Exploring normative models of the visual and auditory systems

Sebastian Klavinskis-Whiting
Christ Church

Thesis submitted for the degree of DPhil in Neuroscience

Nuffield Department of Clinical Neurosciences,
University of Oxford

Hilary 2025

Table of Contents

Acknowledgements

Firstly, I would like to thank my supervisors Nicol, Andy, and David for their continual help, feedback and expertise throughout my DPhil. Thanks also to the Nuffield Department of Clinical Neurosciences for their financial support which made this research possible.

I am also deeply grateful to my friends in the lab and out who made my time in Oxford so enjoyable. Finally, thank you especially to Caiban and to my parents whose experience and patience helped me at every stage.

Abstract

The brain's sensory systems transform the incoming stimuli which impinge on the periphery into useful representations that can guide an organism's behaviour. Indeed, understanding the complex neural computations which underpin these transformations across sensory pathways remains an enduring goal of systems neuroscience. Normative modelling provides one approach to tackling this complexity, by describing neural systems from an optimization perspective. In this way, a normative approach can answer to what extent the diverse structural and functional properties of the sensory brain might emerge by optimizing for a few more fundamental principles of neural function. This normative approach forms the core of this thesis, which I explore across the visual and auditory systems.

To that end, in the first two results chapters, I investigate how optimizing recurrent artificial neural networks for predictive information – termed *temporal prediction* – can capture both the structural and functional properties of the mammalian visual system. Specifically, in chapter two, I describe how a shallow recurrent model trained for temporal prediction can recapitulate the functional connectivity motifs of mouse primary visual cortex (V1). In chapter three, I extend this V1 model into a hierarchical recurrent model of the dorsal visual pathway. There, I demonstrate how feedback connectivity in the network captures many of the known functional properties of higher-order feedback to V1. Finally, in chapter four, I move from the visual system to the auditory system where I take a broader view across the wider landscape of normative models. In this chapter, I investigate which properties determine how well normative models can predict neural activity and how this relates to the models' learned

representations. In particular, I show that networks which learn more general representations are better able to model auditory neural activity.

Overall, this thesis demonstrates the utility of normative networks as models of the brain and shows how the complexity sensory systems might emerge by optimizing for much simpler principles such as temporal prediction.

1

Introduction

The process of sensory perception involves extracting meaningful representations from the otherwise undifferentiated inputs which impinge upon our periphery. As we are constantly immersed in the sensory world, it is easy to take this capacity for granted. Yet our sensory systems are nothing short of remarkable. As we navigate our environment, we effortlessly decompose the changes in sound pressure at the ear into rich auditory scenes and with ease are able to interpret the patterns of light which fall onto the back of the retina as three-dimensional objects embedded in space. Unsurprisingly, the neural computations supporting these abilities are complex and far from completely understood. Indeed, characterizing the mechanisms and means by which our sensory systems operate remains a major goal of systems neuroscience.

In this thesis, taking a modelling approach, I will describe several studies investigating the underlying organizational principles of the visual and auditory systems, with the goal of taking a small step in this direction to better understand their function. In this chapter, I will begin by giving an overview of these two sensory systems – focussed on primates, mice and ferrets as the primary animal models used as comparison experimental datasets in this thesis. I will then describe the modelling approach taken in the remaining chapters in more depth, before outlining existing models which have been

applied to the auditory and visual systems. Finally, I will conclude by briefly summarizing the chapters which make up the remainder of this thesis.

Overview of sensory systems

The visual system

The first stage of perception begins with transduction. In the retina, photoreceptors – rod and cone cells – transduce light into receptor potentials via a biochemical cascade triggered by the absorption of light by the photoreceptor's light-sensitive opsin proteins.[1] From there, visual information is further transformed by the retina across a number of circuits and specialized cell-types. The retina's output cells – retinal ganglion cells (RGCs) – are hugely diverse,[2] but the parvocellular (P-cells), magnocellular (M-cells) and koniocellular (K-cells) RGCs constitute three broad classes which segregate visual information along distinct vision-forming pathways. P-cells are generally insensitive to motion information and predominantly encode fine spatial details and chromatic information, while M-cells are primarily sensitive to luminance changes and motion and are largely achromatic.[3,4] Finally, K-cells constitute a third pathway which also conveys chromatic information, albeit with low spatial resolution.[3,4] The RGCs then project, via the optic nerve, to the lateral geniculate nucleus (LGN) of the thalamus. In primates, the LGN comprises a layered structure with M-type RGCs projecting to layers one and two, P-type RGCs projecting to layers three to six, and K-type RGCs forming interleaved projections between the P- and M-cell layers.[3] Finally, the LGN itself projects to the primary visual cortex (V1).

Earlier models of the primate visual system suggested that this pronounced separation between the P- , M- and K-pathways continued through V1 into secondary visual cortex (V2),[5–7] though more recent research indicates that there is considerable mixing of these streams across V1 and V2.[5] In particular, early histological work established the existence of macroscopic regions in V1 and V2 according to the staining patterns for various

metabolic enzymes such as cytochrome oxidase (CO).[8] In V1, layer 2/3 of the cortex can be divided into areas known as CO-blobs as well as non-staining so-called interblob regions.[9] Similarly, V2 can also be divided into different anatomical subregions based on the density of CO staining, termed the thick, thin and pale stripes.[9,10] In terms of the P-, M- and K-pathways described earlier, the M- and P-cells of the LGN project to layers 4Cα and 4Cβ of V1 respectively, before projecting to the blob and interblob regions of layer 2/3 and, for the M-pathway, to layer 4B in addition.[11] Conversely, the K-pathway is less promiscuous and primarily projects to the layer 2/3 blobs as well as layer 1.[12] Initial evidence suggested a simple correspondence between V1 and V2, where layer 4B V1 neurons projected to V2 thick stripes, whereas blob and interblob inputs projected to the thin and pale stripes of V2.[13] However, more recent evidence suggests at least four distinct pathways, with different relative contributions of the P-, M- and K-pathways.[5,10] Thus, at the level of V1 and V2, there is extensive integration of the different retinal and thalamic visual streams.

Beyond V2, there is more conclusive anatomical and functional evidence for a re-emergence of two distinct pathways.[5,14] Specifically, the dual-stream model proposes two parallel pathways: a ventral pathway sensitive to form and colour information and a dorsal pathway more sensitive to motion information.[14] In particular, chapter three of this thesis will focus on the dorsal pathway up to the middle temporal area (area MT) as a locus of motion processing.[15]

In contrast, the mouse visual system, although widely used as a model of vision, differs from primates along a number of dimensions.[16] From the retina to V1, there are homologous processing streams to the P- and M-pathways of the primate visual system.[17] However, compared with primates, the mouse visual system beyond V1 is significantly less hierarchical in structure and better conceptualized as a network of interacting higher visual areas (HVAs).[18] Nevertheless, as for primates, there is increasing evidence for two parallel – albeit interacting – streams in the mouse visual cortex.[19,20] Analogous to the

primate visual system, dorsal stream regions show greater sensitivity to higher temporal frequencies and project to motor areas, while ventral stream regions are relatively more sensitive to higher spatial frequencies and tend to project more to temporal lobe regions.[19] However, the functional properties of the mouse dorsal and ventral streams – as well the precise relationship to its primate counterpart – is still incompletely understood and remains an active area of research.[21–23]

The auditory system

The cochlea represents the first stage of auditory processing, where the incoming sound pressure wave is transduced into a neural signal and decomposed into its constituent frequency components. Different sound frequencies produce deformations along different regions of the cochlea's basilar membrane where they stimulate inner hair cells. Changes in the membrane potential of these inner hair cells in turn modulate the firing rate of spiral ganglion neurons, whose axons comprise the auditory nerve fibres which carry auditory information onwards to the brainstem. In this way, the cochlea provides the basis for the tonotopic representation of sound throughout the rest of the auditory system, where neurons are spatially organized according to their frequency preference.

Unlike the visual system, where a disynaptic connection links the retinal ganglion cells with cortex, the auditory pathway has a significantly more complex arrangement of subcortical nuclei. The auditory nerve first terminates at the cochlear nucleus which itself projects to the nuclei of the lateral lemniscus, superior olivary complex and the inferior colliculus (IC). The IC is an important relay for ascending auditory information to the thalamus and cortex, and is divided into a shell region comprising the dorsal and lateral nuclei and a tonotopically organized core region termed the central nucleus.[24] In addition to subcortical inputs, corticofugal projections to the IC play an important role in shaping receptive field tuning properties[25,26] and modulating auditory behaviour.[27] While the central nucleus primarily projects to the auditory thalamus, the shell of the IC projects to a broader array of regions including the thalamus and superior colliculus, and

has descending projections to the auditory brainstem.[28–30] Finally, the medial geniculate body (MGB) of the thalamus is the last subcortical processing stage along the auditory pathway prior to the cortex. The MGB can be divided into three subregions – namely, the ventral, dorsal and medial MGB. Of these, the ventral MGB forms the lemniscal pathway which projects to the primary auditory cortex and maintains a tonotopic representation. In contrast, the medial and dorsal MGB make up the non-lemniscal pathway and project to non-primary auditory cortex, among other regions.[31]

The auditory cortex is typically divided into a tonotopically-organized core surrounded by a belt of secondary auditory fields.[32,33] For example, in the ferret – the animal model of the auditory pathway used in chapter four of this thesis – the auditory cortex is situated in the ectosylvian gyrus where it is divided into the middle (MEG), posterior (PEG) and anterior (AEG) ectosylvian gyri.[34] The primary cortical fields – the primary auditory cortex (A1) and the anterior auditory field (AAF) – are found in the MEG with neurons organized tonotopically with short latency responses. Of the secondary areas, the posterior pseudosylvian field (PPF) and posterior suprasylvian field (PSF), both located in the PEG, are explored more in the fourth chapter of the thesis, with neurons in these regions showing weaker tonotopy and distinct temporal characteristics compared with A1 and AAF.[34]

Sensory cortex

The precise role of the cortex in sensory processing is somewhat contested. Although the cortex is often characterized as the site of higher reasoning, more recent work has refined this view given an increasing appreciation for the role of subcortical structures in a wide variety of functions including the integration of contextual cues and statistical learning.[35–37] Nevertheless, at a broad level, sensory cortex has a definite role in processing more complex stimulus attributes, abstracted away from their low-level perpetual features.[38]

For example, in the case of the auditory system, neurons in earlier regions of the auditory pathway such as the IC are well approximated by a single filter.[39] In contrast, auditory cortical neurons respond across multiple stimulus dimensions and are best characterized by a model consisting of multiple filters.[39–41] More generally, auditory cortical responses are modulated by cognitive variables such as reward,[42] attention[43,44], learning,[45] and decision making processes,[46,47] and auditory cortex plays a role in representing distinct acoustic features together as coherent auditory objects.[48–50] This trend continues in secondary auditory areas where task- and behavioural-relevant variables are extracted to produce categorical representations of auditory stimuli.[51]

While the visual cortex as a whole plays a similar role in processing increasingly complex stimulus features, inputs to V1 are arguably less pre-processed than to A1 due to the additional subcortical processing stages of the auditory system.[52] Nevertheless, V1 neurons are similarly modulated by contextual and task demands,[53–55] while higher-order regions of the visual cortex become increasingly selective for invariant representations such as of object identity[56,57] and global motion direction.[58]

Modelling the brain

The role of modelling in biology and neuroscience

What is the role of theory and modelling in neuroscience? In physics, modelling forms a core element in the discipline's scientific practice – though, in biology, its role is more uncertain.[59,60] To be sure, certain models – for example, the Hodgkin-Huxley model of action potential generation – have been hugely influential in neuroscience and have had a fruitful and reciprocal relationship with experimental work. However, modelling results are often relegated to post-hoc explanations of experimental data[61] and purely theoretical work can struggle to be perceived a result in its own right.[59]

This perspective is unfortunate because theoretical and modelling work can be enormously useful. Models can provide precise explanations for experimental data, unify otherwise unrelated findings by bringing them into a single framework and make novel predictions.[59,62,63] At its best, modelling has a generative and reciprocal relationship with experimental work, where model predictions can produce new hypotheses which can then be tested and so iteratively refine and expand the model further. Indeed, in light of the ever increasing quantities of data provided by modern neuroscientific recording techniques,[64] there is arguably a greater need than ever to apply a modelling perspective to the often atheoretical world of big data in biology.[65]

Modelling is, of course, not a unitary activity and takes a variety of different forms which can be grouped into descriptive, mechanistic and normative approaches (Figure 1.1).[63,66] Descriptive models – also known as phenomenological models – can be thought of as 'what' models. That is, they provide a surface level description of the behaviour of a system without attempting to describe its underlying properties.[63] A visual neuron's orientation tuning curve parametrized as a Gaussian provides a classic example of a descriptive model. In particular, the model describes the behaviour of the system – the visual neuron – as a function of stimulus orientation but provides no insight into how the tuning curve arises or why a particular form of tuning curve might be used to represent the stimulus.

In contrast, mechanistic modelling takes a bottom-up perspective and seeks to characterize the system in terms of a set of interacting lower-level assumptions.[63] In this way, mechanistic modelling is a reductionist approach where the explanatory power of a model is taken as its ability to describe some phenomenon in terms of its constituent parts. Returning to the example of the visual neuron, a mechanistic model might describe how orientation tuning emerges as a result of the specific arrangement of retinotopic inputs into V1 from the LGN.[67] In this way, mechanistic models may in fact be highly abstract – a model of neural tuning need not rely on a low-level cellular

explanation – though there is an in-built assumption that the constituent parts of one model can themselves be described mechanistically by another low-level set of assumptions in a nested manner.[68]

Finally, normative modelling takes a top-down approach where phenomena are described in terms of the overall function of the system.[63,69] This higher-level approach is often described as an appeal to optimality to provide an answer for 'why' some phenomenon is as it is. For example, considering the visual neuron again, the precise arrangement of orientation tuning curves has been argued to emerge as a result of optimizing for a metabolically efficient representation of the stimulus.[70] Thus, efficient coding is provided as an explanation for 'why' tuning curves exist in the form they do. The invocation of optimality implies that the target system has been optimized in response to some set of constraints. In the case of neuroscience, this is often explained as a result of evolutionary pressures over longer timescales or through learning processes at shorter timescales.[63,71] Indeed, when the observed solution is *not* judged as optimal by some normative criteria, normative modelling provides a powerful approach to investigate the role of different constraints in shaping neural systems.

Taken as a whole, these different approaches can provide complementary perspectives for a single phenomenon given the inherent trade-offs embodied by each kind of model. In particular, highly abstract normative models may struggle to generate precise quantitative predictions. In contrast, highly detailed mechanistic models – though potentially of high predictive value – may 'lose the forest for the trees' and struggle to provide a wider intuitive understanding of the system.[62] Thus, modelling can benefit from a holistic perspective which considers the value of these different approaches.

As normative modelling comprises the primary emphasis of this thesis, I will now focus on one instantiation of normative models in the form of task-optimized neural networks.

