | **Slope** | **t-statistic** | **p value** |
| MAM | 3.352 | -0.200 | -4.63 | 2.3e-05 |
| WC | 0.345 | -0.033 | -2.38 | 0.0207 |
| BC | -0.064 | 0.007 | 2.87 | 0.0058 |
| CP | 0.385 | -0.033 | -2.33 | 0.0234 |
| SVM | 29.730 | -2.563 | -3.51 | 0.0009 |
| AMA | 0.091 | 0.020 | 2.02 | 0.0478 |
| AMS | 0.054 | 0.026 | 2.68 | 0.0097 |
| IP | 0.130 | 0.018 | 1.12 | 0.2661 |

**TABLE S4.2 Statistics from empirical fMRI data analysis in LOC.**

Shown are the results of the analysis of fMRI data from LOC per outcome metric (MAM, WC, BC, CP, SVM, AMA, AMS, IP; see *fMRI outcome metrics* for details). For each outcome metric the intercept and slope are reported, as well as the associated t-statistic and p value.

| | Statistic | | | |
|---|---|---|---|---|
| **Outcome metric** | **Intercept** | **Slope** | **t-statistic** | **p value** |
| MAM | 4.031 | -0.218 | -5.15 | 3.6e-06 |
| WC | 0.107 | -0.007 | -0.67 | 0.5083 |
| BC | -0.024 | 0.004 | 2.17 | 0.0342 |
| CP | 0.130 | -0.011 | -0.94 | 0.3534 |
| SVM | 19.140 | -0.209 | -0.31 | 0.7599 |
| AMA | -0.006 | 0.042 | 3.81 | 0.0004 |
| AMS | -0.087 | 0.056 | 4.27 | 7.8e-05 |
| IP | 0.223 | -0.006 | -0.42 | 0.6752 |

**TABLE S4.3 Statistics from empirical fMRI data analysis in TOFC.**

Shown are the results of the analysis of fMRI data from TOFC per outcome metric (MAM, WC, BC, CP, SVM, AMA, AMS, IP; see *fMRI outcome metrics* for details). For each outcome metric the intercept and slope are reported, as well as the associated t-statistic and p value.

| | Statistic | | | |
|---|---|---|---|---|
| **Outcome metric** | **Intercept** | **Slope** | **t-statistic** | **p value** |
| MAM | 2.259 | -0.240 | -6.96 | 4.3e-09 |
| WC | 0.035 | -0.003 | -0.41 | 0.6804 |
| BC | -0.015 | 0.006 | 3.87 | 0.0003 |
| CP | 0.050 | -0.009 | -1.05 | 0.2992 |
| SVM | 16.904 | -0.287 | -0.49 | 0.6274 |
| AMA | 0.003 | 0.044 | 4.77 | 1.4e-05 |
| AMS | -0.067 | 0.057 | 6.30 | 5.1e-08 |
| IP | 0.173 | 0.016 | 1.72 | 0.0914 |

**TABLE S4.4 Feature spaces explain neural variance in their target ROIs.**

Shown are the results of one-sample t-tests, and associated effect sizes (Cohen's *d*), comparing the obtained correlation coefficients of neural and model RDM against zero. Results show that the designed feature spaces explain significant neural variance in their target ROI (V1 = orientation energy; LOC = shape complexity; TOFC = semantic similarity).

| | ROI | | |
|---|---|---|---|
| **Feature space** | **V1** | **LOC** | **TOFC** |
| Orientation energy | $t_{(55)}$ = 6.23, p = 6.9e-8, $d_z$ = 0.83 | $t_{(55}$) = 2.14, p = 0.037, $d_z$ = 0.29 | $t_{(55)}$ = -1.61, p = 0.112, $d_z$ = -0.22 |
| Shape complexity | $t_{(55)}$ = -0.71, p = 0.483, $d_z$ = -0.09 | $t_{(55)}$ = 5.21, p = 2.9e-6, $d_z$ = 0.70 | $t_{(55)}$ = 2.84, p = 0.006, $d_z$ = 0.38 |
| Semantic similarity | $t_{(55)}$ = -0.18, p = 0.858, $d_z$ = -0.02 | $t_{(55)}$ = 4.19, p = 1.0e-4, $d_z$ = 0.56 | $t_{(55)}$ = 3.03, p = 0.004, $d_z$ = 0.40 |

**TABLE S4.5 Contrasting feature space models.**

Shown are the results of paired t-tests, and associated effect sizes (Cohen's d), comparing the obtained correlation coefficients of neural and model RDMs between different feature space models. Results show that the designed feature space in V1 (orientation energy) explains more neural variance than any other feature space model. In LOC the designated feature space, shape complexity, reliably outperforms the orientation energy feature space model. However, while numerically larger, it does not significantly explain more variance than the semantic feature space. Similarly, in TOFC semantic similarity significantly explains more neural variance than the orientation energy feature space model. Again while numerically larger, it does not reliably explain more variance than the shape complexity feature space.

| | ROI | | |
|---|---|---|---|
| **Feature space** | **V1** | **LOC** | **TOFC** |
| Orientation energy vs Shape complexity | $t_{(55)} = 6.29$, $p = 5.5e\text{-}8$,<br>$d_z = 0.84$ | $t_{(55)} = -3.47$, $p = 0.001$,<br>$d_z = -0.46$ | $t_{(55)} = -3.06$, $p = 0.004$,<br>$d_z = -0.41$ |
| Shape complexity vs Semantic similar. | $t_{(55)} = 5.44$, $p = 1.2e\text{-}6$,<br>$d_z = 0.73$ | $t_{(55)} = -2.13$, $p = 0.037$,<br>$d_z = -0.29$ | $t_{(55)} = -3.26$, $p = 0.002$,<br>$d_z = -0.44$ |
| Semantic similarity vs Semantic similar. | $t_{(55)} = -0.30$, $p = 0.763$,<br>$d_z = -0.04$ | $t_{(55)} = 1.93$, $p = 0.060$,<br>$d_z = 0.26$ | $t_{(55)} = -0.48$, $p = 0.630$,<br>$d_z = -0.06$ |

## SUPPORTING TEXT S4.1 Alternative implementations of expectation suppression.

In our simulations the neural response was only computed using a modulated response function when the expected stimulus was presented (red curve in Figure 4.3B). By contrast, when a different (unexpected) stimulus was presented, the unmodulated response function was used to compute the response (black curve in Figure 4.3B). As such, the response function was conditional on the identity of the stimulus. While such stimulus-conditional response modulations can be conceptualized in biological terms as reflecting a top-down modulatory effect (i.e., a modulation happening after the initial feedforward sweep), it is not the only logically possible way to formalize a modulation. An alternative way would be to formalize response modulations conditional on the expectation. This way, expectations would affect responses not only to expected but also unexpected stimuli. In this case, different stimuli would be affected differently by virtue of their location along feature space, rather than because of the use of a different response function for expected compared to unexpected stimuli. While this alternative formulation aligns with how response modulations were conceptualized for other modulatory effects, such as attention [148], it cannot provide a coherent mechanism that implements both of the main theoretical accounts of expectation suppression; sharpening and dampening. Here, we demonstrate this using a toy simulation.

Figure S4.1A, illustrates the alternative model definitions in which the responses to both, expected and unexpected stimuli are modulated by expectations. Thus, in these models the modulation by expectations is not conditional on the stimulus being expected, thereby representing not a consequence of recurrent message passing following stimulus presentation (as implemented in the main simulation), but for instance a prestimulus expectation and subsequent suppression of the responses. Crucially, as can be seen in Figure S4.1A and Figure S4.1B, only local models can reliably result in expectation suppression in this implementation. In other words, the response to an expected compared to an unexpected stimulus is exclusively suppressed (lower summed response in Figure S4.1A) for local models. Indeed, population sharpening (remote modulations) can only result in expectation *enhancement* in this model definition. This is not only the case in the depicted example in Figure S4.1A, but true for any combination of expected and unexpected stimuli; i.e., for remote models in S4.1A, the summed response to the unexpected stimulus will always be lower than to the expected stimulus, precisely because the remote modulation is affecting neural populations tuned away from the expected stimulus. In fact, the only case in which sharpening, or any other remote model, under this alternative implementation can account for expectation suppression, is for expected compared to expectation-free stimuli; i.e., stimuli for which no prediction is instantiated. Indeed, for the specific case of expectation-free stimuli this alternative implementation is identical to our implementation. That said, the vast majority of studies exploring expectation suppression contrast unexpected with expected stimuli, thus requiring a model that can account for expectation suppression of expected relative to unexpected (not only expectation-free) stimuli. Thus, we considered these alternative model implementations of little relevance in arbitrating between accounts underlying expectation suppression, because only local modulations (dampening) can reliably result in the phenomenon of interest (expectation suppression), thereby categorically ruling out sharpening, and any other remote or global model.

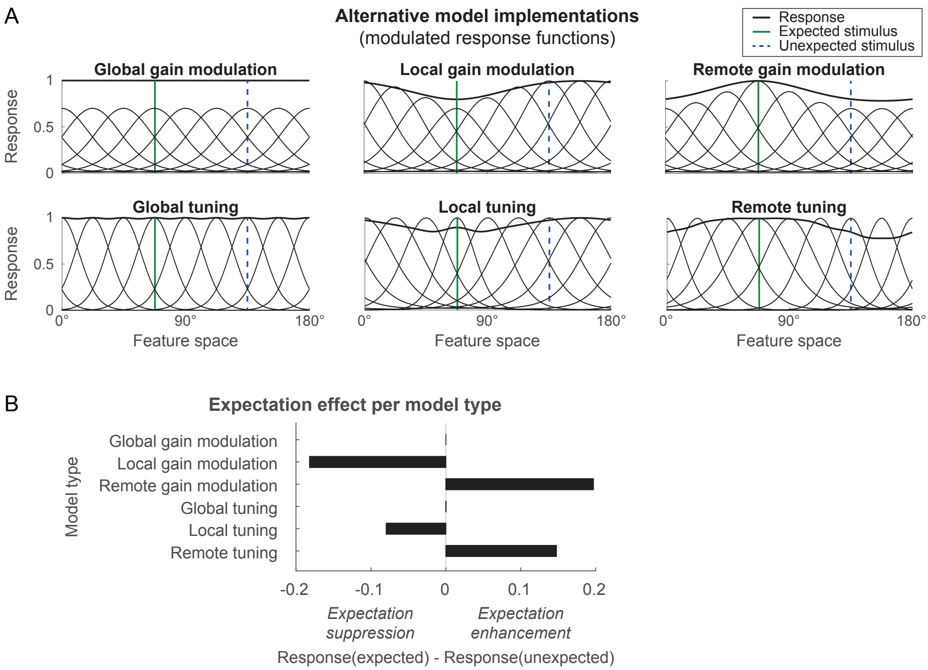


**FIGURE S4.1 Alternative model formulation.**

(**A**) Depicts the modulated neural response functions, with an example of the six modulation models. Thin black lines denote individual neural response functions across feature space, while thick lines indicated the summed response (normalized to one). There is no distinction between expected and unexpected stimuli in terms of the response functions, because a modulation is not conditional on the stimulus being expected in this implementation, but rather always occurs based on the expectation elicited by the leading image. Green shows the position of an expected stimulus in feature space (e.g., ~75 degrees orientation), according to which the responses are modulated. Blue dashed line shows a possible unexpected stimulus in feature space. (**B**) Shows the expectation effect (expectation suppression or enhancement) in response to the expected (solid green line) and unexpected (dashed blue line) stimulus shown in A. The contrast expected – unexpected is shown; i.e., negative values indicate expectation suppression, positive values expectation enhancement. As evident, only local models result in expectation suppression in this model implementation. Global models do not result in amplitude differences and remote models result in expectation enhancement. Notice, based on the summed response in A (thick line) that, the results in B are not dependent on the chosen unexpected or expected stimulus, but qualitatively would remain identical for each chosen stimulus. For example, for each position in feature space (stimulus) remote gain modulations would result in expectation enhancement, an increased response to expected (green solid line) compared to unexpected (blue dashed line) stimuli. In other words, sharpening accounts, and in fact any remote model, under this implementation cannot result in expectation suppression; similarly global models cannot reliably produce expectation suppression either.

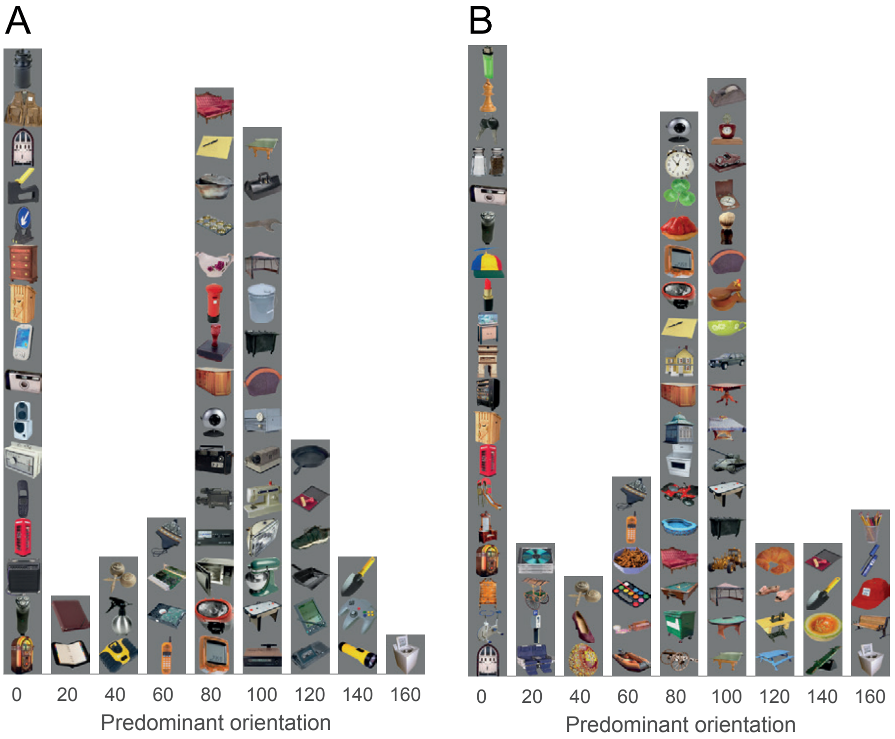


**FIGURE S4.2 Stimuli in orientation feature space (V1).**

Utilized object stimuli arranged by their predominant orientation in steps of 20 degrees. (A) Stimuli from Richter and de Lange [145]. (B) Stimuli from Richter et al. [113].

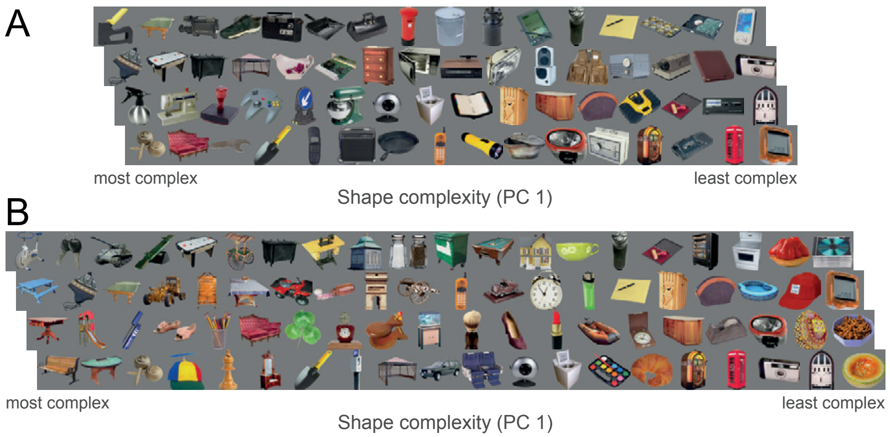


**FIGURE S4.3 Stimuli in shape complexity feature space (LOC).**

Object stimuli arranged by shape complexity; first principle component (PC 1) on shape complexity measures. The most complex stimuli (e.g., irregular shapes with many protrusions) are displayed on the left, while the least complex objects (simple, squared or circular objects) are found on the right. Positions on the horizontal axis are meaningful, while stimuli are stacked vertically only for display purposes. (A) Stimuli from Richter and de Lange [145]. (B) Stimuli from Richter et al. [113].

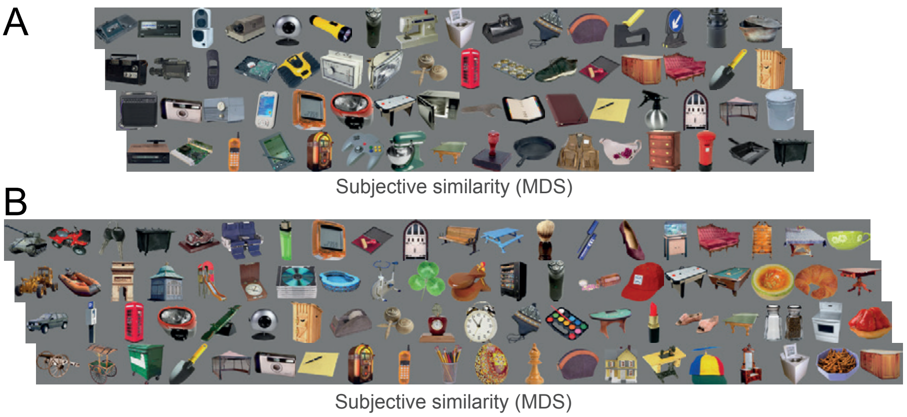


**FIGURE S4.4 Stimuli in semantic similarity feature space (TOFC).**

Object stimuli arranged by semantic similarity (after multidimensional scaling; MDS), based on human ratings. For example, note that vehicles on the left in panel (B) are clustered together, in close proximity to structures and other objects found outside, but very distant to food and kitchen items on the right. Only relative distances on the horizontal axis are meaningful, not absolute positions. Stimuli are stacked vertically for display purposes. (A) Stimuli from Richter and de Lange [145]. (B) Stimuli from Richter et al. [113].

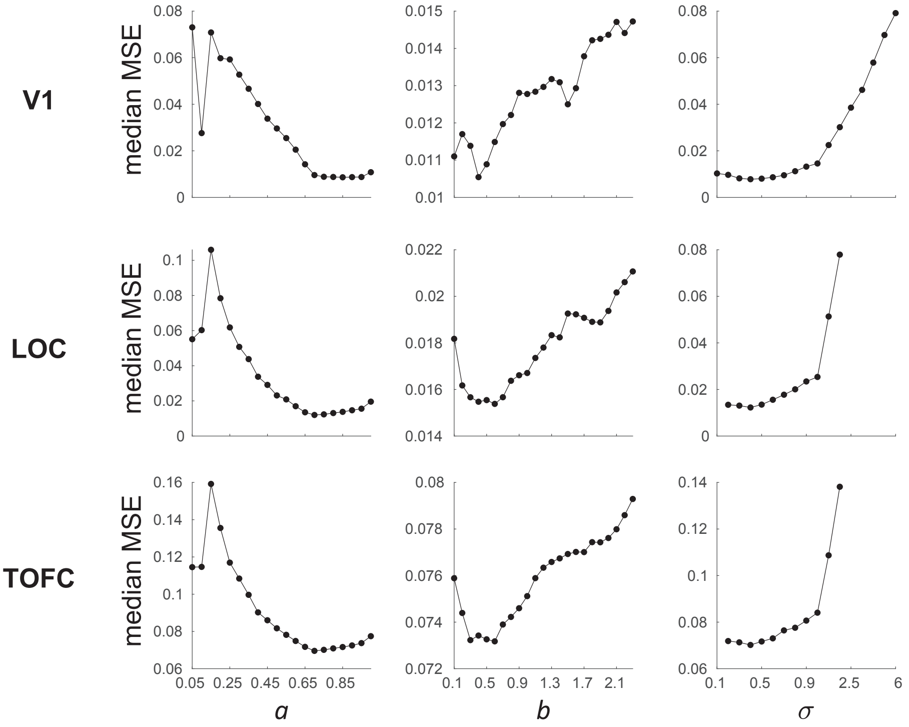
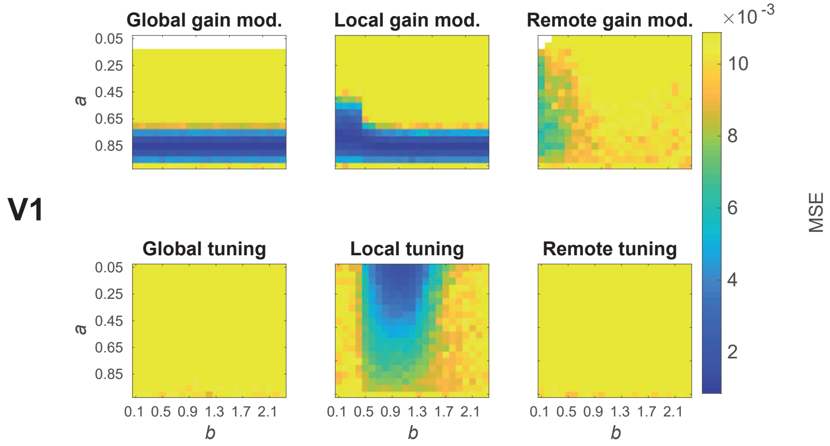
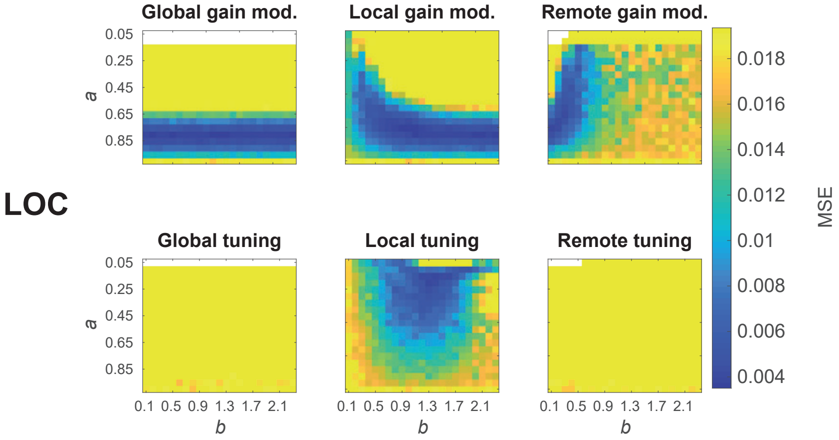


**FIGURE S4.5 Median MSE values across parameter space are smooth and well contain the minimum.**

We used a large parameter grid to ensure that we thoroughly explore the relevant parameter space. Additionally, we performed several tests to ensured that we succeeded in exploring the parameter space in sufficient detail. Shown above are median MSE values, averaged over all model types and parameterizations, for each ROI (V1 top, LOC middle, TOFC bottom) and free parameter (a left, b center, right) separately. Two characteristics of all curves should be noted: (1) the minimum MSE is well contained within a valley in the parameter range, and (2) MSE curves, particularly close to the minimum value, are smooth. Combined these characteristics suggest a well sampled parameter space. Moreover, all three parameter have theoretical value boundaries. For instance, $a = 0$ results in no neural response at all after suppression, and $a = 1$ resulting in no expectation suppression, therefore resulting in the sampled range of $a = 0.05$ to 1. For non-global models, the $b$ parameter value of $b = 0$ results in no expectation suppression, and on the other extreme, b was limited to $b <= 2.3$ (i.e., approximately ¾ π of the π sized feature space; also see *Results, Locality vs globality*). Thus, $b$ values ranged from 0.1 to 2.3. Similarly, $\sigma = 0.1$ as lower limit is bound by the requirement of $\sigma > 0$ for any Gaussian or von Mieses distribution. On the other end, large σ resulted in particularly poor fits, as evident above. Combined these considerations, and the results shown here, suggest that parameter space was sampled exhaustively and in sufficient detail. This in turn, boosts confidence in the reliability and validity of the presented results. Additional, more fine-grained analyses of the smoothness of MSE values over parameter spaces, which provide additional evidence that parameter space was thoroughly sampled, are presented in Figures S4.6, S4.7, and S4.8.


V1
Global gain mod.
Local gain mod.
Remote gain mod.
Global tuning
Local tuning
Remote tuning
a
0.05
0.25
0.45
0.65
0.85
b
0.1 0.5 0.9 1.3 1.7 2.1
×10 -3
10
8
6
4
2
MSE


LOC
Global gain mod.
Local gain mod.
Remote gain mod.
Global tuning
Local tuning
Remote tuning
a
0.05
0.25
0.45
0.65
0.85
b
0.1 0.5 0.9 1.3 1.7 2.1
0.018
0.016
0.014
0.012
0.01
0.008
0.006
0.004
MSE

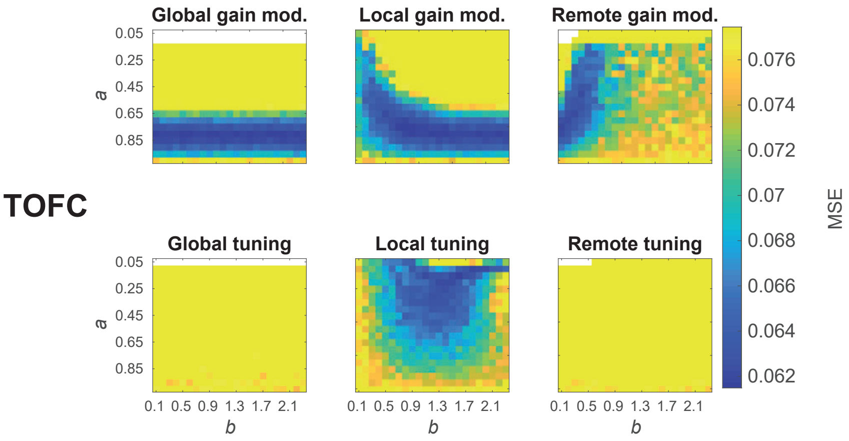
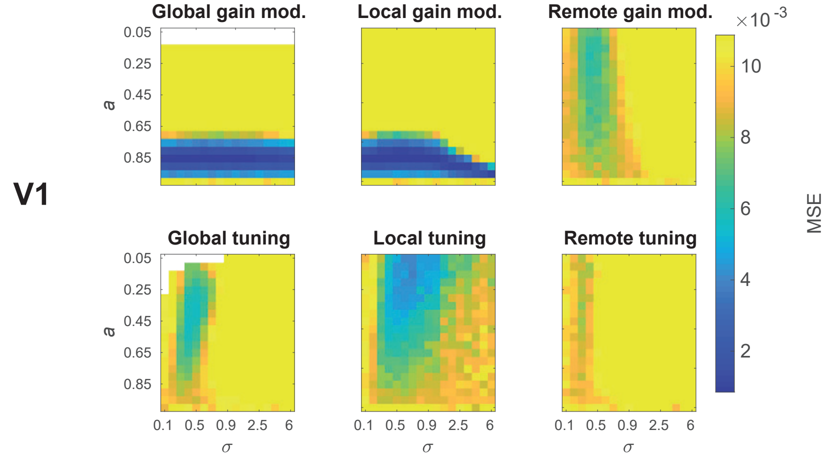
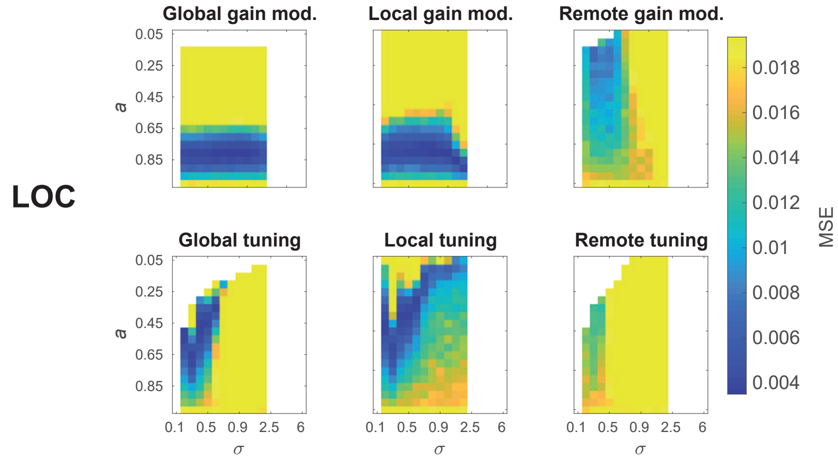

TOFC
Global gain mod.
Local gain mod.
Remote gain mod.
Global tuning
Local tuning
Remote tuning
a
0.05
0.25
0.45
0.65
0.85
b
0.1 0.5 0.9 1.3 1.7 2.1
0.076
0.074
0.072
0.07
0.068
0.066
0.064
0.062
MSE

**FIGURE S4.6 MSE across parameter grid of parameters a and *b*.**

Shown are median MSEs, averaged over σ, per model type and ROI (V1 top, LOC middle, TOFC, bottom). For each model type the median MSE for all parameter combinations of *a* and *b* are shown. Colors are thresholded such that the median MSE of $a = 1$ (i.e., no suppression by expectation) is yellow, and the lowest MSE is dark blue. White cells denote rejected parameter combinations (also see: *Response requirements in Materials and Methods*). Results show that the change of MSEs, particularly for low MSE values, is smooth, suggesting that parameter space has been well sampled in all ROIs. The error landscape for each model type and ROI is smooth and well contains the respective minima, unless the parameter was bound by theoretical reasons, within the sampled range, indicating a well sampled parameter space.