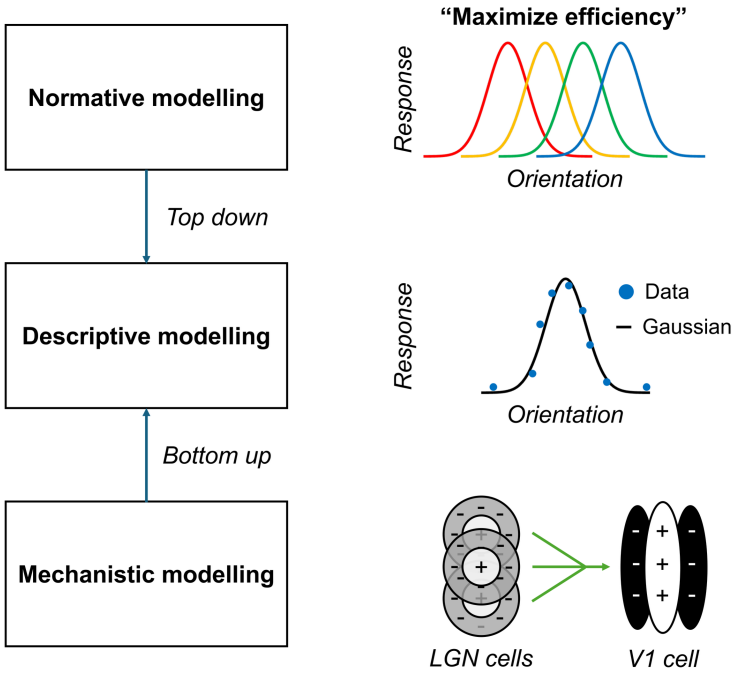


Figure 1.1. Schematic of normative, descriptive and mechanistic modelling exemplified through the orientation tuning properties of a visual neuron.
Modelling approaches in neuroscience can be described in terms of a hierarchy from low-level mechanistic modelling, through to simple descriptive models, up to high-level normative models. As illustrated here through the example of orientation tuning, these approaches can provide complimentary lenses to analyse a given neural system of interest.

Neural networks as normative models

In recent years, artificial neural networks have emerged as a dominant modelling approach in neuroscience[72–74]. At their core, neural networks in their most common instantiation consist of simple repeated units which take in some inputs, apply a set of weightings, summate the result and apply a nonlinear transfer function. In their earliest instantiations, such as in the perceptron,[75] these networks were limited to a single layer and as such were limited in their expressive capacity and incapable of learning non-linearly separable mappings.[76] The computational power and flexibility of neural networks arise when units are stacked across several layers, or recurrently within a layer over time, in the form of deep neural networks.[74]

The advent of deep learning brought about models consisting of many layers, made possible through the use of GPUs to accelerate training.[77,78] This advance made large models computationally feasible and unlocked a key tenet in the power of scale in deep learning.[77] In particular, a general finding from deep learning has been that hand-crafted solutions based on domain expertise are generally outperformed in the long run by general purpose methods – the so-called "bitter lesson".[79] This idea underscores the premise of deep learning where, rather than building an algorithm from first principles, the algorithm is learned implicitly through an iterative trial-and-error approach that maps input to output via the backpropagation algorithm. In this way, deep learning has been enormously successful across many domains of engineering where it has paved the way for effective computer vision,[78] natural language processing,[80] and time series forecasting,[81] among other uses.[82]

From the onset of deep learning, these networks have been rapidly adopted in neuroscience as models of the brain.[73,74] Conceptually, we can decompose their use in neuroscience into two distinct approaches. The first is a non-normative approach which uses deep networks to directly model neural data. In this case, the goal is to build an end-to-end encoding model which maps a stimulus onto the firing rate of individual neurons or populations of neurons.[83–85] The second is a normative approach which treats deep neural networks as task-optimized models of the brain in their own right.[63,72] In this case, networks are analysed as optimizable neural-like systems which can be used to test hypotheses about neural function. While these categories are often blurred in practice – task-optimized networks are frequently used to predict neural activity as a measure of their representational alignment to the brain, I will focus primarily on this second normative approach using task-optimized networks for the rest of this discussion.

These task-optimized networks possess several features which make them attractive as normative models of the brain. Firstly, they provide a means of precisely testing computational theories about neural function. Research in experimental neuroscience

often makes an appeal to the 'computations' which a given neural circuit might support. However, these claims are often justified in an informal manner. Using task-optimized networks, we can invert the question by precisely specifying the computation, optimizing a network for that function, and measuring whether neural-like representations emerge. For example, supporting the proposed role of place cells in path-integration,[86] recurrent networks trained for path-integration can develop place-like cells in their artificial units.[87,88] Secondly, as a computational model, researchers have complete access to probe and manipulate the learned weight structure and hidden activity in these networks. In this way, questions about neural architecture and connectivity can be more directly interrogated compared with the brain. Thus, where experimental neuroscience is often limited in the capacity to record and control neural activity, ablation and loss-of-function experiments can be performed with ease in these models to gain a causal understanding of network function.

Normative modelling principles

With these considerations in mind, I will now outline the normative principle guiding much of the work in this thesis – temporal prediction – and situate it within the wider space of normative principles that have been applied to sensory systems.

Efficient coding and temporal prediction

The principle of temporal prediction argues that sensory systems are optimized to represent those features which are predictive of the immediate sensory future.[71,89,90] This principle forms the core of the first two research chapters of this thesis, where I analyse how recurrent networks optimized for temporal prediction can recapitulate local connectivity within V1 as well as the functional properties of visual neurons along the dorsal visual pathway.

Conceptually, temporal prediction can be understood as an extension of efficient coding, which itself comprises a family of related normative principles which have been applied

to sensory systems (Figure 1.2).[91,92] First proposed by Horace Barlow in 1961, efficient coding proposes that sensory systems are optimized to represent stimuli *efficiently*, in the information-theoretic sense, by recoding highly redundant sensory inputs to increase the entropy of the neural representation.[93] Barlow justifies this principle on two grounds. Firstly, a less redundant neural code better conserves spiking activity to minimize metabolic costs. Secondly, this coding scheme has the effect of minimizing spiking activity for more probable (and hence metabolically expensive) sensory inputs, while maximizing spiking activity for rare or unusual events. Accordingly, an efficient coding scheme can highlight novel events in the environment and in turn signal the need for new learning to guide future behaviour.[93,94] This framework has since proven capable of explaining a wide range of physiological properties across many sensory systems. For example, within the retina, the centre-surround organization of retinal ganglion cell receptive fields has been explained as a means of whitening visual inputs to produce a less redundant neural code.[95] Similarly, within the auditory system, efficient coding of natural sounds in the form of independent component analysis can recapitulate a wide range of auditory nerve fibre tuning properties.[96]

Since Barlow's original paper, the idea of efficient coding has been expanded into a range of related ideas depending on what sensory information is considered relevant and the constraints applied to the coding scheme.[91,97] For example, in the case of sparse coding, redundancy is reduced via a sparsity constraint which minimizes the number of units which are active for any given input.[98] Conversely, in predictive coding, predictions conveyed via feedback connections 'cancel out' bottom-up stimulus driven activity such that only the residual, non-predicted components of the input signal are conveyed to downstream areas.[99] In both cases, despite differences in the implementation, the result is a system optimized for an efficient code.

However, under higher noise conditions, reducing redundancy in the signal can come at the expense of the robustness of the neural coding scheme. That is, the sparse spiking

activity predicted by efficient coding could be 'drowned out' where the signal-to-noise ratio is unfavorable.[91,97] Under the constraint of high-noise, so-called robust coding has been used to explain the distribution of oriented, visual cortical-like receptive fields versus the retinal-like centre-surround organization of receptive fields when moving from an efficient to a robust model of neural coding.[100]

In contrast, where efficient coding seeks to encode *all* incoming sensory information, alternative neural coding schemes have been proposed which preferentially encode only a behaviourally relevant subset.[91,101,102] Returning to the idea of metabolic efficiency, encoding a subset of inputs has the benefit of only expending resources for behaviourally useful information. Indeed, there is evidence that sensory systems are optimized to preferentially encode information based on its behavioural relevance. For example, the optimal stimuli for locust auditory receptors have a greater overlap with those frequency regions common for insect communication than would be expected if they were merely adapted to the statistics of their auditory environment in general.[103] However, since the behavioural relevance of a given stimulus is clearly contingent on the specifics of an organism and its environment, a general principle which prioritizes behavioural relevance must depend on some more fundamental property. Optimizing for information which is predictive of the future provides one approach, since arguably only information which is predictive of the future state of the world can usefully guide future behaviour.[91,101,102] In support of this idea, experimental work has demonstrated that the activity of both retinal and cortical neural populations are informative of their immediate futures.[102,104] Indeed, within the information bottleneck paradigm – which defines an optimal mapping based on the amount of information a system can have about the future given its information about the past[92,105] – these neural populations have been shown to operate close the maximal limits given their input statistics.[102,104]

In one instantiation of this idea, under a limited set of assumptions, optimizing for predictive information is equivalent to slow feature analysis.[106] Slow feature analysis

proposes a coding scheme which is optimized to extract slowly varying features from the input which exhibit greater temporal invariance compared with potentially rapidly changing sensory inputs.[107] Intuitively, those features which vary slowly are likely to also predict the immediate future.[90] Further, when trained on naturalistic spatiotemporal inputs, slow feature analysis produces units which recapitulate many feature of cortical complex cells.[108] However, as a key limitation, not all predictive features are slowly varying.[71]

Accordingly, temporal prediction provides a further generalization of this approach, optimizing for temporally predictive features in general, compared with the subset of slowly varying features. Beyond providing a behaviourally useful filter, optimizing to encode predictive features in the input can compensate for the transduction and conduction delays inherent to sensory systems. In the same vein, a predictive approach could facilitate complex action where motor output depends on the capacity to predict the future state of the world.[105,109] Finally, temporal prediction could facilitate the extraction of underlying variables by encouraging a general world model to be learned.[105,110] Temporal prediction thus inherits the basic ideas from the efficient coding paradigm while incorporating the wider benefits of a temporally predictive approach.

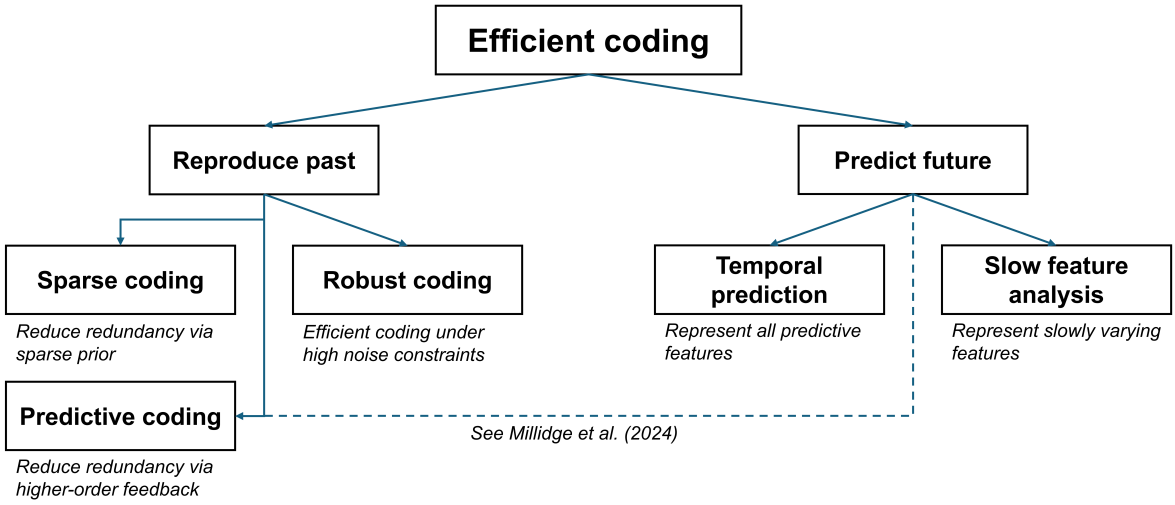


Figure 1.2. Typology of different modelling approaches within efficient coding.
Efficient coding has given rise to several different approaches based on optimizing for different kinds of 'efficient' representations. Note that, while predictive coding has often been conceptualized as reproducing past information, other approaches have also implemented

predictive coding networks optimized for future prediction (see Millidge et al.[110]).

Other normative principles

While influential, efficient coding-type models are only one among many which have been employed to explain neural function.[72] In contrast to the principled justifications for efficient coding, many normative principles are decidedly atheoretical and largely co-opted from computer science where they were initially developed without a role in neuroscience in mind. For example, object-recognition, speech-recognition and large language models have been remarkably successful at modelling different regions of the sensory cortex despite their origins outside of computational neuroscience.[111,112] While a detailed exposition of all these models is beyond the scope of this discussion, these approaches can be broadly categorized into unsupervised (or self-supervised) and supervised machine learning.[77]

Unsupervised learning refers to a broad set of techniques which learn from intrinsic patterns in data without reference to labelled target examples. From an engineering perspective, unsupervised algorithms have the advantage that they can make use of much larger datasets as they are unconstrained by the need for time-consuming and expensive hand-labelled training examples. From a neuroscience perspective, unsupervised principles are easier to integrate into a general theory of the brain as the absence of labelled examples makes them more biologically plausible than their supervised counterparts. Although far from exhaustive, autoassociative and contrastive learning are two forms of unsupervised learning which have been widely explored in sensory neuroscience. Autoassociative networks – typically realized in the form of an autoencoder – are optimized to reconstruct their inputs under some set of constraints which encourage the network to learn useful representations. Autoencoders have been widely employed as models of early regions of sensory cortex,[89,113,114] and – when combined with a sparse penalty – can be recast as a kind of efficient coding model akin to sparse coding.[98] Conversely, contrastive networks are trained to produce invariant

representations by learning to bring similar stimuli closer together and dissimilar stimuli further apart in the network's latent space.[115] Contrastive approaches have proven particularly influential in engineering contexts – for example, the speech recognition system *wav2vec*[116] Here again, the capacity to learn useful representations over large datasets via unsupervised learning can be leveraged to reduce the quantities of labelled data needed for downstream fine tuning.[115]

Finally, supervised learning comprises those methods which depend on labelled targets – and therefore constitutes a broad set of machine learning objectives and algorithms. In general, it is hard to reconcile this need for labelled data in supervised techniques with the kinds of learning that underpin neural function. In the context of object recognition, for example, animals do not have access to labelled targets in any sense like those used to train these models.[72,74] Accordingly, supervised approaches are less amenable as a scaffold on which to build a general framework to describe neural function. However, to reject them out of hand on these grounds would be premature, and supervised approaches have been usefully applied to answer many specific normative hypotheses about the brain.[72] For example, deep neural networks trained to localize sounds are capable of replicating the characteristics of human auditory behaviour only when trained on a dataset that matches the statistics of the human auditory environment.[117] In this way, supervised approaches can reveal some of the constraints which shape sensory behaviour and neural representations.

Summary

Temporal prediction sits within an extensive body of literature exploring how different normative principles can be used to explain neural function. As a result, it can be challenging at times to make sense of how existing studies relate to each other and what wider conclusions can be drawn from these. That many different normative principles can produce brain-like representations does not necessarily imply these results are in conflict. Often, different principles are manifestations of a common idea – as emphasized

by the discussion relating different principles to efficient coding – and therefore may tap into a common latent factor. Equally, differing constraints across the brain or across cell types may have resulted in neural populations that are optimized for different principles.[91]

Nevertheless, it is possible to give too harmonious account, and an important element of model building is falsification: producing models which can better explain a given phenomenon to improve our understanding of the brain. Throughout this thesis, I have tried to balance these concerns by advancing an argument for temporal prediction as a unifying principle that can explain many elements of sensory systems, while situating temporal prediction within a wider framework to better understand how different normative models might relate to each other.

Thesis overview

Chapter two: Prediction of future input explains lateral connectivity in primary visual cortex

Neurons in V1 show a remarkable functional specificity in their pre- and postsynaptic partners. Recent work has revealed a variety of wiring biases describing how the short- and long-range connections of V1 neurons relate to their tuning properties. However, it is less clear whether these connectivity rules are based on some underlying principle of cortical organization. Here, we show that the functional specificity of V1 connections emerges naturally in a recurrent neural network optimized to predict upcoming sensory inputs for natural visual stimuli. This temporal prediction model reproduces the complex relationships between the connectivity of V1 neurons and their orientation and direction preferences, the tendency of highly connected neurons to respond more similarly to natural movies, and differences in the functional connectivity of excitatory and inhibitory V1 populations. Together, these findings provide a principled explanation for the functional and anatomical properties of early sensory cortex.

Chapter three: Hierarchical recurrent temporal prediction as a model of the mammalian dorsal visual pathway

A major goal of neuroscience is to identify whether there are generalized principles that can explain the diverse structures and functions of the brain. The principle of temporal prediction provides one approach, arguing that the sensory brain is optimized to represent stimulus features that efficiently predict the immediate future. Previous work has demonstrated that feedforward hierarchical temporal prediction models can capture the tuning properties of neurons along the visual pathway, and that recurrent temporal prediction models can explain local functional connectivity within primary visual cortex. However, the visual system is also characterized by extensive inter-areal feedback recurrency, which these existing models lack. We aimed to better account for the dynamic features of neurons in the visual cortex by incorporating both local recurrency and inter-areal feedback connectivity into a hierarchical temporal prediction model. The resulting model captured tuning for pattern motion, surround suppression and elements of inter-areal functional connectivity in visual cortex. Moreover, compared with several alternative normative models, the hierarchical recurrent temporal prediction model provided the closest fit to these tuning properties and was best able to explain neuronal response properties across the visual cortex. Accordingly, temporal prediction may account for the emergence of information processing along the visual pathway.

Chapter four: Exploring the impacts of modelling choices on predicting neural responses across the auditory pathway

Why do neurons along the auditory pathway show the particular response properties that they do? Normative modelling offers one approach by providing an optimization perspective to answer principled 'why' questions about neural function. The gold standard for comparing these normative models is in their capacity to explain neural responses to natural stimuli. However, existing work often compares models without systematically controlling for different modelling choices, making it difficult to determine how the model's architecture, training data and normative objective influence

neural prediction performance. Here, we systematically varied different normative models to explore which factors impacted the models' capacity to predict single-unit responses to natural sounds across the ferret auditory pathway, from the midbrain to higher auditory cortex. We found that the network architecture and normative objective had the largest impacts on neural prediction performance compared with the natural sound datasets used to train the models. Moreover, we observed that models which learned more biologically realistic receptive fields also tended to better predict neural responses. Most notably, models which were better able to generalize to novel tasks were also better at predicting neural responses. These results highlight the impact of different modelling choices in the context of normative modelling and suggest that neural representations along the auditory pathway may be governed by normative principles that promote this capacity for task generalization.

Chapter five: General discussion and conclusions

An overview of the thesis where I discuss the broader picture of the results chapters and how they sit within the existing literature, including limitations and potential future directions to extend this thesis.

Thesis declaration

Chapter two has now been published at Current Biology (https://doi.org/10.1016/j.cub.2024.11.073). This work was completed in collaboration with several co-authors. In particular, the project was initiated by a previous student (Emil Fristed). However, I further developed the project substantially in terms of the project's overall direction and in terms of new analyses, and the project which was rewritten from scratch such that all code, analyses, figures, and text which make up the chapter are my own.

Chapter three was conceptually based on an earlier project I began with my current supervisor during my MSc in Neuroscience. This current chapter describes a novel

model of the visual system based on the idea of temporal prediction. During my MSc, I began a limited exploration based on this approach, which was later built upon and significantly expanded into the current DPhil chapter. All current figures are entirely original to the DPhil thesis, all analyses and models have been run and trained from scratch, and the work has been significantly extended and built upon.

2

Prediction of future input explains lateral connectivity in primary visual cortex

Introduction

An increasing number of studies – mostly focusing on mouse primary visual cortex (V1) – have begun to uncover the underlying rules specifying how cortical neurons connect.[118–121] Some findings, such as the tendency of V1 neurons to synapse with other neurons that show similar orientation selectivity, follow a simple like-for-like pattern.[119-121] In contrast, other results such as the spatial organization of synaptic inputs to orientation- and direction-tuned visual neurons, appear more complex and less amenable to a unifying theoretical explanation.[118,122] An outstanding question, then, is how to understand these putative connectivity rules and whether they can be explained by a single general principle.

By taking a normative approach, we can ask whether the patterns of structure and function observed at the level of individual neurons are optimized for achieving particular goals from a behavioural or evolutionary perspective. One such promising normative framework is that of temporal prediction, which posits that sensory systems are optimized to represent those features in the recent past which are predictive of the immediate future sensory inputs.[71,90] Why might temporal prediction serve as a useful objective for an organism's sensory systems? Firstly, predictive processing is valuable in

compensating for neural conduction and processing delays to facilitate complex actions which depend on estimating the future state of the world.[109] Secondly, extracting only predictive features acts as a behaviourally useful filter to reduce the vast amount of information that sensory systems would otherwise need to process. Finally, temporal prediction requires no explicit teaching signal beyond the sensory input itself, making it more biologically plausible as an unsupervised principle than other supervised counterparts.[123,124]

When optimized for temporal prediction, feedforward networks have been shown to capture many of the receptive field characteristics and response properties of V1 neurons, as well as motion processing along the visual pathway.[71,90] Nevertheless, existing feedforward temporal prediction models neglect the role of recurrency, which experimental and theoretical studies have implicated in a range of key brain functions.[12-127]

Here, we show that a recurrent network optimized for temporal prediction on dynamic natural visual scenes can capture many motifs of local connectivity in visual cortex. Furthermore, when we compared network models optimized for different normative objectives, temporal prediction stood out in its capacity to explain these connectivity motifs. Hence, both the relationship between the connectivity patterns of V1 neurons and their response characteristics appears to be optimized to support the predictive processing of dynamic stimuli.

Results

Model response properties

We trained a recurrent network model to predict the upcoming visual input (40 ms ahead) based on the recent stimulus history (Figure 2.1A). The model was trained on a diverse dataset of movies of natural scenes, including wildlife and panning over natural

environments, which were bandpass-filtered to approximate the filtering properties of the retina.[98] After training, we first compared the model's hidden unit response properties to those of neurons in mouse V1.

We estimated the model units' receptive fields by means of the response-weighted average (Figure 2.1B). Like simple cells in V1 (Figure 2.1C), model units generally had well-defined receptive fields with a Gabor-like structure consisting of oriented excitatory and inhibitory subfields. To probe the tuning properties of the model units, we recorded the model's responses to oriented full-field drifting gratings. Model units were generally orientation tuned (24%) or direction tuned (57%; Figures 2.1D, E), with a similar distribution of orientation selectivity indices (OSI) and direction selectivity indices (DSI) as found in mouse V1 (Figure S2.1). As in visual cortex, model units varied in their phase responsiveness, characterized by their modulation ratio (Figure 2.1F). Highly modulated units were classed as simple-cell-like, while units that displayed little or no phase modulation were classed as complex-cell-like.[128,129] At the population level, there was a bimodal distribution similar to that found in mouse V1,[130,131] indicating the existence of distinct populations of simple-cell-like and complex-cell-like model units (Figure 2.1F). Both complex and simple cells thus emerged naturally in the temporal prediction network. Notably, temporal contiguities across naturalistic visual stimuli – albeit in a feedforward network – has previously been shown to be crucial for the development of complex cells, demonstrating the important role of temporal information in developing these complex-like invariant representations.

We further analysed the circuit-level basis for the simple-cell-like and complex-cell-like behaviour of model units. In support of previous work, we found that simple-cell-like units had a significantly greater ratio of feedforward to recurrent inputs (Figure S2.2A; $t(685)=-18.3$, $p<.001$), which increased as a function of the unit's modulation ratio (Figure S2.2B). Thus, complex-cell-like responses were supported by greater recurrent

connectivity which could maintain activity over time to produce a more invariant response to the moving gratings.

Some neurons in V1 are thought to encode prediction errors that represent the difference between the brain's internally generated predictions and the true sensory data.[132,133] While the model hidden units do not explicitly represent prediction errors by default, they could emerge spontaneously during the inference process in response to 'unexpected' stimuli that violate a recurring pattern (Figure S2.3A). We found a small number of units that responded selectively to these unexpected stimuli (omitted stimuli: 0.2% of units, deviant stimuli: 5.7% of units), while a slightly larger number of model units showed mixed-sensitivity and responded to both standard and unexpected stimuli (omitted stimuli: 0.5% of units, deviant stimuli: 9% of units). Prediction-error responses have been reported in some neurons in mouse V1,[133–135] and to varying degrees in other primary sensory areas.[136–140] Moreover, this mixed-sensitivity aligns with the biology where neurons rarely exist in a strictly error-coding capacity.[135,141]

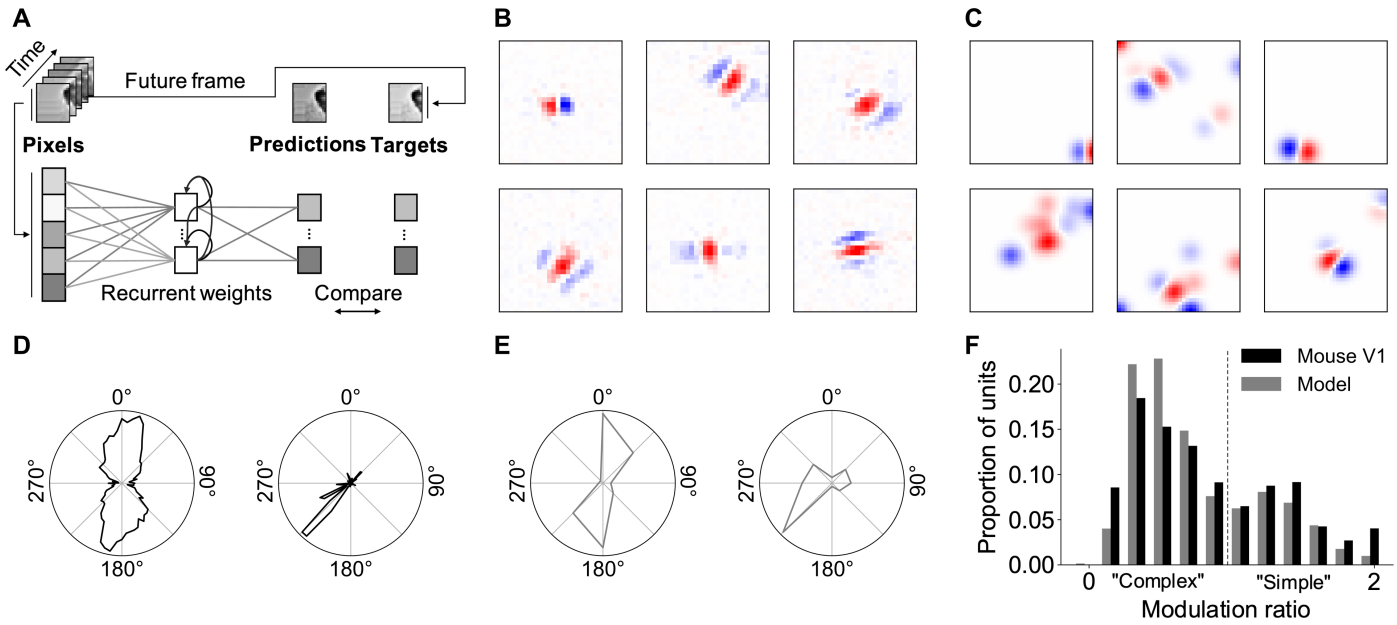


Figure 2.1. The recurrent temporal prediction model captures basic tuning properties of V1 neurons.

(A) Schematic of the recurrent temporal model. 2,332 hidden units (90%) were excitatory with non-negative outgoing recurrent weights, the remaining 260 hidden units (10%) were inhibitory with non-positive outgoing recurrent weights.

(B) Response-weighted-average receptive field estimates of model units (units in pixels).

(C) Mouse V1 receptive fields from publicly available recordings,[142] pre-processed for visualization by thresholding and smoothing with a Gaussian filter.
(D) Example tuning curves for orientation (left) and direction (right) tuned model units. Response is the normalized hidden unit activity as a function of the stimulus value.
(E) Example tuning curves for orientation and direction tuned cells in V1.[142] Response is the normalized firing rate as a function of the stimulus value.
(F) Distribution of modulation values across model and pooled excitatory and inhibitory mouse V1 units.[131] Typically, a modulation ratio less than one is taken as a complex cell, while a modulation ratio greater than one is taken as a simple cell.
See also Figures S2.1-3.

Short- and long-range functional connectivity

Short-range functional connectivity between excitatory units in the model resembled that of mouse V1 neurons (Figure 2.2A).[118,120] Model units were more likely to synapse with other units with the same orientation preference (Figure 2.2B), with the connection probability monotonically decreasing as the difference in orientation preference increased (Cochran-Armitage, $p<.001$). Likewise, connectivity between direction-tuned excitatory model units resembled that of mouse V1 neurons.[120] Model units were most likely to connect when tuned to either the same or opposite direction of motion (Cochran-Armitage, both $p<.001$), with connection probability decreasing as the presynaptic unit's tuning became more orthogonal to the postsynaptic unit's preferred direction (Figure 2.2C).

For inhibitory model units, co-tuning with the post-synaptic unit was much weaker (Figure 2.2D), as has been previously reported in V1.[122,143,144] Neither inhibitory-to-excitatory (Cochran-Armitage, $p=.102$) nor inhibitory-to-inhibitory (Cochran-Armitage, $p=.606$) model unit connections showed a significant linear dependence of connection probability on the difference in orientation preference. For direction-tuned inhibitory-to-excitatory model units, the model predicts a weak but significant monotonic trend of increasing connection probability as the difference in preferred direction increases (Figure S2.4; Cochran-Armitage, $p=.032$), distinct from the u-shaped trend observed for excitatory model units and excitatory V1 neurons. No significant trend was found for inhibitory-to-inhibitory direction-tuned model units (Cochran-

Armitage, p=.484). Finally, for orientation- and direction-tuned excitatory-to-inhibitory model units, the model predicts a similar trend as for excitatory-to-excitatory model units and V1 neurons (Cochran-Armitage, p<.001), albeit with the minimum connection probability shifted to the 135° bin for direction-tuned model units (Figure S2.4).

We further investigated how connectivity differs among units with simple-cell-like versus complex-cell-like responses (Figures S2.3E, F). Overall, we found few differences between these two populations. For the short-range connectivity motifs investigated (Figures S2.3E, F), the connectivity pattern was qualitatively similar to the population distributions described above (Figures 2.2B, C), though the aggregate connection probability was higher for simple-cell-like compared with complex-cell-like units.

Across long-range connections, the model also replicated the dependence of connectivity on neuronal orientation preference and receptive field location found in visual cortex.[118] We measured the connection probability between pre- and postsynaptic model units as a function of their difference in preferred orientation and the presynaptic unit's receptive field location in visual space relative to that of the post-synaptic unit (Figure 2.2E). As for mouse V1, model units were more likely to project to the post-synaptic unit if their receptive field aligned along the axis of the post-synaptic unit's receptive field. To quantify this effect, we divided visual space relative to the postsynaptic unit into four quadrants. Those quadrants that aligned with the postsynaptic unit's orientation tuning were referred to as 'co-axial' space (green regions in Figure 2.2A), while those quadrants orthogonal to the unit's preferred orientation were referred to as 'co-orthogonal' space (pink regions in Figure 2.2A). As with the biology, orientation-tuned model units were more likely to synapse with other units when they had similar orientation selectivity and their receptive fields were located in co-axial visual space (Figure 2.2F; permutation test, p<.001). Although a similar effect was found for co-orthogonal units (Figure 2.2G; permutation test, p<.001), this relationship was much weaker. In particular, there was a significantly higher proportion of model units in the 0° orientation difference bin and a

significantly lower proportion in the 90° bin for receptive fields in co-axial versus co-orthogonal space (permutation test, both $p<.001$).