V1
Global gain mod.
Local gain mod.
Remote gain mod.
Global tuning
Local tuning
Remote tuning
a
0.05
0.25
0.45
0.65
0.85
σ
0.1
0.5
0.9
2.5
6
×10 -3
10
8
6
4
2
MSE


LOC
Global gain mod.
Local gain mod.
Remote gain mod.
Global tuning
Local tuning
Remote tuning
a
0.05
0.25
0.45
0.65
0.85
σ
0.1
0.5
0.9
2.5
6
0.018
0.016
0.014
0.012
0.01
0.008
0.006
0.004
MSE

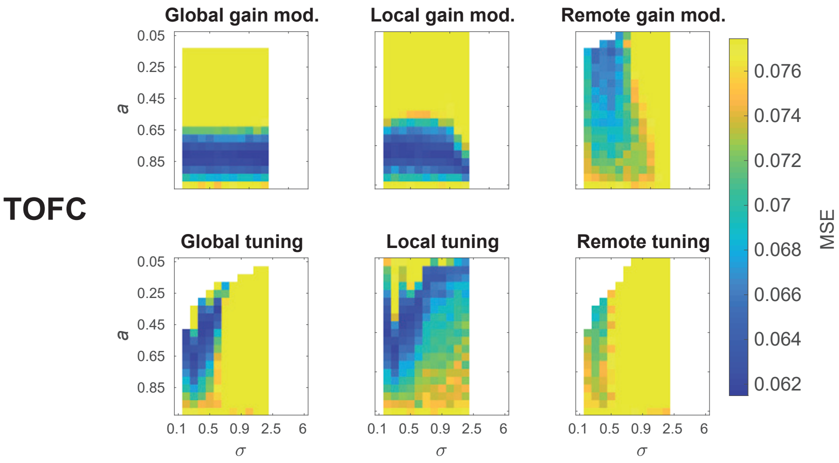

TOFC
Global gain mod.
Local gain mod.
Remote gain mod.
Global tuning
Local tuning
Remote tuning
a
0.05
0.25
0.45
0.65
0.85
σ
0.1
0.5
0.9
2.5
6
0.076
0.074
0.072
0.07
0.068
0.066
0.064
0.062
MSE

**FIGURE S4.7 MSE across parameter grid of parameters *a* and σ.**

Shown are median MSEs, averaged over *b*, per model type and ROI (V1 top, LOC middle, TOFC, bottom). For each model type the median MSE for all parameter combinations of *a* and σ are shown. Colors are thresholded such that the median MSE of $a = 1$ (i.e., no suppression by expectation; in *a*, *b* parameter space) is yellow, and the lowest MSE is dark blue. White cells denote rejected parameter combinations (also see: *Response requirements in Materials and Methods*). It is apparent that very small and large sigma values in non-circular feature spaces (LOC and TOFC) resulted in rejections of the neural response spaces, compared to circular feature spaces (V1). Moreover, results show that the change of MSE values, particularly for low MSE values, is smooth, suggesting that parameter space has been well sampled in all ROIs.

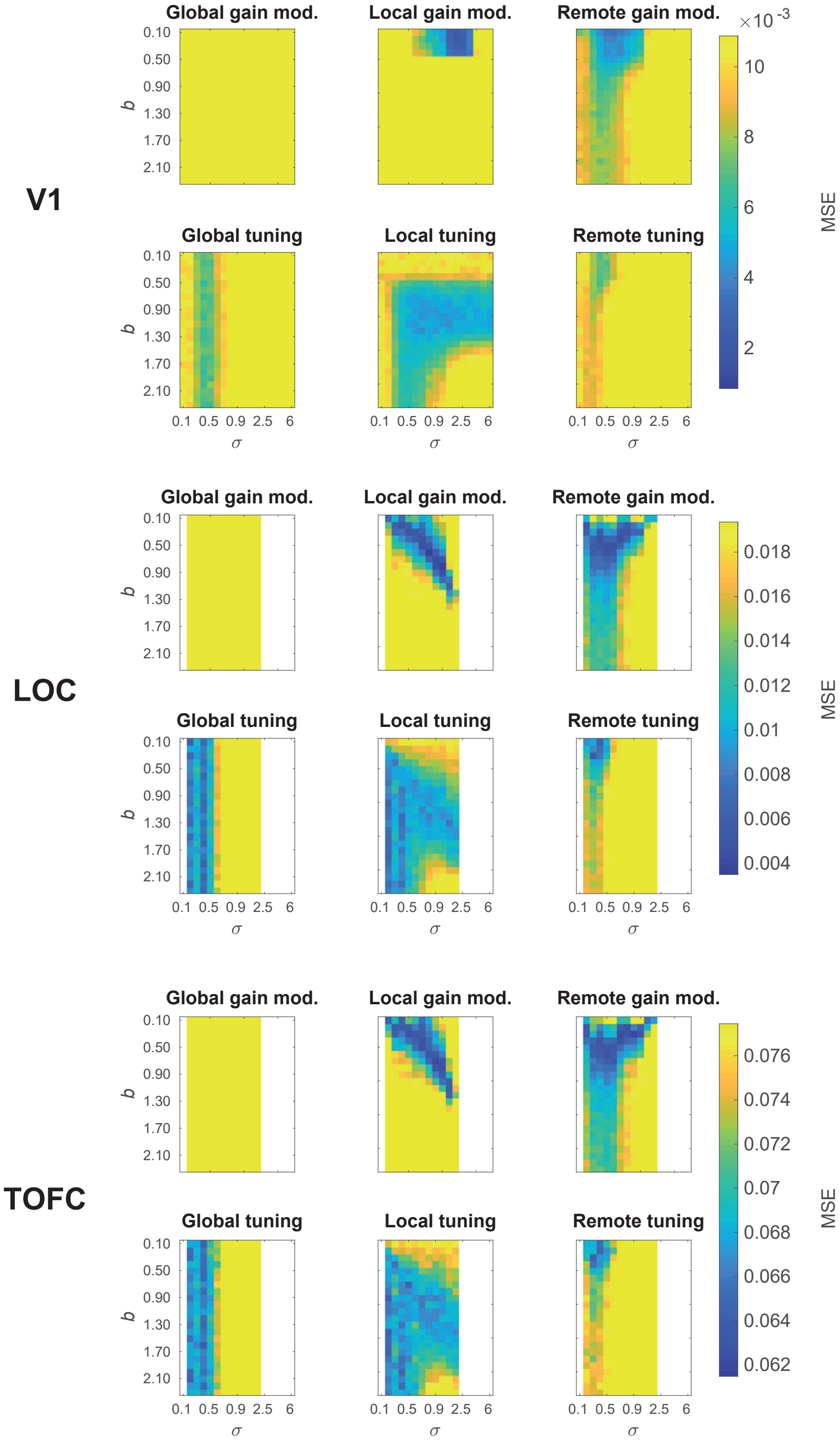
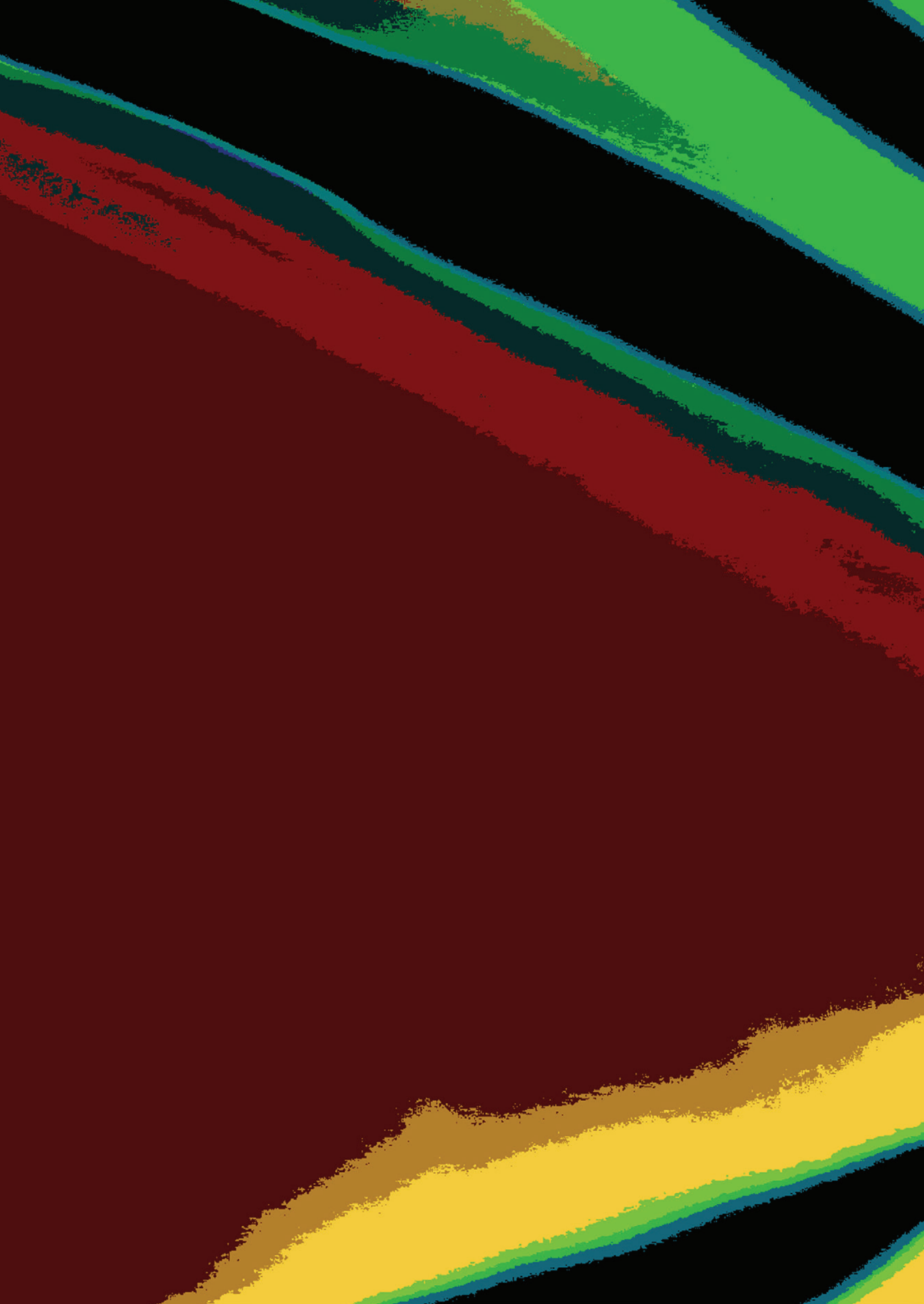
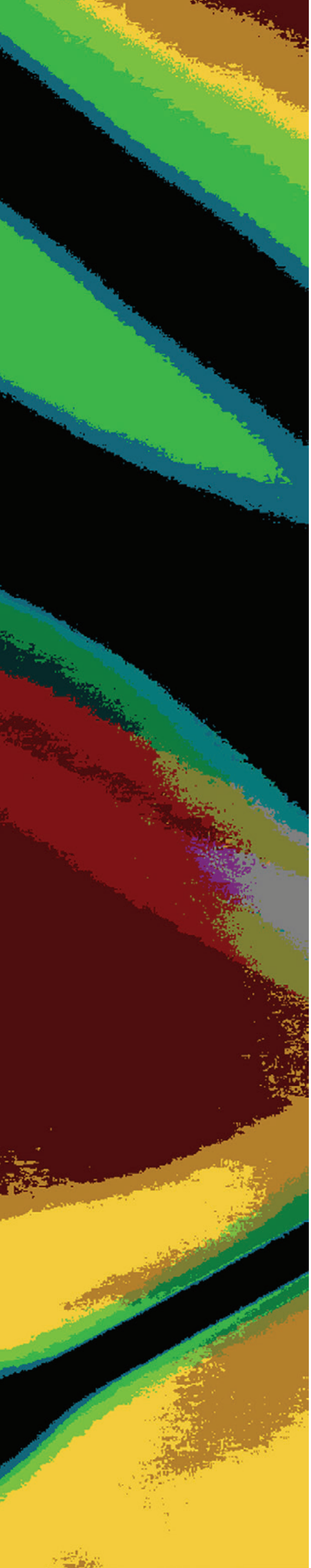

V1
LOC
TOFC
Global gain mod.
Local gain mod.
Remote gain mod.
Global tuning
Local tuning
Remote tuning
b
0.10
0.50
0.90
1.30
1.70
2.10
σ
0.1
0.5
0.9
2.5
6
×10 -3
10
8
6
4
2
MSE
0.018
0.016
0.014
0.012
0.01
0.008
0.006
0.004
0.076
0.074
0.072
0.07
0.068
0.066
0.064
0.062

**FIGURE S4.8 MSE across parameter grid of parameters *b* and σ.**

Shown are median MSEs, averaged over *a*, per model type and ROI (V1 top, LOC middle, TOFC, bottom). For each model type the median MSE for all parameter combinations of *b* and σ are shown. Colors are thresholded such that the median MSE of $a = 1$ (i.e., no suppression by expectation; in *a*, *b* parameter space) is yellow, and the lowest MSE is dark blue. White cells denote rejected parameter combinations (also see: *Response requirements in Materials and Methods*). Results show that the change of MSEs, particularly for low MSE values, is smooth, suggesting that parameter space has been well sampled in all ROIs.





## CHAPTER 5

# Incidental statistical learning of unimodal but not cross-modal statistical regularities

*This chapter is based on:*
Richter, D., Voorrips, E.S., Degutis, J.K., Spaak, E., and de Lange, F.P. (in preparation). Incidental statistical learning of unimodal but not cross-modal statistical regularities

# Abstract

Incidental statistical learning (SL) refers to the acquisition of statistical regularities without intention or instruction to learn. While the consequences of incidental SL have been extensively investigated, the characteristics and evolution of the underlying learning process itself remain poorly understood. As such, it is unclear to which extent humans can incidentally learn probabilistic, compared to deterministic, associations. In addition, it is unknown whether cross-modal statistical regularities, involving cue stimulus associations of at least two modalities, are equally well acquired as unimodal regularities. In fact, whether SL crucially depends on modality-specific or domain general mechanisms remains debated. A modality-specific account of SL predicts that unimodal associations should be substantially easier and faster to learn than cross-modal associations. On the other hand, if domain general mechanisms are central to SL, cross-modal and unimodal SL should be largely comparable. In the present study we investigated the trajectory of incidental SL in cross-modal and unimodal association paradigms, while participants categorized object stimuli. Surprisingly, participants did not learn cross-modal (audio-visual) statistical regularities, with no evidence of SL in either behavioral or neural markers of SL. Indeed, even when both auditory cues and visual stimuli were task-relevant no SL was apparent. However, reliable learning was found for unimodal (visual-visual) associations. Data also show that the acquisition of these unimodal regularities occurred exclusively during blocks with deterministic, but not probabilistic associations, even though statistical information was available for learning during both blocks. In sum, our results suggest incidental SL of unimodal associations but not cross-modal regularities, suggesting that modality-specific mechanisms are critical for incidental SL, and that the acquisition of such regularities primarily occurs during exposure to strong, reliable associations.

# Introduction

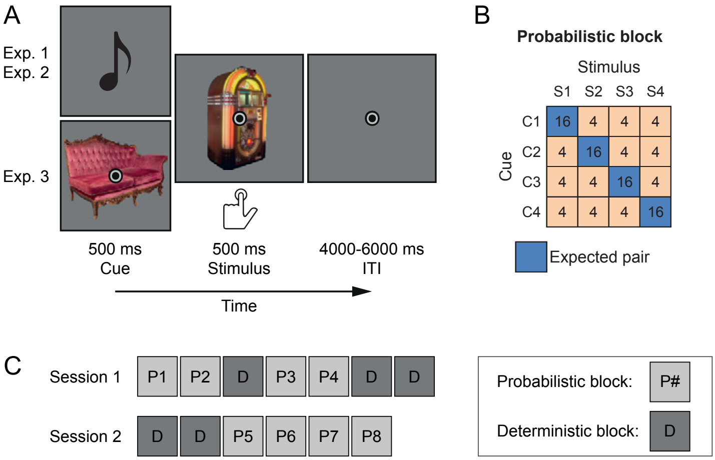
The world is marked by statistical regularities. For instance, sensory information often unfolds in predictable sequences, thus allowing prior information to predict future input. Accordingly, numerous studies show that prior knowledge can be used to respond faster and more accurately to expected stimuli [9,50,51,58,59,145]. Indeed, the formation of expectations can occur without any intention or instruction to learn [50,51,58,113], as evident during incidental statistical learning (SL). Once acquired, expectations can modulate sensory processing throughout cortex (review: [19]). For instance, neural responses to expected compared to unexpected stimuli are often attenuated in sensory areas. This phenomenon, expectation suppression, has been demonstrated both in vision [23,26,113,145] and audition [21,22,159].

While the consequences of SL have been explored in numerous studies, the evolution and mechanisms of the learning process itself remain largely unclear. A central debate in the SL literature concerns domain-general and modality-specific contributions to SL [47,49,63]. Domain-generality suggests that SL in different modalities and contexts relies on similar neural mechanisms and computations, possibly comprising a unitary learning system [36,56,68–70]. In contrast, modality-specificity holds that SL is subject to modality specific constraints and crucially depends on neural changes and computations within sensory areas [63–67]. Moreover, how the trajectory of incidental SL is affected by the reliability of the underlying statistical regularities is largely unexplored, particularly in the case of cross-modal associations. Indeed, recent work suggests that humans may treat deterministic (rule based) regularities as categorically different from probabilistic (statistical) associations [160] and nonlinearities may characterize the learning of different association strengths [161,162].

Here we conducted a series of experiments to explore the development of incidental SL and to contrast predictions of domain-generality and modality-specificity. In particular, we investigated the extent and the evolution of learning during exposure to cross-modal (audio-visual) and unimodal (visual-visual) statistical regularities in probabilistic and deterministic contexts. If SL relies on modality-specific mechanisms, cross-modal SL, requiring information integration across modalities, should be more limited in scope and/or slower in its development than unimodal learning. On the other hand, if SL mainly depends on domain-general processes, cross-modal learning should be largely comparable to unimodal SL. Surprisingly, we found no evidence of cross-modal SL in either behavioral or fMRI markers of learning (Experiment 1). We also show that even when auditory cues and visual stimuli were task-relevant, and thus attended, there was no evidence of SL (Experiment 2). In contrast, we observed reliable behavioral facilitation by expectations for unimodal (visual-visual) statistical regularities (Experiment 3). Combined, our results suggest that, unlike for unimodal associations, cross-modal statistical regularities are not readily acquired during incidental SL, even when both cue and stimulus are task-relevant. Thereby the present data imply that incidental SL may primary rely on modality-specific mechanisms, which favor the formation of predictions within sensory modalities.

## Results

We set out to investigate the development and circumstances under which incidental SL occurs. In particular, we explored SL in a cross-modal paradigm, with both task-irrelevant (experiment 1) and task-relevant auditory cues (experiment 2) predicting visual stimuli. We also investigated SL in an identical paradigm with unimodal (visual-visual) cue-stimulus pairs (experiment 3). A single trial is depicted in Figure 5.1A. Statistical regularities, shown in Figure 5.1B, determined the association between cue-stimulus pairs. In all three experiments it was not necessary for participants to learn the underlying associations, albeit predictions could aid in performing the task performance, the classification of the object stimulus as (non-)electronic. Participants were exposed to the statistical regularities throughout 13 blocks, over the course of two sessions, including eight blocks with probabilistic (896 trials) and five with deterministic associations (480 trials), as illustrated in Figure 5.1C.



**Probabilistic block**

| Cue | S1 | S2 | S3 | S4 |
|---|---|---|---|---|
| C1 | 16 | 4 | 4 | 4 |
| C2 | 4 | 16 | 4 | 4 |
| C3 | 4 | 4 | 16 | 4 |
| C4 | 4 | 4 | 4 | 16 |

**Deterministic block**

| Cue | S1 | S2 | S3 | S4 |
|---|---|---|---|---|
| C1 | 24 | 0 | 0 | 0 |
| C2 | 0 | 24 | 0 | 0 |
| C3 | 0 | 0 | 24 | 0 |
| C4 | 0 | 0 | 0 | 24 |

**FIGURE 5.1 Experimental paradigm.**

(A) Illustration of a single trial, showing a cue (500 ms), followed by an object stimulus (500 ms) and a variable ITI (4000-6000 ms). In all three experiments participants classified the stimulus object as electronic or non-electronic by button press. Thus, expectations were not necessary for performing the task, but could be beneficial. For experiment 1 cues were task-irrelevant and consisted of sine-wave tones of different frequencies. For experiment 2, on 20% of trials white noise was added to the sine-wave tones, indicating a no-go trial, requiring participants to withhold the object classification response. Experiment 3 used task-irrelevant visual object cues instead of auditory cues. (B) Depicts the cue-stimulus association matrices for probabilistic blocks (left) and deterministic blocks (right). In deterministic blocks only expected cue-stimulus pairs were shown, while probabilistic blocks also contained unexpected (i.e., less likely) pairs. However, even during probabilistic blocks the expected stimulus was nonetheless four times more likely than each unexpected stimulus. The stimuli and their associations remained the same throughout the experiment for each participant. Association matrices were identical in the three experiments. (C) Illustrates the order of blocks over two sessions. Probabilistic blocks (P#) were of interest for analysis of the behavioral data, as these allow for a quantification of the expectation induced behavioral facilitation. Deterministic blocks (D) were added to promote additional SL, and assess learning from deterministic compared to probabilistic associations. Thus, particularly changes of expectation induced effects from P2 to P3, as well as from P4 to P5 are interesting, as these blocks were separated by deterministic blocks.

Data was analyzed in terms of the expectation benefit on reaction times (RT) for each block ($RT_{benefit} = RT_{unexpected} - RT_{expected}$), as well as the magnitude of the expectation benefit between chronologically adjacent blocks. Moreover, to investigate the development of RT benefits over the course of the experiment we used Bayesian model comparison to arbitrate between six distinct models (Figure 5.2). In particular, we created models that embodied the hypothesis that expectations may either have no effect on RT, a constant effect, or a gradually emerging linear effect. Additionally, we modeled a one switchpoint and a two switchpoint model, which implemented learning as taking place exclusively during the blocks with deterministic associations. A full model was added that comprised of two switchpoints with an additional linear learning effect. Finally, expectation effects on response accuracies were analyzed per block, and the magnitude of the effect compared between chronologically adjacent blocks.

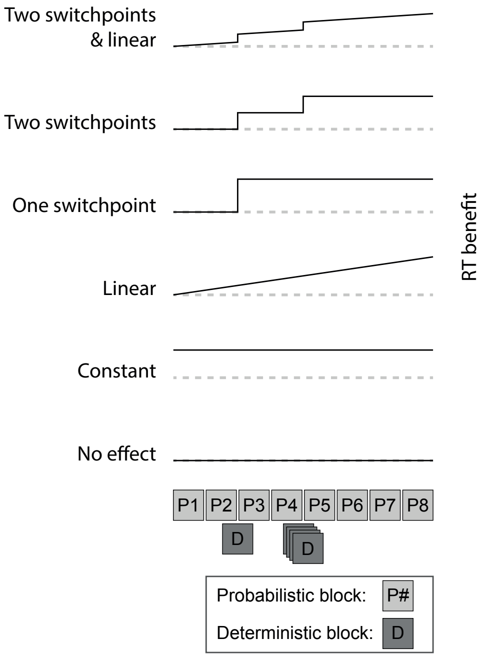


**FIGURE 5.2 Reaction time models.**

Illustration of the RT benefit modeled by six different reaction time models. First, RTs were preprocessed, including log transformation and then modeled per trial. RTs were allowed to vary between blocks in order to account for generic task-learning effects. Grey dashed lines indicate no RT benefit. Black lines depict the development of the RT benefit according to the respective model. In the bottom, the order of blocks is displayed, starting with two blocks with probabilistic association blocks (P1, P2), followed by one deterministic block (D), etc. Only trials of the probabilistic blocks were modeled, as deterministic blocks did not contain any unexpected stimulus pairs (i.e., allowing no estimation of the expectation benefit). Starting from the bottom, the *no effect* model does not include an RT benefit by expectations. The *constant* model adds a constant offset by expectations. A linear increase of RT benefits is modeled in the *linear* model. The *one switchpoint* model initially assumes no RT benefit, followed by a constant offset starting with block P3 or block P5 (i.e., one of the two blocks following the deterministic blocks). The *two switchpoint* model is identical to the one switchpoint model, except for that both switchpoints change the RT benefit. The full model, *two switchpoints & linear*, combines the two switchpoint model with an additional linear modulation of the RT benefit. Models were fit for each participant separately.

## No behavioral facilitation by cross-modal statistical regularities

In experiment 1, we investigated the behavioral and neural consequences of perceptual expectations, as well as their development during cross-modal SL. To this end, we exposed participants to auditory cues, which probabilistically predicted the identity of visually presented object stimuli. During fMRI scanning participants performed a classification task on the predictable object stimuli. Thus, responses to expected compared to unexpected objects should be faster, if participants made use of the statistical regularities.

Results, depicted in Figure 5.3A-B, show that there was no reliable RT benefit due to expectations ($RT_{benefit} = RT_{unexpected} - RT_{expected}$). Indeed, no block yielded a statistically significant RT benefit, except for block 6 ($t_{(23)} = 3.15$, $p = 0.004$, $d_z = 0.64$; all other blocks $p > 0.05$. Note: all p-values are uncorrected for multiple comparisons). Similarly, no reliable effect on response accuracy was evident (Figure 5.3D-E), with again only one block showing a weakly significant accuracy benefit ($t_{(23)} = 2.11$, $p = 0.046$, $d_z = 0.43$; all other blocks $p > 0.05$). Moreover, the magnitude of the expectation benefit was not statistically different between any chronologically adjacent blocks in either response accuracy (all $p > 0.05$) or RT (all $p > 0.05$). In agreement with these results, Bayesian model comparison provided evidence for the absence of any RT benefit by expectations, with the best model fit (lowest WAIC) for the 'no effect' model (Figure 5.3C). Indeed the average WAIC of the 'no effect' model was reliably lower than that of all other models (all $p < 0.05$), except the 'one switchpoint' model. In sum, there was no effect of cross-modal expectations on behavioral responses. Detailed results of all statistical tests are summarized in Table S5.1 and Table S5.2. Thus, the present data suggest that participants either did not make use of the statistical regularities to facilitate behavioral responses, or statistical regularities were not acquired, even after extensive exposure (i.e., 248 repetitions of each expected pair).

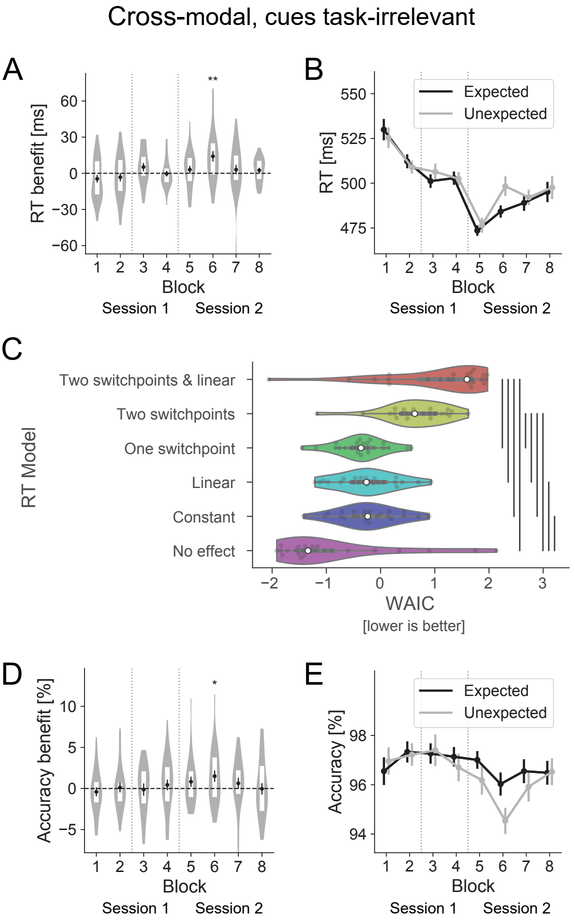


**FIGURE 5.3 No behavioral facilitation by cross-modal statistical regularities.**

(**A**) and (**D**) show the absence of an RT benefit ($RT_{benefit} = RT_{unexpected} - RT_{expected}$) and accuracy benefit (Accuracybenefit = Accuracyexpected – Accuracyunexpected) during the probabilistic blocks. All RT and accuracy benefits are (near) zero, and only one block statistically differed from zero in terms of RT benefit. Moreover, the magnitudes of the RT and accuracy benefits were not reliably different between any adjacent blocks. The horizontal dashed line denotes zero (i.e., no benefit of expectations). Vertical dotted lines indicate when a block with deterministic associations took place between the displayed probabilistic blocks (i.e., before block 3 and 5). Error bars indicate within-subject SEM. White bars indicate first and third quartile ranges. * $p < 0.05$, ** $p < 0.01$ (uncorrected p-values). (**B**) and (**E**), show RTs and response accuracies to expected and unexpected stimuli. Over time RTs become noticeably faster, suggesting a generic task-familiarity effect, but without a reliable differences between expected unexpected occurrences of the stimuli. Overall accuracies are near ceiling (mean: 96.6%) and stable across the experiment, indicating reliable task performance. (**C**) Depicts results of the Bayesian model fits (WAIC). The ‘no effect’ model fit the data best; i.e., lower WAIC than all other models. Vertical lines denote statistically significant differences between WAICs ($p < 0.05$). This suggests that perceptual expectations were not acquired or did not influence RTs.

## No modulation of sensory processing by cross-modal statistical regularities

If participants did not make use of the statistical regularities to facilitate behavioral responses, we may nonetheless find evidence for SL in neural modulations of sensory responses. Indeed, previous studies reported suppressed sensory responses even for task-irrelevant predictions or during passive fixation [23,113]. To investigate whether there is evidence for a neural modulation by expectations, we averaged BOLD responses in each block for all ROIs separately; primary visual cortex (V1), lateral occipital complex (LOC), and temporal occipital fusiform cortex (TOFC). Figure 5.4A and Figure 5.4B, show that visual responses to expected and unexpected stimuli were highly similar in object-selective LOC – i.e., expectation suppression was (near) zero for all blocks. Indeed, only one block showed a statistically reliable difference in BOLD responses to expected compared to unexpected stimuli ($t_{(23)} = -2.31$, $p = 0.031$, $d_z = -0.49$), while all other blocks did not yield a difference (all $p > 0.05$). Moreover, no reliable difference in the magnitude of expectation suppression (unexpected – expected) between adjacent blocks was found (all $p > 0.05$). Again, as for behavioral analyses, Bayesian model comparison (Figure 5.4C) favored the 'no effect' model (all $p < 0.05$), suggesting that cross-modal SL did not modulate perceptual processing, even after excessive exposure to the statistical regularities. Detailed statistics are reported in Table S5.3 and Table S5.4. Results in V1 and TOFC, representing early and higher visual areas, were qualitatively identical to LOC and are depicted in Figure S5.1 and Figure S5.2 respectively. Thus, throughout the ventral visual stream there was no indication of cross-modal SL in terms of attenuation of sensory processing for expected stimuli.