As in V1,[119,121] model units whose responses to natural movies were highly correlated were also more likely to be connected. In both cases, the distribution of correlation values between pairs of connected model units was skewed, with the majority of pairs of units showing a relatively low response correlation, and a smaller proportion that were highly correlated (Figure 2.2H). A qualitatively similar relationship was seen for both model units and V1 neurons between the response correlation and both the connection probability (Figure 2.2I; Cochran-Armitage, $p<.001$) and the average strength of those connections (Figure 2.2J).

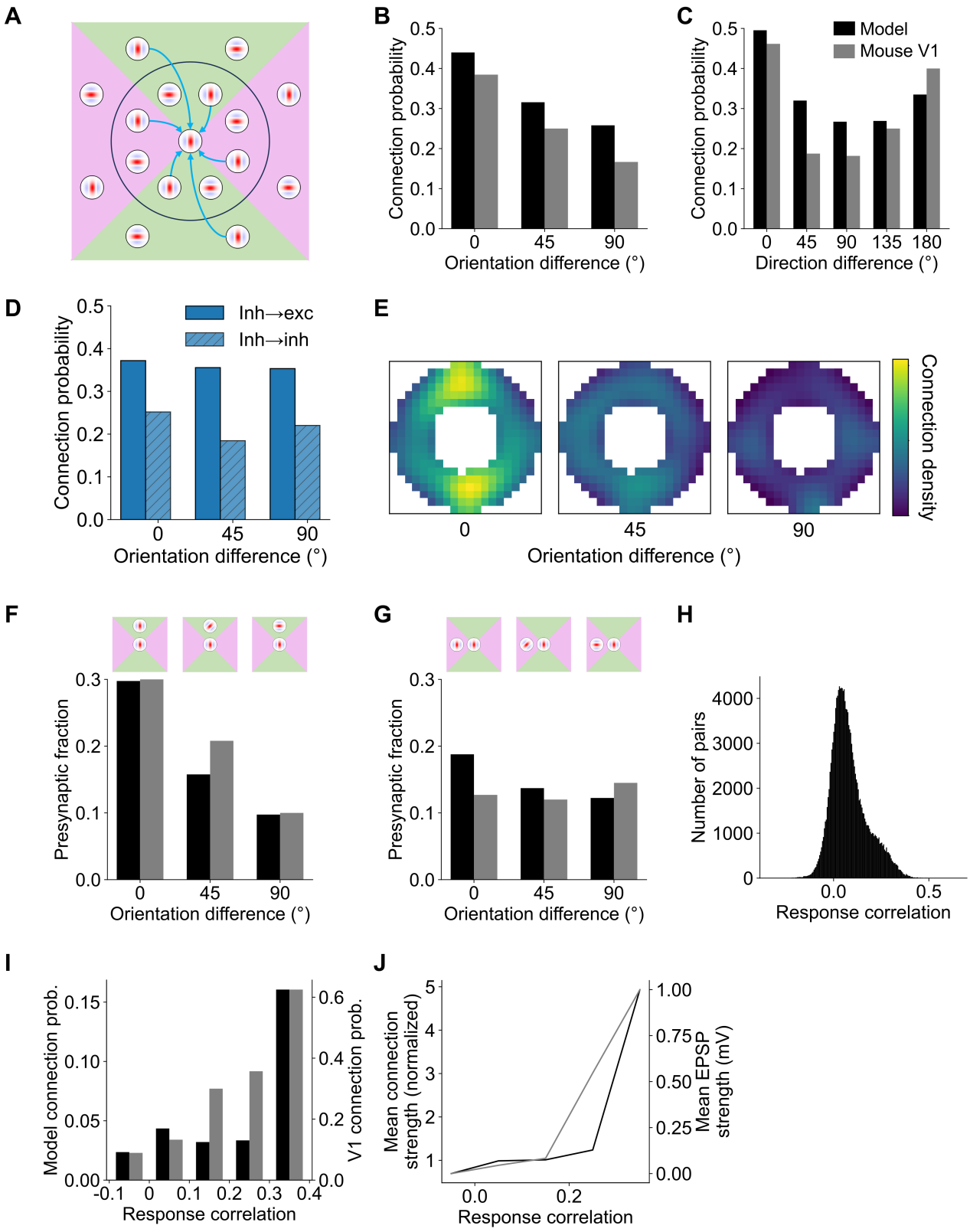


Figure 2.2. The model captures short- and long-range functional connectivity in V1.
(A) Schematic of short- and long-range connectivity in V1. Short-range connections are more prevalent for similarly tuned V1 neurons, whereas the probability of long-range connections is

greater for V1 neurons with similar orientation tuning when their receptive fields are located in co-axial space.
(B, C) Short-range connections are more prevalent when excitatory model units have similar orientation tuning (B) and for direction-tuned units that have similar or opposite preferred directions of motion (C), as is also the case in V1.[120]
(D) As in B, but for inhibitory-to-inhibitory and inhibitory-to-excitatory connections in the model.
(E-G) In both the model and V1,[118] long-range connection probability is higher for presynaptic model units with similar orientation preferences when their receptive fields are located in co-axial (F) than in co-orthogonal (G) locations relative to the receptive field of the post-synaptic unit. For B, C, F and G, data are binned using the same convention as in Ko et al.[120] with orientation bins of 0-22.5°, 22.5-67.5° and 67.5-90° and motion direction bins of 0-22.5°, 22.5-67.5°, etc. Heatmap (E) shows the normalized connection probability over visual space across differences in orientation tuning for model units. Heatmap is smoothed for display purposes with a Gaussian filter (standard deviation, σ=2 pixels).
(H) Histogram of the response correlation distribution across pairs of connected model units for natural stimuli. The distribution is right skewed, indicating that a minority of model units have highly correlated responses.
(I, J) As for mouse V1,[119] response correlation for model units co-varies with the connection probability (I) as well as the input connection strength (J).
These results were abolished after randomly shuffling the recurrent weights between the units when measuring connectivity, resulting in uniform distributions (Figure S2.5). Accordingly, the model connectivity biases cannot be explained by the underlying distribution of orientation and direction tuning preferences among model units.
See also Figures S2.3-4.

Excitatory-inhibitory functional connectivity of direction-tuned units

As for orientation-tuned cells, synaptic inputs to direction-tuned cells in V1[122] are not uniformly distributed in visual space (Figure 2.3A). In particular, direction-tuned excitatory cells preferentially receive inputs from other excitatory cells whose receptive fields are situated in the opposite sector of visual space (i.e., behind) the postsynaptic cell's preferred direction. In contrast, the opposite effect is observed for inhibitory cells, which preferentially synapse with excitatory cells if the location of their receptive fields is ahead of the post-synaptic cell's preferred direction of motion. This connectivity motif provides a plausible circuit basis for direction selectivity, whereby a spatial offset combined with a conductance delay for inhibitory cells could facilitate the detection of moving stimuli.[122]

In line with this evidence from V1, excitatory presynaptic ensembles in the model were more numerous in the opposite sector relative to the postsynaptic unit's preferred motion direction (Figures 2.3B-E; t-test, $t(113)=2.93$, $p=.004$). Conversely, the opposite

pattern was found for the inhibitory model units, whose connection probability with direction-tuned excitatory units was higher if their receptive fields were located in the sector of visual space ahead of the preferred direction of motion (Figures 2.3B-E; t-test, $t(113)=-2.80$, $p=.006$). This effect was specific to direction-tuned excitatory model units, with no significant difference in the spatial locations of excitatory and inhibitory presynaptic ensembles synapsing with weakly direction-selective units, as defined by a DSI<0.8 (see Methods; Figure 2.3E; t-test, $t(113)=225$, $p=.337$).

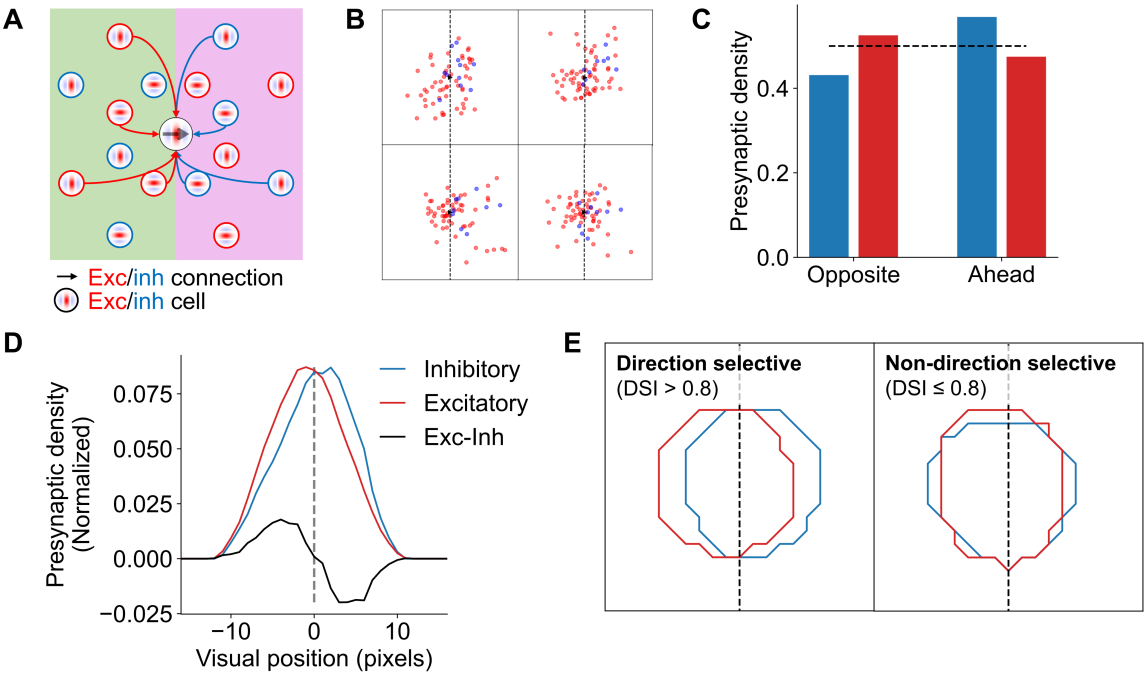


Figure 2.3. The model captures direction-dependent differences in functional connectivity between excitatory and inhibitory populations in V1.

(A) Schematic of connectivity biases in excitatory and inhibitory inputs to direction-tuned cells in V1.[122]

(B) Exemplar excitatory and inhibitory presynaptic ensembles for direction-tuned excitatory model units.

(C) Model unit presynaptic density for excitatory and inhibitory subpopulations in the sectors of visual space opposite to and ahead of the postsynaptic unit's preferred direction of motion. Dashed line represents equal density (0.5).

(D) Profile of model presynaptic unit location density across horizontal visual space for excitatory and inhibitory inputs. Profiles smoothed with a 5-pixel moving average.

(E) Pooled location density contours over visual space across all excitatory (red) and inhibitory (blue) model units, for direction and non-direction selective post-synaptic excitatory units. These results were abolished after randomly shuffling the recurrent weights between the units when measuring connectivity, with no difference in density across either sector of visual space for excitatory or inhibitory model units. See also Figure S2.5.

We further analysed whether the spatial displacement in unit receptive fields across excitatory and inhibitory subpopulations would be accompanied by a corresponding temporal displacement, as found in mouse V1. At the level of individual units, (Figures 2.4Ai, Bi), we found that the average activity of the excitatory presynaptic ensemble rose faster than that of the inhibitory presynaptic ensemble when a moving bar was presented in the preferred direction of the postsynaptic unit. In contrast, the opposite effect occurred when the bar was presented in the non-preferred direction, where there was a faster and larger rise in the average inhibitory presynaptic ensemble's activity compared to that of the excitatory ensemble. Thus, the spatial displacement (Figures 2.4Aii, Bii) was matched in the time course of presynaptic activity. These results were similarly replicated across the population of postsynaptic units on average (Figure 2.4C), which qualitatively matched the behaviour of excitatory and inhibitory neurons in mouse V1 (Figure 2.4D). However, as model units were all alike excepting the sign of their projections, the time course of these dynamics occurred over a slower timescale in the model compared to mouse V1. That is, in the model, the temporal displacement in presynaptic activity across excitatory and inhibitory ensembles was realised by recurrent network dynamics alone, whereas in the mouse V1 units it likely resulted in part from

the intrinsic properties of excitatory and inhibitory neurons.

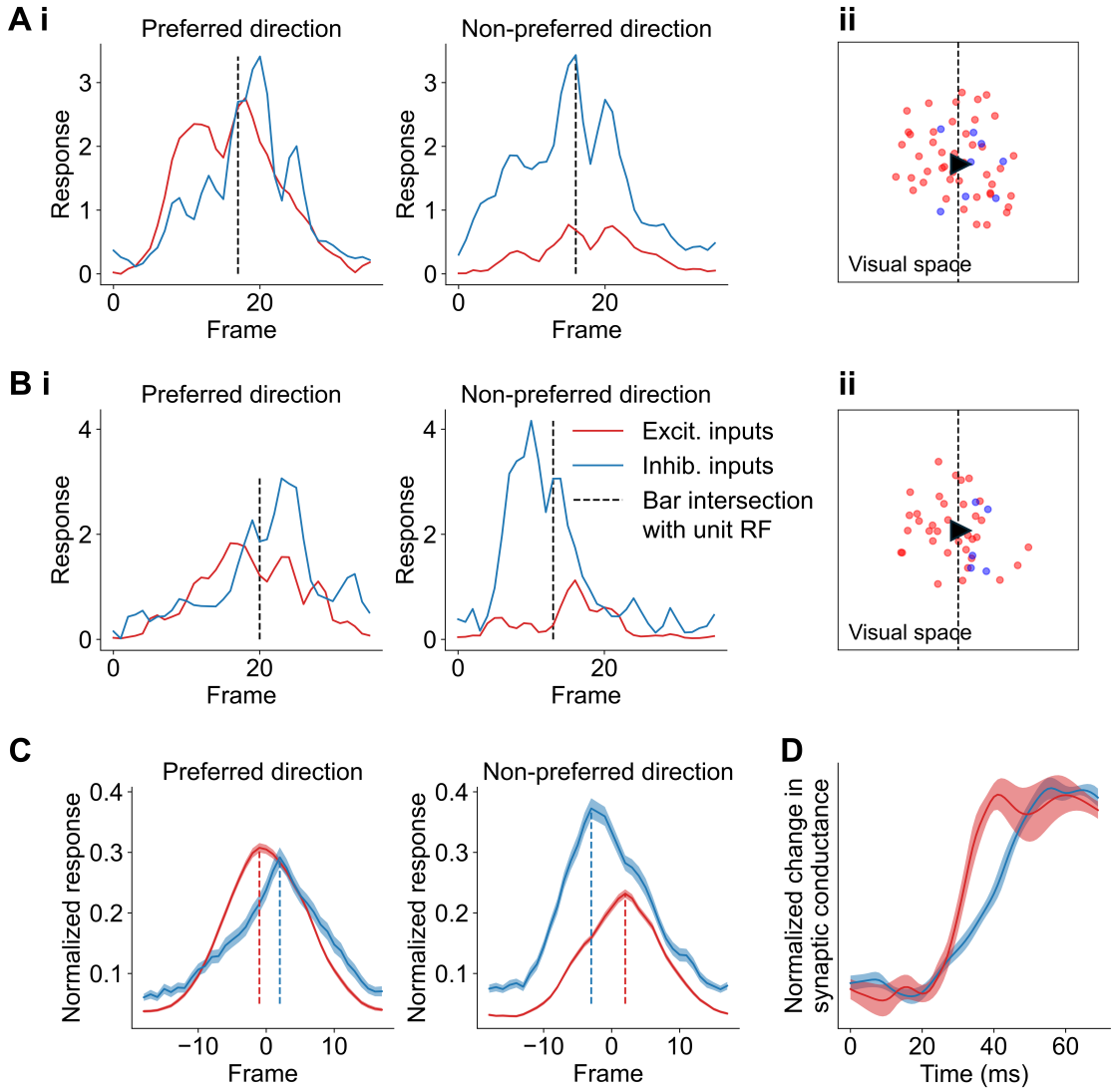


Figure 2.4. The spatial offset in inhibition is matched by a corresponding temporal delay inhibition.

(A, B) Mean activity of the inhibitory and excitatory presynaptic ensembles for exemplar units in response to a bar moving in the preferred or non-preferred direction of the postsynaptic unit (i). Black dashed line indicates the timestep where the moving bar maximally overlapped with the postsynaptic unit's receptive field. The location in visual space of each presynaptic unit's receptive field centres for each exemplar unit, normalized such that the preferred direction points rightwards (ii).

(C) Mean activity of the inhibitory and excitatory presynaptic ensembles in response to a bar moving in the preferred or non-preferred direction of the postsynaptic unit, averaged across all postsynaptic units. The x-axis is relative to the timestep of maximal overlap between the moving bar and the postsynaptic unit's receptive field. The dashed lines denote the maximal value of the average excitatory and inhibitory response, indicating the temporal displacement in response.

(D) Normalized change in synaptic currents in response to a bar moving the preferred direction of the postsynaptic for mouse V1 neurons. Data reproduced from Rossi et al.[122]

Comparing V1 response prediction and connectivity across models

To assess how well temporal prediction performed as a normative model of mouse V1, we examined how well it captured the properties of V1 neurons compared to several other commonly used models (see Methods). Firstly, to directly compare each model's learned representations to V1, we predicted the responses of awake mouse V1 neurons from each model's hidden unit activity for two natural movie clips[142] (Figure 2.5A). Secondly, we quantitatively compared model unit properties and connectivity biases to those found in V1. Overall, the recurrent temporal prediction model predicted the recorded neural responses well (mean CC_{norm}=0.242) compared with the other models tested (Figure 2.5B, C), and best accounted for V1 connectivity motifs (Figure 2.5D).

Firstly, considering the neural prediction performance measure, the linear-nonlinear baseline model consisted of the same fitting procedure as the other models surveyed, but was applied directly to the input stimuli, and performed significantly worse than the temporal prediction model in predicting V1 responses (CC_{norm}=0.189, paired t-test, $t(738)=-6.03$, $p<.001$).

VGG-19 and PredNet are both published models that capture elements of V1 responses. VGG-19 is a deep feedforward network trained for object recognition,[145] while PredNet implements a form of predictive coding.[146] PredNet was similarly trained for next-frame prediction but, unlike temporal prediction, explicitly uses prediction errors during the inference process.[99] The temporal prediction model also exceeded the performance of PredNet (mean CC_{norm}=0.113, paired t-test, $t(738)=-8.97$, $p<.001$), but performed significantly worse than VGG-19 (mean CC_{norm}=0.296; paired t-test, $t(738)=5.40$, $p<.001$). However, as a supervised model, VGG-19 is fitted using human annotated image labels, providing it with constraints not available to the other unsupervised models.

We also considered two variants of the temporal prediction model, comparing an untrained network with random weights and a feedforward temporal prediction model.

In both cases, there was a significant benefit from both training (untrained mean CC_{norm}=0.175, paired t-test, t(738)=-7.03, p<.001) and the addition of recurrency (feedforward mean CC_{norm}=0.184, paired t-test, t(738)=-5.74, p<.001), implying the importance of both the network's training and architecture in modelling V1 neural responses.

Finally, we considered three models based on the same recurrent network architecture but trained using alternative learning objectives. The denoising network aimed to recover the original current frame from one with Gaussian noise injected, the inpainting network aimed to predict the complete current frame given inputs with patches blanked out, and the sparse autoencoder aimed to reproduce the current frame while minimizing hidden unit activity. The recurrent temporal prediction model performed significantly better than the inpainting network (mean CC_{norm}=0.187, paired t-test, t(738)=-4.09, p<.001) and the sparse autoencoder network (mean CC_{norm}=0.206, paired t-test, t(738)=-2.31, p=.021), and non-significantly better than the denoising network (mean CC_{norm}=0.222, paired t-test, t(738)=-1.63, p=.110), highlighting the importance of the training objective over model architecture alone.

The distribution of orientation- and direction-selective model units varied substantially among these models (Figure 2.5C). The temporal prediction model most closely reproduced the overall distribution of unit types found in mouse V1,[142] with comparable proportions of orientation-selective (V1=31%, temporal prediction=24%), direction-selective (V1=39%, temporal prediction=57%) and non-selective units (V1=30%, temporal prediction=19%), albeit with an overrepresentation of direction-selective units. In contrast, the denoising and inpainting networks had a clear overrepresentation of orientation-selective units (62% and 97%, respectively) relative to mouse V1. Similarly, far fewer units met the criteria for direction-selectivity across these alternative models, implying that their learned representations had no or only a weak temporal component.

In tandem, we compared how well each model recapitulated the connectivity biases found in mouse V1 (Figure 2.5D). For each model, we calculated a model connectivity score as the average correlation between the model and V1 connectivity distributions (Figures 2.2B, C, F, G, I, J). Overall, the temporal prediction model (mean=0.68) had a much closer correspondence to the connectivity profiles found in mouse V1 compared with the other models (inpainting mean=0.48, denoising mean=0.49, sparse autoencoder=0.25). Notably, we found no significant correlation between each model's connectivity score and its neural prediction performance (CC_{norm}; Pearson's *r*, r=.602, p=.398). Thus, the capacity of a model to predict neural responses in V1 does not imply that it can accurately capture the underlying organization of cortical connectivity.

We additionally compared the learned connectivity of the PredNet model where possible (Figure S2.6). Although the model's architecture precluded analysis of connectivity motifs dependent on visual space and excitatory or inhibitory subpopulations, we were able to analyse for the motifs described by Ko et al.[120] and found that these were not well captured by PredNet. However, we were not able to look at lateral connectivity in the VGG model as it lacks recurrency. Hence, due to the limited or lacking lateral

connectivity of the VGG and Prednet models we could not calculate their connectivity scores.

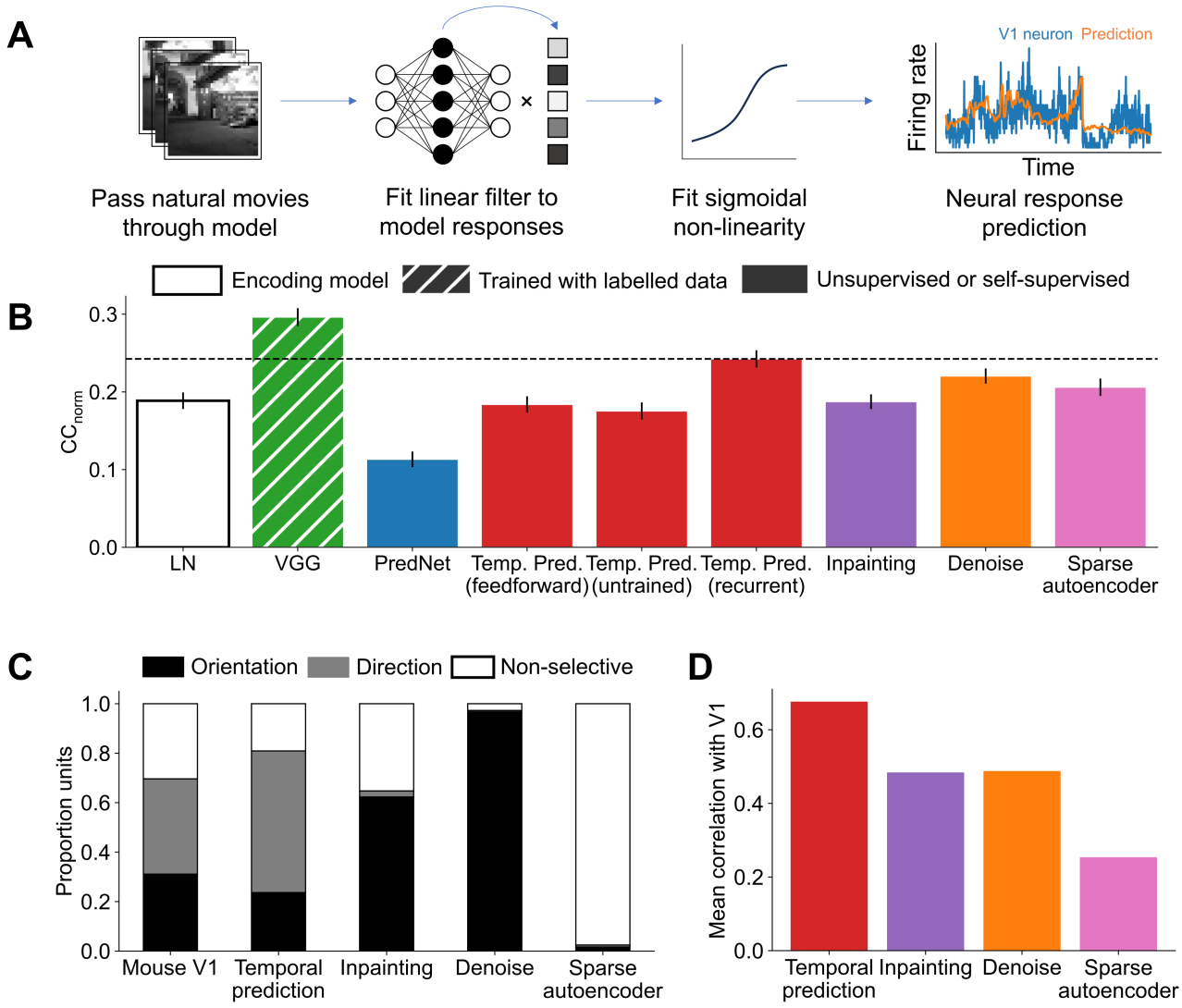


Figure 2.5. Comparison of neural prediction performance and model connectivity to V1 across alternative models.

(A) Schematic of the neural response fitting procedure.

(B) Performance (average CC_{norm}) of the recurrent temporal prediction model (dashed line) relative to other comparison models in predicting neural responses to natural movies. Error bars indicate s.e.m.

(C) Distribution of orientation-, direction- or non-selective units across mouse V1 and each model.

(D) Comparison for each model of the average correlation of the functional connectivity distributions (Figures 2.2B, C, F, G, I, J) with those found in mouse V1. A higher correlation indicates an overall better fit with the V1 data. Note that VGG-19 and PredNet could not be included in the functional connectivity analysis.

See also Figure S2.6.

Variants of the temporal prediction model

We further explored how different modelling parameters impacted the capacity of the temporal prediction model to predict V1 connectivity.

We first investigated how the future prediction offset affected the model's connectivity motifs. Using a high frame-rate (120 Hz) dataset,[147] we produced a continuum of models trained to predict a single frame for 0-10 frames (0-83 ms) into the future. As expected, we found that the aggregate connectivity score increased smoothly as a function of the future offset, to a maximum of 0.62 at 33 ms (Figure 2.6A). Similarly, if we trained the model to predict a span of offsets (i.e., predicting multiple future frames at 0 ms, 0-25 ms, 0-45 ms, etc.), we found that the connectivity score similarly increased with the future offset, with a maximum score of 0.60 when predicting 0-58 ms into the future (Figure 2.6B). Notably, for the multi-frame offset-span model, we found that units that integrated information further back into the past tended to project further into the future (Pearson's r, r=-.15, p<.001; Figure S2.7). Thus, across these networks, as the prediction target was shifted away from an intermediate temporal offset, the connectivity score declined.

Finally, we investigated the role of the model's wiring constraint (L1 regularization). Varying L1 regularization had a marked effect on receptive field structure (Figures 2.6C, D). At the highest regularization strength, receptive fields were reminiscent of the centre-surround organization of retinal receptive fields, while at weaker regularization strengths, receptive fields were only weakly spatially localized. Only at the optimal regularization strength for next-frame prediction did receptive fields resemble the Gabor-like structure of V1 neurons. A similar dependence on regularization strength was found for the aggregate connectivity score (Figure 2.6E) where the connectivity score peaked at the optimal L1 regularization strength, declining across neighbouring strengths. That the most future-predictive hyperparameter setting is also the most brain-like provides further evidence for temporal prediction as a principle of neural sensory processing.

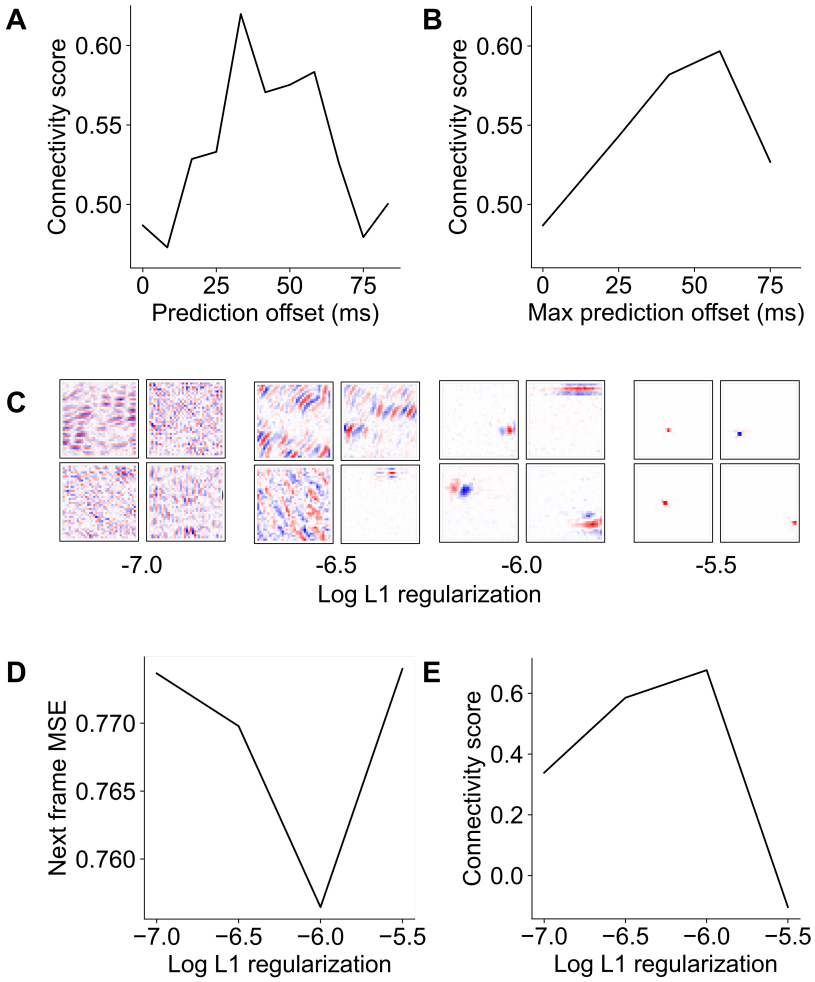


Figure 2.6. Variants of the temporal prediction model.
(A) Connectivity score as a function of the future prediction offset, with a maximum value reached when predicting the next frame at 33 ms into the future.
(B) Connectivity score as a function of increasing future prediction span, with a maximum value when predicting across a span of frames 0-58 ms into the future.
(C, D) Example unit receptive fields (C) and next-frame prediction mean squared error (D) as a function of L1 regularization. The optimal V1-like receptive fields coincide with the L1 regularization value that minimizes the mean squared error in the test set.
(E) Connectivity score as a function of the L1 regularization value, with the maximum again occurring at the optimal L1 regularization value that minimizes the mean squared error.
See also Figure S2.7.