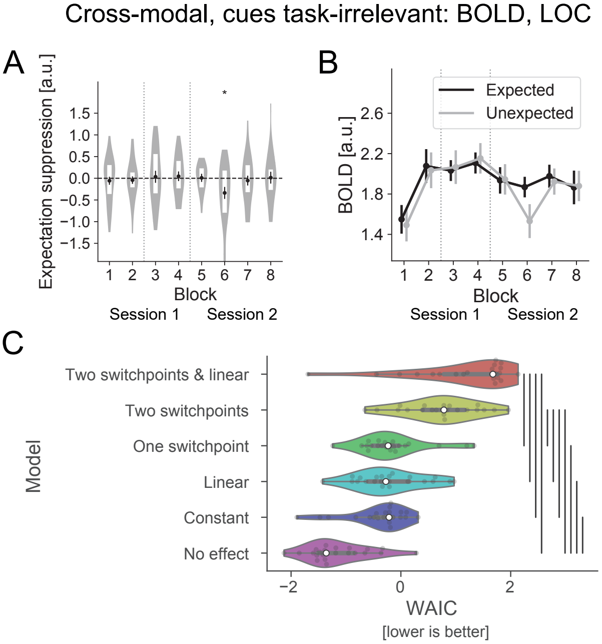


**FIGURE 5.4 No modulation of sensory responses by cross-modal statistical regularities.**

Cross-modal perceptual expectations do not modulate sensory responses in object-selective LOC. (**A**) Displays expectation suppression ($BOLD_{unexpected} - BOLD_{expected}$) for each block. Again, no reliable evidence for an influence of expectation status on sensory processing is evident; i.e., expectation suppression is (near) zero in almost all blocks and did not statistically deviate from zero in any, but one block. Similarly, the magnitudes of expectation suppression was not reliably different between any adjacent blocks. The dashed line indicates zero; i.e., no difference in the response between expected and unexpected stimuli. Vertical dotted lines indicate when a block with deterministic associations took place between the displayed blocks with probabilistic associations (i.e., before block 3 and 5). Error bars indicate within-subject SEM. White bars indicate first and third quartile ranges. * $p < 0.05$ (uncorrected p-values). (**B**) BOLD response to expected and unexpected stimuli for each block and session. (**C**) shows model fits in terms of the WAIC. The 'no effect' model outperformed all other models, suggesting that expectations did not influence sensory processing. Vertical lines denote statistically significant differences between WAICs ($p < 0.05$).

To ensure that the absence of SL is not specific to the a-priori selected ROIs, we performed a whole-brain analysis, contrasting responses to unexpected with responses to expected stimuli per session. Results of this whole-brain analysis are shown in Figure 5.5. We did not find evidence for a reliable modulation of BOLD responses by cross-modal SL anywhere in the brain.

In sum, there was no evidence for any modulation of responses by expectations following extensive exposure to cross-modal statistical regularities. In fact, there was considerable evidence for the absence of a modulation of neural and behavioral responses by perceptual expectations. Given that neither a behavioral nor neural modulation by expectations was found, it is likely that no cross-modal SL occurred.

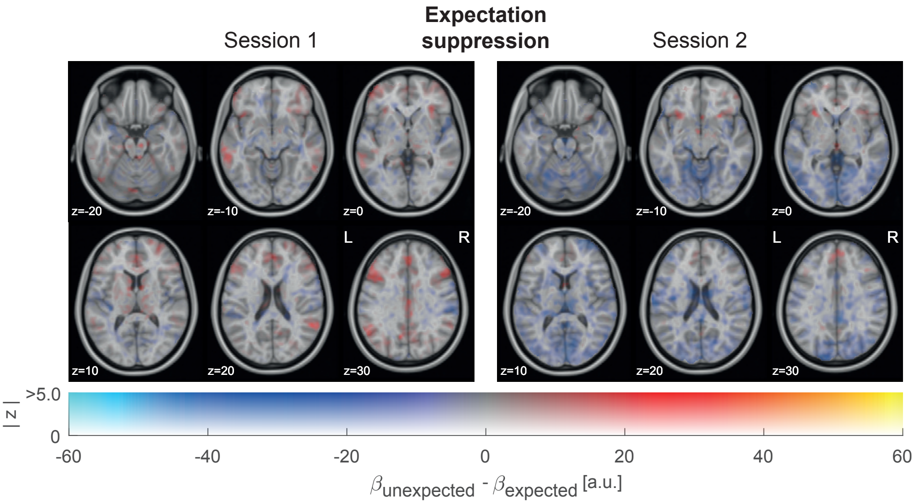


**FIGURE 5.5 No expectation suppression or expectation enhancement anywhere in cortex.** Depicted are results of a whole-brain analysis, contrasting BOLD responses to unexpected – expected stimuli (i.e., expectation suppression). Color indicates the parameter estimates, opacity the associated unthresholded z statistic. The panel on the left shows results for session 1, the right panel for session 2. As evident by the absence of any significant clusters (no outlines around clusters), there was no evidence for reliable expectation suppression or expectation enhancement in any brain area.

## No behavioral facilitation by cross-modal statistical regularities with task-relevant cues

Given that previous studies showed reliable modulations of behavioral and neural responses by perceptual expectations following SL [23,26,50,58,113,145], the question arises why experiment 1 failed to show any evidence for SL. One potential explanation is that auditory cues were task-irrelevant, and thus unattended, or even ignored, thereby precluding learning. This hypothesis is supported by previous studies suggesting that attention on the cue-stimulus pairs may be necessary for expectation effects to arise [62,145]. In order to investigate this possibility, we performed experiment 2, in which we exposed participants to the same paradigm as in experiment 1, but modified the task. While participants still performed the electronic object classification task, following an auditory cue, they were now instructed to withhold any response if an auditory cue was paired with white noise (20% of trials), thereby making both auditory cue and visual stimulus task-relevant and hence attended.

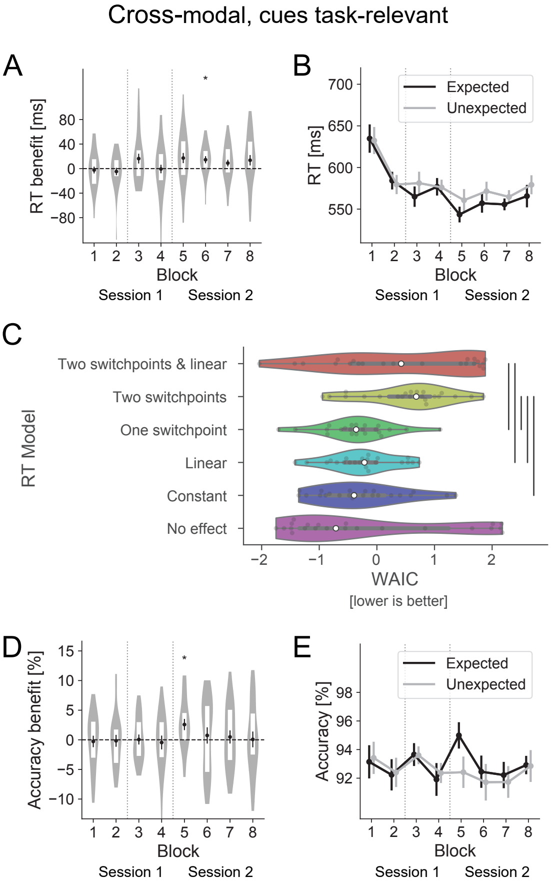


**FIGURE 5.6 No behavioral facilitation by cross-modal statistical regularities when cue and stimulus are task-relevant.**

Cross-modal perceptual expectations do not result in behavioral facilitation. (**A**) and (**B**) show the RT benefit ($RT_{benefit} = RT_{unexpected} - RT_{expected}$) and accuracy benefit ($Accuracy_{benefit} = Accuracy_{expected} - Accuracy_{unexpected}$) for each block. Only go trials were analyzed to ensure comparability with the other experiments. All RT and accuracy benefits are (near) zero, and did not significantly differ from zero. Moreover, the magnitudes of the RT and accuracy benefits were not reliably different between any adjacent blocks. The dashed horizontal line denotes zero (i.e., no benefit of expectations). Vertical dotted lines indicate when a block with deterministic associations took place between the displayed blocks with probabilistic associations (i.e., before block 3 and 5). Error bars indicate within-subject SEM. White bars indicate first and third quartile ranges. * $p < 0.05$ (uncorrected p-values). (**C**) Depicts results of Bayesian model fits (WAIC). The 'no effect' model fit the data best (i.e., lower WAIC than all other models), suggesting that perceptual expectations did not influence RTs. Vertical lines denote statistically significant differences between WAICs ($p < 0.05$). (**D**) and (**E**), show RTs and response accuracy to expected and unexpected stimuli. Over time RTs become noticeably faster, suggesting a generic task-familiarity effect, but without a reliable differences between expected and unexpected occurrences of the stimuli. Accuracies are stable and near ceiling across the experiment.

Results, depicted in Figure 5.6, show that there was again no evidence for a modulation of behavioral responses by expectations, even when both cue and stimulus were task-relevant. Figure 5.6C shows that the ‘no effect’ model again had the lowest WAIC, and thus fit the data better, than all other models – albeit the difference was less reliable than in experiment 1. Figure 5.6A additionally illustrates that there was no evidence for an effect of expectations on RT in any but one block ($t_{(22)} = 2.64$, $p = 0.015$, $d_z = 0.55$; all other blocks $p > 0.05$). Moreover, the magnitude of the RT difference ($RT_{unexpected} - RT_{expected}$) did not reliably differ between any chronologically adjacent blocks (all $p > 0.05$), showing that participants did not benefit from expectations in terms of faster responses to expected stimuli. Similarly, there was no reliable benefit of expectations on response accuracy in any block (all $p > 0.05$), except for one ($t_{(22)} = 2.72$, $p = 0.013$, $d_z = 0.57$), nor was there a difference in accuracy benefit between adjacent blocks (Figure 5.6D; all $p > 0.05$). Results of all statistical tests are summarized in Table S5.5 and Table S5.6. Overall RTs were slower compared to experiment 1, as evident in Figure 5.6B, which was expected given the increased difficulty of the task due to the additional go/no-go signal. Response accuracies were marginally lower than in experiment 1, however still near ceiling (mean accuracy = 92.8%), indicating reliable task performance (Figure 5.6D). In sum, there was no behavioral facilitation by cross-modal SL, even when both cue and stimulus were task-relevant.

## Statistical learning facilitates behavioral responses to unimodal cue-stimulus pairs

After ruling out that a lack of attention to the cue modality caused the absence of behavioral facilitation by expectations, we hypothesized that the cross-modal nature of the cue-stimulus pairs may have prevented SL. Moreover, it is possible that the amount of exposure to the statistical regularities, or the strength of the association between cue-stimulus pairs, was not sufficient to yield reliable SL. To address these possibilities, we performed the same experiment as before, but replaced the auditory cues with visually presented object cues, while retaining all other parameters of experiment 1; hence resulting in a unimodal (visual-visual) associations paradigm with task-irrelevant cues.

Results, depicted in Figure 5.7, show behavioral facilitation effects following unimodal SL. As evident in Figure 5.7A-B, participants responded faster to expected than unexpected stimuli, starting with block 3. In fact, all blocks from block 3 onward showed significantly faster responses to expected stimuli (blocks 3-8: all $p < 0.001$, $d_z > 0.8$; blocks 1 and 2: $p > 0.05$). Interestingly, the magnitude of the expectation induced RT benefit did only reliably increase between blocks 2-3 ($t_{(24)} = 3.95$, $p = 6e\text{-}4$, $d_z = 0.79$) and blocks 4-5 ($t_{(24)} = 4.09$, $p = 4e\text{-}4$, $d_z = 0.82$), precisely the blocks between which additional blocks with deterministic associations took place (all other $p > 0.05$, or a significant decrease in RT benefit between blocks 5-6). Results of all statistical tests are summarized in Table S5.7 and S5.8. In accordance with these results, Bayesian model comparison (Figure 5.7C) yielded the best model fit for the 'two switchpoints' model (lowest WAIC), closely followed by the 'one switchpoint' model. Interestingly, the 'two switchpoint' model reliably fit the data better than even the 'two switchpoints & linear' model ($W$=37, $p$=7e-04). Thus, combined these results suggest that participants learned to benefit from expectations at distinct switchpoints instead of gradually increasing in RT benefits. In fact, the switchpoints were modeled such that learning took place only from block 2 to 3, and from block 4 to 5, thus at the time that the deterministic association blocks were performed.

We conducted an additional analysis, quantifying how much participants benefitted from each exposure to an expected cue-stimulus pair. In line with results presented above, data show that participants improved in RT benefit during deterministic blocks, with an average RT gain per exposure to an expected pair of 0.43ms ($t_{(24)} = 4.64$, $p$ = 1e-4, $d_z = 0.93$), but decreased in RT benefit during probabilistic blocks, with an average RT loss per exposure to an expected pair of -0.27ms ($t_{(24)} = -2.35$, $p = 0.027$, $d_z = -0.47$), likely due to the presence of intermittently presented unexpected pairs. To avoid potential ceiling effects following learning in deterministic blocks, we also analyzed only the initial two probabilistic blocks (i.e., before exposure to deterministic associations). No reliable effect on RT benefit per exposure to expected pairs was found during the initial probabilistic blocks (RT loss per exposure to an expected pair of -0.34ms; $t_{(24)} = -1.23$, $p = 0.230$, $d_z = -0.25$). Thus, results suggest that participants exclusively learned the statistical regularities during blocks with deterministic associations. In fact, the acquire statistical regularities may have even been unlearned, or at least the reliance on these regularities decreased, when exposed to probabilistic associations, even though expected images were four times more likely than each unexpected image.

A similar albeit less reliable pattern of results is also evident in terms of response accuracies, depicted in Figure 5.7D and Figure 5.7E. Again, behavioral facilitation due to expectations (i.e., more accurate responses to expected than unexpected stimuli) arose in block 3 and block 5, with most blocks following block 3 showing a response accuracy facilitation by expectations (4 of 6 blocks: $p < 0.05$). A reliable increase in accuracy benefits between adjacent blocks was only found from block 4 to block 5 ($t_{(24)} = 2.2$, $p = 0.037$, $d_z = 0.44$). Detailed results for response accuracy analyses are summarized in Table S5.8.

Finally, we asked whether the RT benefits reported above reflect perceptual surprise and/or response preparation. That is, an unexpected stimulus is not only perceptually surprising, but may also require subsequent adjustments to the predicted motor response, if the unexpected stimulus is of a different category than the expected object (e.g., expected an electronic object, but saw an unexpected non-electronic object). Thus, we performed an analysis of the RT benefit per block, but only analyzed expected images, and unexpected images which required the same response as the expected stimulus would have required. Results of this analysis are depicted in Figure 5.7G-H, and are qualitatively similar to the analysis including all unexpected trials (Figure 5.7A-B). Detailed results of the statistical tests are summarized in Table S5.8E-F. In brief, there is a reliable RT benefit of prediction from block 3 onward (with the exception of block 4), even after accounting for possible effects of response preparation. Therefore, perceptual expectations facilitate behavioral response speed alongside any potential motor response adjustments. Indeed, note that the magnitude and reliability of the RT effect appears to be reduced, thereby suggesting that motor preparations also contributed to the overall RT benefit induced by valid expectations.

In sum, results from experiment 3 demonstrate that the amount of exposure to the statistical regularities, as well as the reliability of the associations, were sufficient for robust SL to emerge. In particular, data show that behavioral responses in terms of RTs, and to a lesser degree response accuracies, are improved by unimodal perceptual expectations. Moreover, these behavioral benefits of expectations exclusively emerged after blocks with deterministic associations, suggesting that participants may particularly depend on strong, reliable associations during incidental SL.

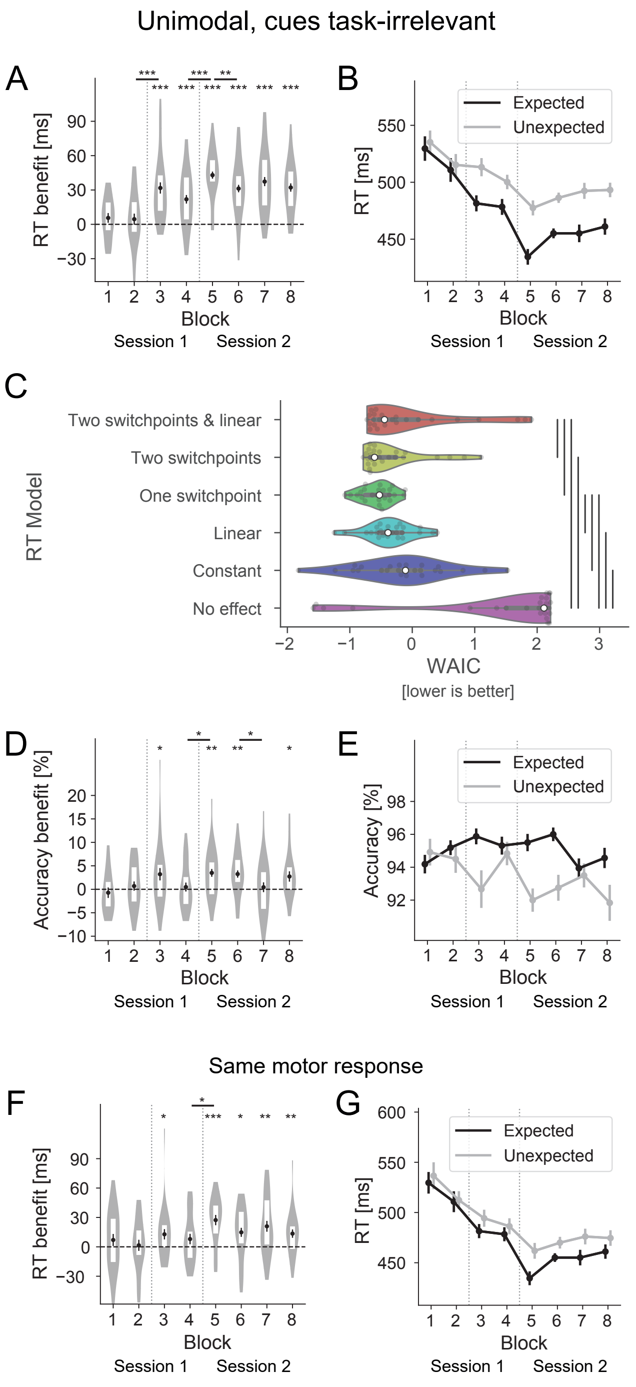

Unimodal, cues task-irrelevant
A
RT benefit [ms]
90
60
30
0
−30
***
***
***
**
***
***
***
***
***
***
1 2 3 4 5 6 7 8
Block
Session 1
Session 2
B
Expected
Unexpected
RT [ms]
550
500
450
1 2 3 4 5 6 7 8
Block
Session 1
Session 2
C
RT Model
Two switchpoints & linear
Two switchpoints
One switchpoint
Linear
Constant
No effect
−2 −1 0 1 2 3
WAIC
[lower is better]
D
Accuracy benefit [%]
20
15
10
5
0
−5
−10
*
*
**
**
*
*
1 2 3 4 5 6 7 8
Block
Session 1
Session 2
E
Expected
Unexpected
Accuracy [%]
98
96
94
92
1 2 3 4 5 6 7 8
Block
Session 1
Session 2
Same motor response
F
RT benefit [ms]
90
60
30
0
−30
*
*
***
*
**
**
1 2 3 4 5 6 7 8
Block
Session 1
Session 2
G
Expected
Unexpected
RT [ms]
600
550
500
450
1 2 3 4 5 6 7 8
Block
Session 1
Session 2

**FIGURE 5.7 Statistical learning facilitates behavioral responses to unimodal cue-stimulus pairs.**

Valid perceptual expectations result in behavioral facilitation. **(A)** shows the expectation induced RT benefit ($RT_{benefit} = RT_{unexpected} - RT_{expected}$) for each block. Significant RT benefits due to valid expectations are evident from block 3 onward, following the first block with deterministic associations (vertical gray dotted lines). All subsequent blocks show a strong and reliable RT benefit. The dashed horizontal line denotes zero (i.e., no benefit of expectations). Vertical dotted lines indicate when a block with deterministic associations took place between the displayed blocks with probabilistic associations (i.e., before block 3 and 5). RT benefits between adjacent blocks increased from block 2 to 3, and from block 4 to 5. Error bars indicate within-subject SEM. White bars indicate first and third quartile ranges. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (uncorrected p-values). **(B)** shows RTs to expected and unexpected stimuli. As in experiment 1 and 2, RTs become noticeably faster over time, indicating the generic task-familiarity effect. However, at the same time responses to expected stimuli are reducing in RTs more than to unexpected stimuli (RT benefit). **(C)** depicts results of Bayesian model fits (WAIC). The 'two switchpoints' model fit the data best (lowest WAIC), while the 'no effect' model performed worse than all other models. These results suggest that perceptual expectations did significantly influence RTs, following two distinct learning events, in particular during the blocks with deterministic associations. Vertical lines denote statistically significant differences between WAICs ($p < 0.05$). **(D)** shows expectation induced accuracy benefits ($Accuracy_{benefit} = Accuracy_{expected} - Accuracy_{unexpected}$). **(E)** shows response accuracies to expected and unexpected stimuli per block, indicating stable response accuracies near ceiling (mean: 94.2%). **(F)** RT benefit by valid predictions without effects of motor response preparation. Analyzed were expected trials and unexpected trials which required the same response as the expected object. Results are similar to those reported in (A), albeit of a smaller magnitude and less reliable. Thus, expectations facilitate behavioral responses even when accounting for motor response adjustments. **(G)** shows RTs to expected stimuli and unexpected stimuli requiring the same response as the expected object. A similar pattern of results is evident as in (B).

## Discussion

In the present study we investigated the evolution of cross-modal and unimodal incidental statistical learning (SL). Participants were exposed to probabilistically and deterministically associated cue-stimulus pairs, while performing a categorization task on the stimulus image. While participants were informed about the presence of statistical regularities and knowledge of these regularities could aid in task performance, learning was not requested, nor required to perform the task, thus constituting incidental learning. Data showed no facilitation of reaction times (RT) or response accuracies, nor a modulation of neural responses, as measured by fMRI BOLD, by statistical regularities during presentation of cross-modal (audio-visual) pairs (experiment 1). Moreover, no facilitation of behavioral responses was evident even when both cue and stimulus modalities were task-relevant (experiment 2). However, RTs were faster and responses more accurate to expected stimuli when unimodal (visual-visual) pairs were presented (experiment 3). Moreover, we demonstrated that learning of unimodal pairs primarily took place during exposure to deterministic, but not probabilistic associations. Thus, our results provide evidence for robust unimodal, but not cross-modal, incidental SL from deterministic associations.

## Incidental statistical learning depends on strong statistical regularities

The present results demonstrate that participants readily learn and use incidentally acquired unimodal statistical regularities (experiment 3), shown here in terms of faster and more accurate responses to expected compared to unexpected stimuli. These results echo conclusions of previous studies showing reliable behavioral facilitation following incidental SL [50,51,54,58,60,73,145]. Learning was incidental in the present study, as it took place without any instructions to learn, nor the need to acquire the statistical regularities for performing the task. Nevertheless, valid expectations were associated with improvements in behavioral performance. In an additional analysis we also showed that expectations improved response speed, even if the required motor response was identical for expected and unexpected stimuli. However, the magnitude of the response speeding appeared to be smaller and less reliable than when unexpected trials with different motor responses were included in the analysis. Thus, these result suggest that perceptual expectations benefit behavioral responses by facilitating perception and inducing response preparations.

Interestingly, results showed that reliable learning appeared to take place exclusively during exposure to deterministic associations, while no learning was apparent during blocks with probabilistic associations. In particular, there was no evidence of SL during the initial two (probabilistic) blocks, even though 16 repetitions of each expected cue-stimulus pair were shown per block, compared to only 4 repetition of each unexpected pair. Given the number of trials with expected pairs and the fairly simple transition matrix (four cue-stimulus pairs), one may have expected at least some evidence for SL following the initial two blocks, particularly in light of the demonstration of reliable learning following limited exposure to statistical regularities in classic studies of SL; e.g. only 24 repetitions of each of four triplets [50,54], or six repetitions of eight image pairs, even with interference by two unpaired images [73]. Moreover, learning following such limited exposure was demonstrated if triplets were interrupted by unattended triplets within the same modality [60]. Additionally, even learning of differences in conditional probability between triplets following limited exposure has been shown [50]. Compared to these studies, the number of exposure to the expected pairs was larger in the first two probabilistic blocks alone in the present study. However, a crucial difference is that the pairs and triplets in the above cited studies were deterministically associated. In our data, no SL occurred during the initial two blocks with probabilistic associations (32 repetitions per expected pair), even though expected trailing images were four times more likely than each unexpected image. Following a mere 24 exposures per image pair in the first deterministic block, substantial behavioral facilitation by expectations emerged. Similarly, following the second set of deterministic blocks, another significant increase in behavioral benefits was evident. In contrast, no evidence of additional SL was found during any probabilistic block. Indeed, there was even evidence that participants decreased in RT benefits during exposure to probabilistic associations. The observation that learning took exclusively place during blocks with deterministic associations was further supported by Bayesian model comparison, showing that the model implementing learning exclusively during the two sets of deterministic blocks explained the RT data best. In fact, this model reliably fit the data better than an identical model with two learning points and an additional linear learning effect during probabilistic blocks, thus implying that no additional SL occurred due to exposure to probabilistic associations.

Our data also show that once acquired, statistical regularities appear to be maintained during probabilistic blocks, as evident by the fact that even the last block still yielded a reliable RT benefit. A similar robustness of learned priors has been shown in contextual cueing experiments [163]. However, we did observe evidence that the RT benefit may have decreased slightly during probabilistic blocks, suggesting that decreases in statistical reliability are tracked even if people are not instructed to learn the underlying probabilities. That said, it should be noted that this decrease does not necessarily reflect a decrease of the learned association, but may also reflect generic processes, such as an increase in fatigue, and thus reduced exploitation of the perceptual priors, over the course of the experiment.

In brief, the present results suggest that incidental SL may primarily occur when humans are exposed to particularly strong statistical associations. Our data show that learning of weaker, probabilistic associations is comparably negligible – indeed, there was no evidence for learning from probabilistic associations. These results are surprising and we do not believe that incidental SL of probabilistic associations is categorically impossible [162]. However, the required exposure for reliable learning may substantially exceed the amount required for learning from deterministic regularities [161]. We address additional explanations and interpretational concerns later in the discussion (see: *Limitations*).

## Actively attending cue-stimulus pairs does not suffice for cross-modal incidental statistical learning

Surprisingly, we did not find any evidence of cross-modal SL during experiment 1 and 2. In particular, we showed that the 'no effect' model had a better fit to the RT data than any other model. Moreover, there was no reliable RT or accuracy benefit due to expectations, nor any increase in the expectation benefit between adjacent blocks. One plausible explanation for the absence of cross-modal SL in experiment 1 is that auditory cues were ignored by participants, because the auditory modality was task-irrelevant. Indeed, previous studies have suggested that attention gates the effects of visual SL [62,145]. We ruled out this explanation by showing in experiment 2 that, even when both cue and stimulus were task-relevant, and thus actively attended, no SL was evident for cross-modal associations. Therefore, attending and actively processing both cue and stimulus does not suffice for cross-modal incidental SL to arise, unlike in unimodal SL [113,145]. The present results support and extend results of previous studies using explicit assessments of SL. Indeed, at least one study of artificial grammar learning concluded that cross-modal learning may be substantially more difficult than unimodal learning [164]. We show an absence of cross-modal SL using a different paradigm and additionally provide fMRI data demonstrating that cross-modal SL did not only fail to produce behavioral facilitation, but also did not result in neural modulations anywhere in cortex. Thus, combined the data suggest that no cross-modal SL occurred irrespective of whether cue and stimulus were task-relevant.

## Modality-specific vs domain general contributions to incidental statistical learning

Recent proposals have attempted to unify arguments for modality-specific and domain general mechanisms of SL. For example, Frost et al. [63] hypothesize that local computations within unimodal sensory areas, as well as a multi-domain network, operating on modality specific representations, both contribute to detect statistical regularities. Our results show reliable incidental SL within, but not across modalities in behavioral measures or BOLD responses. In line with previous studies showing extensive expectation suppression in sensory areas following unimodal SL [21–23,26,113,145,159], the present results support the notion that local computations within sensory modalities may be crucial in detecting statistical regularities and pose one possible mechanism by which SL can occur [63,165]. In contrast, cross-modal SL would depend on an integration of information in multisensory areas and regularity detection in the hypothesized domain general network [63]. Following this account, potential explanations for the absence of cross-modal SL in the present data are a failure of the information integration in multisensory areas and/or lack of regularity detection in the domain general network. While we cannot distinguish between these two explanations, the following section will discuss possible explanations why cross-modal SL did not arise in the present study.

## Cross-modal statistical learning may depend on explicit learning

While the present data fail to provide any evidence for incidental SL of cross-modal regularities, there is ample previous evidence that cross-modal associations can in principle be learned and can affect behavior [18,25,166–168]. This apparent contradiction raises the question under which circumstances cross-modal learning may occur and how these conditions differ from the present study. A crucial difference may be the type of SL that was investigated here compared to previous studies, namely implicit and incidental learning compared to explicit learning.

A rich research tradition has linked implicit learning to non-declarative memory, and explicit learning to declarative memory, each associated with different neural mechanisms. The medial temporal lobe (MTL) is thought to be a key area in the formation of declarative memory [169], and thus explicit learning. Moreover, MTL has been suggested to be involved in perceptual predictions as well [170,171]. Notably, the hippocampus, part of the MTL, receives and sends input to all sensory modalities [172], thus constituting a plausible candidate to support cross-modal SL. Moreover, conscious awareness of a stimulus is associated with wide-spread information broadcasting across a network of distributed cortical areas [173,174], which analogously may hold for explicit learning, allowing the broadcasting of representations of the associations between stimuli from different modalities across a broad network. Thus, cross-modal SL may particularly depend on explicit learning processes, possibly involving the MTL and information integration across a distributed network, thereby enabling domain-general learning. On the other hand, unimodal SL may not necessarily depend on the MTL (although see: [31]) or wide-spread broadcasting. This in turn, would allow unimodal SL to unfold implicitly, relying on local computations in sensory cortex, thus giving rise to the widely reported sensory suppression of expected stimuli [21–23,26,113,145,159], thereby resulting in modality-specific SL [63–67].

This account proposes that previous studies showing cross-modal SL should involve (more) explicit learning, and hence explicit knowledge, compared to the present experiments. While difficult to assess formally, in line with this suggestion is that studies reporting cross-modal SL frequently used passive exposure during familiarization [167,168], with some providing explicit instructions that a test will follow after familiarization [166]. It is plausible that participants actively and explicitly learned statistical regularities during passive familiarization, given the absence of any competing task and particularly given the instruction of a subsequent test. Of the three above cited studies only one reported testing for explicit knowledge of the underlying regularities and did indeed find evidence of explicit knowledge following passive exposure [168]. Thus, these results support the hypothesis that explicit learning, and as a consequence explicit knowledge, may underlie previous reports of cross-modal SL. Moreover, other studies reporting cross-modal SL used substantially simpler association matrices consisting of only four cue-stimulus combinations [18,25], compared to the present 16 cue-stimulus combinations. Given that simple cue-stimulus combinations are more likely to be noticed explicitly (i.e., consciously keeping track of four compared to 16 probabilities), it is possible that cross-modal SL in these studies may also reflect a result of explicit learning. In contrast, in the present study we used comparably complex statistical regularities while participants were engaged in an active task, which did not explicitly require, nor request from them, the learning of the underlying associations. As a consequence learning may have been less explicit than in previous studies. Indeed, during debriefing participants frequently reported not noticing any statistical regularities. Therefore, learning in the present study may have relied largely on non-declarative memory systems in the specific sensory areas, without MTL involvement or broadcasting of the representations across multi-domain areas, thereby precluding cross-modal, but not unimodal SL.

## Limitations

Finally, it is worth considering alternative explanations for the absence of cross-modal statistical learning. For instance, while we show that attending both cue and stimulus is not sufficient to yield cross-modal SL, our task did, strictly speaking, not require participants to attend the identity of the auditory cues to perform the task; i.e., only detection of simultaneously presented white noise was necessary, hence identity cue may have been ignored. Therefore, we cannot rule out that cross-modal learning would have arisen, if cue identity would have been task-relevant. However, given that cue identity was also task-irrelevant in the unimodal experiment (Experiment 3), and that unimodal SL has been demonstrated in both humans and non-human primates during passive fixation and a range of cover-tasks [23,26,31,50,58–60,73,113,145], this explanation still suggests a qualitative difference between cross-modal and unimodal statistical learning, as argued here.

Next, it is possible that learning in the unimodal (visual-visual) case was aided by the fact that visual stimuli were naturalistic objects, unlike the utilized artificial auditory cues (sine-wave sounds). Naturalistic objects have semantic content, fitting into readily available categories and possessing clear linguistic labels. On the other hand, sine-wave tones may be categorized based on low level characteristics, such as pitch, and an according verbal label assigned (e.g. 'the high pitched sound'). While such sounds may evoke some semantic associations, they certainly do not possess the rich and reliable semantic content and categorization as naturalistic object images afford. Thus, these differences between visual and auditory stimuli may have made SL easier in the unimodal case, because category representations and linguistic labels may help structuring and acquiring statistical regularities. However, given that previous studies show unimodal SL in audition using artificial stimuli [22], it seems unlikely that a lack of semantic content accounts for the complete absence of cross-modal SL. Nonetheless additional research is required to conclusively rule out alternative explanations for the present results, for instance by showing reliable learning of artificial auditory-auditory associations following the same amount of exposure, while utilizing a comparable task.

Finally, we consider two limitations which may explain the absence of learning during probabilistic blocks. It is possible that the presence of blocks with deterministic associations may have prevented reliable learning from probabilistic associations due to changes in the underlying statistics; i.e., volatility in the environment. In other words, participants may have noticed that the statistical regularities are more reliable in deterministic blocks and thus, discarded information during probabilistic blocks. While theoretically sound, we do not find any evidence for SL during (probabilistic) blocks 1-2, before any exposure to the deterministic associations, even though each expected pair was presented 32 times. Thus, if any learning occurred during these first probabilistic blocks, it must be so limited in scope, that it was not detectable in the present data. That said, without more extensive exposure to only probabilistic associations, we cannot conclusively rule out this explanation.

A related limitation, possibly occluding learning during probabilistic blocks, is that deterministic blocks may have led to a ceiling effect for the RT benefit before any learning from probabilistic associations could be observed. Irrespective of the degree of learning, there is an upper limit on the magnitude of the RT benefit expectations can provide in the object categorization task. Thus, if learning from probabilistic associations is indeed very slow, therefore not evident during initial probabilistic blocks, it is possible that an effect could emerge in later blocks, if the maximum RT benefit was not already reached following the deterministic blocks. However, it should be noted that no learning was observed during any of the initial four probabilistic block, while learning still occurred during the second set of deterministic blocks (i.e., two switchpoints). Nonetheless, a study with extensive exposure to only probabilistic association could resolve this limitation. Moreover, it should be noted that learning from probabilistic associations must be possible, because the reliability of probabilistic associations can be (almost) arbitrarily increased. However, surprisingly we show that no learning at all took place during the here implemented associations; i.e., an expected stimulus which is four times more likely than each one of three unexpected stimuli. This drastic absence of any learning suggests the possibility that nonlinearities may exist in how people learn from increasingly reliable associations; indeed, recent work even suggests that, at least if assessed explicitly, humans treat deterministic and probabilistic regularities differently [160]. Thus, an interesting question is how the potentially nonlinear learning profile for incidental SL may look like.

### Conclusion

In sum, we found that incidental SL was not observed during exposure to cross-modal statistical regularities, but was robustly present during exposure to unimodal statistical regularities. These results suggest that incidental SL may crucially depend on modality-specific mechanisms, favoring the acquisition of unimodal over cross-modal associations. We speculate that cross-modal SL may depend on broadcasting of multimodal stimulus representations into a domain general network, possibly triggered by explicit learning of the underlying regularities. Finally, we demonstrate that the acquisition of unimodal statistical regularities depends on particularly strong (here deterministic) associations, and is negligible during exposure to weaker, probabilistic associations.

## Materials and Methods

### Participants and Data Exclusion

For all experiments we recruited adult healthy, right-handed volunteers from Radboud University research participation system. All three experiments followed the guidelines for ethical treatment of research participants by CMO region Arnhem-Nijmegen, The Netherlands. Sample sizes were based on obtaining 80% power to detect an effect size of Cohen's *d* >= 0.6, resulting in a desired n = 24. For experiment 1, we acquired MRI data from 26 healthy volunteers. All participants were right handed and had normal or corrected-to-normal vision. Data from two participants were excluded from all analysis due to incomplete datasets, resulting in a final sample of 24 analyzed datasets (16 females; age 23.8 ± 4.5 years, mean ± SD). Additionally, two participants were rejected from fMRI data analysis due to due excessive motion during MRI scanning, formalized as showing significantly more (3 SD above the group mean) relative motion events exceeding half a voxel size (i.e., 1 mm).

We collected behavioral data from two additional samples of 24 and 25 healthy volunteers for experiments 2 and 3 respectively. For experiment 2, data from one participant was rejected due to an excessive amount of incorrect responses (accuracy 3 SD below the group mean). Thus the analyzed sample sizes were n = 23 for experiment 2 (16 females; age 23.5 ± 4.0 years, mean ± SD) and n = 25 for experiment 3 (12 females; age 24.6 ± 3.5 years, mean ± SD).

## Stimuli and experimental paradigm, Experiment 1: multimodal, cue task-irrelevant

The experimental design was similar in the three experiments. Thus, in the following we first describe experiment 1, an fMRI experiment with multimodal associations and task-irrelevant cues. Next, we illustrate the differences of experiments 2 (multimodal associations with task-relevant cues) and experiment 3 (unimodal associations with task-irrelevant cues) with respect to experiment 1. All experiments consisted of two sessions on two consecutive days.

### Paradigm

On each trial participants were presented with an auditory cue (~300 ms), an ISI of 200 ms, then an object image at fixation (500 ms), followed by a variable ITI of 3000-6000 ms; Figure 5.1A illustrates a single trial. Participants were instructed to indicate by button press whether the image showed an electronic or non-electronic object. Both accuracy and response speed were emphasized in the instructions. Each participant was presented with four different auditory cues and four object images. Cues were identical for all participants, while the four objects were randomly sampled from a database of 56 object images, with the only constraint that each participant saw two electronic and two non-electronic objects. We ensured that participants recognized the objects as clearly (non-)electronic. Crucially, the auditory cues were predictive of the identity of the object image, thus making an object (un-)expected given the cue. Each cue was associated with one object, with 57.1% reliability of the associations; i.e., each expected stimulus was four times more likely than each unexpected stimulus given its associated cue. Figure 5.1B shows the transition matrices. All objects and cues occurred equally often throughout the experiment and objects served both as expected and unexpected stimuli. Each block (run) consisted of 112 trials (64 expected, 48 unexpected) taking ~14 minutes, including instruction and seven null events. Null events were 11 second events with blank screens only showing a fixation bull's-eye. The stimuli and their association remained identical throughout the experiment, except for a deterministic variation of the main task, which used the same cue-stimulus pairings, but only consisted of expected cue-stimulus pairs, and shorter ITIs (1000-3000 ms). The rationale of these deterministic blocks was to promote additional SL. Figure 5.1C illustrates the order of blocks. Participants were not informed of the different versions of the task, and were usually unable to notice any difference between the blocks, except for the shorter ITI. However, to reduce between-subject variability, participants were informed of the presence of statistical regularities governing the relationship between the cue-stimulus pairs, without indicating which particular associations are present. Throughout the entire block a fixation bull's-eye (outer-circle, 0.7° visual angle) was presented. Participants were instructed to maintain fixation on the bull's-eye. Performance, both in terms of accuracy and reaction time were displayed at the end of each block.

### *Localizer*

In addition to the main task blocks participants performed a localizer run used to define ROIs and obtain expectation-neutral representations of each object image. During the localizer run the four object images were shown, one at a time, flashing at 2Hz (300 ms on, 200 ms off) for 11 seconds. Each image was repeated ten times. Additionally, the phase-scrambled version of each image was also repeated five times in order to contrast responses to intact compared to scrambled objects. Trial order was pseudo-randomized, excluding direct repetitions of the same stimulus. Participants were instructed to press a button whenever a low contrast version of the stimulus was shown. Targets occurred once during each 11 second trial, and were shown for one cycle (i.e., 300 ms). This task ensured that participants attended the object stimuli. The duration of the localizer run was ~14 min.

### *Procedure*

On each day participants performed four blocks (runs) of the probabilistic main task, as described above, in the MRI scanner. Additionally, participants performed a localizer run and practice block at the beginning of the first day. The practice block used different stimuli, and ensured that participants understood the task. Before performing the practice block, the sound volume was adjusted for each participant to ensure that they could hear the stimuli over the noise of the MRI scanner. On day one, after two blocks of the main task, participants performed one block of the deterministic version of the main task during which a T1-weighted image was acquired. Next, participants performed another two blocks of the probabilistic main task. Finally, outside of the MRI scanner, participants performed two more blocks of the deterministic version. In total day one took ~2.15 hours, of which 1.30 hours were in the scanner. Day two started with two blocks of the deterministic version of the task, outside the MRI scanner. Next, as during day one, two blocks of the probabilistic main task were performed in the MRI, followed by one deterministic block, and two more probabilistic main task blocks. Day two concluded with a final run of the localizer and took ~2 hours in total. Across the entire experiment participants performed 896 trials of the probabilistic and 480 trials of the deterministic version of the task, thereby resulting in 248 repetitions per expected cue-stimulus pair and 32 repetitions per unexpected pair.

### *Stimuli*

Auditory stimuli were designed to be expectation neutral (i.e., avoiding naturalistic tones), distinct from one another, and well audible in the MRI scanner (i.e., avoiding frequencies and timbre of the MRI noise). Thus, we used pure tones (sine waves) at 450 and 1000 Hz, as well as two sliding sine tones linearly modulating from 800-1250 and 1200-750 Hz. Similar tones, but at different frequencies were used during the practice block (725, 1275, 500-950, and 1100-650 Hz). All sounds were sampled at 48kHz, lasted 300 ms, and were padded with a 5 ms fade-in and fade-out. During MRI scanning auditory stimuli were presented using in-ear, MR-compatible headphones (Insert Earphones Model S14; Sensimetrics Corporation, MA, USA). Outside the scanner over-ear headphones (Sennheiser HD-202) were used.

The 56 object images were a subset of the stimuli used in one of our previous experiments (Richter and de Lange, 2019), taken from a larger stimulus set from Brady et al. (2008). Images were 5° x 5° in visual angle, presented on a mid-gray (128,128,128 RGB) background. During MRI scanning stimuli were back-projected using an EIKI LC-XL1000 projector at 1024 x 768 pixel resolution and 60 Hz refresh rate. The screen was visible using an adjustable mirror. Outside the scanner stimuli were presented on a LCD screen (BenQ XL2420T, 1920 x 1080 pixel resolution, 60 Hz refresh rate), while keeping the visual angle the as similar as possible as during MRI scanning.

## Experiment 2: multimodal, cue task-relevant

Experiment 2 was identical to experiment 1, except for the addition of a go/no-go condition based on the auditory cue. Participants still performed the classification of the electronic items, however only when a standard auditory cue was played (go trials). On 20% of trials the auditory cue was played, but with white noise superimposed to the sine wave sound, thereby indicating a no-go trial. Participants were instructed to withhold any response to objects on no-go. Thus, during experiment 2 both, the auditory cue and visual stimulus were task-relevant and presumably attended. Besides this modification experiment 2, was purely behavioral and therefore did not include the localizer runs. However, the ITI duration and general procedure was kept identical to experiment 1, including a 10 minute break to mimic the transitions from inside the MRI to the behavioral part of experiment 1.

## Experiment 3: unimodal, cue task-irrelevant

Experiment 3 was identical to experiment 1, except for that the auditory cues were replaced with visual cues. Visual cues were pseudo-randomly sampled from the same database that also supplied the object stimuli in experiment 1. Thus, experiment 3 consisted of object-object cue-stimulus pairs, with each object being presented for 500 ms, without ISI. Participants performed the same electronic item identification task, as during experiment 1, in this case only on the trailing (second) object on each trial. The same statistical regularities as in experiment 1 governed the transitions between objects, including an identical number of trials and blocks. As experiment 2, experiment 3 was purely behavioral, and thus the localizer runs were omitted. The transition matrix, ITI and general procedure were kept identical to experiment 1 and 2.

## fMRI data acquisition

MRI data was acquired on a 3T Skyra scanner (Siemens, Erlangen, Germany), using a 32-channel head coil. For the acquisition of anatomical images a T1-weighted magnetization prepared rapid gradient echo sequence (MP-RAGE; GRAPPA acceleration factor = 2, TR/TE = 2300/3.03 ms, voxel size 1 mm isotropic, 8° flip angle) was used. A whole-brain T2*-weighted multiband-4 sequence (time repetition [TR] / time echo [TE] = 1500/33.4 ms, 68 slices, voxel size 2 mm isotropic, 75° flip angle, A/P phase encoding direction, FOV = 213 mm, BW = 1850 Hz/Px) was used to acquire functional images. The first three volumes of each run were discarded to allow for signal stabilization.

## Data analysis

### *Behavioral data analysis*

Behavioral data was analyzed in terms of accuracy and response time (RT). Outliers, which were arbitrarily defined as trials with RT < 200 ms or RT > 1500 ms, were rejected from analysis. Moreover, trials with no response (misses) were not included in RT analyses. No-go trials in experiment 2 were removed from the main analyses. Because only the probabilistic version of the main task contained unexpected and expected cue-stimulus pairs, blocks with deterministic associations were not included in the following analyses. In addition, participants with poor performance, indicating a lack of attention to the task, formalized as a mean accuracy of 3 SD below the group mean, were excluded from all analyses.

We analyzed RT and accuracy data from the probabilistic blocks, comparing responses to expected and unexpected object stimuli. First we averaged, for each subject separately, RTs across trials within each block for expected and unexpected stimulus pairs. In addition, we calculated the amount of expectation induced response benefits as $RT_{benefit} = RT_{unexpected} - RT_{expected}$, and $Accuracy_{benefit} = Accuracy_{expected} - Accuracy_{unexpected}$. We then averaged across subjects per block as well as calculated the within-subject normalized standard error of the mean [85] with bias correction [86]. For each block we analyzed the behavioral benefit induced by expectations using two-sided, one-sample t-tests, contrasting the observed benefit against zero (no benefit). Additionally, we contrasted the obtained RT and accuracy benefits between chronologically adjacent blocks using two-sided, paired t-tests. Thus, these tests assess whether the expectation benefit changed significantly from one to the other block. For all t-tests effect sizes were calculated as Cohen's $d_z$ [84].

If a reliable RT benefit was evident, we quantified in an additional analysis the magnitude of this RT benefit for each exposure to an expected cue-stimulus pair. To this end, we calculated the difference in RT benefit between chronologically adjacent blocks, divided by the number of exposures to expected pairs. We separately averaged the resulting 'RT benefit per expected pair' for those blocks separated by blocks with deterministic associations and those without deterministic associations. This yields how much participants improved in RT benefit for each exposure to an expected depending on whether these exposures took place during deterministic or probabilistic blocks. We also calculated the same metric for the first two probabilistic blocks separately, as these blocks precede the first deterministic block. This avoids the concern that changing probabilities in the underlying statistics, or a ceiling effect following learning in deterministic blocks, prevented learning in probabilistic blocks. Finally, we compared the obtained RT benefit per expected pair exposure against zero for deterministic, probabilistic and the first two probabilistic blocks.

Additionally, if we observed a reliable RT benefit by prediction, we performed an analysis in which we tested whether motor response preparations and/or perceptual surprise account for the RT benefit. To this end we repeated the analysis of the RT benefit per block, as outlined above, but only analyzed unexpected trials which required the same response (button press) as the expected stimulus. Thus, in this analysis responses to expected and unexpected stimuli are identical in terms of the motor response, including potential priming of the button press induced by expectation. Hence, only perceptual surprise differed in this analysis between expected and unexpected trials.

### *Data modelling and model comparisons*

In order to assess in more detail how and when expectation effects develop, as well as to provide evidence for the possible absence of an expectation effect, we performed Bayesian model comparison. For each participant, we modeled the log-transformed RTs per trial ($t$) as drawn from a normal distribution, with a given standard deviation ($\sigma$) and mean ($\mu$). The mean was modeled separately for each block ($r$) to allow for generic improvements in RT over time. RTs were modulated by an effect of expectations ($\Delta\mu$).

$$y_{unexp} \sim Normal(\mu_{unexp,r}, \sigma^2)$$
$$y_{exp}(r) \sim Normal(\mu_{exp,r}, \sigma^2)$$
$$\sigma \sim HalfNormal(2 \times S_{obs})$$
$$\mu_{unexp,r} \sim Normal(\overline{x}_{obs}, 2 \times S_{obs})$$
$$\mu_{exp,r} \sim \mu_{unexp,r} - \Delta\mu(t)$$

The precise modulation by expectations was modeled according to one of six models:
**No effect:** responses to expected and unexpected pairs do not differ; i.e., no benefit of expectations.
$\Delta\mu = 0$

**Constant:** responses to expected and unexpected pairs differ by a constant amount.
$\Delta\mu \sim Normal(0, 2 \times S_{obs})$

**Linear:** differences between expected and unexpected pairs change in a linear, trial-wise fashion; i.e., modelling a gradual learning process over time.
$\Delta\mu(t) = \lambda t$
$\lambda \sim Normal(0, 2 \times S_{obs})$

**One switchpoint:** initially no difference between expected and unexpected pairs until a change point occurs, after which conditions differ by a constant amount. Two possible switchpoints were fit, either block 3, or block 5, because these blocks follow the deterministic blocks, during which learning may particularly occur. Thus, this model implements a learning process taking place exclusively in one deterministic block.

$$\Delta\mu(r) \cong \begin{cases} Normal(0, 2 \times S_{obs}), & r > y * 2 + 3 \\ 0, & r \leq y * 2 + 3 \end{cases}$$

$$y \sim DiscreteUniform(0, 1)$$

**Two switchpoints:** similar to the one switch points model, but modelling two switchpoints, both block 3, and block 5. Thereby this model allows for learning to occur during both deterministic blocks.

$$\Delta\mu(r) = \Delta 1(r) + \Delta 2(r)$$

$$\Delta 1(r) \cong \begin{cases} Normal(0, 2\times S_{obs}), & r>3 \\ 0, & r\leq 3 \end{cases}$$

$$\Delta 2(r) \cong \begin{cases} Normal(0, 2\times S_{obs}), & r>5 \\ 0, & r\leq 5 \end{cases}$$

**Two switchpoints & linear:** maximal model, including both two switchpoints and the trial-wise linear effect of expectations. Thus, this model implements a gradual learning process, with two additional learning steps during the deterministic blocks.

$$\Delta\mu(r) = \lambda t + \Delta 1(r) + \Delta 2(r)$$

$$\lambda \sim Normal(0, 2\times S_{obs})$$

$$\Delta 1(r) \cong \begin{cases} Normal(0, 2\times S_{obs}), & r>3 \\ 0, & r\leq 3 \end{cases}$$

$$\Delta 2(r) \cong \begin{cases} Normal(0, 2\times S_{obs}), & r>5 \\ 0, & r\leq 5 \end{cases}$$

Finally, model fit was compared in terms of the Watanabe-Akaike information criterion; WAIC [175]. We calculated the median WAIC for each model type, and additionally compared average WAIC between model types using Wilcoxon signed-rank tests.

### *fMRI data preprocessing*

fMRI data was preprocessed using FSL 6.0 (FMRIB Software Library; Oxford, UK; www.fmrib.ox.ac.uk/fsl; [87], RRID:SCR_002823). The following steps were carried out for all fMRI data: brain extraction (BET), motion correction (MCFLIRT), temporal high-pass filtering (128 seconds), spatial smoothing (Gaussian kernel with full-width at half-maximum of 5 mm). Functional images were registered to each subject's anatomical image using FSL FLIRT's boundary-based registration (BBR). Anatomical image were linearly registered (12 degrees of freedom) to the MNI152 T1 2mm template.

### *fMRI data analysis*

First level fMRI data analysis was performed using the least squares separate approach as outlined in [89,153]. In this analysis a separate voxel-wise general linear model (GLM) is fit for each trial, with one regressor of interest, the current trial (onset and duration of the regressor corresponding to that of the (un-)expected stimulus). Additionally, regressors of no interest are fit, modelling all stimulus classes (minus the current trial), as well a generic nuisance regressors, including instruction screens and 24 motion regressors (FSL's standard + extended set of motion parameters; i.e., 6 standard motion parameters, the 6 temporal derivatives, and the squares of the standard and temporal derivatives). Stimulus events were convolved with a double gamma haemodynamic response function to account for the shape of the haemodynamic response. This analysis yields a parameter estimate map of the BOLD response for each trial. From these parameter estimate maps we extracted data using three ROI masks (V1, LOC, TOFC). Within each ROI, parameter estimates were averaged across voxels, thus yielding one parameter estimate for each trial and ROI. Subsequently, we modelled the trial-by-trial BOLD response for each subject separately, using the same Bayesian modelling procedure used for the RT data (for details see: *Data modelling*). Before Bayesian modelling of BOLD responses, we removed trials with incorrect behavioral responses, as well as too fast (< 200 ms) or too slow (> 1500 ms) responses. Moreover, for each participant outlier BOLD responses were removed, defined as parameter estimates exceeding 3 SD below or above the average parameter estimate. Finally, before performing Bayesian modelling, the parameter estimates were z-scored. In addition to the trial-wise modelling, we also calculated the average amount of expectation suppression ($BOLD_{expectation\ suppression} = BOLD_{unexpected} - BOLD_{expected}$) per block (run). As for behavioral data, we again compared the obtained expectation effect against zero (no effect) using two-sided, one-sample t-tests. We also compared average expectation suppression between all adjacent blocks using two-sided, paired t-tests (for details see: *Behavioral data analysis*).

In addition to the ROI analyses, we performed a whole-brain analysis, contrasting responses to unexpected compared to expected stimuli for the two sessions separately. To this end we fit voxel-wise GLMs to each subject's data using FSL FEAT. For each run the model consisted of a regressor for expected and unexpected occurrences of the object stimuli. Each stimulus event was modeled with 500 ms duration (presentation duration and onset of the object image), and convolved with a double gamma haemodynamic response function. Additionally, a regressor of no interest, modelling instruction screens was added to the model. Moreover, the temporal derivatives of these three regressors were added. Finally, FSL's standard + extended set of motion parameters were added to the model to account for head motion. Across runs data was averaged using FSL's fixed effects analysis within each session. Finally, for whole-brain analyses, data was averaged across participants using FSL's mixed effects analysis, FLAME 1. The contrast of interest was unexpected minus expected (yielding expectation suppression), averaged across runs within each session, as well as the difference in expectation suppression between the sessions. Multiple comparison correction was performed using Gaussian random-field cluster thresholding; cluster forming threshold $p < 0.001$ and cluster significance threshold $p < 0.05$.

### *ROI definition*

All ROIs were defined a-priori, following a similar procedure as in [145], using independent data. In brief, primary visual cortex (V1) masks were individually extracted from V1 labels using Freesurfer 6.0 cortex segmentation [97]; RRID:SCR_001847. To establish object selective lateral occipital complex (LOC) masks, a GLM was fit to each participant's localizer data, modelling intact and scrambled objects separately, plus nuisance regressors for motion and instruction events. Object-selective LOC masks were subsequently defined as voxels with significant preferential responses to intact objects compared to scrambled objects [95], within anatomically defined LOC (Harvard-Oxford cortical atlas, as distributed by FSL; RRID:SCR_001476). A default threshold for significant preferential responses of $z > 5$ was used, which was adjusted individually if resulting the LOC mask contained less than 300 voxels. The anatomical mask for temporal occipital fusiform cortex (TOFC) from the Harvard-Oxford cortical atlas was used to define TOFC. This anatomical mask was further constrained to voxels, which showed significant expectation suppression in both [113] and [145], thus resulting in a TOFC mask, which contains voxels which were previously shown to be sensitive to statistical regularities. All three ROI masks were further constraint, for each participant individually, to the 300 most active voxels during the localizer run, using the contrast intact object compared to baseline (i.e., no visual stimulation).

## Software

For MRI data preprocessing and analysis FSL 6.0 (FMRIB Software Library; Oxford, UK; www.fmrib.ox.ac.uk/fsl; [87] RRID:SCR_002823) was used. Additionally, custom Python 3.7.4 (Python Software Foundation, RRID:SCR_008394) scripts were employed for additional analyses and data visualization, making use of NumPy 1.17.2 [98] RRID:SCR_008633, SciPy 1.3.1 [176] RRID:SCR_008058, Matplotlib 3.1.1 [100] RRID:SCR_008624, Statsmodels 0.10.1 [177], Pandas 0.25.2 [178], PyMC3 3.7 [179]. Slice Display [102], implemented in Matlab 2017a (The MathWorks, Inc., Natick, Massachusetts, United States, RRID:SCR_001622), was used for whole-brain result visualization. Experiments were programmed in Presentation® software (version 20.2, Neurobehavioral Systems, Inc., Berkeley, CA, RRID:SCR_002521).

## Data and code availability

Data and code will be made available upon publication in the Donders Repository: http://hdl.handle.net/11633/aadhrw2q

# Supplemental Information

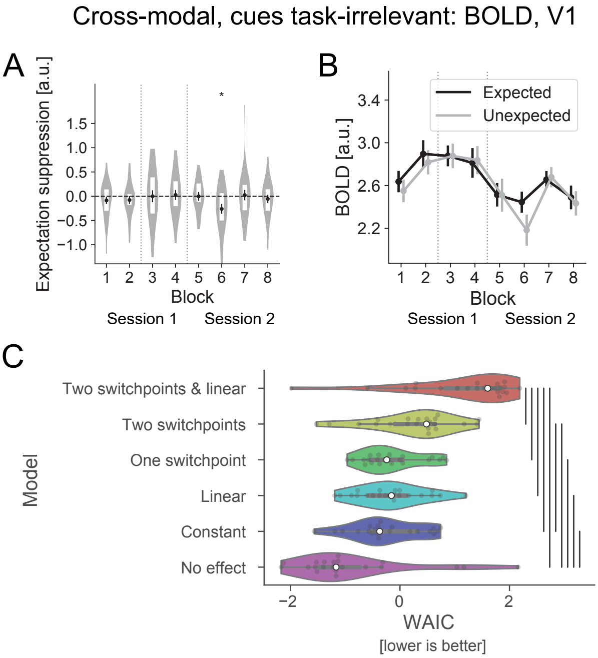


**FIGURE S5.1 No modulation of sensory processing by cross-modal statistical regularities in early visual areas.**

Cross-modal perceptual expectations do not modulate sensory responses in primary visual cortex (V1). (**A**) Displays expectation suppression ($\text{BOLD}_{\text{unexpected}} - \text{BOLD}_{\text{expected}}$) for each block. Again no evidence for an influence of expectation status on sensory processing is evident – i.e., expectation suppression is (near) zero in each block for both sessions and, in fact, did not statistically deviate from zero for any block. Moreover, expectation suppression did not differ between any adjacent blocks. The dashed lines indicates zeros; i.e., no difference in the response between expected and unexpected stimuli. Vertical dotted lines indicate when a block with deterministic associations took place between the displayed blocks with probabilistic associations (i.e., after session 1, block 2 and 4). Error bars indicate within-subject SEM. White bars indicate first and third quartile ranges. * $p < 0.05$ (uncorrected p-values). (**B**) BOLD response to expected and unexpected stimuli for each block and session. (**C**) Shows model fits in terms of WAIC. The 'no effect' model outperformed (lower WAIC) all other models, suggesting that expectations did not influence sensory processing in V1. Vertical lines denote statistically significant differences between WAICs ($p < 0.05$).