Model connectivity supports temporal prediction

To determine how the observed motifs are related to temporal prediction, we perturbed the network's orientation- and direction-dependent connectivity while measuring the next-frame prediction performance (Figures 2.7A, B). To that end, we ablated (set to the

median weight) an increasing number of connections in the model across different functional classes of units while measuring next frame prediction performance. Given the finding in both V1 and the model that similarly tuned units are more likely to be connected, we hypothesized that ablating connections between units with similar orientation or direction tuning ("co-tuned" units for orientation-selective units; "co/anti-tuned" units for direction-selective units) would result in a larger impairment in prediction performance than for units with a large orientation difference or more orthogonal difference in direction preference ("orthogonal" units).

In support of our hypothesis, as the number of ablated connections increased, there was a relatively greater increase in the error when ablating connections between co-tuned and co/anti-tuned versus orthogonally-tuned units (orientation: t-test, $t(1,998)=290$, $p<.001$; direction: t-test, $t(1,998)=382$, $p<.001$). Thus, ablating the connectivity motifs in the model described by Ko et al.[120] specifically impaired temporal prediction.

From a more mechanistic perspective, we investigated how the learned connectivity motifs might influence the network's internal representations. Considering the finding that units with similar orientation preferences are more likely to be connected, we hypothesized that this connectivity motif might help improve the discriminability of network representations. Within both the visual pathway and deep neural networks, one hypothesised outcome of processing is to transform representations to facilitate linear readout to downstream regions.[148] In the case of next-frame prediction, transforming these representations – for example, to increase their separability – could facilitate model performance by helping to decompose visual inputs and thereby better capture their underlying causes.

To test this hypothesis, we presented oriented grating stimuli with different levels of noise while measuring how distinctly the network represented these stimulus classes in low-dimensional space. We used the silhouette score to measure clustering in the model's hidden activity,[149] and compared these scores between the default and ablated networks

(Figures 2.7C, D). Under low-noise conditions (0 dB SNR), we did not find any difference in the silhouette score between the full and ablated networks (co-tuned: t-test, $t(46)=0.240$, $p=.814$; orthogonal: t-test, $t(46)= -0.186$, $p=.853$). However, under high-noise conditions (-9.5 dB SNR), a higher silhouette score (indicating improved clustering) was found for the full network compared to the co-tuned but not orthogonal ablated networks (co-tuned: t-test, $t(46)=2.94$, $p=.005$; orthogonal: t-test, $t(46)=1.09$, $p=.282$). Thus, under conditions of high noise, connections between similarly orientation-tuned units may help disentangle the internal representations of different stimuli.

Finally, we examined the potential roles of the connectivity motifs described by Iacaruso et al.[118] in temporal prediction. We hypothesized that the tendency of V1 neurons to synapse with neurons of similar orientation tuning in co-axial space – that is, when their receptive fields are aligned in visual space – might facilitate the detection of contiguous moving edges, which should aid the network in next-frame prediction.

Under this hypothesis, we predicted that ablating connections between these model units would impair the next-frame prediction of contiguous-edge-like-features (i.e., a moving bar). Furthermore, this impairment should increase with the bar length, under the assumption that this motif detects elongated features. As a control, we compared the bar stimulus to moving random dot stimuli, equating the total area of both (Figure 2.7E). Since random dot stimuli are non-contiguous, increasing the stimulus area (unlike bar length) should not affect the degree of impairment produced by ablating model connectivity. In line with this hypothesis, we found that stimulus area was correlated with the impact of ablating network connections on mean squared error for the moving bar stimuli (Figure 2.7F; Pearson's r, $r=.778$, $p<.001$) but not random dot stimuli (Figure 2.7G; Pearson's r, $r=.180$, $p=.400$).

Together, these results suggest that the identified connectivity motifs have specific functions in supporting temporal prediction.

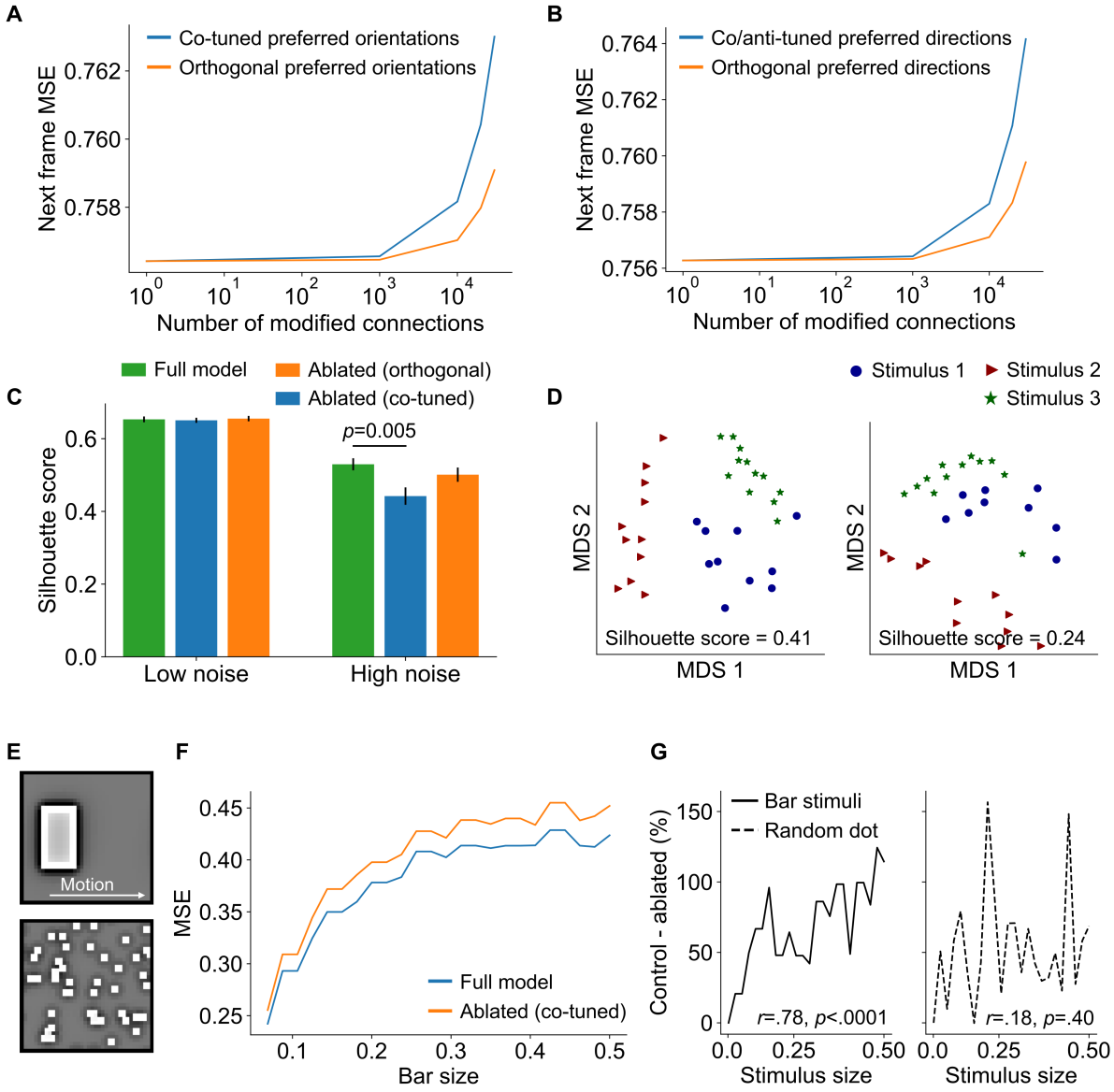


Figure 2.7. Model connectivity motifs support temporal prediction.
(A, B) Ablating connections between connected units with similar preferences for stimulus orientation ("co-tuned" units) or similar and opposite preferences for direction of motion ("co/anti-tuned" units) impairs next-frame prediction more than for units with orthogonal orientation or direction tuning ("orthogonal" units).
(C) Clustering between network representations (measured via the silhouette score) is significantly lower under high noise when connections between pairs of units that are co-tuned (for orientation) are ablated. This finding suggests that this connectivity motif may improve the robustness of cortical representations of visual stimuli in noise. Error bars indicate s.e.m.
(D) Example plots of unit network activity projected onto two dimensions using multidimensional scaling. The silhouette scores illustrate the clustering across different stimulus types in the default network (left) or the network with ablated connections between

co-tuned units (right).
(E) Illustration of the moving bar or moving random dot stimuli used to probe the network's next-frame prediction performance across different stimuli when disrupting specific connectivity motifs.
(F) The mean squared error for next frame prediction in the ablated network increases relative to the control network as a function of bar length.
(G) Stimulus size, equated across both stimulus types predicts the ablation deficit (the percentage increase in mean squared error for the ablated versus control networks) for bar but not random dot stimuli.

Discussion

The recurrent temporal prediction model exhibits response properties and functional connectivity patterns remarkably akin to those found in mouse V1, providing a unifying normative explanation for these wiring biases. In particular, the model captured the relationship between both short- and long-range connectivity patterns and neuronal preferences for stimulus orientation and direction of motion, as well as spatial differences in the inputs from excitatory and inhibitory cells to direction-selective cortical neurons.

The extent to which cortical circuits are fundamentally stereotyped remains an enduring question in systems neuroscience. The concept of a canonical microcircuit proposes that cortical networks follow the same basic organization in which functional differences are defined primarily by their inputs and outputs, rather than by idiosyncratic, local circuits.[150,151] In support of this hypothesis, recurring cortico-thalamic and cortico-cortical loop motifs, as well as cell-type- and layer-specific patterns of connectivity, have been found to be consistent across many cortical areas.[152,153] However, the extent to which cortex-wide connectivity motifs might relate to the functional properties of cortical neurons is still emerging.[154]

Within mouse V1, this chapter demonstrates that a plausible computational principle – temporal prediction – can account for these functional connectivity patterns. Crucially, the network model was not optimized for specific response properties of visual neurons (e.g., particular receptive field characteristics). Instead, the resulting patterns of connectivity arose naturally as an emergent function of optimizing for the more general

objective of predicting the neurons' future inputs. These results suggest that wiring biases found in mouse V1 are not arbitrary but rather that they underpin an important cortical function.

Comparison with other normative models

Despite the clear functional importance of recurrent connectivity in V1, there are comparatively few normative modelling studies addressing this topic. The key contribution of this chapter is in uniting different aspects of both short-range and long-range V1 connectivity with neuronal feature preferences under a single unsupervised learning objective.

Sparse coding networks have been widely employed in modelling receptive fields and response properties,[91,98,155] and more recently have been applied to local connectivity in visual cortex. Sparse coding argues that the brain is optimized to represent stimuli efficiently such that only a small number of neurons are strongly activated at a given time. When trained on static images, sparse coding models have been found to replicate the like-for-like connectivity pattern among units with similar orientation tuning.[91,156,157] Where motion has been included in these models, they can capture the asymmetry in excitatory and inhibitory inputs for direction tuning.[158] However, these sparse coding models have been shown to replicate simple-cell responses only, but not complex-cell responses. Similarly, such models have not been shown to reproduce other distinct connectivity profiles across separate excitatory and inhibitory cell populations or the long-range tuning biases reported in the present study.

Finally, local recurrent connectivity in V1 has also been approached from a Bayesian perspective, where the dependence of cortical connectivity on the similarity in orientation tuning is argued to represent an optimal means of integrating contextual information.[159] However, this Bayesian model depends on hard-coded basis functions derived from V1 simple cells and unlike our approach, can neither be said to be truly unsupervised nor learned exclusively from natural stimulus statistics.

In the context of the current study, we found that only the temporal prediction model could closely reproduce the observed relationships between V1 neurons and their functional connectivity. Thus, the results cannot be accounted for by the choice of dataset or model architecture but are specific to the temporal prediction model's training objective. The temporal prediction model therefore provides a more complete explanation than the other models for the relationship between the connectivity of visual cortical neurons and the stimulus features to which they are tuned. In turn, these results suggest that the functional specificity of connections in V1 enables the brain to process dynamic stimuli by facilitating the prediction of upcoming sensory information.

Comparison with the biology of V1

While the current temporal prediction model is trained using backpropagation through time, the principle of temporal prediction itself is largely agnostic to the underlying learning mechanisms. Indeed, novel and more biologically-plausible learning algorithms are being developed that could, in principle, be applied to the current temporal prediction network.[110,160] In this sense, the present work does not preclude either a hard-wired or learned origin for the connectivity patterns found in visual cortex.[71]

From an evolutionary perspective, temporal prediction is likely to confer several advantages. By encoding only those features that are efficiently predictive of future sensory inputs, temporal prediction provides a principled way of extracting underlying variables and discarding non-predictive, and therefore less behaviourally relevant, information.[90,105] Furthermore, given the inherent delays due to neural conduction and processing in sensory pathways and in preparing motor outputs, some form of predictive processing may be essential to accurately guide an animal's actions.[109]

Given that the model's structure is learned from an initial random state, such a configuration can, at least in theory, emerge from the interplay of some optimization principle and the natural statistics of visual inputs. Following the onset of vision, the connectivity of mouse V1 neurons that respond to similar visual features progressively

increases.[161] These response-specific connectivity patterns still develop in dark-reared mice, indicating that the emergence of like-for-like wiring biases is not dependent on visual experience.[162] Nevertheless, the relationship between connection probability and the similarity of V1 responses to natural movies (but not the similarity of their orientation preferences) was found to be weaker in dark-reared mice than in animals reared with normal visual inputs. Thus, it is likely that these biases in functional connectivity result from an interplay of innate developmental programs which, at least to some extent, are later fine-tuned by sensory experience.[163]

In conclusion, we show that many aspects of functional connectivity in mouse V1 can be parsimoniously described by a single framework – temporal prediction. By optimizing a recurrent network for temporal prediction, model units naturally recapitulate both structural and functional properties of mouse visual cortex. Consequently, even seemingly disparate examples of connectivity rules may be united by a simple underlying principle of cortical organization.

Methods

Dataset

Training data consisted of natural wildlife videos using the same dataset as described previously.[90] Videos were taken from the repository http://www.arkive.org/species and contributed by: BBC Natural History Unit, http://www.gettyimages.co.uk /footage/bbcmotiongallery; BBC Natural History Unit & Discovery Communications Inc, http://www.bbcmotiongallery.com; Granada Wild, http://www.itnsource.com; Mark Deeble & Victoria Stone Flat Dog Productions Ltd., http://www.deeblestone.com; Getty Images, http://www.gettyimages.com; National Geographic Digital Motion, http://www.ngdigitalmotion.com. In brief, videos were converted to grayscale, bandpass filtered then down sampled to 180x180 pixels. Finally, each video was cropped into non-overlapping 36x36 pixel patches of 50 frames each,

leading to a total of 40,000 clips for the training dataset and 4000 clips for the validation dataset used for hyperparameter selection. In addition, to mimic the effects of noise present in the nervous system, Gaussian noise was added to each video clip during training with a signal-to-noise ratio of 6 dB.[90]

For the models in which the temporal offset was varied, training data consisted of a set of 120 Hz naturalistic videos.[147] For each temporal offset, we shifted the target to be predicted 0-10 frames (0-83 ms) into the future. Both the input and target datasets were then down sampled to 24 Hz to approximately match the 25 Hz frame rate of the original temporal prediction model, before they were clipped into 50-frame long segments. This setup ensured that the input dataset was the same across temporal offsets, while only the frame targets varied as a function of temporal offset.