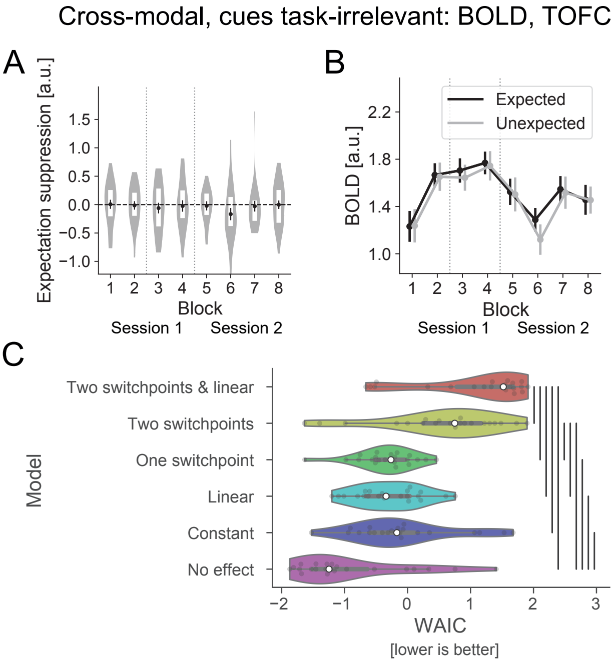


**FIGURE S5.2 No modulation of sensory processing by cross-modal statistical regularities in higher visual areas.**

Cross-modal perceptual expectations do not modulate sensory responses in higher visual areas (TOFC). (A) Displays expectation suppression ($\text{BOLD}_{\text{unexpected}} - \text{BOLD}_{\text{expected}}$) for each block. Again no evidence for an influence of expectation status on sensory processing is evident – i.e., expectation suppression is (near) zero in each block for both sessions and, in fact, did not statistically deviate from zero for any block. Moreover, expectation suppression did not differ between any adjacent blocks. The dashed lines indicates zeros; i.e., no difference in the response between expected and unexpected stimuli. Vertical dotted lines indicate when a block with deterministic associations took place between the displayed blocks with probabilistic associations (i.e., after session 1, block 2 and 4). Error bars indicate within-subject SEM. White bars indicate first and third quartile ranges. (B) BOLD response to expected and unexpected stimuli for each block and session. (C) Shows model fits in terms of WAIC. The 'no effect' model outperformed (lower WAIC) all other models, suggesting that expectations did not influence sensory processing in TOFC. Vertical lines denote statistically significant differences between WAICs ($p < 0.05$).

**TABLE S5.1 Model comparison for RT effects during cross-modal SL with task-irrelevant cues.**

Results of all pair-wise Wilcoxon signed-rank tests comparing model WAICs. SP = switchpoint. Redundant information omitted.

| | | Model | | | | |
|---|---|---|---|---|---|---|
| | | **2 SP** | **1 SP** | **Linear** | **Constant** | **No effect** |
| **Model** | **2 SP & linear** | W=85<br>p=0.063 | W=25<br>p=4e-04 | W=36<br>p=0.001 | W=32<br>p=7e-04 | W=18<br>p=2e-04 |
| | **2 SP** | - | W=3<br>p=3e-05 | W=29<br>p=5e-04 | W=26<br>p=4e-04 | W=42<br>p=0.002 |
| | **1 SP** | - | - | W=116<br>p=0.331 | W=108<br>p=0.23 | W=82<br>p=0.052 |
| | **Linear** | - | - | - | W=149<br>p=0.977 | W=77<br>p=0.037 |
| | **Constant** | - | - | - | - | W=60<br>p=0.01 |

**TABLE S5.2 Results of statistical tests evaluating RT and response accuracy benefits during cross-modal SL with task-irrelevant cues.**

(**A**) Results of one-sample t-tests comparing the observed RT benefit for each block against zero (no expectation benefit). Positive effects indicate RT benefits. (**B**) Results of paired-sample t-tests comparing the RT benefit between chronologically adjacent blocks. Positive effects indicate that RT benefits increased from one to the other block. (**C**) and (**D**) show the corresponding results for response accuracy. All presented p-values are uncorrected. Effect sizes are reported in terms of Cohen's *d*.

**A**

**RT benefit per block against zero**

| | | | |
|---|---|---|---|
| Block 1: | t(23)=-1.26 | p=0.219 | d=-0.26 |
| Block 2: | t(23)=-0.84 | p=0.408 | d=-0.17 |
| Block 3: | t(23)=1.73 | p=0.098 | d=0.35 |
| Block 4: | t(23)=-0.13 | p=0.900 | d=-0.03 |
| Block 5: | t(23)=0.9 | p=0.377 | d=0.18 |
| Block 6: | t(23)=3.15 | p=0.004 | d=0.64 |
| Block 7: | t(23)=0.69 | p=0.497 | d=0.14 |
| Block 8: | t(23)=1.2 | p=0.243 | d=0.24 |

**B**

**RT benefits between sequential blocks**

| | | | |
|---|---|---|---|
| Block 1 vs 2: | t(23)=0.24 | p=0.815 | d=0.05 |
| Block 2 vs 3: | t(23)=1.82 | p=0.081 | d=0.37 |
| Block 3 vs 4: | t(23)=-1.45 | p=0.161 | d=-0.3 |
| Block 4 vs 5: | t(23)=0.85 | p=0.404 | d=0.17 |
| Block 5 vs 6: | t(23)=1.96 | p=0.062 | d=0.4 |
| Block 6 vs 7: | t(23)=-1.75 | p=0.093 | d=-0.36 |
| Block 7 vs 8: | t(23)=-0.12 | p=0.903 | d=-0.03 |

**C**

**Accuracy benefit per block against zero**

| | | | |
|---|---|---|---|
| Block 1: | t(23)=-0.73 | p=0.470 | d=-0.15 |
| Block 2: | t(23)=0.19 | p=0.854 | d=0.04 |
| Block 3: | t(23)=-0.21 | p=0.837 | d=-0.04 |
| Block 4: | t(23)=0.6 | p=0.557 | d=0.12 |
| Block 5: | t(23)=1.28 | p=0.214 | d=0.26 |
| Block 6: | t(23)=2.11 | p=0.046 | d=0.43 |
| Block 7: | t(23)=0.97 | p=0.342 | d=0.2 |
| Block 8: | t(23)=-0.06 | p=0.954 | d=-0.01 |

**D**

**Accuracy benefits between sequential blocks**

| | | | |
|---|---|---|---|
| Block 1 vs 2: | t(23)=0.69 | p=0.494 | d=0.14 |
| Block 2 vs 3: | t(23)=-0.29 | p=0.777 | d=-0.06 |
| Block 3 vs 4: | t(23)=0.52 | p=0.607 | d=0.11 |
| Block 4 vs 5: | t(23)=0.6 | p=0.557 | d=0.12 |
| Block 5 vs 6: | t(23)=0.67 | p=0.510 | d=0.14 |
| Block 6 vs 7: | t(23)=-0.87 | p=0.394 | d=-0.18 |
| Block 7 vs 8: | t(23)=-0.64 | p=0.529 | d=-0.13 |

**TABLE S5.3 Model comparison for expectation suppression (BOLD) effects during cross-modal SL with task-irrelevant cues.**

Results of all pair-wise Wilcoxon signed-rank tests comparing model WAICs. SP = switchpoint. Redundant information omitted.

| | | Model | | | | |
|---|---|---|---|---|---|---|
| | | 2 SP | 1 SP | Linear | Constant | No effect |
| Model | 2 SP & linear | W=33<br>p=0.002 | W=30<br>p=0.002 | W=23<br>p=8e-04 | W=26<br>p=0.001 | W=27<br>p=0.001 |
| | 2 SP | - | W=74<br>p=0.088 | W=72<br>p=0.077 | W=62<br>p=0.036 | W=40<br>p=0.005 |
| | 1 SP | - | - | W=126<br>p=0.987 | W=107<br>p=0.527 | W=48<br>p=0.011 |
| | Linear | - | - | - | W=111<br>p=0.615 | W=58<br>p=0.026 |
| | Constant | - | - | - | - | W=63<br>p=0.039 |

**TABLE S5.4 Results of statistical tests evaluating expectation suppression during cross-modal SL with task-irrelevant cues.**

(**A**) Results of one-sample t-tests comparing expectation suppression ($\text{BOLD}_{\text{unexpected}}$ – $\text{BOLD}_{\text{expected}}$) for each block (run) against zero (no expectation effect on BOLD). Positive effects would indicate expectation suppression. (**B**) Results of paired-sample t-tests comparing expectation suppression between chronologically adjacent blocks. Positive effects would indicate that expectation suppression increased from one to the other block. All presented p-values are uncorrected. Effect sizes are reported in terms of Cohen's *d*.

A

Expectation suppression per block against zero

| | | | |
|---|---|---|---|
| Block 1: | t(21)=-1.03 | p=0.314 | d=-0.22 |
| Block 2: | t(21)=-1.01 | p=0.325 | d=-0.22 |
| Block 3: | t(21)=-0.0 | p=0.997 | d=-0.0 |
| Block 4: | t(21)=0.26 | p=0.798 | d=0.06 |
| Block 5: | t(21)=-0.04 | p=0.966 | d=-0.01 |
| Block 6: | t(21)=-2.54 | p=0.019 | d=-0.54 |
| Block 7: | t(21)=0.19 | p=0.851 | d=0.04 |
| Block 8: | t(21)=-0.64 | p=0.53 | d=-0.14 |

B

Expectation suppression between sequential blocks

| | | | |
|---|---|---|---|
| Block 1 vs 2: | t(21)=0.1 | p=0.923 | d=0.02 |
| Block 2 vs 3: | t(21)=0.58 | p=0.568 | d=0.12 |
| Block 3 vs 4: | t(21)=0.15 | p=0.884 | d=0.03 |
| Block 4 vs 5: | t(21)=-0.22 | p=0.829 | d=-0.05 |
| Block 5 vs 6: | t(21)=-1.99 | p=0.059 | d=-0.43 |
| Block 6 vs 7: | t(21)=1.94 | p=0.066 | d=0.41 |
| Block 7 vs 8: | t(21)=-0.49 | p=0.631 | d=-0.1 |

**TABLE S5.5 Model comparison for RT effects during cross-modal SL with task-relevant cues.**

Results of all pair-wise Wilcoxon signed-rank tests comparing model WAICs. SP = switchpoint. Redundant information omitted.

| | | Model | | | | |
|---|---|---|---|---|---|---|
| | | 2 SP | 1 SP | Linear | Constant | No effect |
| Model | 2 SP & linear | W=129<br>p=0.784 | W=61<br>p=0.019 | W=70<br>p=0.039 | W=74<br>p=0.052 | W=89<br>p=0.136 |
| | 2 SP | - | W=12<br>p=1e-04 | W=31<br>p=0.001 | W=41<br>p=0.003 | W=97<br>p=0.212 |
| | 1 SP | - | - | W=113<br>p=0.447 | W=134<br>p=0.903 | W=127<br>p=0.738 |
| | Linear | - | - | - | W=128<br>p=0.761 | W=135<br>p=0.927 |
| | Constant | - | - | - | - | W=134<br>p=0.903 |

**TABLE S5.6 Results of statistical tests evaluating RT and response accuracy benefits during cross-modal SL with task-relevant cues.**

(A) Results of one-sample t-tests comparing the observed RT benefit for each block against zero (no expectation benefit). Positive effects indicate RT benefits. (B) Results of paired-sample t-tests comparing the RT benefit between chronologically adjacent blocks. Positive effects indicate that RT benefits increased from one to the other block. (C) and (D), show the corresponding results for response accuracy. All presented p-values are uncorrected. Effect sizes are reported in terms of Cohen's *d*.

A

**RT benefit per block against zero**

| | | | |
|---|---|---|---|
| Block 1: | t(22)=-0.35 | p=0.729 | d=-0.07 |
| Block 2: | t(22)=-0.68 | p=0.501 | d=-0.14 |
| Block 3: | t(22)=1.81 | p=0.084 | d=0.38 |
| Block 4: | t(22)=-0.07 | p=0.944 | d=-0.01 |
| Block 5: | t(22)=1.82 | p=0.083 | d=0.38 |
| Block 6: | t(22)=2.64 | p=0.015 | d=0.55 |
| Block 7: | t(22)=1.41 | p=0.172 | d=0.29 |
| Block 8: | t(22)=1.5 | p=0.147 | d=0.31 |

B

**RT benefits between sequential blocks**

| | | | |
|---|---|---|---|
| Block 1 vs 2: | t(22)=-0.41 | p=0.683 | d=-0.09 |
| Block 2 vs 3: | t(22)=2.05 | p=0.052 | d=0.43 |
| Block 3 vs 4: | t(22)=-1.46 | p=0.158 | d=-0.31 |
| Block 4 vs 5: | t(22)=1.97 | p=0.062 | d=0.41 |
| Block 5 vs 6: | t(22)=-0.27 | p=0.789 | d=-0.06 |
| Block 6 vs 7: | t(22)=-0.69 | p=0.495 | d=-0.14 |
| Block 7 vs 8: | t(22)=0.53 | p=0.598 | d=0.11 |

C

**Accuracy benefit per block against zero**

| | | | |
|---|---|---|---|
| Block 1: | t(22)=-0.27 | p=0.787 | d=-0.06 |
| Block 2: | t(22)=-0.18 | p=0.862 | d=-0.04 |
| Block 3: | t(22)=0.06 | p=0.951 | d=0.01 |
| Block 4: | t(22)=-0.43 | p=0.673 | d=-0.09 |
| Block 5: | t(22)=2.72 | p=0.013 | d=0.57 |
| Block 6: | t(22)=0.55 | p=0.585 | d=0.12 |
| Block 7: | t(22)=0.43 | p=0.673 | d=0.09 |
| Block 8: | t(22)=0.05 | p=0.958 | d=0.01 |

D

**Accuracy benefits between sequential blocks**

| | | | |
|---|---|---|---|
| Block 1 vs 2: | t(22)=0.1 | p=0.925 | d=0.02 |
| Block 2 vs 3: | t(22)=0.18 | p=0.859 | d=0.04 |
| Block 3 vs 4: | t(22)=-0.34 | p=0.735 | d=-0.07 |
| Block 4 vs 5: | t(22)=1.81 | p=0.084 | d=0.38 |
| Block 5 vs 6: | t(22)=-1.19 | p=0.245 | d=-0.25 |
| Block 6 vs 7: | t(22)=-0.13 | p=0.897 | d=-0.03 |
| Block 7 vs 8: | t(22)=-0.24 | p=0.809 | d=-0.05 |

**TABLE S5.7 Model comparison for RT effects during unimodal SL with task-irrelevant cues.**

Results of all pair-wise Wilcoxon signed-rank tests comparing model WAICs. SP = switchpoint. Redundant information omitted.

| | | **Model** | | | | |
|---|---|---|---|---|---|---|
| | | **2 SP** | **1 SP** | **Linear** | **Constant** | **No effect** |
| **Model** | **2 SP & linear** | W=34<br>p=5e-04 | W=61<br>p=0.006 | W=146<br>p=0.657 | W=157<br>p=0.882 | W=75<br>p=0.019 |
| | **2 SP** | - | W=102<br>p=0.104 | W=152<br>p=0.778 | W=125<br>p=0.313 | W=26<br>p=2e-04 |
| | **1 SP** | - | - | W=88<br>p=0.045 | W=83<br>p=0.032 | W=10<br>p=4e-05 |
| | **Linear** | - | - | - | W=112<br>p=0.174 | W=14<br>p=6e-05 |
| | **Constant** | - | - | - | - | W=32<br>p=4e-04 |

**TABLE S5.8 Results of statistical tests evaluating RT and response accuracy benefits during unimodal SL with task-irrelevant cues.**

(**A**) Results of one-sample t-tests comparing the observed RT benefit for each block against zero (no expectation benefit). Positive effects indicate RT benefits. (**B**) Results of paired-sample t-tests comparing the RT benefit between chronologically adjacent blocks. Positive effects indicate that RT benefits increased from one to the other block. (**C**) and (**D**) show the corresponding results for response accuracy. (**E**) Results of one-sample t-tests comparing RT benefits for each block against zero, including only expected trials and unexpected trials requiring the same button press as the expected object. (**F**) Paired-sample t-test results comparing RT benefits between adjacent blocks, again only including unexpected trials with the same response as the expected stimulus would have required. All presented p-values are uncorrected. Effect sizes are reported in terms of Cohen's *d*.

**A**

**RT benefit per block against zero**

| | | | |
|---|---|---|---|
| Block 1: | t(24)=1.52 | p=0.142 | d=0.3 |
| Block 2: | t(24)=0.9 | p=0.379 | d=0.18 |
| Block 3: | t(24)=5.31 | p=2e-05 | d=1.06 |
| Block 4: | t(24)=4.17 | p=3e-04 | d=0.83 |
| Block 5: | t(24)=9.23 | p=2e-09 | d=1.85 |
| Block 6: | t(24)=5.72 | p=7e-06 | d=1.14 |
| Block 7: | t(24)=6.48 | p=1e-06 | d=1.3 |
| Block 8: | t(24)=6.7 | p=6e-07 | d=1.34 |

**B**

**RT benefits between sequential blocks**

| | | | |
|---|---|---|---|
| Block 1 vs 2: | t(24)=-0.19 | p=0.854 | d=-0.04 |
| Block 2 vs 3: | t(24)=3.95 | p=6e-04 | d=0.79 |
| Block 3 vs 4: | t(24)=-1.77 | p=0.089 | d=-0.35 |
| Block 4 vs 5: | t(24)=4.09 | p=4e-04 | d=0.82 |
| Block 5 vs 6: | t(24)=-2.99 | p=0.006 | d=-0.6 |
| Block 6 vs 7: | t(24)=1.5 | p=0.146 | d=0.3 |
| Block 7 vs 8: | t(24)=-1.13 | p=0.270 | d=-0.23 |

C

**Accuracy benefit per block against zero**

| | | | |
|---|---|---|---|
| Block 1: | t(24)=-0.8 | p=0.431 | d=-0.16 |
| Block 2: | t(24)=0.69 | p=0.498 | d=0.14 |
| Block 3: | t(24)=2.17 | p=0.040 | d=0.43 |
| Block 4: | t(24)=0.46 | p=0.646 | d=0.09 |
| Block 5: | t(24)=3.13 | p=0.005 | d=0.63 |
| Block 6: | t(24)=3.28 | p=0.003 | d=0.66 |
| Block 7: | t(24)=0.36 | p=0.725 | d=0.07 |
| Block 8: | t(24)=2.68 | p=0.013 | d=0.54 |

D

**Accuracy benefits between sequential blocks**

| | | | |
|---|---|---|---|
| Block 1 vs 2: | t(24)=1.47 | p=0.155 | d=0.29 |
| Block 2 vs 3: | t(24)=1.56 | p=0.132 | d=0.31 |
| Block 3 vs 4: | t(24)=-1.54 | p=0.137 | d=-0.31 |
| Block 4 vs 5: | t(24)=2.2 | p=0.037 | d=0.44 |
| Block 5 vs 6: | t(24)=-0.22 | p=0.825 | d=-0.04 |
| Block 6 vs 7: | t(24)=-2.15 | p=0.042 | d=-0.43 |
| Block 7 vs 8: | t(24)=1.47 | p=0.155 | d=0.29 |

E

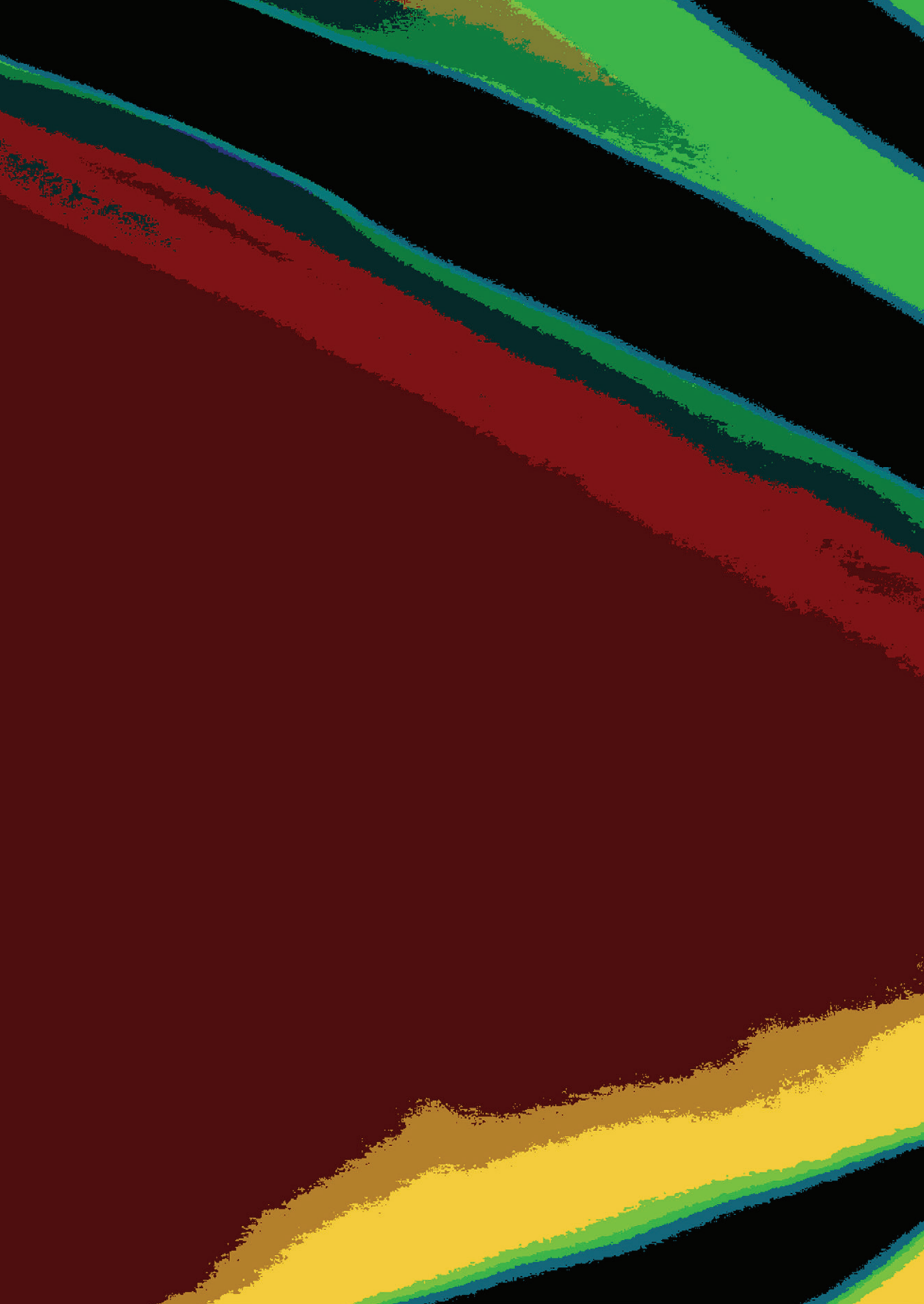
**RT benefit per block against zero (only same response trials)**

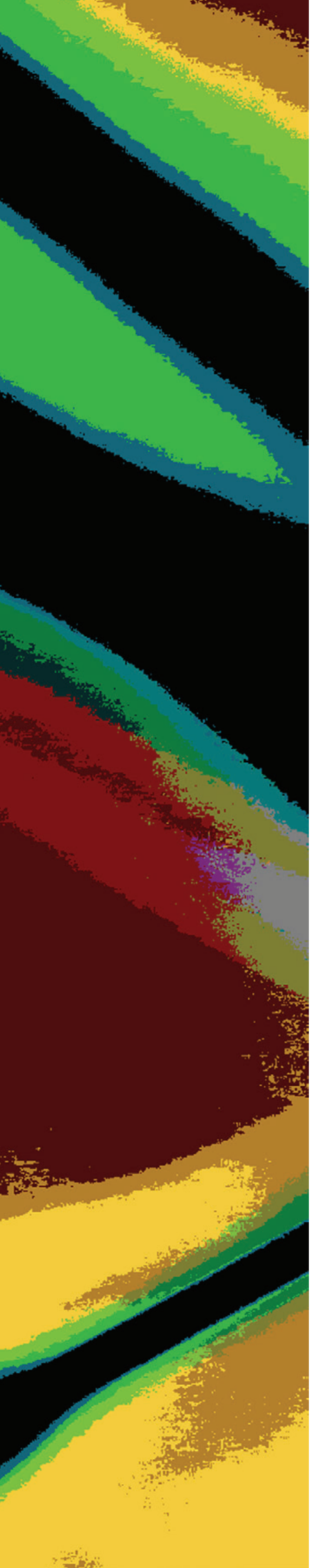
| | | | |
|---|---|---|---|
| Block 1: | t(24)=1.12 | p=0.275 | d=0.22 |
| Block 2: | t(24)=0.28 | p=0.782 | d=0.06 |
| Block 3: | t(24)=2.23 | p=0.035 | d=0.45 |
| Block 4: | t(24)=1.56 | p=0.132 | d=0.31 |
| Block 5: | t(24)=5.41 | p=1e-05 | d=1.08 |
| Block 6: | t(24)=2.64 | p=0.015 | d=0.53 |
| Block 7: | t(24)=3.42 | p=0.002 | d=0.68 |
| Block 8: | t(24)=2.82 | p=0.010 | d=0.56 |

F

**RT benefits between sequential blocks (only same response trials)**

| | | | |
|---|---|---|---|
| Block 1 vs 2: | t(24)=-0.6 | p=0.554 | d=-0.12 |
| Block 2 vs 3: | t(24)=1.56 | p=0.131 | d=0.31 |
| Block 3 vs 4: | t(24)=-0.61 | p=0.546 | d=-0.12 |
| Block 4 vs 5: | t(24)=2.48 | p=0.020 | d=0.50 |
| Block 5 vs 6: | t(24)=-1.93 | p=0.066 | d=-0.39 |
| Block 6 vs 7: | t(24)=0.92 | p=0.369 | d=0.18 |
| Block 7 vs 8: | t(24)=-1.23 | p=0.231 | d=-0.25 |





## CHAPTER 6

# Discussion

In Figure 1.1A of the introduction you were confronted with an initially unintelligible, ambiguous stimulus. The subsequent expectation to see a cat in the stimulus likely resolved the ambiguity, phenomenologically demonstrating that expectations can shape perception. Throughout this thesis I have asked the question, how expectations modulate perceptual processing, and whether this modulation may constitute a general operating principle of the sensory brain. In several experiments I used incidental statistical learning, the unsupervised extraction of statistical regularities from the environment across time and space [46–48], to induce perceptual expectations. As a consequence we saw an attenuated sensory response to expected compared to unexpected stimuli throughout the ventral visual stream, also known as expectation suppression (**chapters 2-3**; [19,20]). In the following discussion I will attempt to answer key questions, raised in the introduction.

First, I will discuss whether the here presented evidence supports the proposition that predictions are a fundamental principle of sensory processing. In this context, I will review how wide-spread and task-independent expectation suppression is. Additionally, I will discuss the feature-specificity of complex object predictions across different levels of the visual processing hierarchy, thereby exploring the properties of expectation suppression in more detail. Next, I will revisit the limits of statistical learning and the automaticity of its sensory consequences, thereby further characterizing the underlying neural mechanism by placing important constrains on its ubiquity and automaticity. Furthermore, I will evaluate an alternative account, casting expectation suppression as an effect of attention instead of prediction, which fundamentally questions prediction error coding in sensory cortex. Then, I will briefly discuss whether expectation suppression reflects a suppression of neural responses or an enhanced response to surprising input. Before concluding, I will further explore the neural modulation underlying expectation suppression by reviewing when expectations may sharpen or dampen neural representations. Finally, I will conclude with a condensed synthesis of the results presented here and the wider literature, demonstrating how perception is fundamentally influenced by expectations.

## Statistical learning and expectation suppression

In the introduction of my thesis, I raised the question whether prediction constitutes a fundamental operating principle of the sensory brain. I proposed several characteristics this supposition may entail. First, if prediction is a core principle of sensory processing, its effect should be evident across the visual hierarchy. In **chapter 2 and 3**, I reported expectation suppression, an attenuated sensory responses to expected compared to unexpected stimuli, throughout the ventral visual stream. Thus, expectations appear to modulate sensory processing from early to late stages of the cortical visual hierarchy, including key processing areas such as primary visual cortex (V1), object selective lateral occipital cortex (LOC), and temporal occipital fusiform cortex (TOFC). Second, the sensory consequences of prediction should be evident for predictions of stimuli that are common and behaviorally relevant in everyday life (e.g., objects). Accordingly, I demonstrated that complex associations between arbitrarily paired objects can be learned incidentally (**chapters 2-5**) and subsequently modulate sensory processing (**chapters 2-4**), thus further supporting the ubiquity of predictions in shaping perception. Third, sensory modulations should arise without intention to learn or to use the underlying predictions. Indeed, once acquired, perceptual predictions appear to affect neural processing irrespective of whether the specific priors are task-relevant (**chapter 3**) or task-irrelevant (**chapter 2**), thereby suggesting that predictions modulate perception even without any intention or behavioral benefit to predict. Moreover, it is worth noting that while the effects of priors tend to be stronger when stimuli are noisy (review: [19]), the prediction induced modulations observed here were evident even though unambiguous stimuli were presented without additional noise, thus further supporting the pervasiveness of prediction induced modulations of sensory processing.

These results corroborate and extent previously reported sensory suppression for simple, possibly explicitly learned predictions [18,24,25,28], by demonstrating that complex perceptual priors, such as associations between arbitrarily paired naturalistic object images, can be extracted incidentally and subsequently suppress sensory responses throughout the ventral visual stream. Moreover, these results also align with data from non-human primates, demonstrating expectation suppression in terms of spike rates, following incidental statistical learning of complex associations during passive exposure [23,26,27]. Thus, my results also bridge a gap between studies in non-human primates and human volunteers by demonstrating comparable suppression of sensory responses using experimental paradigms based on studies in non-human primates. Combined with prior studies, the available evidence therefore supports the hypothesis that prediction constitutes a general operational principle of sensory processing [12,19,30], evident for simple and complex associations, affecting responses across the ventral visual stream, following intentional and incidental statistical learning (**chapters 2-5**; [18,23–28]). Expectation suppression thus appears to be a pervasive and general neural phenomenon [19]. But, how can we account for this mismatch response? In **chapters 2 and 3**, I argued for an interpretation of expectation suppression in line with predictive coding accounts [11–13], as reflecting smaller prediction errors for stimuli conforming to prior expectations compared to unexpected stimuli. Next, I will discuss what the present results imply for the characteristics of prediction errors across the visual hierarchy.