Network model

The model was implemented as a single-layer recurrent network with a linear readout layer to project the hidden activity to the network's output predictions. The network's input consisted of a 50-frame video clip, where the model was trained to predict each subsequent frame given the preceding video frames in the clip. More formally, the network receives a 1296 length vector $\mathbf{u}[t]$ at each time step t, consisting of the flattened 36x36 pixel video frame. The 2592 length hidden state vector $\mathbf{s}[t]$ at each time step t is then given by:

$$\mathbf{s}[t] = f(\mathbf{W}_{\text{in}}\mathbf{u}[t] + \mathbf{W}_{\text{rec}}\mathbf{s}[t-1] + \mathbf{b}_{\text{rec}})$$

where f is the ReLU function, $\mathbf{W}_{\text{in}}$ is the weight matrix which describes the input weights to the network, $\mathbf{W}_{\text{rec}}$ is the weight matrix which describes the hidden, recurrent weights mapping the previous state $\mathbf{s}[t-1]$ to the new hidden state $\mathbf{s}[t]$, and $\mathbf{b}_{\text{rec}}$ is the bias term.

The hidden activity vector $\mathbf{s}[t]$ at each time step t is then mapped to the output predictions $\hat{\mathbf{v}}[\mathbf{t}]$ by:

$$\hat{\mathbf{v}}[t] = \mathbf{W}_{\text{out}}\mathbf{s}[t] + \mathbf{b}_{\text{out}}$$

where $\mathbf{W}_{\text{out}}$ is the weight matrix describing the linear mapping from the hidden state to the output prediction and $\mathbf{b}_{\text{out}}$ is the bias term for the output weights.

In addition, to enforce Dale's Law, whereby hidden units make exclusively excitatory or inhibitory connections, each recurrent weight w was constrained during the forward pass as:

$$w \leftarrow \begin{cases} -|w| & \text{if inhibitory} \\ +|w| & \text{if excitatory} \end{cases}$$

with a total of 2,332 (90%) units set as excitatory and the remaining 260 (10%) as inhibitory units.[164,165]

The network was then optimized via backpropagation through time, 'unrolled' across the preceding 50 frames, to minimize the loss function *E*:

$$E = \sum_{n=1}^{N}\sum_{t=1}^{T} \|\hat{\mathbf{v}}_n[t] - \mathbf{v}_n[t+1]\|_2^2 + \lambda(\|\mathbf{W}_{\text{in}}\|_1 + \|\mathbf{W}_{\text{rec}}\|_1 + \|\mathbf{W}_{\text{out}}\|_1)$$

where n is the clip number, N is the total number of clips in a minibatch, T is the total number of time steps, and $\hat{\mathbf{v}}_n[t] - \mathbf{v}_n[t+1]$ is the difference between the predicted pixel values $\hat{\mathbf{v}}_n[t]$ and true future pixel values $\mathbf{v}_n[t+1]$. Finally, L1 regularization is included as the sum of absolute values of all weights in the network, weighted by the λ hyperparameter.

Implementation

The temporal prediction model was implemented in PyTorch, with gradient descent performed using the ADAM optimizer set at a learning rate of 10^{-4}. Unless otherwise noted, the regularization strength hyperparameter λ was set at 10^{-6} after a hyperparameter search across lambda values (λ range = $10^{-5.5}$ - 10^{-7}) to minimize the mean squared error on the held-out validation set.

Comparison models

The inpainting, denoising and sparse autoencoder networks consisted of the same network architecture as the recurrent temporal prediction model but with modified datasets and training objectives. For the inpainting network, the input dataset was masked with 8 randomly placed 8x8 pixel patches on each frame. For the denoising network, the input was combined with Gaussian noise with a signal-to-noise ratio of 3 dB. Finally, the sparse autoencoder was trained on the same dataset as the temporal prediction model but was trained to recover the current frame under a sparsity constraint. In all these networks, the models were optimized to produce the current unmodified frame (rather than subsequent frame, as for the temporal prediction model) by minimizing the mean squared error between the predicted and actual current frame:

$$E = \sum_{n=1}^{N} \sum_{t=1}^{T} \|\hat{\mathbf{v}}_n[t] - \mathbf{v}_n[t]\|_2^2 + \lambda(\|\mathbf{W}_{\text{in}}\|_1 + \|\mathbf{W}_{\text{rec}}\|_1 + \|\mathbf{W}_{\text{out}}\|_1)$$

For the sparse autoencoder, an additional regularization term λ_{act} was included as the absolute sum of activity across all units, to encourage sparsity in the network's representations:

$$E = \sum_{n=1}^{N} \sum_{t=1}^{T} \|\hat{\mathbf{v}}_n[t] - \mathbf{v}_n[t]\|_2^2 + \lambda(\|\mathbf{W}_{\text{in}}\|_1 + \|\mathbf{W}_{\text{rec}}\|_1 + \|\mathbf{W}_{\text{out}}\|_1) + \lambda_{\text{act}} \sum_{n=1}^{N} \sum_{t=1}^{T} \|\mathbf{s}_n[t]\|_1$$

For the inpainting and denoising networks, the L1 weight regularization hyperparameter was chosen as for the temporal prediction network to minimize the mean squared error on the validation set across a range of values. For the sparse autoencoder, where no such comparable selection criterion exists, the hyperparameter was set qualitatively to produce the most biologically realistic receptive fields.[71]

For VGG-19, we used a publicly available model from the PyTorch 'torchvision.models' package, pre-trained on the ImageNet dataset for object recognition.[145] For each layer of the network, we took the hidden activity as the concatenated, flattened feature maps of each filter in the layer.

For PredNet, we reimplemented the model described by Lotter et al.,[146] trained for next-frame prediction on the same dataset as the main temporal prediction model. However, because the resolution halves at each layer, we used a slightly larger input size of 40x40 pixels. Again, as for VGG-19, we took the hidden activity across layers as the concatenated, flattened feature maps of each filter in that layer.

The LN model consisted of the same basic fitting procedure as the other models. However, where the other models regressed the neural responses on each network's hidden activity, the output of the LN model was used to directly predict neural responses.

Neural response prediction

Neural data were taken from the Allen Brain Institute's Neuropixels Visual Coding dataset.[142] For each model, a linear-nonlinear mapping was fitted to predict the response of V1 units to natural movie stimuli ("Natural Movie One" and "Natural Movies Three", 150 seconds total) from the pre-trained model's hidden unit activity. We divided the dataset into 15x10 second clips, taking 3 representative clips (30s total) as the final held-out test set, with the remaining 120 seconds taken as the training and validation sets with hyperparameter selection by k-fold cross-validation. We included all recorded V1 units from wildtype mice whose noise-to-signal power ratio[166] in response to the natural movies was below 60.

For each model, the neural fitting process consisted of first fitting a linear mapping using Lasso regression before fitting a rectified sigmoidal non-linearity.[167] Prior to fitting, the dimensionality of the model unit activity that was regressed on was reduced to the first 200 components of a principal component analysis (PCA) fitted on the training-set

model responses. PCA was used to equate the number of parameters across models and increase the efficiency of model fitting.[71] The non-linearity $f(x)$ was defined as:

$$f(x) = \mathrm{ReLU}\left(\frac{a}{1 + e^{\frac{c-x}{b}}} + d\right)$$

where the parameters a, b, c and d were optimized to minimize the mean squared error between the true and predicted neural firing rate using the SciPy "curve_fit" function.[168] The L1 regularization strength (α) of the Lasso was chosen via cross-validation from 40 values log-spaced between 10^{1} and 10^{-5} to maximize the average normalized correlation coefficient (CC_{norm})[169,170] across each fold's validation set for the combined linear-nonlinear mapping. The reported values are finally taken as the performance on the held-out test set. CC_{norm} is defined as:

$$CC_{norm} = \frac{Cov(y, \hat{y})}{Var(y)Var(\hat{y})}$$

where y and $\hat{y}$ are the true and predicted firing rates, respectively.

Receptive field and model tuning analyses

Receptive field mapping. Model unit receptive fields were estimated using their response-weighted average. In brief, the responses of model units to 25,000 frames of random Gaussian noise (μ=0, σ=1) were recorded. Each noise frame was then weighted by the unit's response to give the receptive field estimate. Model receptive fields were subsequently parameterized by a Gabor function to extract the receptive field centres and 2D extent. We calculated the unit's spatiotemporal response-weighted average, similarly to the standard response-weighted average, but included the past 7 frames. The temporal power was then taken as the mean squared value over space for each time step, normalized by the total power. Similarly, to calculate the projective centre of mass, we took the mean squared weight for each unit for each predicted future frame time step.

For each of these, we then determined the centre of mass as the average of the time values weighted by that time step's corresponding normalized power.

Unit inclusion criteria. To maintain consistency across analyses, only those units whose receptive fields that were spatially well defined and which could be well modelled as Gabors were included for analysis. To that end, units whose receptive fields were less than 0.5 pixels in size and therefore had little spatial extent (19% of total units) or which were poorly fitted by the Gabor function were excluded ($r<0.7$, 12% of total units; 30% including both criteria). Short-range connections were defined as those less than 15° (2.5 pixels) and long-range as greater than 30° (5 pixels).[118] Connections greater than 9.17 pixels were excluded because of the experimental constraints imposed by screen size. In the case of Ko et al.,[120] connections were not explicitly defined according to the distance between receptive fields, but we use the same short-range convention as in Iacaruso et al.[118] that, under the assumption of retinotopy, physically short-range connections (<50 μm) are likely to be close in visual space.

Unit tuning characteristics. To measure the model units' tuning properties, each unit's response to sinusoidal gratings was recorded. Gratings varied in temporal frequency (0.02-0.25 cycles/frame), spatial frequency (0.03-0.5 cycles/pixel), and orientation (0-360 degrees) with an amplitude of ±1. Each unit's preferred temporal frequency, spatial frequency and orientation were taken as the parameter combination that maximized the unit's mean response across 50 frames. For those analyses dependent on the unit's spatial location, only units within the central 16x16 pixel bounds of the visual fields were included to avoid edge effects.

Orientation and direction selectivity. These indices (OSI and DSI) were defined as:

$$OSI = \frac{R^{\mathrm{Or}}_{\mathrm{pref}} - R^{\mathrm{Or}}_{\mathrm{orth}}}{R^{\mathrm{Or}}_{\mathrm{pref}} + R^{\mathrm{Or}}_{\mathrm{orth}}}$$

$$DSI = \frac{R^{\text{Dir}}_{\text{pref}} - R^{\text{Dir}}_{\text{opp}}}{R^{\text{Dir}}_{\text{pref}} + R^{\text{Dir}}_{\text{opp}}}$$

where $R^{\text{Or}}_{\text{pref}}$ and $R^{\text{Or}}_{\text{orth}}$ are the unit responses at the preferred and orthogonal orientations, and $R^{\text{Dir}}_{\text{pref}}$ and $R^{\text{Dir}}_{\text{opp}}$ are the unit responses at the preferred and opposite (+180 degrees) directions. For Figure 2.2, we take the same thresholds as Ko et al.,[120] where direction selective units were defined as those with OSI values exceeding 0.4 and DSI values exceeding 0.3. For Figure 2.3, where no threshold is given in Rossi et al.,[122] we take the more stringent threshold of 0.8 for direction-selective units.

Modulation ratio. Model units were classified as simple- or complex-like based on their phase-responsiveness to drifting grating stimuli. Quantitatively, units with a modulation ratio $F > 1$ were classed as simple-cell-like or as complex-cell-like for a modulation ratio < 1. The modulation ratio was defined as $F = \frac{F_1}{F_0}$ where F_0 is the mean response of the neuron to its preferred stimulus and F_1 is the amplitude of the fitted sinusoid to the neuron's response to its preferred stimulus. Where the correlation between the fitted sinusoid and the true response was < 0.9, we did not calculate the modulation ratio for that unit.

Prediction error analyses

We used two paradigms to assess the presence of sensitivity to prediction errors in the network hidden units: oddball stimuli and omission stimuli. In the oddball paradigm, a deviant stimulus interrupts the pattern generated by a preceding set of standard stimuli. In the omission paradigm, the violating stimulus consists of an omission – i.e., the absence of an expected stimulus presentation.

In the odd-ball paradigm, we presented the model with stimuli consisting of two full-field gratings of orientations A and B that each spanned 0°, 45°, 90° or 135° (Figure S2.3A). The omission paradigm was similar but consisted of a single full-field grating A and a blank stimulus B. A and B stimuli alternated for 25 frames, while the omission or

deviant position was varied to occur after 5-25 standard frames. We constructed two sets of control stimuli to ensure that prediction-error responses could not be explained by differences in stimulus tuning or unrelated network dynamics. First, we compared the response to the violating stimulus with that generated by the same set of stimuli using the standard stimulus (i.e., ABAB**B** vs. ABAB**A**). Second, we constructed a 'shifted' set of stimuli, where the deviant position was matched but without violating the pattern (i.e., ABAB**B** vs. BABA**B**).

We chose a relatively conservative set of criteria to avoid miscategorising non-prediction-error responses. Specifically, a unit was defined as prediction-error-like only if for a particular orientation or orientation pair it had no response in either control condition and responded to the violating stimulus across at least 5 different deviant positions. We also adopted slightly looser criteria (Figure S2.3C), whereby prediction error-like responses were defined when responses to deviant or omission stimuli exceeded three times the response to control stimuli.

Unit connectivity analyses

Unit connectivity. Units were defined as connected if their connection strength exceeded the 95th percentile of connection weights ($\mathbf{W}_{\text{rec}}$) across all pairs of units. Due to the sparse nature of the recurrent weight connectivity matrix, this threshold equated to rejecting the very low or zero weight connections, while retaining the smaller subset of highly connected units. Thus, varying this threshold across a range of values (92.5-99 percentile) did not qualitatively change the results.

Visual space-dependent connectivity. For each presynaptic ensemble, visual space was normalized according to the receptive field centre and preferred orientation of the postsynaptic unit. Receptive field centres were first translated such that the postsynaptic unit receptive field was centred at the origin:

$$\begin{bmatrix} x'_{\text{pre}} \\ y'_{\text{pre}} \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & -x_{\text{post}} \\ 0 & 1 & -y_{\text{post}} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_{\text{pre}} \\ y_{\text{pre}} \\ 1 \end{bmatrix}$$

Next, receptive field centres were rotated according to the postsynaptic unit's preferred orientation θ:

$$\begin{bmatrix} x'_{\text{pre}} \\ y'_{\text{pre}} \\ 1 \end{bmatrix} = \begin{bmatrix} \cos(\theta - \frac{\pi}{2}) & -\sin(\theta - \frac{\pi}{2}) & 0 \\ \sin(\theta - \frac{\pi}{2}) & \cos(\theta - \frac{\pi}{2}) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_{\text{pre}} \\ y_{\text{pre}} \\ 1 \end{bmatrix}$$

For the comparison with Iacaruso et al.,[118] presynaptic units were binned into co-orthogonal and co-axial receptive field centres according to whether they fell in one of the four quadrants orthogonal to or parallel with the postsynaptic unit's preferred orientation (i.e., defined by $y = -x$ and $y = x$). For the comparison with Rossi et al.,[122] presynaptic units were binned according to whether they fell opposite to or ahead of the postsynaptic unit's preferred direction of motion (i.e., defined by $x = 0$). Presynaptic densities were computed by binning the presynaptic receptive field centres (taken as the centre of the Gabor fit to the unit's receptive field, as described above) across visual space.

Natural movie response correlations. One hundred 50-frame clips were randomly selected from the validation set and the response recorded for each model unit. The response correlation was then taken as the average correlation across the set of clips for each pair of units in the network.

Ablation experiments

For the ablation experiments, we randomly selected n connections belonging to the relevant class of units. For each of these units, the connection was 'ablated' by setting its value to the median connection weight for recurrent weights between the excitatory subpopulation of the network. This was repeated 1,000 times and the average taken to obtain the MSE values in Figure 2.7.