## Feature-specific and feature-unspecific prediction (errors) across the ventral visual stream

Hierarchical predictive coding [12] suggests that predictions and prediction error calculations occur iteratively at every level of the sensory hierarchy. However, as noted in the introduction, this raises the question how feature-specific predictions, and consequently prediction error calculations, are at the different levels of the visual hierarchy. In other words, do complex predictions, such as expected faces or objects, translate into feature-specific predictions at the level of V1, such as local contrasts and orientations?

My results from **chapters 2-4** speak to these questions. In each study I manipulated predictions of object images and subsequently showed expectation suppression throughout the ventral visual stream. Thus, on first sight these results suggest that complex object predictions are relayed down the visual hierarchy such that prediction errors arise at the early and late levels in visual cortex. However, in **chapter 2 and 3** I also showed that stimulus-driven and non-stimulus-driven voxels in V1 are equally suppressed by expectations. This suggests that expectation suppression in early visual cortex may be stimulus-unspecific, as voxels (neural populations) appear to be suppressed irrespective of their responsiveness to a stimulus. Moreover, in **chapter 4** I moved beyond the voxel level, using forward models, and demonstrated that expectation suppression in early visual cortex was best explained by a feature-unspecific global gain modulation of neural responses. In other words, expectation suppression appears to affect neural populations equally in V1, irrespective of feature tuning (**chapter 4**) and stimulus responsiveness (**chapters 2 and 3**). In contrast, in higher visual areas, including object-selective LOC and TOFC, I did demonstrate stimulus-specific and feature-specific expectation suppression (**chapters 2-4**). Based on these results one can speculate that two partially distinct mechanism may underlie expectation suppression. One mechanism is feature-specific, and may directly reflect the hypothesized prediction error signals [15] in line with hierarchical predictive coding [12]. This prediction error computation may underlie the detection of prediction violations. The second mechanism is feature-unspecific, and may reflect the consequence of the previous prediction error computation in higher visual areas, but not constitute a cause for the detection of the expectation violation itself. That is, detecting an expected stimulus depends on feature-specificity, because specific stimuli were predicted in **chapters 2-4**. Hence, a feature-unspecific modulation is unlikely to reflect a cause, but merely a consequence of the prediction (dis-)confirmation.

What mechanism may underlie this feature-unspecific suppression? One can speculate that feature-unspecific suppression may reflect arousal and attention disengagement from the expected stimuli. That is, if a stimulus is well-predicted no new information is gained, hence attention may be disengaged, resulting in less weight on the ascending prediction error units [16]. As this disengagement is not specific to any stimulus features (e.g. spatial attention may disengage from the whole stimulus area) all signals in prediction error units might be reduced, thus resulting in the observed unspecific suppression. This interpretation is supported in **chapter 3** by enhanced pupil dilations in response to unexpected stimuli, which in turn have been associated with signifying arousal changes [115,116]. Indeed, pupil dilation and expectation suppression magnitude in V1 were correlated. Thus, feature-unspecific expectation suppression may reflect global changes, such as fluctuations in arousal and general attention disengagement from well-predicted stimuli.

Curiously, predictions capitalizing on neural tunings in V1, such as orientation predictions of grating stimuli, do result in feature-specific prediction effects in early visual areas [18,136]. Thus, the lack of feature-specificity observed here, does not constitute a general limitation of V1. Rather, combined these results suggest that the type of prediction determines where stimulus-specific prediction (error) signals emerge. That is, if an object is predicted, the whole object and its large scale features, such as its shape appear to be predicted, thereby accounting for the feature-specific suppression in object-selective visual areas. However, specific low level features, such as local contrasts, are not necessarily predicted, thus resulting in unspecific expectation suppression in V1. In contrast, stimuli whose primary predicted feature is orientation (e.g. a grating stimulus) do evoke feature-specific predictions in early visual cortex [18,136]. Thus, these results suggest a remarkable flexibility of the visual system, with the localization of feature-specific expectation suppression dependent on the type of prediction.

In sum, two separate mechanisms may underlie expectation suppression. One, feature-specific ‘true’ prediction error computation, here shown in higher visual areas, and one feature-unspecific global arousal or attention signal, likely reflecting a consequence of the preceding prediction error computation. This hypothesis predicts differences in the time courses for feature-specific and feature-unspecific expectation suppression, which future research could investigate using neuroimaging methods with high temporal resolution, such as magnetoencephalography. That is, if feature-unspecific expectation suppression does reflect a consequence of feature-specific prediction errors, the here observed feature-specific suppression should precede the unspecific modulations in early visual cortex. Finally, these results also caution that superficially similar expectation suppression may not necessarily represent the same underlying neural mechanism, thus suggesting that the consequences of prediction can be multifaceted across the ventral visual stream.

## Limits of incidental statistical learning and the gating of its sensory consequences by attention

Thus far I have primarily discussed evidence for the wide-spread influence of expectations on sensory processing. However, the work presented in this thesis also shines light on crucial limitations of both the learning process and the sensory consequences of statistical learning, thereby constraining the ubiquity and automaticity of sensory modulations by predictions.

**Chapter 3** showed that attention may gate the sensory consequences of statistical learning. In particular, only when stimuli were attended modulations by expectations were found. These results complement another study, showing that pupils are dilated following surprising input, but only if the stimuli were attended [180]. The precise nature of this gating by attention remains to be investigated further, but one can speculate that it may be tied to the generation or modulation of the prior [114]. For example, flexibly gating the generation of predictions, such that unattended stimuli do not necessarily instantiate predictions, may conserve processing resources. However, also note that similar attention gating may not generalize to all types of sensory priors. That is, simpler, more fundamental priors, such as perceptual fill-in during the Kanizsa illusion, may modulate processing also for unattended stimuli [127]. Moreover, as previously seen, predictions do arise from task-irrelevant priors (**chapter 2**). Therefore, at least for the complex sensory priors investigated here, the consequences of statistical learning appear to be limited to attended stimuli (**chapter 3**), but not necessarily task-relevant predictions (**chapter 2**). Thus, while perceptual inference is largely an unconscious process, with priors influencing perception irrespective of intentional usage or general utility of a prediction, not all modulations by expectations are necessarily pre-attentive.

I additionally demonstrated limits of statistical learning, the process by which priors are extracted from statistical regularities. In **chapter 5**, statistical learning was evident for unimodal, but not cross-modal statistical regularities, thereby implying crucial modality-specific contributions to statistical learning. Indeed, such results suggest that the incidental formation of perceptual priors may depend on modulations within the sensory processing streams, such as local changes in synaptic efficacy [12,112]. On the other hand, cross-modal learning likely depends on the integration of information across sensory streams and multisensory areas. It is plausible that such long-distance integration may require broadcasting of associated stimulus representations into a domain general network [63], or global neuronal workspace [173,174]. This hypothesis suggests that the acquisition of cross-modal regularities may depend more on intentional learning, while unimodal learning can occur incidentally. Additionally, learning was only evident for deterministic, but not probabilistic associations, questioning the frequently proposed reliability and rapidity of learning from statistical regularities [46,47]. While more worked is required to assess how sensitive humans are to statistical regularities, the present results do suggest limits, both in terms of cross-modal learning and probabilistic learning, at least during the incidental acquisition of regularities. Whether these limitations are specific to incidental compared to intentional statistical learning poses an intriguing avenue for future research. For instance, one can speculate that incidental statistical learning strongly depends on local changes within sensory areas, while intentional learning may recruit (sub-)cortical areas associated with other types of memory formation, such as episodic memory, hence posing alternative routes towards the acquisition of statistical regularities [19]. Moreover, whether the sensory consequences of these two routes towards learning overlap remains to be investigated as well.

In sum, while statistical learning can indeed operate incidentally across different contexts, task-demands, and stimuli (**chapters 2, 3 and 5**; reviews: [46–49]), the automaticity, reliably and rapidity of learning, as well as its generalization across modalities, may be less robust than previously thought (**chapter 5**). While communalities are evident across different studies employing statistical learning (e.g., expectation suppression and behavioral facilitation), it is crucial to note that, because of the broad definition of statistical learning, different results in learning might be expected. Therefore, I believe that it is important for future studies of statistical learning to adequately define what type of learning is investigated, as well as how the consequences of learning are assessed. Crucial distinctions may include unimodal vs. cross-modal statistical regularities, incidental vs. intentional learning, the reliability of the underlying associations, and explicit vs. implicit knowledge of the regularities.

## Prediction vs attention based accounts of expectation suppression

Up to this point I have shown that incidental statistical learning can serve as a basis for generating expectations to guide perception, but also highlighted crucial limits to the automaticity and ubiquity of statistical learning and its sensory consequences. Moreover, I have demonstrated that expectation suppression is evident across the ventral visual stream, thereby suggesting that this phenomenon may reflect a general operating principle of the brain. Throughout this thesis, I have argued that expectation suppression may reflect sensory prediction errors, in line with predictive processing accounts [11–13]. Next, I will discuss an alternative account, casting expectation suppression as a direct effect of attention, and contrast it to the here prevalent prediction based account. Evaluating this attention based explanation is important, as a central question of this thesis is how expectations modulate sensory processing. This question critically entails whether expectation suppression does reflect prediction errors or modulations by attention. First, I will introduce the attention based account and then review empirical evidence which supports and challenges it.

### *Attention account*

The attention account, as summarized by Alink and Blank [40], starts by noting that surprise has been shown to attract attention [35–37]. A functional explanation for this observation is that stimuli which are not well-predicted by internal models are potentially valuable sources of information and may require important adjustments to behavior. If surprise attracts attention, unexpected stimuli will be attended more than expected ones. Given that attention modulates the gain of neural responses [38,39], this disproportionate allocation of attention towards unexpected stimuli will appear as expectation suppression, when responses to unexpected and expected stimuli are contrasted as in **chapters 2-3**. Crucially, this account does not require any direct modulation of sensory responses by expectations, thereby questioning the coding of prediction errors in sensory cortex in favor of a gain modulation by attention. An attention based account also explains the results in **chapter 3** by proposing that when attention is directed towards the unpredictable alphanumeric characters, surprise elicited by the predictable, but task-irrelevant object stimuli would not result in a reallocation of attention, because these stimuli are behaviorally irrelevant. Thereby, this account elegantly explains expectation suppression with well-known gain modulations by attention [38,39], and the absence of expectation suppression when objects are unattended by a lack of behavioral relevance.

### *Challenges for the attention based account*

**Task-irrelevant predictions.** A core feature of the attention based explanation [40] is that unexpected stimuli are particularly relevant for behavior, as they may require modifications to our responses. This contention justifies the automatic capture of attention by surprising stimuli [35–37]. However, expectation suppression has also been observed during exposure to task-irrelevant predictions (e.g., **chapter 2**; [32,113]) and during passive fixation [23,26,27]. Therefore, the attention based account would have to assert that surprise always attracts attention, irrespective of the behavioral relevance of the stimuli. While this extension is certainly plausible, it complicates the explanation why diverting attention away from the predictable stimuli abolishes attentional capture entirely (**chapter 3**), because relying on behavioral relevance alone cannot account for the results of **chapters 2 and 3** combined. In contrast, prediction based accounts inherently accommodate predictions for task-relevant and task-irrelevant priors, as expectation and behavioral relevance, the latter being related to attention, are considered to be separate influences on sensory processing [12,16].

**Surprise calculations.** While the attention based account can explain the observed gain modulations in sensory cortex without positing prediction error calculations within sensory areas, it does nonetheless require a computation of surprise elsewhere in cortex. Following this surprise calculation, attention would be reallocated, if a stimulus was surprising. Thus, two separate mechanisms are required to account for the observed expectation suppression phenomenon: a surprise calculation and a subsequent attention allocation resulting in the observed gain modulation. In comparison, predictive processing accounts explain the observed expectation suppression effects by proposing that prediction errors are computed at each stage of the visual hierarchy, directly reflecting the mismatch between top-down predictions and bottom-up inputs [11–13]. Hence, on the prediction account no additional mechanism for the detection of surprise is required.

**Omission responses and pre-stimulus templates.** Finally, previous studies demonstrated that neural responses to omissions of expected stimuli can carry stimulus-specific information [181] and that expectations can induce sensory templates before stimulus onset [182]. In other words, pre-stimulus expectation effects, and unexpected stimulus omissions, suggest that predictions can modulate sensory responses in a stimulus-specific fashion even in the absence of a stimulus. A prediction based account explains such pre-stimulus effects by prospective prediction and subsequent prediction errors due to unexpected stimulus omission [183,184]. In contrast, on an attention based account it is difficult to accommodate omission and pre-stimulus responses, particularly feature-specific ones, because attention allocation is proposed to occur after a stimulus has been evaluated to be surprising.

### *Prediction based explanations are more parsimonious*

In sum, the attention based account is challenged by task-irrelevant predictions yielding expectation suppression (**chapter 2**; [32,113]), as well as by omission and pre-stimulus effects of predictions [181,182,184]. Moreover, the necessity of proposing an additional surprise calculation elsewhere in cortex makes the attention account more complex. Thus, while the currently available data does not conclusively rule out either account, it does appear that a prediction based explanation, along the lines of predictive coding, does explain the results reported in **chapters 2-4**, as well as in the wider literature (e.g., [23,26,27,32,143,181,182,184]) in a more parsimonious fashion. However, as argued by Alink and Blank [40], both explanations should be considered in future studies, and ideally explicitly plotted against one another. For example, the attention based account holds that a prediction error calculation, likely computed outside of sensory cortex, has to occur before subsequent attention allocation. In contrast, on a prediction based account, expectation suppression does directly reflect prediction error computation in sensory areas. Hence, the two accounts make different predictions whether a prediction error computation outside of sensory cortex is necessary before any modulations in sensory areas are evident. Thus, future work could specifically target the distinct localizations and time courses for prediction error computations and sensory modulations predicted by the two accounts.

## Expectation suppression vs. surprise enhancement

Another question which frequently arises when discussing expectation suppression is whether it does constitute a suppression of neural responses to expected stimuli, or in fact an enhanced response to surprising stimuli. The data I have presented throughout this thesis do not directly speak to this question, as I compared only responses to expected and unexpected stimuli, without any expectation-free, neutral stimuli. That said, it is worth elaborating on this question. First, what could a neutral stimulus be in order to serve as an expectation-free reference? We could propose that a previously unseen stimulus does not involve any expectations. However, an unfamiliar stimulus is in fact very surprising and results in a significantly upregulated sensory response [185]. Alternatively, in an experiment similar to those reported here (**chapters 2-5**), we could suggest that a stimulus which follows all leading stimuli equally often could serve as a neutral stimulus. However, such a stimulus is clearly not expectation-free either, but rather the conditional probability is simply lower than for expected stimuli, and higher than for unexpected stimuli. Assuming we would observe expectation suppression following the probability of these stimuli (i.e., $\text{BOLD}_{\text{unexpected}} > \text{BOLD}_{\text{neutral}} > \text{BOLD}_{\text{expected}}$, which is plausible given previous work [108]), we would still not know whether we see a suppression of the neutral stimuli relative to unexpected ones, or an enhanced response to unexpected stimuli relative to neutral ones. In other words, expectations are relative, only defined by the relative probability of the different possible outcomes. Therefore, arguably there is no 'absolute' neutral stimulus against which to compare whether expectations induce a suppression, or surprise enhances responses. Therefore, while the question of expectation suppression vs. surprise enhancement may initially sound relevant, it may not reflect any meaningful properties of the underlying neural modulation. That said, above I have argued that expectation suppression may be best explained by prediction error computations. Given that on a predictive coding account predictions explain away bottom-up activity, the term expectation suppression appears to more accurately reflect our current understanding of the neural modulations underlying perceptual predictions.

## Dampening vs. sharpening of neural representations

After establishing that expectation suppression is wide-spread throughout sensory cortex and reaffirming that it likely reflects prediction error calculations, I will now turn towards the question of what type of neural modulation underlies expectation suppression. That is, even if we assume that expectation suppression indeed reflects prediction errors, as argued above and throughout this thesis, we still do not know how and what neural responses are suppressed by expectations. Two leading accounts of expectation suppression, debated in the literature [19] and outlined in the introduction, are a sharpening and dampening of the population representation.

To recapitulate, using conventional fMRI analyses (**chapter 2**) and forward models (**chapter 4**), I provided evidence that perceptual expectations dampen sensory representations by suppressing neurons tuned *towards* expected stimulus features. These results are in agreement with previous studies [23,43,79,143], and suggest that expectations reduce redundancy in sensory cortex, and highlight surprising input. However, a number of other studies showed the opposite modulation, a sharpening of representations [18,41,142]. On this account, expectations suppress neurons tuned *away* from the expected stimulus features, hence allowing for faster and more accurate representations of expected stimuli. What are possible explanations for these, prima facie, incompatible results, and which explanations can be ruled out by synthesizing the results of previous studies and the work presented here?

I start by briefly mentioning potential explanations, some of which have been brought forward in the literature, that can now be ruled out. Both sharpening and dampening have been shown in humans using fMRI BOLD (**chapters 2 and 4**; [18,43]), and in non-human primates using electrophysiological recordings [23,41], hence ruling out systematic differences in species or recording modalities [79]. Perceptual demands of the utilized tasks are also unlikely to account for the opposite results, as both sharpening and dampening have been demonstrated using perceptually simple (**chapters 2 and 4**; [126]) and challenging tasks [18,43]. Moreover, differences in the stages of the processing hierarchy [79] are unlikely to account for the discrepancy in results, as both modulations have been shown in similar object-selective visual areas (**chapters 2 and 4**; [142]).

### *When expectations sharpen and when they dampen representations*

Next, I will discuss factors which, given the currently available evidence, may account for why some studies report a sharpening, while others report a dampening of representations.

**Recently vs. well-established priors.** Dampening results are usually associated with paradigms employing extensive exposure to statistical regularities (**chapters 2-4**; [23,79]), or even building on lifelong experience of congruency [43]. On the other hand, sharpening results are commonly found while the underlying associations are, or just have been, acquired [18,41]. A similar argument for recent learning can be made for the correspondence of self-generated motion and an computer-generated avatar mirroring this motion, resulting a sharpening of representations, reported by Yon et al. [142]. This suggests that sharpening and dampening may be associated with two separate stages of utilizing prior information – an initial learning stage, during which representations are sharpened, and a subsequent exploitation stage, during which representations are dampened (although see: [143]).

**Concurrent attention manipulations.** Another possibility is that attention may account for representational sharpening. For example, in Kok et al. [18] predicting the orientation of a grating stimulus may lead to prospective feature-based attention allocation for the expected orientation. Similarly, in Yon et al. [142], congruent finger motion may coincide with spatial attention allocation to the finger that is expected to move. Given that attention boosts the gain of attended stimulus features and spatial locations (review: [186]), it is possible that expectation sharpening reflects a gain modulation due to selectively shifting attention to the predicted stimulus features and locations. That is, neurons selective for the attended features are upregulated, while expectation suppression results in the overall attenuated response. In contrast, in studies supporting the dampening account, only expectation suppression may arise, without sharpening by prospective attention allocation. In these studies, reallocation of attention may not be task-relevant (e.g., passive fixation), or not feasible given the paradigm, because stimuli were presented at the same spatial location and the predicted features (object stimuli) were too complex to reliably pre-allocate attention in a feature-specific fashion (**chapters 2-4**; [23,79]).

### *Future directions*

In sum, the work presented in **chapters 2-4** contributes to ruling out potential explanations for the discrepancy in the expectation literature supporting sharpening and dampening. The evidence suggests that systematic differences in recording modality, task-relevance of the predictions, studied species or cortical areas cannot consistently account for the opposite results. However, the amount of exposure to the investigated priors (well-established vs. just acquired), and concurrent manipulations of attention may account for how the balance between sharpening and dampening may be tipped in favor of one or the other process. Thus, future work could explicitly contrast the effects of using well-established vs. recently acquired priors, or better yet, investigate the development of predictions across the learning process. For example, a modified version of the paradigm used in **chapter 2** could be used, while recording fMRI already during the initial learning of probabilistic associations. However, see **chapter 5** for potential limitations of using probabilistic associations during learning. Moreover, careful orthogonalization of attention and expectation manipulations may shed additional light on whether expectation sharpening may reflect representation modulations by attention, as hypothesized above. Future studies could assess this hypothesis by manipulating predictions such that one type of prediction does, and one does not allow for prospective attention allocation – e.g., using the paradigm in **chapter 2** with an additional, orthogonal spatial prediction manipulation (although see: [32]).

In sum, the here presented evidence supports the dampening account, suggesting that expectations reduce the gain of neural populations tuned towards expected stimulus features. Moreover, while I have outlined some possible explanations why other studies have reported a sharpening of representations, these hypotheses remain to be tested. Additionally, beyond the representation modulations underlying expectation suppression following statistical learning, as discussed here, the broader question of how these modulations relate to other sensory modulations by predictions requires investigation as well. For example, contextual priors [187], such as word contexts, have been shown to sharpen sensory representations [188]. Resolving the discrepancy in the literature outlined above, may also provide new insight into whether these different types of priors rely on similar or distinct neural mechanism, and hence ultimately elucidate how unitary or multifaceted the modulations of sensory processing by predictions are.

## Forms of predictions

Perceptual predictions are frequently cast as a unitary phenomenon, with the accompanying implicit assumption that the underlying computations also constitute one common neural modulation, involving the same circuits across different priors. However, predictions are diverse in nature. For example, predictions can involve different classes of stimuli, which, as discussed here, may directly affect at which level of the visual hierarchy feature-specific effects arise (see: *Feature-specific and feature-*

*unspecific prediction (errors) across the ventral visual stream*). Moreover, predictions can operate on different timescales, from short term expectations in volatile environments [25,41] to priors formed on phylogenetic or ontogenetic timescales [126]. Indeed, as also shown in the discussion of statistical learning, predictions can be acquired from statistical regularity following extensive (**chapters 2-3**) or limited exposure [54]. As argued above, such differences may in fact account for apparently contradictory results in the expectation suppression literature (e.g., recently vs. well-established priors; see: *Dampening vs sharpening of neural representations*). Moreover, predictions can be anticipatory (**chapter 2-5**; [28]) or contextual [187,188], top-down or bottom-up [189]. In short, given the multitude of predictions and their distinct characteristics, it remains unclear whether one neural mechanism can account for all conceivable predictions. Establishing common ground between distinct phenomena is frequently the aim of scientific investigation and overlapping modulations underlying different perceptual predictions have certainly been found. Consider, for example, similarities in expectation suppression in vision and audition [19,20], or that both task-irrelevant and task-relevant predictions, in **chapters 2 and 3** respectively, resulted in comparable expectation suppression and a dampening of representations (**chapter 4**). However, even the same expectation may result in two partially distinct types of expectation suppression across different levels of the sensory hierarchy (i.e., feature-specific and feature-unspecific suppression; see: *Feature-specific and feature-unspecific prediction (errors) across the ventral visual stream*). Thus, assuming that all predictions necessarily involve one unitary neural mechanism appears ill-fated. Therefore, I believe future work ought to explicitly acknowledge the diversity of perceptual priors, and carefully consider how and what type of predictions are induced. Over time this approach will elucidate the commonalities and differences between different predictions the sensory brain employs to guide perception. Indeed, since its early days, investigations and theories of perceptual inference have matured and shown that predictive processes likely constitute a fundamental operating principle of the brain [12,19,30], as supported by the results presented throughout this thesis. Therefore, appreciating the multifaceted nature of predictions is arguably not a limitation, but an important feature.

## Conclusion

In conclusion, results reported in my thesis support that perceptual priors fundamentally modulate sensory processing, as evident by wide-spread expectation suppression, the reduced neural response to expected compared to unexpected stimuli, throughout the ventral visual stream. Moreover, expectations appear to modulate sensory responses irrespective of the behavioral relevance of a prediction, further supporting the general role of predictions in guiding perception. However, the precise neural modulation underlying this suppression of sensory responses appears to depend on the prediction and cortical area in question. Expectations of objects result in a feature-specific dampening in object selective visual areas, but a feature-unspecific modulation in early visual cortex, potentially reflecting two distinct mechanisms. My results also support that statistical learning constitutes a crucial source for the acquisition of sensory priors, evident by wide-spread expectation suppression following incidental learning of associations between arbitrarily paired objects. However, I also showcased some limits of statistical learning and its sensory consequences, such as a lack of incidental cross-modal statistical learning and a gating of prediction by attention.

Combined these results extent and support theories conceptualizing perception as fundamentally relying on prediction [11–13,16]. The here presented evidence is well accommodated by casting perception as an inferential process, inferring the most likely sources for sensory input using a combination of sensory evidence and prior knowledge, derived from statistical regularities in the sensory world. Such inference appears to occur across the sensory hierarchy, involving feature-specific predictions and error computations, as well as subsequent feature-unspecific modulations. While a common neural signature, expectation suppression, appears to underlie this hierarchical inference, its characteristics depend on the specific predictions and level in the processing hierarchy, and may in fact reflect two distinct mechanisms. Therefore, predictions appear to be a core principle of perceptual processing, but the precise characteristics of the underlying modulation appear to be multifaceted, echoing the multifaceted nature of predictions, and the multitude of response properties throughout visual cortex.

# References

1 Riesenhuber M, Poggio T. Hierarchical models of object recognition in cortex. Nat Neurosci. 1999;2: 1019–1025. doi:10.1038/14819

2 DiCarlo JJ, Zoccolan D, Rust NC. How does the brain solve visual object recognition? Neuron. 2012;73: 415–434. doi:10.1016/j.neuron.2012.01.010

3 Oliva A, Torralba A. The role of context in object recognition. Trends Cogn Sci. 2007;11: 520–527. doi:10.1016/j.tics.2007.09.009

4 Helmholtz von H. Handbuch der physiologischen Optik. Leipzig Leopold Voss Publ parts from 1856 to 1866, then Publ toto 1867 as Vol Nine Allg Encycl der Phys ed Gustav Karsten. 1867;9.

5 Kersten D, Mamassian P, Yuille A. Object perception as Bayesian inference. Annu Rev Psychol. 2004;55: 271–304. doi:10.1146/annurev.psych.55.090902.142005

6 Aitchison L, Lengyel M. With or without you: predictive coding and Bayesian inference in the brain. Curr Opin Neurobiol. 2017;46: 219–227. doi:10.1016/j.conb.2017.08.010

7 Knill DC, Pouget A. The Bayesian brain: The role of uncertainty in neural coding and computation. Trends Neurosci. 2004;27: 712–719. doi:10.1016/j.tins.2004.10.007

8 Kok P, Brouwer GJ, van Gerven MAJ, de Lange FP. Prior expectations bias sensory representations in visual cortex. J Neurosci. 2013;33: 16275–16284. doi:10.1523/JNEUROSCI.0742-13.2013

9 Wyart V, Nobre AC, Summerfield C. Dissociable prior influences of signal probability and relevance on visual contrast sensitivity. PNAS. 2012;109: 3593–3598. doi:10.1073/pnas.1120118109

10 Stein T, Peelen M V. Content-specific expectations enhance stimulus detectability by increasing perceptual sensitivity. J Exp Psychol Gen. 2015;144: 1089–1104. doi:10.1037/xge0000109

11 Rao RPN, Ballard DH. Predictive coding in the visual cortex : a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci. 1999;2: 79–87.

12 Friston K. A theory of cortical responses. Philos Trans R Soc Lond B Biol Sci. 2005;360: 815–36. doi:10.1098/rstb.2005.1622

13 Spratling MW. Predictive coding as a model of response properties in cortical area V1. J Neurosci. 2010;30: 3531–3543. doi:10.1523/JNEUROSCI.4911-09.2010

14 Walsh KS, McGovern DP, Clark A, O'Connell RG. Evaluating the neurophysiological evidence for predictive processing as a model of perception. Ann N Y Acad Sci. 2020;1464: 242–268. doi:10.1111/nyas.14321

15 Keller GB, Mrsic-Flogel TD. Predictive Processing: A Canonical Cortical Computation. Neuron. 2018;100: 424–435. doi:10.1016/j.neuron.2018.10.003

16 Feldman H, Friston KJ. Attention, uncertainty, and free-energy. Front Hum Neurosci. 2010;4: 1–23. doi:10.3389/fnhum.2010.00215

17 Spratling MW. A review of predictive coding algorithms. Brain Cogn. 2017;112: 92–97. doi:10.1016/j.bandc.2015.11.003

18 Kok P, Jehee JFM, de Lange FP. Less Is More: Expectation Sharpens Representations in the Primary Visual Cortex. Neuron. 2012;75: 265–270. doi:10.1016/j.neuron.2012.04.034

19 De Lange FP, Heilbron M, Kok P. How Do Expectations Shape Perception? Trends Cogn Sci. 2018;22: 764–779. doi:10.1016/j.tics.2018.06.002

20 Heilbron M, Chait M. Great Expectations: Is there Evidence for Predictive Coding in Auditory Cortex? Neuroscience. 2018;389: 54–73. doi:10.1016/j.neuroscience.2017.07.061

21 Todorovic A, Ede F Van, Maris E, Lange FP De. Prior Expectation Mediates Neural Adaptation to Repeated Sounds in the Auditory Cortex : An MEG Study. 2011;31: 9118–9123. doi:10.1523/JNEUROSCI.1425-11.2011

22 Todorovic A, de Lange FP. Repetition Suppression and Expectation Suppression Are Dissociable in Time in Early Auditory Evoked Fields. J Neurosci. 2012;32: 13389–13395. doi:10.1523/JNEUROSCI.2227-12.2012

23 Meyer T, Olson CR. Statistical learning of visual transitions in monkey inferotemporal cortex. Proc Natl Acad Sci U S A. 2011;108: 19401–6. doi:10.1073/pnas.1112895108

24 Egner T, Monti JM, Summerfield C. Expectation and Surprise Determine Neural Population Responses in the Ventral Visual Stream. J Neurosci. 2010;30: 16601–16608. doi:10.1523/JNEUROSCI.2770-10.2010

25 Den Ouden HEM, Daunizeau J, Roiser J, Friston KJ, Stephan KE. Striatal Prediction Error Modulates Cortical Coupling. J Neurosci. 2010;30: 3210–3219. doi:10.1523/JNEUROSCI.4458-09.2010

26 Kaposvari P, Kumar S, Vogels R. Statistical Learning Signals in Macaque Inferior Temporal Cortex. Cereb Cortex. 2018;28: 250–266. doi:10.1093/cercor/bhw374

27 Ramachandran S, Meyer T, Olson CR. Prediction Suppression in Monkey Inferotemporal Cortex Depends on the Conditional Probability between Images. J Neurophysiol. 2016;115: 355–362. doi:10.1152/jn.00091.2015

28 Utzerath C, St John-Saaltink E, Buitelaar J, De Lange FP. Repetition suppression to objects is modulated by stimulus-specific expectations. Sci Rep. 2017;7: 1–8. doi:10.1038/s41598-017-09374-z

29 Wacongne C, Labyt E, Van Wassenhove V, Bekinschtein T, Naccache L, Dehaene S. Evidence for a hierarchy of predictions and prediction errors in human cortex. Proc Natl Acad Sci U S A. 2011;108: 20754–20759. doi:10.1073/pnas.1117807108

30 Summerfield C, De Lange FP. Expectation in perceptual decision making: Neural and computational mechanisms. Nat Rev Neurosci. 2014;15: 745–756. doi:10.1038/nrn3838

31 Turk-Browne NB, Scholl BJ, Chun MM, Johnson MK. Neural Evidence of Statistical Learning: Efficient Detection of Visual Regularities Without Awareness. 2009;21: 1934–1945. doi:10.1162/jocn.2009.21131.Neural

32 Kok P, Rahnev D, Jehee JFM, Lau HC, De Lange FP. Attention reverses the effect of prediction in silencing sensory signals. Cereb Cortex. 2012;22: 2197–2206. doi:10.1093/cercor/bhr310

33 Hubel DH, Wiesel TN. Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex. J Physiol. 1962;160: 106–154.

34 Vernon RJW, Gouws AD, Lawrence SJD, Wade AR, Morland AB. Multivariate Patterns in the Human Object-Processing Pathway Reveal a Shift from Retinotopic to Shape Curvature Representations in Lateral Occipital Areas, LO-1 and LO-2. J Neurosci. 2016;36: 5763–5774. doi:10.1523/JNEUROSCI.3603-15.2016

35 Itti L, Baldi P. Bayesian surprise attracts human attention. Vision Res. 2009;49: 1295–1306. doi:10.1016/j.visres.2008.09.007

36 Kirkham NZ, Slemmer JA, Johnson SP. Visual statistical learning in infancy: evidence for a domain general learning mechanism. Cognition. 2002;83: 4–5.

37 Howard CJ, Holcombe AO. Unexpected changes in direction of motion attract attention. Attention, Perception, Psychophys. 2010;72: 2087–2095.

38 Reynolds JH, Chelazzi L. Attentional Modulation of Visual Processing. Annu Rev Neurosci. 2004;27: 611–647. doi:10.1146/annurev.neuro.26.041002.131039

39 Williford T, Maunsell JHR. Effects of spatial attention on contrast response functions in macaque area V4. J Neurophysiol. 2006;96: 40–54. doi:10.1152/jn.01207.2005

40 Alink A, Blank H. Comment on "Statistical learning attenuates visual activity only for attended stimuli."

41 Bell AH, Summerfield C, Morin EL, Malecek NJ, Ungerleider LG. Encoding of Stimulus Probability in Macaque Inferior Temporal Cortex. Curr Biol. 2016;26: 2280–2290. doi:10.1016/j.cub.2016.07.007

42 Blakemore S, Wolpert DM, Frith CD. Central cancellation of self-produced tickle sensation. Nat Neurosci. 1998;1: 635–640.

43 Blank H, Davis MH. Prediction Errors but Not Sharpened Signals Simulate Multivoxel fMRI Patterns during Speech Perception. PLoS Biol. 2016;14: 1–32. doi:10.1371/journal.pbio.1002577

44 Press C, Kok P, Yon D. The Perceptual Prediction Paradox. Trends Cogn Sci. 2020;24: 13–24. doi:10.1016/j.tics.2019.11.003

45 Simoncelli EP, Olshausen BA. Natrual Image Statistics and Neural Representation. Annu Rev Neurosci. 2001;24: 1193–1216.

46 Sherman BE, Graves KN, Turk-Browne NB. The prevalence and importance of statistical learning in human cognition and behavior. Curr Opin Behav Sci. 2020;32: 15–20. doi:10.1016/j.cobeha.2020.01.015

47 Batterink LJ, Paller KA, Reber PJ. Understanding the Neural Bases of Implicit and Statistical Learning. Top Cogn Sci. 2019;11: 482–503. doi:10.1111/tops.12420

48 Schapiro A, Turk-Browne N. Statistical Learning. Brain Mapp An Encycl Ref. 2015;3: 501–506. doi:10.1016/B978-0-12-397025-1.00276-1

49 Frost R, Armstrong BC, Christiansen MH. Statistical Learning Research : A Critical Review and Possible New Directions. Psychol Bull. 2019;145: 1128–1153.

50 Fiser J, Aslin RN. Statistical learning of higher-order temporal structure from visual shape sequences. J Exp Psychol Learn Mem Cogn. 2002;28: 458–467. doi:10.1037/0278-7393.28.3.458

51 Kim R, Seitz A, Feenstra H, Shams L. Testing assumptions of statistical learning: Is it long-term and implicit? Neurosci Lett. 2009;461: 145–149. doi:10.1016/j.neulet.2009.06.030

52 Dotsch R, Hassin RR, Todorov A. Statistical learning shapes face evaluation. Nat Hum Behav. 2017;1: 1–6. doi:10.1038/s41562-016-0001

53 Kaposvari P, Kumar S. Statistical Learning Signals in Macaque Inferior Temporal Cortex. Cereb Cortex. 2016; 1–17. doi:10.1093/glycob/cwx064

54 Turk-Browne NB, Scholl BJ. Flexible Visual Statistical Learning: Transfer Across Space and Time. J Exp Psychol Hum Percept Perform. 2009;35: 195–202. doi:10.1037/0096-1523.35.1.195

55 Leung Y, Dean RT. Learning unfamiliar pitch intervals: A novel paradigm for demonstrating the learning of statistical associations between musical pitches. PLoS One. 2018;13. doi:10.1371/journal.pone.0203026

56 Thiessen ED. Effects of Visual Information on Adults' and Infants' Auditory Statistical Learning. Cogn Sci. 2010;34: 1093–1106. doi:10.1111/j.1551-6709.2010.01118.x

57 Saffran JR, Johnson EK, Aslin RN, Newport EL. Statistical learning of tone sequences by human infants and adults. Cognition. 1999;70: 27–52. doi:10.1016/S0010-0277(98)00075-4

58 Bertels J, Franco A, Destrebecqz A. How implicit is visual statistical learning? J Exp Psychol Learn Mem Cogn. 2012;38: 1425–1431. doi:10.1037/a0027210

59 Hunt RH, Aslin RN. Statistical learning in a serial reaction time task: Access to seperable statistical cues by individual learners. J Exp Psychol Gen. 2001;130: 658–680. doi:10.1037/0096-3445.130.4.658

60 Turk-Browne NB, Junge JA, Scholl BJ. The Automaticity of Visual Statistical Learning. 2005;134: 552–564. doi:10.1037/0096-3445.134.4.552

61 Ouden HEM Den, Friston KJ, Daw ND, Mcintosh AR, Stephan KE. A Dual Role for Prediction Error in Associative Learning. Cereb Cortex. 2009;19. doi:10.1093/cercor/bhn161

62 Larsson J, Smith AT. fMRI Repetition Suppression: Neuronal Adaptation or Stimulus Expectation ? Cereb Cortex. 2012; 567–576. doi:10.1093/cercor/bhr119

63 Frost R, Armstrong BC, Siegelman N, Christiansen MH. Domain generality versus modality specificity: the paradox of statistical learning. Trends Cogn Sci. 2015;19: 117–125. doi:10.1016/j.tics.2014.12.010

64 Reber PJ. Neuropsychologia The neural basis of implicit learning and memory : A review of neuropsychological and neuroimaging research. Neuropsychologia. 2013;51: 2026–2042. doi:10.1016/j.neuropsychologia.2013.06.019

65 Emberson LL, Conway CM, Christiansen MH. Timing is everything: Changes in presentation rate have opposite effects on auditory and visual implicit statistical learning. Q J Exp Psychol. 2011;64: 1021–1040. doi:10.1080/17470218.2010.538972

66 Redington M, Chater N. Transfer in Artificial Grammar Learning: A Reevaluation. J Exp Psychol Gen. 1996;125: 123–138. doi:10.1037/0096-3445.125.2.123

67 Conway CM, Christiansen MH. Modality-constrained statistical learning of tactile, visual, and auditory sequences. J Exp Psychol Learn Mem Cogn. 2005;31: 24–39. doi:10.1037/0278-7393.31.1.24

68 Saffran JR, Thiessen ED. Domain-General Learning Capacities. 2007.

69 Bulf H, Johnson SP, Valenza E. Visual statistical learning in the newborn infant. Cognition. 2011;121: 127–132. doi:10.1016/j.cognition.2011.06.010

70 Endress AD, Mehler J. The surprising power of statistical learning : When fragment knowledge leads to false memories of unheard words. J Mem Lang. 2009;60: 351–367. doi:10.1016/j.jml.2008.10.003

71 Schapiro AC, Kustner L V., Turk-Browne NB. Shaping of object representations in the human medial temporal lobe based on temporal regularities. Curr Biol. 2012;22: 1622–1627. doi:10.1016/j.cub.2012.06.056

72 Schapiro AC, Gregory E, Landau B, McCloskey M, Turk-Browne NB. The Necessity of the Medial Temporal Lobe for Statistical Learning. J Cogn Neurosci. 2014;26: 1736–1747. doi:10.1162/jocn

73 Turk-Browne NB, Scholl BJ, Johnson MK, Chun MM. Implicit Perceptual Anticipation Triggered by Statistical Learning. J Neurosci. 2010;30: 11177–11187. doi:10.1523/JNEUROSCI.0858-10.2010

74 Bubic A, von Carmon YD, Schubotz RI. Prediction, cognition and the brain. Front Hum Neurosci. 2010;4: 1–15. doi:10.3389/fnhum.2010.00025

75 Brady TF, Oliva A. Statistical learning using real-world scenes: extracting categorical regularities without conscious intent. Psychol Sci. 2008;19: 678–685. doi:10.1111/j.1467-9280.2008.02142.x

76 Garrido MI, Teng CLJ, Taylor JA, Rowe EG, Mattingley JB. Surprise responses in the human brain demonstrate statistical learning under high concurrent cognitive demand. npj Sci Learn. 2016;1: 16006. doi:10.1038/npjscilearn.2016.6

77 Denys K, Vanduffel W, Fize D, Koen N, Peuskens H, Van Essen D, et al. The Processing of Visual Shape in the Cerebral Cortex of Human and Nonhuman Primates: A Functional Magnetic Resonance Imaging Study. J Neurosci. 2004;24: 2551–2565. doi:10.1523/JNEUROSCI.3569-03.2004

78 John-Saaltink ES, Utzerath C, Kok P, Lau HC, De Lange FP. Expectation suppression in early visual cortex depends on task set. PLoS One. 2015;10: 1–14. doi:10.1371/journal.pone.0131172

79 Kumar S, Kaposvari P, Vogels R. Encoding of Predictable and Unpredictable Stimuli by Inferior Temporal Cortical Neurons. J Cogn Neurosci. 2017;29: 1445–1454. doi:10.1162/jocn

80 Coggan DD, Liu W, Baker DH, Andrews TJ. Category-selective patterns of neural response in the ventral visual pathway in the absence of categorical information. Neuroimage. 2016;135: 107–114. doi:10.1016/j.neuroimage.2016.04.060

81 Dobbins IG, Schnyer DM, Verfaellie M, Schacter DL. Cortical activity reductions during repetition priming can result from rapid response learning. Nature. 2004;428: 316–9. doi:10.1038/nature02400

82 Horner AJ, Henson RN. Priming, response learning and repetition suppression. Neuropsychologia. 2008;46: 1979–1991. doi:10.1016/j.neuropsychologia.2008.01.018

83 Brady TF, Konkle T, Alvarez GA, Oliva A. Visual long-term memory has a massive storage capacity for object details. Proc Natl Acad Sci. 2008;105: 14325–14329. doi:10.1073/pnas.0803390105

84 Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. Front Psychol. 2013;4: 1–12. doi:10.3389/fpsyg.2013.00863

85 Cousineau D. Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. Tutor Quant Methods Psychol. 2005;1: 42–45. doi:10.20982/tqmp.01.1.p042

86 Morey RD. Confidence Intervals from Normalized Data: A correction to Cousineau (2005). Tutor Quant Methods Psychol. 2008;4: 61–64. doi:10.20982/tqmp.04.2.p061

87 Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TEJ, Johansen-Berg H, et al. Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage. 2004;23: 208–219. doi:10.1016/j.neuroimage.2004.07.051

88 Mumford J. A guide to calculating percent change with featquery. Unpubl Tech Rep http//mumford …. 2007; 1–6. Available: http://mumford.fmripower.org/perchange_guide.pdf

89 Mumford JA, Turner BO, Ashby FG, Poldrack RA. Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. Neuroimage. 2012;59: 2636–2643. doi:10.1016/j.neuroimage.2011.08.076

90 Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12: 2825–2830. doi:10.1007/s13398-014-0173-7.2

91 Schwarz G. Estimating the dimension of a model. Ann Stat. 1978;6: 461–464.

92 JASP Team. JASP (Version 0.8.1.1) [Computer Software]. 2017.

93 Wagenmakers EJ, Love J, Marsman M, Jamil T, Ly A, Verhagen J, et al. Bayesian inference for psychology. Part II: Example applications with JASP. Psychon Bull Rev. 2017; 1–19. doi:10.3758/s13423-017-1323-7

94 Lee MD, Wagenmakers E-J. Bayesian cognitive modeling: A practical course. Cambridge Univ Press. 2013.

95 Kourtzi Z, Kanwisher N. Representation of Perceived Object Shape by the Human Lateral Occipital Complex. Science (80- ). 2001;293. doi:10.1093/cercor/13.9.911

96 Haushofer J, Livingstone MS, Kanwisher N. Multivariate patterns in object-selective cortex dissociate perceptual and physical shape similarity. PLoS Biol. 2008;6: 1459–1467. doi:10.1371/journal.pbio.0060187

97 Dale AM, Fischl B, Sereno MI. Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction. Neuroimage. 1999;9: 179–194. doi:10.1006/nimg.1998.0395

98 Walt S Van Der, Colbert SC, Varoquaux G. The NumPy array: a structure for efficient numerical computation. Comput Sci Eng. 2011.

99 Jones E, Oliphant E, Peterson P, Et Al. SciPy Open Source Scientific Tools for Python. 2001. Available: www.scipy.org

100 Hunter BJD. Matplotlib: A 2D Graphics Environment. Comput Sci Eng. 2007; 90–95.

101 Ramachandran P, Varoquaux G. Mayavi: 3D Visualization of Scientific Data. IEEE Comput Sci Eng. 2011;13: 40–51.

102 Zandbelt B. Slice Display. figshare 106084/m9.figshare4742866. 2017.

103 Allen EA, Erhardt EB, Calhoun VD. Data Visualization in the Neurosciences: Overcoming the Curse of Dimensionality. Neuron. 2012;74: 603–608. doi:http://dx.doi.org/10.1016/j.neuron.2012.05.001

104 Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage. 2006;31: 968–980. doi:10.1016/j.neuroimage.2006.01.021

105 Alink A, Abdulrahman H, Henson RN. Forward models demonstrate that repetition suppression is best modelled by local neural scaling. Nat Commun. 2018;9. doi:10.1038/s41467-018-05957-0

106 Alink A, Schwiedrzik CM, Kohler A, Singer W, Muckli L. Stimulus Predictability Reduces Responses in Primary Visual Cortex. J Neurosci. 2010;30: 2960–2966. doi:10.1523/JNEUROSCI.3730-10.2010

107 Brockmole JR, Boot WR. Should I stay or should I go? Attentional disengagement from visually unique and unexpected items at fixation. J Exp Psychol Hum Percept Perform. 2009;35: 808–815.

108 Kimura M, Takeda Y. Automatic prediction regarding the next state of a visual object: Electrophysiological indicators of prediction match and mismatch. Brain Res. 2015;1626: 31–44. doi:10.1016/j.brainres.2015.01.013

109 Kaliukhovich DA, Vogels R. Neurons in Macaque Inferior Temporal Cortex Show No Surprise Response to Deviants in Visual Oddball Sequences. 2014;34: 12801–12815. doi:10.1523/JNEUROSCI.2154-14.2014

110 Kaliukhovich DA, Vogels R. Stimulus Repetition Probability Does Not Affect Repetition Suppression in Macaque Inferior Temporal Cortex. Cereb Cortex. 2011; 1547–1558. doi:10.1093/cercor/bhq207

111 Serences JT, Saproo S, Scolari M, Ho T, Muftuler LT. Estimating the influence of attention on population codes in human visual cortex using voxel-based tuning functions. Neuroimage. 2009;44: 223–231. doi:10.1016/j.neuroimage.2008.07.043

112 Den Ouden HEM, Kok P, de Lange FP. How prediction errors shape perception, attention, and motivation. Front Psychol. 2012;3: 1–12. doi:10.3389/fpsyg.2012.00548

113 Richter D, Ekman M, de Lange FP. Suppressed Sensory Response to Predictable Object Stimuli throughout the Ventral Visual Stream. J Neurosci. 2018;38: 7452–7461. doi:10.1523/JNEUROSCI.3421-17.2018

114 Rao RPN. Bayesian inference and attentional modulation in the visual cortex. 2005;16: 3–8.

115 Reimer J, Froudarakis E, Cadwell CR, Yatsenko D, H DG, Tolias AS. Pupil fluctuations track fast switching of cortical states during quiet wakefulness. Neuron. 2014;84: 355–362. doi:10.1016/j.neuron.2014.09.033.Pupil

116 Vinck M, Batista-brito R, Knoblich U, Cardin JA. Arousal and locomotion make distinct contributions to cortical activity patterns and visual encoding. Neuron. 2015;86: 740–754. doi:10.1016/j.neuron.2015.03.028.Arousal

117 Damsma A, Rijn H Van. Brain and Cognition Pupillary response indexes the metrical hierarchy of unattended rhythmic violations. Brain Cogn. 2017;111: 95–103. doi:10.1016/j.bandc.2016.10.004

118 Kloosterman NA, Meindertsma T, Loon AM Van, Lamme VAF, Yoram S, Donner TH. Pupil size tracks perceptual content and surprise. 2015;41: 1068–1078. doi:10.1111/ejn.12859

119 Preuschoff K, Marius B, Einhäuser W, Nieuwenhuis S. Pupil dilation signals surprise : evidence for noradrenaline ' s role in decision making MATERIALS AND METHODS. 2011;5: 1–12. doi:10.3389/fnins.2011.00115

120 Summerfield C, Trittschuh EH, Monti JM, Mesulam MM, Egner T. Neural repetition suppression reflects fulfilled perceptual expectations. Nat Neurosci. 2008;11: 1004–1006. doi:10.1038/nn.2163

121 Brass M, Haggard P. The hidden side of intentional action: the role of the anterior insular cortex. Brain Struct Funct. 2010;214: 603–610. doi:10.1007/s00429-010-0269-6

122 Nelson SM, Dosenbach NUF, Cohen AL, Wheeler ME, Schlaggar BL, Petersen SE. Role of the anterior insula in task-level control and focal attention. Brain Struct Funct. 2010;214: 669–680. doi:10.1007/s00429-010-0260-2

123 Aron AR, Robbins TW, Poldrack RA. Inhibition and the right inferior frontal cortex. Trends Cogn Sci. 2004;8: 170–177. doi:10.1016/j.tics.2004.02.010

124 Aron AR, Fletcher PC, Bullmore ET, Sahakian BJ, Robbins TW. Stop-signal inhibition disrupted by damage to right inferior frontal gyrus in humans. Nat Neurosci. 2003;6: 115–116. doi:10.1038/nn1003

125 Horstmann G, Herwig A. Surprise attracts the eyes and binds the gaze. 2015; 743–749. doi:10.3758/s13423-014-0723-1

126 Yon D, Lange FP De, Press C. The Predictive Brain as a Stubborn Scientist. Trends Cogn Sci. 2019;23: 6–8. doi:10.1016/j.tics.2018.10.003

127 Kok P, Bains LJ, Mourik T Van, Norris DG, Lange FP De, Kok P, et al. Selective Activation of the Deep Layers of the Human Primary Visual Cortex by Top-Down Feedback Report Selective Activation of the Deep Layers of the Human Primary Visual Cortex by Top-Down Feedback. Curr Biol. 2016;26: 371–376. doi:10.1016/j.cub.2015.12.038

128 Ekman M, Kok P, De Lange FP. Time-compressed preplay of anticipated events in human primary visual cortex. Nat Commun. 2017;8: 1–9. doi:10.1038/ncomms15276

129 Jack AI, Shulman GL, Snyder AZ, Mcavoy M. Separate Modulations of Human V1 Associated with Spatial Attention and Task Structure. 2006; 135–147. doi:10.1016/j.neuron.2006.06.003

130 Donner TH, Sagi D, Bonneh YS, Heeger DJ. Opposite Neural Signatures of Motion-Induced Blindness in Human Dorsal and Ventral Visual Cortex. J Neurosci. 2008;28: 10298–10310. doi:10.1523/JNEUROSCI.2371-08.2008

131 Aston-Jones G, Cohen JD. An Integrative Theory of Locus Coeruleus-Norepinephrine Function: Adaptive Gain and Optimal Performance. Annu Rev Neurosci. 2005;28: 403–450. doi:10.1146/annurev.neuro.28.061604.135709
132 Yu AJ, Dayan P. Uncertainty, neuromodulation, and attention. Neuron. 2005;46: 681–692. doi:10.1016/j.neuron.2005.04.026
133 Reimer J, Mcginley MJ, Liu Y, Rodenkirch C, Wang Q, Mccormick DA, et al. Pupil fluctuations track rapid changes in adrenergic and cholinergic activity in cortex. Nat Commun. 2016;7: 1–7. doi:10.1038/ncomms13289
134 Berridge CW, Schmeichel BE, Espana RA. Noradrenergic Modulation of Wakefulness/Arousal. Sleep Med Rev. 2012;16: 187–197. doi:10.1016/j.brainresrev.2007.10.013
135 Haynes J, Lotto RB, Rees G. Responses of human visual cortex to uniform surfaces. 2004.
136 Gavornik JP, Bear MF. Learned spatiotemporal sequence recognition and prediction in primary visual cortex. Nat Publ Gr. 2014;17: 732–737. doi:10.1038/nn.3683
137 Yu AJ, Dayan P. Inference, attention, and decision in a bayesian neural architecture. Adv Neural Inf Process Syst. 2005;17.
138 Stokes M, Anderson M, Chandrasekar S, Motta R. A Standard Default Color Space for the Internet – sRGB. 1996. Available: www.w3.org/Graphics/Color/sRGB
139 JASP Team. JASP (Version 0.9.0.1) [Computer Software]. 2018.
140 Mathôt S. A simple way to reconstruct pupil size during eye blinks Blink detection Pupil-size reconstruction. 2013. Available: https://doi.org/10.6084/m9.figshare.688001
141 Mathôt S. Pupillometry: Psychology, Physiology, and Function. J Cogn. 2018;1: 1–23. doi:10.5334/joc.18
142 Yon D, Gilbert SJ, De Lange FP, Press C. Action sharpens sensory representations of expected outcomes. Nat Commun. 2018;9: 1–8. doi:10.1038/s41467-018-06752-7
143 Han B, Mostert P, De Lange FP. Predictable tones elicit stimulus-specific suppression of evoked activity in auditory cortex. Neuroimage. 2019;200: 242–249. doi:10.1016/j.neuroimage.2019.06.033
144 Weiner KS, Sayres R, Vinberg J, Grill-spector K. fMRI-Adaptation and Category Selectivity in Human Ventral Temporal Cortex: Regional Differences Across Time Scales. J Neurophysiol. 2010;103: 3349–3365. doi:10.1152/jn.01108.2009.
145 Richter D, De Lange FP. Statistical learning attenuates visual activity only for attended stimuli. Elife. 2019; 1–27.
146 Wiggs CL, Martin A. Properties and mechanisms of perceptual priming. Curr Opin Neurobiol. 1998;8: 227–233.
147 Henson RNA, Rugg MD. Neural response suppression, haemodynamic repetition effects, and behavioural priming. Neuropsychologia. 2003;41: 263–270.

148 Martinez-Trujillo JC, Treue S. Feature-Based Attention Increases the Selectivity of Population Responses in Primate Visual Cortex. Curr Biol. 2004;14: 744–751. doi:10.1016/j

149 Kriegeskorte N, Mur M, Ruff DA, Kiani R, Bodurka J, Esteky H, et al. Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey. Neuron. 2008;60: 1126–1141. doi:10.1016/j.neuron.2008.10.043. Matching

150 Boynton GM. Imaging orientation selectivity: decoding conscious perception in V1. Nat Neurosci. 2005;8: 541–542.

151 Freeman J, Brouwer GJ, Heeger DJ, Merriam EP. Orientation Decoding Depends on Maps, Not Columns. J Neurosci. 2011;31: 4792–4804. doi:10.1523/JNEUROSCI.5160-10.2011

152 Roth ZN, Heeger DJ, Merriam EP. Stimulus vignetting and orientation selectivity in human visual cortex. Elife. 2018; 1–19.

153 Turner BO, Mumford JA, Poldrack RA, Ashby FG. Spatiotemporal activity estimation for multivoxel pattern analysis with rapid event-related designs. Neuroimage. 2012;62: 1429–1438. doi:10.1016/j.neuroimage.2012.05.057. Spatiotemporal

154 Haxby J V, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P. Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. Science (80- ). 2001;293: 2425–2431.

155 Jaini AK, Farrokitniaj F. Unsupervised texture segmentation using Gabor filters. Pattern Recognit. 1991;24: 1167–1186.

156 Kriegeskorte N, Mur M. Inverse MDS: inferring dissimilarity structure from multiple item arrangements. Front Psychol. 2012;3: 1–13. doi:10.3389/fpsyg.2012.00245

157 Bonds AB. An “Oblique Effect” in the Visual Evoked Potential of the Cat. Exp Brain Res. 1982;46: 151–154.

158 Furmanski CS, Engel SA. An oblique effect in human primary visual cortex. Nat Neurosci. 2000;3: 535–536.

159 Rubin J, Ulanovsky N, Nelken I, Tishby N. The Representation of Prediction Error in Auditory Cortex. 2016; 1–28. doi:10.5061/dryad.3m5v5

160 Maheu M, Meyniel F, Dehaene S. Rational arbitration between statistics and rules in human sequence learning. 2020.

161 Siegelman N, Bogaerts L, Kronenfeld O, Frost R. Redefining “Learning” in Statistical Learning: What Does an Online Measure Reveal About the Assimilation of Visual Regularities? Cogn Sci. 2018;42: 692–727. doi:10.1111/cogs.12556

162 Bogaerts L, Siegelman N, Frost R. Splitting the variance of statistical learning performance: A parametric investigation of exposure duration and transitional probabilities. Psychon Bull Rev. 2016;23: 1250–1256. doi:10.3758/s13423-015-0996-z

163 Jungé JA, Scholl BJ, Chun MM. How is spatial context learning integrated over signal versus noise? A primacy effect in contextual cueing. Vis cogn. 2007;15: 1–11. doi:10.1038/jid.2014.371

164 Walk AM, Conway CM. Cross-Domain Statistical – Sequential Dependencies Are Difficult to Learn. 2016;7: 1–9. doi:10.3389/fpsyg.2016.00250

165 Conway CM, Christiansen MH. Statistical Learning Within and Between Modalities. Pitting Abstract Against Stimulus-Specific Representations. Psychol Sci. 2006;17: 905–912.

166 Mitchel AD, Christiansen MH, Weiss DJ, Andrews M, Trent N. Multimodal integration in statistical learning : evidence from the McGurk illusion. 2014;5: 1–6. doi:10.3389/fpsyg.2014.00407

167 Seitz AR, Kimô R, Wassenhoveô V Van, Shamsô L. Simultaneous and independent acquisition of multisensory and unisensory associations. 2007;36: 1445–1454. doi:10.1068/p5843

168 Piazza EA, Denison RN, Silver MA. Recent cross-modal statistical learning influences visual perceptual selection. 2018;18: 1–12.

169 Squire LR, Zola SM. Structure and function of declarative and nondeclarative memory systems. Proc Natl Acad Sci U S A. 1996;93: 13515–13522. doi:10.1073/pnas.93.24.13515

170 Davachi L, DuBrow S. How the hippocampus preserves order: The role of prediction and context. Trends Cogn Sci. 2015;19: 92–99. doi:10.1016/j.tics.2014.12.004

171 Hindy NC, Ng FY, Turk-Browne NB. Linking pattern completion in the hippocampus to predictive coding in visual cortex. Nat Neurosci. 2016;19: 665–667. doi:10.1038/nn.4284

172 Lavenex P, Amaral DG. Hippocampal-Neocortical Interaction: A Hierarchy of Associativity Pierre. Hippocampus. 2000;10: 420–430.

173 Dehaene S, Naccache L. Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. Cognition. 2001;79: 1–37. doi:10.1016/S0010-0277(00)00123-2

174 Dehaene S, Changeux JP. Experimental and Theoretical Approaches to Conscious Processing. Neuron. 2011;70: 200–227. doi:10.1016/j.neuron.2011.03.018

175 Watanabe S. Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. 2010;11: 3571–3594.

176 Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nat Methods. 2020. doi:https://doi.org/10.1038/s41592-019-0686-2

177 Seabold S, Perktold J. statsmodels: Econometric and statistical modeling with python. 9th Python in Science Conference. 2010.

178 Mckinney W. Data Structures for Statistical Computing in Python. 2010;1697900: 51–56.

179 Salvatier J, Wiecki T V, Fonnesbeck C. Probabilistic programming in Python using PyMC3. PeerJ Comput Sci. 2016; 1–24. doi:10.7717/peerj-cs.55

180 Alamia A, VanRullen R, Pasqualotto E, Mouraux A, Zenon A. Pupil-linked arousal responds to unconscious surprisal. J Neurosci. 2019;39: 5369–5376. doi:10.1523/JNEUROSCI.3010-18.2019

181 Fiser A, Mahringer D, Oyibo HK, Petersen A V., Leinweber M, Keller GB. Experience-dependent spatial expectations in mouse visual cortex. Nat Neurosci. 2016;19: 1658–1664. doi:10.1038/nn.4385

182 Kok P, Failing MF, de Lange FP. Prior Expectations Evoke Stimulus Templates in the Primary Visual Cortex. J Cogn Neurosci. 2014;26: 1546–1554.

183 Schröger E, Marzecová A, Sanmiguel I. Attention and prediction in human audition: A lesson from cognitive psychophysiology. Eur J Neurosci. 2015;41: 641–664. doi:10.1111/ejn.12816

184 Bendixen A, Schröger E, Winkler I. I heard that coming: Event-related potential evidence for stimulus-driven prediction in the auditory system. J Neurosci. 2009;29: 8447–8451. doi:10.1523/JNEUROSCI.1493-09.2009

185 Manahova ME, Mostert P, Kok P, Schoffelen J, Lange FP De. Stimulus Familiarity and Expectation Jointly Modulate Neural Activity in the Visual Ventral Stream. J Cogn Neurosci. 2018;30: 1366–1377. doi:10.1162/jocn

186 Maunsell JHR. Neuronal Mechanisms of Visual Attention. Annu Rev Vis Sci. 2015;1: 373–391. doi:10.1146/annurev-vision-082114-035431

187 Bar M. Visual objects in context. Nat Rev Neurosci. 2004;5: 617–629. doi:10.1038/nrn1476

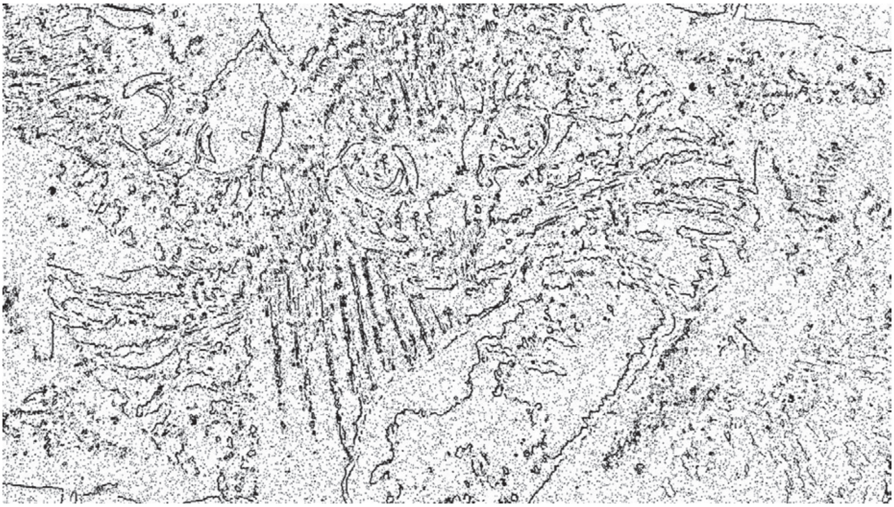
188 Heilbron M, Richter D, Ekman M, Hagoort P, de Lange FP. Word contexts enhance the neural representation of individual letters in early visual cortex. Nat Commun. 2020;11: 1–11. doi:10.1038/s41467-019-13996-4

189 Teufel C, Fletcher PC. Forms of prediction in the nervous system. Nat Rev Neurosci. 2020;21: 231–242. doi:10.1038/s41583-020-0275-5

# Nederlandse Samenvatting

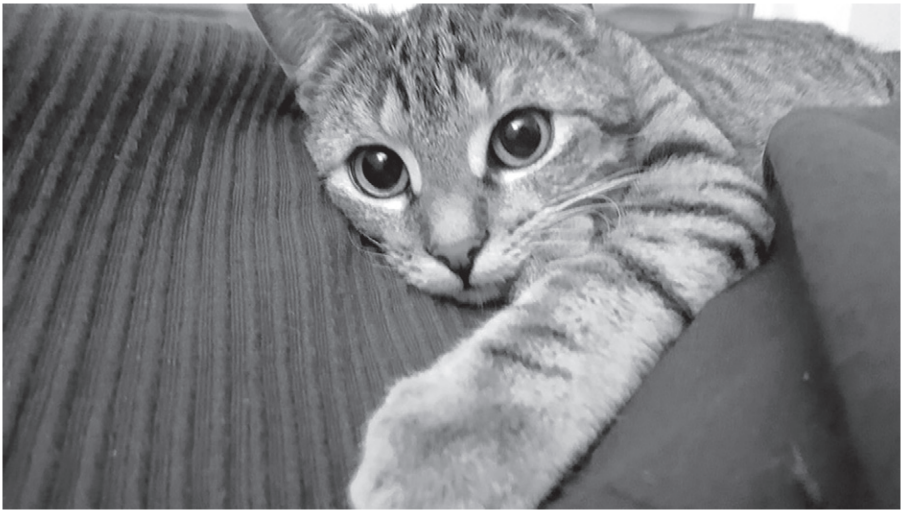
Als we naar de wereld kijken, gebruiken we onze voorkennis om te begrijpen wat we zien. Meestal worden we ons pas bewust van dit proces als onze voorspellingen onjuist blijken. Denk bijvoorbeeld aan de situatie waarin je een hoek omloopt en bijna tegen iemand aanbotst. De verrassing en de schrik die je voelde, is het gevolg van een voorspellingsfout. In mijn proefschrift heb ik onderzocht hoe het brein voorspellingen gebruikt om onze visuele perceptie te sturen. Een eenvoudig voorbeeld hiervan wordt geïllustreerd in Figuur 1 hieronder. Als je naar de afbeelding kijkt zie je in eerste instantie waarschijnlijk alleen onsamenhangende vormen en lijnen. Ga nu eerst naar de volgende pagina om Figuur 2 te bekijken en ga daarna terug en kijk opnieuw naar Figuur 1.

Nadat je hebt gezien dat er een kat op de afbeelding in Figuur 1 staat, is je waarneming waarschijnlijk drastisch veranderd: van een onsamenhangend geheel bestaande uit lijnen en vormen naar de perceptie van een kat. Het beeld zelf is natuurlijk niet veranderd, alleen je voorkennis is veranderd en daarmee ook je bewuste waarneming. Deze ervaring suggereert dat onze kennis vormgeeft aan wat we zien.



**FIGUUR 1 Illustratie van het effect van kennis op perceptie.**
Aanvankelijk lijkt het of de afbeelding alleen uit willekeurige vormen bestaat. Echter, als je naar de volgende pagina gaat en naar Figuur 2 kijkt, dan kun je zien dat er een kat in de afbeelding verborgen zit. Nu je weet wat je van de afbeelding kunt verwachten, is je waarneming van willekeurige vormen veranderd in die van een kat, ook al blijft de afbeelding zelf identiek.

In **hoofdstuk 1** introduceer ik de kernvraag van mijn proefschrift: hoe beïnvloedt voorkennis onze waarneming? Uit het bovenstaande voorbeeld heb je kunnen ervaren dat kennis inderdaad kan veranderen wat je ziet. Echter, gedurende het grootste deel van ons leven merken we niet bewust dat we onze voorkennis met zintuiglijke informatie combineren om onze waarnemingen te vormen. Het proces dat ik bestudeer lijkt dus een onbewuste perceptuele gevolgtrekking te zijn – dat wil zeggen, het brein voert deze operatie automatisch uit. Ook kunnen we uit het eerdergenoemde voorbeeld, waarin je per ongeluk bijna tegen iemand aanbotst terwijl je de hoek omloopt, opmaken dat het brein zintuigelijke waarnemingen lijkt te voorspellen. Hoe het brein voorkennis opdoet en vervolgens gebruikt om sensorische input te voorspellen heb ik onderzocht met behulp van magnetic resonance imaging (kernspintomografie) in hoofdstuk 2-5 van mijn proefschrift.



**FIGUUR 2** **Oorspronkelijke afbeelding van Figuur 1.**

Als je nu opnieuw naar de afbeelding in Figuur 1 kijkt, dan zijn de schijnbaar willekeurige vormen en lijnen in het figuur mogelijk veranderd door de nieuwe kennis die je hebt opgedaan door naar deze originele afbeelding te kijken.

In **hoofdstuk 2** heb ik onderzocht hoe mensen binnenkomende zintuigelijke informatie voorspellen en hoe deze voorspellingen neurale processen moduleren. Daartoe heb ik proefpersonen paren van afbeeldingen laten zien. Zonder dat ze het wisten, waren sommige afbeeldingen voorspellend voor andere afbeeldingen, wat betekent dat het zien van bijvoorbeeld afbeelding A de kans vergrootte dat afbeelding B zou volgen. Ik heb dergelijke statistische regelmatigheden gebruikt om te begrijpen hoe de hersenen reageren op verwachte afbeeldingen in vergelijking met onverwachte beelden. Dat wil zeggen, verandert het correct voorspellen van afbeelding B hoe het beeld wordt verwerkt in het sensorische brein? Mijn resultaten suggereren dat dit inderdaad het geval is. Het zintuiglijk brein reageert minder krachtig en minder duidelijk op correct voorspelde beelden, zelfs zonder enige bewuste intentie om te voorspellen. Met andere woorden, ons brein lijkt zintuigelijke input automatisch te voorspellen op basis van statistische regelmatigheden die we hebben geleerd.

Vervolgens, in **hoofdstuk 3**, heb ik onderzocht hoe automatisch deze voorspellingen zijn. Eerder werk, waaronder het onderzoek uit hoofdstuk 2, suggereerde dat we continu en automatisch voorspellen, zelfs zonder enige intentie om dat te doen. Ik wilde echter weten of aandacht hier mogelijk toch een rol in zou kunnen spelen. Om dit te onderzoeken liet ik mijn proefpersonen weer afbeeldingenparen zien, opnieuw met statistische regelmatigheden die bepaalden of afbeeldingen meer of minder waarschijnlijk zouden verschijnen. Deelnemers kregen de instructie om naar de afbeelding of naar letters te kijken, die direct boven de abeeldingen werden getoond. Tot mijn verbazing bleek dat wanneer de aandacht van deelnemers van de afbeelding was afgeleid, alle voorspellingseffecten waren verdwenen. Met andere woorden, de hersenen reageerden nu op gelijke wijze op verwachte en onverwachte afbeeldingen. Echter, zodra hun aandacht weer naar de afbeeldingen werd geleid, waren de neurale reacties op onverwachte beelden weer sterker dan op verwachte beelden. Mijn resultaten laten dus zien dat hoewel ons brein visuele input automatisch en onbewust lijkt te voorspellen, aandacht voor de voorspelbare input nodig lijkt te zijn.

In **hoofdstuk 4** werd de vraag gesteld hoe we de verminderde neurale reactie bij verwachte input in vergelijking tot onverwachte input kunnen verklaren. Een mogelijkheid is dat het brein neurale reacties op correct voorspelde afbeeldingen vermindert, door de respons van neuronen die sterk reageren op het verwachte beeld te reduceren. Een verwacht beeld is namelijk niet erg informatief, omdat je het al kon voorspellen. Het zou dus logisch zijn als het brein minder energie besteedt aan het reageren op correct voorspelde sensorische input. De reactie wordt hierdoor effectief ‘gedempt’. Een alternatief is dat het brein nog steeds sterk reageert op de correct verwachte beeldkenmerken, maar dat de totale reactie specifieker is en minder ruis bevat, waardoor de gemiddelde reactie lager is. Denk bijvoorbeeld aan een koptelefoon met ruisonderdrukking waardoor je muziek duidelijker kunt horen bij een lager volume omdat het omgevingsgeluid wordt verminderd. Deze vraag heb ik onderzocht door opnieuw de data uit hoofdstuk 2 en 3 te onderzoeken, waar ik proefpersonen paren van (on)verwachte afbeeldingen liet zien. Aan de hand van rekenmodellen laat ik zien dat correct verwachte sensorische input lijken te worden 'gedempt'. Dat wil zeggen, het brein lijkt de correct verwachte informatie te reduceren, mogelijk omdat deze sensorische input minder informatief is. Hierdoor wordt onze aandacht naar potentieel belangrijke onverwachte sensorische input gestuurd.

**Hoofdstuk 5** verkent de grenzen van het leren van statistische regelmatigheden. In de voorgaande hoofdstukken heb ik laten zien dat mensen, automatisch en zonder de bewuste intentie om te leren statistische regelmatigheden kunnen oppikken. Vervolgens heb ik onderzocht of dit leerproces, en de gevolgen daarvan voor gedrag en het zintuiglijke brein, verschillende zintuigen kan overstijgen. Dus in plaats van proefpersonen afbeeldingsparen te laten zien, presenteerde ik geluiden die afbeeldingen voorspelden, waardoor de mensen informatie van twee verschillende zintuigen moesten gebruiken om te voorspellen. Ook hier heb ik, net als in eerdere onderzoeken, deelnemers niet gevraagd om de regelmatigheden te leren, maar vertrouwde ik op het verbazingwekkende vermogen van het brein om statistische structuur te ontdekken. Verrassend genoeg bleek dat de deelnemers deze audiovisuele regelmatigheden niet hadden geleerd: noch in hun gedrag, noch in de hersenen vonden we enig bewijs voor statistische leereffecten. Dit suggereert dat het mechanisme van statistisch leren dat we in de hoofdstukken 2-4 zagen, afhankelijk is van neurale mechanismen binnen een specifieke modaliteit (bijv. veranderingen in de visuele cortex).

In **hoofdstuk 6** integreer ik de resultaten van de voorgaande hoofdstukken tot een samenhangend geheel. Boven in Figuur 1 zag je in eerste instantie een onbegrijpelijk beeld bestaande uit willekeurige lijnen. Maar na het zien van Figuur 2, het oorspronkelijke beeld van de kat, veranderde je perceptie van Figuur 1. Deze verandering illustreerde dat onze voorkennis een directe invloed lijkt te hebben op onze waarnemingen. In mijn proefschrift heb ik laten zien hoe het brein gebruik maakt van voorkennis om waarneming te sturen. Het brein blijkt opmerkelijk goed te zijn in het detecteren en leren van statistische structuren in de zintuiglijke wereld, zelfs zonder enige intentie om te leren. Eenmaal aangeleerd lijkt dergelijke kennis door de hersenen te worden gebruikt om zintuiglijke input te voorspellen, waarbij discrepanties resulteren in voorspellingsfouten; bijvoorbeeld bijna tegen iemand aanbotsen wanneer je een hoek omloopt. In het zintuigelijke brein is zo’n voorspellingsfout duidelijk zichtbaar door een sterkere en helderder neurale reactie. Mijn werk suggereert ook dat het brein voorspellingen kan gebruiken om reacties op voorspelde sensorische input te dempen, mogelijk vanwege het feit dat goed voorspelde zintuigelijke input minder nieuwe informatie bevat dan verrassende sensorische input. In conclusie, voorspelling lijkt dus een fundamenteel aspect te zijn van sensorische informatieverwerking.

# Curriculum Vitae

David Richter was born on the 29th of May 1987 in Düsseldorf, Germany. He obtained his bachelor's degree in Psychology from Utrecht University in 2011. One year later, he completed his masters in Psychology with distinction at Leiden University. David then switched disciplines to cognitive-neuroscience and obtained a second master’s degree in Cognitive Science from the University of Osnabrück in 2016. During his master’s David worked as a research assistant in Peter König’s lab under the supervision of Tim Kietzmann. In his master’s thesis, David investigated perceptual bistability using closed-loop EEG systems under the supervision of Axel Kohler and Gordon Pipa. During his work on perceptual bistablity he became interested in predictive coding and the work of Floris de Lange, under whose supervision David started his PhD in 2016. Currently, David is working as a post-doctoral researcher in Floris de Lange’s lab, contributing to an international consortium investigating the neural correlates of consciousness.

# List of Publications

**Richter, D.**, Ekman M., and de Lange F.P. (2018) Suppressed Sensory Response to Predictable Object Stimuli throughout the Ventral Visual Stream. *The Journal of Neuroscience, 38*, 7452–7461. doi.org/10.1523/JNEUROSCI.3421-17.2018

**Richter, D.**, and de Lange, F.P. (2019) Statistical learning attenuates visual activity only for attended stimuli. *eLife*. 1–27. doi.org/10.7554/eLife.47869

Heilbron, M., **Richter, D.**, Ekman, M., Hagoort, P., and de Lange, F.P. (2020) Word contexts enhance the neural representation of individual letters in early visual cortex. *Nature Communications, 11*, 1-11. doi.org/10.1038/s41467-019-13996-4

He, T., **Richter, D.**, Wang, Z., and de Lange F. P. (2020) Spatial and temporal context jointly modulate the sensory response within the ventral visual stream. *bioRxiv*. doi.org/10.1101/2020.07.24.219709

**Richter, D.**, Heilbron, M., and de Lange, F.P. (*in preparation*). Dampened sensory representations for expected input across the ventral visual stream.

**Richter, D.**, Voorrips, E.S., Degutis, J.K., Spaak, E., and de Lange, F.P. (*in preparation*). Incidental statistical learning of unimodal but not cross-modal statistical regularities.

# Acknowledgements

I would like to thank all the people who have made this thesis possible. Within the first weeks of my time at the Donders I noticed that I had arrived at a special place, where conditions are amazing for doing science, but without neglecting the human element. Over the years, I have come to appreciate the latter more and more, and all the people who make the Donders what it is.

**Floris**, of course I want to thank you in particular. Without your supervision and your support, none of this would have been possible. I recall that already the first interview for the PhD position felt more like a pleasant discussion of neuroscience among colleagues than a job interview. To my surprise, you offered me the position still on the same day – a very positive prediction error. I have since come to know you as an excellent supervisor, who is both knowledgeable and supportive. I could not have asked for a better supervisor. Thank you for everything!

**Matthias E.**, thanks for introducing me to all the nitty gritty details on how to run experiments at the Donders and how to perform fMRI analyses. Within our first project I have learned so many essential skills, many of which I still rely on during my everyday work today. Moreover, having you around throughout my PhD ensured me that there would always be a brilliant scientist who has my back. Thank you for that!

**Lieke-chan**, I cannot thank you enough for everything. Sometimes you were guiding me on how to find my way (figuratively and literally) at the Donders, sometimes you were my office-mate to discuss science and all kinds of gossip with, and now you are my paranimf. You made the office into a second home and I am happy that we could share our time during our PhDs. **Rui-chan**, you, as much as Lieke, made the office into my second home. I recall many late evenings in the office, and your sunny presence certainly made those long evenings at work feel much more positive – even when we were both just working away. Of course, you too **Zarah**, **Christina** and **Vahid** made the office the fun place it is. I was always enjoying our chats about work (e.g. wondering whether SPM is better) and all sorts of random topics.

Of course, I also want to thank all the predictors in the Predictive Brain Lab and the former PredAtters, who came and went over the years. **Sam**, **Alexis**, **Biao**, **Annelinde**, **Alya**, **Ashley**, **Britta**, **Kim**, **Floortje**, **Marisha**, **Ilayda**, **Particia**, **Joey**, **Chris**, **Erik**, **Pim**, **Claudia** thank you all for making the lab an inspiring and hospitable place. **Micha**, thanks for being my paranimf and being awesome – it's always fun to chat with you, whether that's about science or the world at large. I have enjoyed working with you immensely, both in your and my projects; I hope that there are more collaborations to come in the future. **Ambra**, I am glad we are working together now and hope that we will have many more coffee breaks to enjoy together. **Chuyao**, thanks for the already great collaboration; I am looking forward to see what our project will uncover. **Christoph**, I hope we can have a great climbing season this year - it is about time! **Matthias F.**, I am always happy to hear your insightful thoughts, which often helped me to consider another perspective. Also, thanks for indulging all my "how do I organize a PhD defense and symposium?" questions. **Eelke**, thanks for introducing me to Bayesian modelling that was truly a breath of fresh air and made me excited to continue our project ASAP. **Tao**, it was fun working together; I hope we will meet again someday. **Ellen** and **Lola**, I am delighted to have supervised you (partially) during your Masters and wish you the best for your careers in academia. **Karolis** I am glad that you supported me for several of my projects, and that I could always rely on your help. See you in Berlin. **Sarah**, I am happy that you are now supporting me for this giant project, it feels less daunting knowing that you're helping me out. **Bene**, it's crazy how quickly time passes. It seems just like yesterday that you explained mixed models to me over lunch in Osnabrück, using mashed potatoes to draw illustrations. Since then you joined the lab and left again to start your own. I'll miss our chats and the informative illustrations drawn in your lunch. **Ella**, teaching buddy, it was fun and an interesting time teaching and organizing – I was certainly always happy that you were there to support me. Thank you **Eric Maris** for the trust to organize your workgroups and supervise your students, I learned a lot during that time. **Xiaochen**, thanks supporting me in my fight against meetings early in the morning and having the much needed coffee breaks with me.

**Paul**, the Minotaur in the basement, thanks for trusting me with your machines and teaching me a lot about MRI. Also, thanks for indulging all my unusual MRI questions and requests. The same goes to Uriel and the other members of the technical group and the administration! Thank you all for making our work possible at all. **Peter Hagoort** and **David Norris**, thank you both as directors and as scientists for making the Donders the amazing place it is today.

Thank you to my family for supporting me throughout all the years. Thank you **Ingrid** and **Klaus** for giving me the best start possible and providing amazing support for all of my choices. **Moritz**, **Max**, **Helga** and **Stephan**, thank you for being a great reminder of the joy outside of science – the same is true for **Tim**, **Malte** and **Eike**, even though we don't get to see each other as often as we'd like, you nonetheless supported all this work by helping me relax and not loose perspective. Finally, and most of all, thank you **Rosa** for being by my side throughout all of this and making me happy.

# Research Data Management

The research presented in this thesis followed the applicable laws and ethical guidelines. Research data management was conducted according to the FAIR principles (Findable, Accessible, Interoperable, Reusable). The paragraphs below specify this in detail and provide access information to the data.

## Ethics

This thesis uses data from human participants. All studies were conducted in accordance with the principles of the Declaration of Helsinki and were approved by the local ethics committee (CMO region Arnhem-Nijmegen, The Netherlands; CMO2014/288). The research was funded by The Netherlands Organisation for Scientific Research, Vidi Grant 452-13-016, and the EC Horizon 2020 Program, ERC Starting Grant 678286 “Contextvision”, awarded to Floris P. de Lange.

## Findable, Accessible

Data, code and research documentation can be found in the Donders Repository. All data is archived in a Data Acquisition Collection (DAC) or on hard disk drives (HDD). All data, code and documentation necessary to replicate published results are archived and shared in a Data Sharing Collection (DSC). All data will remain available for at least 10 years after termination of the studies.

| Chapter | DAC | RDC | DSC | DSC License |
|---|---|---|---|---|
| 2 | HDD archive:<br>DCCN HDD 1265 | - | - | - |
| 3 | di.dccn.DAC_3018028.03_898 | di.dccn.RDC_3018028.03_272 | di.dccn.DSC_3018028.03_962 | RU-DI-HD-1.0 |
| 4 | di.dccn.DAC_3018028.07_507 | - | di.dccn.DSC_3018028.07_544 | RU-DI-HD-1.0 |
| 5 | di.dccn.DAC_3018028.06_252 | - | di.dccn.DSC_3018028.06_396 | RU-DI-HD-1.0 |

**DSC: hdl handle (public access link)**
Chapter 3: http://hdl.handle.net/11633/aacg3rkw
Chapter 4: http://hdl.handle.net/11633/aaddrwao
Chapter 5: http://hdl.handle.net/11633/aadhrw2q

For chapters 2-5 data were, or in case of ongoing projects are, also stored on project/networkdrives (DCCN: project/3018028.01, project/3018028.03, project/3018028.06, project/3018028.07). These data are/were accessible to all members involved in the project. After finalization of each project, data are removed from the project/networkdrives. Projects presented in chapters 4 and 5 are still ongoing and the associated DSCs will be made available upon publication. For each project, informed consent was obtained following the DCCN informed consent and screening procedures. All associated forms are archived in the central archive for 10 years after termination of the studies.

## Interoperable, Reusable

Raw data are stored in the DAC or on HDD in their original form. Data in DSCs use long-lived file formats (e.g. .nii, .csv, .txt) to improve data access and reusability. DSCs have been organized according to, or similar to, BIDS standards, with concomitant readme files detailing code usage and data organization. Results reported in this thesis are reproducible by following instructions in the documentation and respective chapters, and/or using the raw data and analysis code provided in the DSCs.

## Privacy

The privacy of all participants has been warranted by using pseudonymized subject codes. Linking pseudonymized codes to personal data is only possible using a key file, which was stored on the secured network drive and was only accessible to members of the research project. Key files were deleted after finalization of the projects presented in chapter 2 and 3. Key files of ongoing projects, chapters 4 and 5, will be deleted after the finalization of the respective projects. MRI-data were defaced before being shared in the DSC of chapters 3-5 and are/will be shared under the restricted license RU-DI-HD-1.0, which provides additional statements for the protection of the identity of the participants.

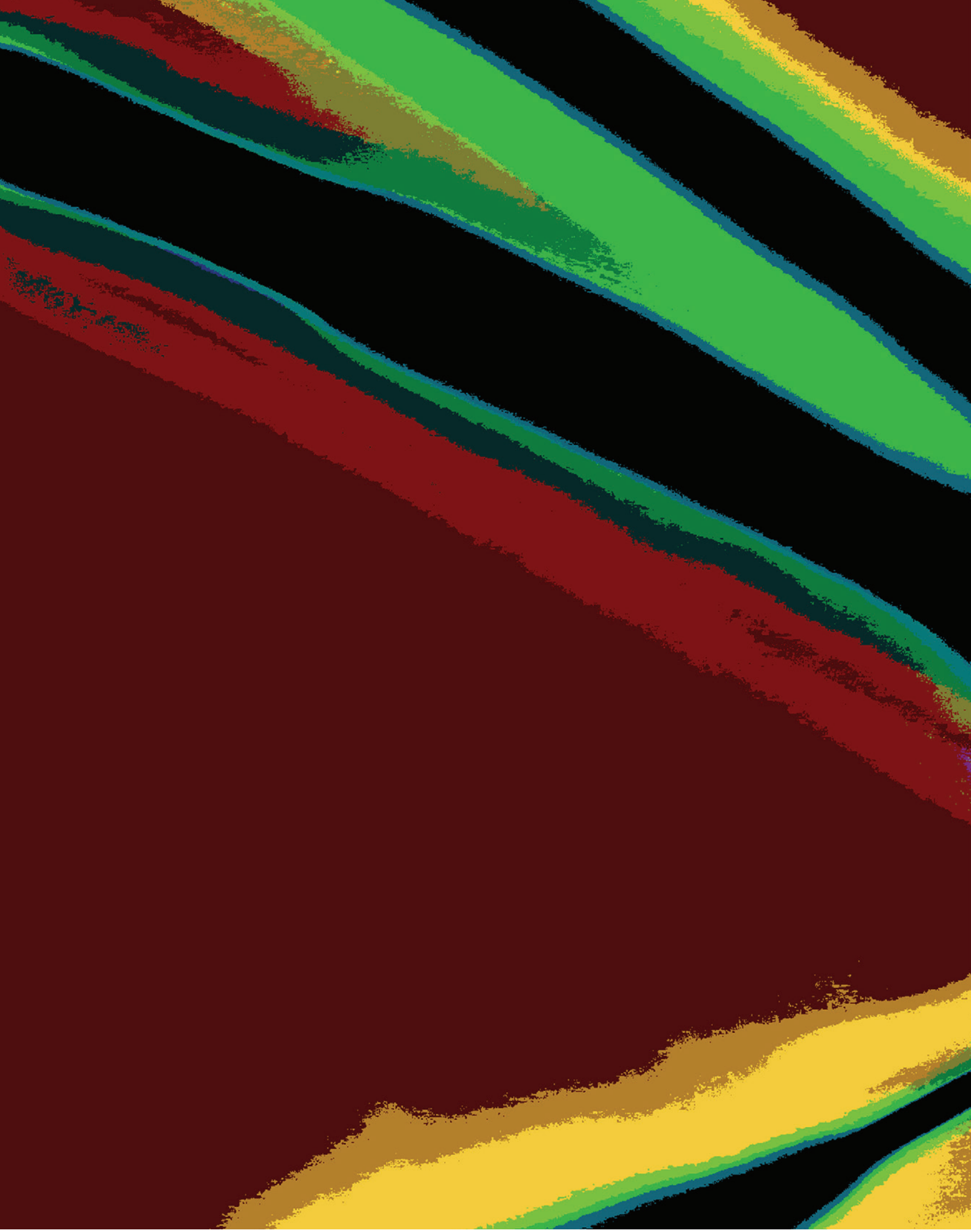
# Donders graduate school for cognitive neuroscience

For a successful research Institute, it is vital to train the next generation of young scientists. To achieve this goal, the Donders Institute for Brain, Cognition and Behaviour established the Donders Graduate School for Cognitive Neuroscience (DGCN), which was officially recognised as a national graduate school in 2009. The Graduate School covers training at both Master's and PhD level and provides an excellent educational context fully aligned with the research programme of the Donders Institute.

The school successfully attracts highly talented national and international students in biology, physics, psycholinguistics, psychology, behavioral science, medicine and related disciplines. Selective admission and assessment centers guarantee the enrolment of the best and most motivated students.

The DGCN tracks the career of PhD graduates carefully. More than 50% of PhD alumni show a continuation in academia with postdoc positions at top institutes worldwide, e.g. Stanford University, University of Oxford, University of Cambridge, UCL London, MPI Leipzig, Hanyang University in South Korea, NTNU Norway, University of Illinois, North Western University, Northeastern University in Boston, ETH Zürich, University of Vienna etc.. Positions outside academia spread among the following sectors: specialists in a medical environment, mainly in genetics, geriatrics, psychiatry and neurology. Specialists in a psychological environment, e.g. as specialist in neuropsychology, psychological diagnostics or therapy. Positions in higher education as coordinators or lecturers. A smaller percentage enters business as research consultants, analysts or head of research and development. Fewer graduates stay in a research environment as lab coordinators, technical support or policy advisors. Upcoming possibilities are positions in the IT sector and management position in pharmaceutical industry. In general, the PhDs graduates almost invariably continue with high-quality positions that play an important role in our knowledge economy.

For more information on the DGCN as well as past and upcoming defenses please visit: http://www.ru.nl/donders/graduate-school/phd/



DONDERS
INSTITUTE

Radboud University

Radboudumc