For the silhouette score analyses, we presented three classes of grating stimuli whose orientations were offset by 0°, 11.25° and 22.5°, repeating the analysis across the complete span of orientations (0-360°). Over these classes of stimuli, we then calculated the silhouette score to measure the level of distinctness of the three classes in the hidden representations. The silhouette score is bounded between -1 and 1, with a high score indicating that the clusters are well separated according to the stimulus (orientation) class.[149]

Control analyses

The distribution of preferred orientations and directions among model units was not uniform (Figure S2.1). To control for the possibility that the observed model connectivity distributions resulted from this overrepresentation of particular orientation and direction tuning preferences, we compared the true model results to those after randomly shuffling across model weights. Specifically, we took the total set of model units fulfilling the relevant criteria for each analysis (e.g., orientation selectivity, receptive field distance, etc.) and randomly shuffled the recurrent weights connecting these units. It is important to note that we used these shuffled weights only for the connectivity analyses, not for any other part of any analysis, such as getting unit responses for response-weighted averages. We repeated this process 1,000 times, taking the mean value of the resulting distribution to compare to the true unshuffled model results.

Mouse V1 comparison data

Mouse data for comparison were either extracted from published figures using WebPlotDigitizer (Figure 2.1F; Figure 2.2J), taken as the exact statistics from the published paper (Figures 2.2B, C, F, G, I) or computed directly from the Neuropixels Visual Coding dataset from the Allen Brain Institute[142] (Figures 2.1C, E; Figure 2.5C; Figure S2.1). For the V1 receptive fields (Figure 2.1C), these were estimated by fitting a

linear filter to predict the responses of V1 single units to natural movies as described above.

To compare how well each model captured the connectivity patterns described for mouse V1, we calculated a model connectivity score as the average of the Pearson correlation coefficients between the described neural connectivity profiles (data in Figures 2.2B, C, F, G, I, J) and the corresponding connectivity profiles of the given model.

Supplemental information

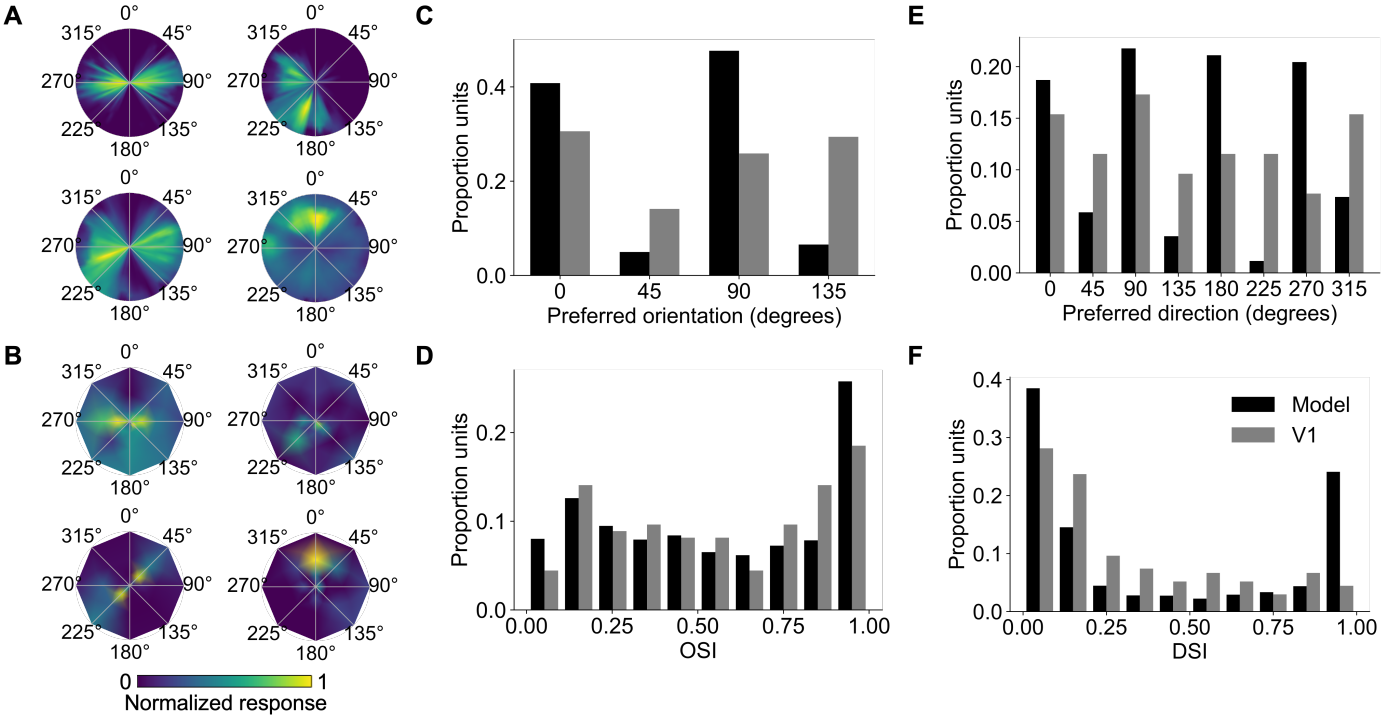


Figure S2.1: Model and mouse V1 tuning properties, Related to Figure 2.1.
(A) Example model unit responses to drifting gratings as a function of temporal frequency (1-8 Hz, radial axis) and direction (polar angle).
(B) Example mouse V1 responses to drifting gratings as a function of temporal frequency (2-15 Hz, radial axis) and direction (polar angle).
(C) Distribution of preferred drifting grating orientations for model units and mouse V1.
(D) Distribution of orientation selectivity indices (OSI) for model units and mouse V1.
(E) Distribution of preferred drifting grating directions for model units and mouse V1.
(F) Distribution of direction selectivity indices (DSI) for model units and mouse V1.
All mouse data is taken from the Allen Brain Institute's visual coding dataset.

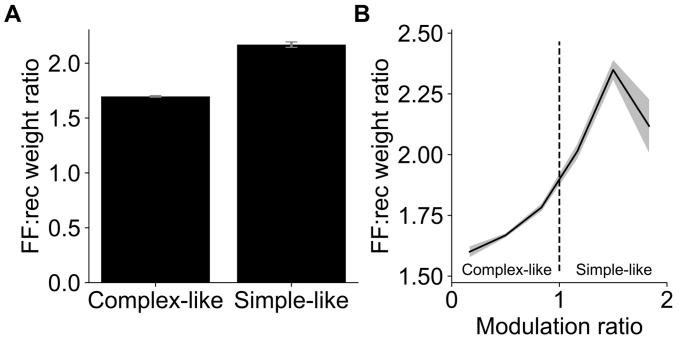


Figure S2.2. Complex-like units are characterised by greater recurrent inputs, Related to Figure 2.1.
(A) Mean feedforward to recurrent weight ratio across recurrent units.
(B) Mean feedforward to recurrent weight ratio as a function of unit modulation ratio.

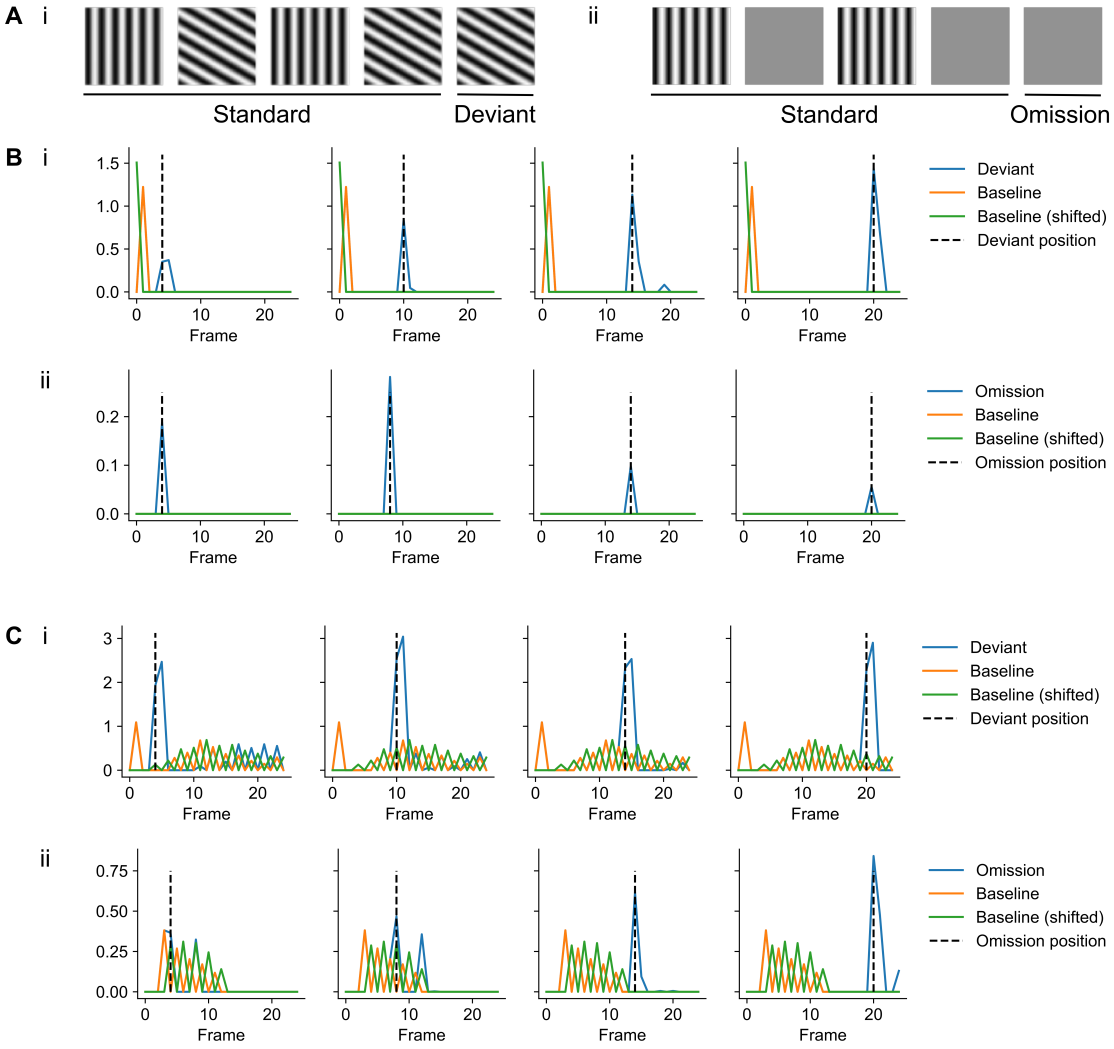


Figure S2.3: The recurrent temporal prediction model generates prediction error responses, Related to Figure 2.1.

(A) Example deviant (i) and omission (ii) stimuli used to probe the network.

(B) Example unit activations for deviant (i) and omission (ii) stimuli. Each subplot reflects a different deviant position in the sequence, with the prediction error response tracking this position. The baseline (control) stimuli for the odd-ball paradigm show only an initial onset response. Baseline stimuli consisted of the default stimulus set without inclusion of the deviant or omitted stimulus, while the shifted baseline stimuli consisted of the same stimulus set at the baseline but shifted forwards by one frame to match the timing without incurring a violation.

(C) Example unit activations for deviant (i) and omission (ii) stimuli using a less conservative set of criteria. Here, these units show mixed-selectivity and are responsive to both stimuli.

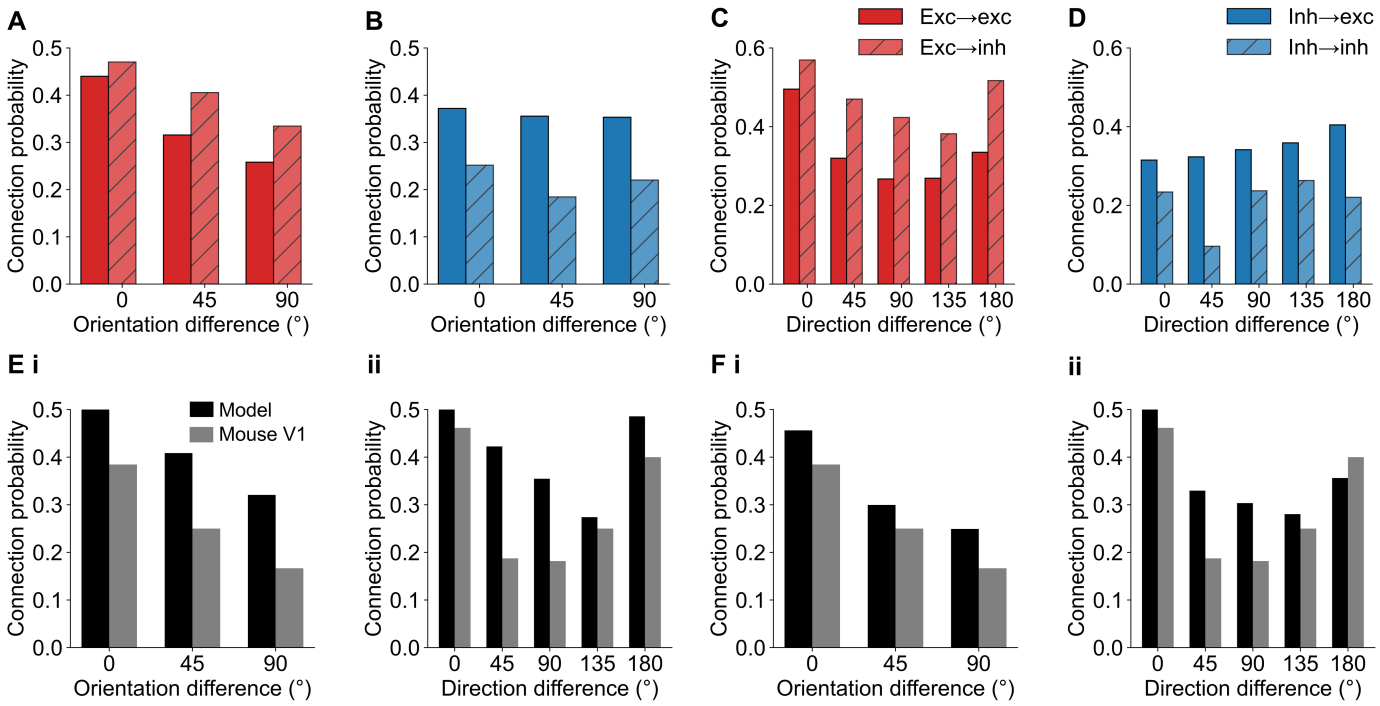


Figure S2.4: Orientation and direction dependence of connection probabilities across model unit types, Related to Figure 2.2.

(A, B) Connection probability as a function of orientation tuning difference for excitatory (A) and inhibitory (B) model units. Excitatory units show a monotonic trend of decreasing connection probability, whereas inhibitory units show a much weaker effect, implying a broader functional spread of inhibitory connections.

(C, D) Connection probability as a function of direction tuning difference for excitatory (C) and inhibitory (D) model units. Excitatory units show a characteristic u-shaped curve, as found in V1, where units with similar or opposite direction tuning are most likely to connect. Inhibitory units synapsing with excitatory units show a weak but monotonically increasing trend, where the closer they are to having opposite direction preferences, the more likely they are to connect. Finally, inhibitory model units synapsing with other inhibitory units show a more heterogenous pattern, with no overall preference for a given difference in direction tuning.

(E, F). Connectivity motifs in the model across simple-cell-like (E) and complex-cell-like (F) populations in the model for short-range orientation-dependent connectivity (i), and short-range direction-dependent connectivity (ii).

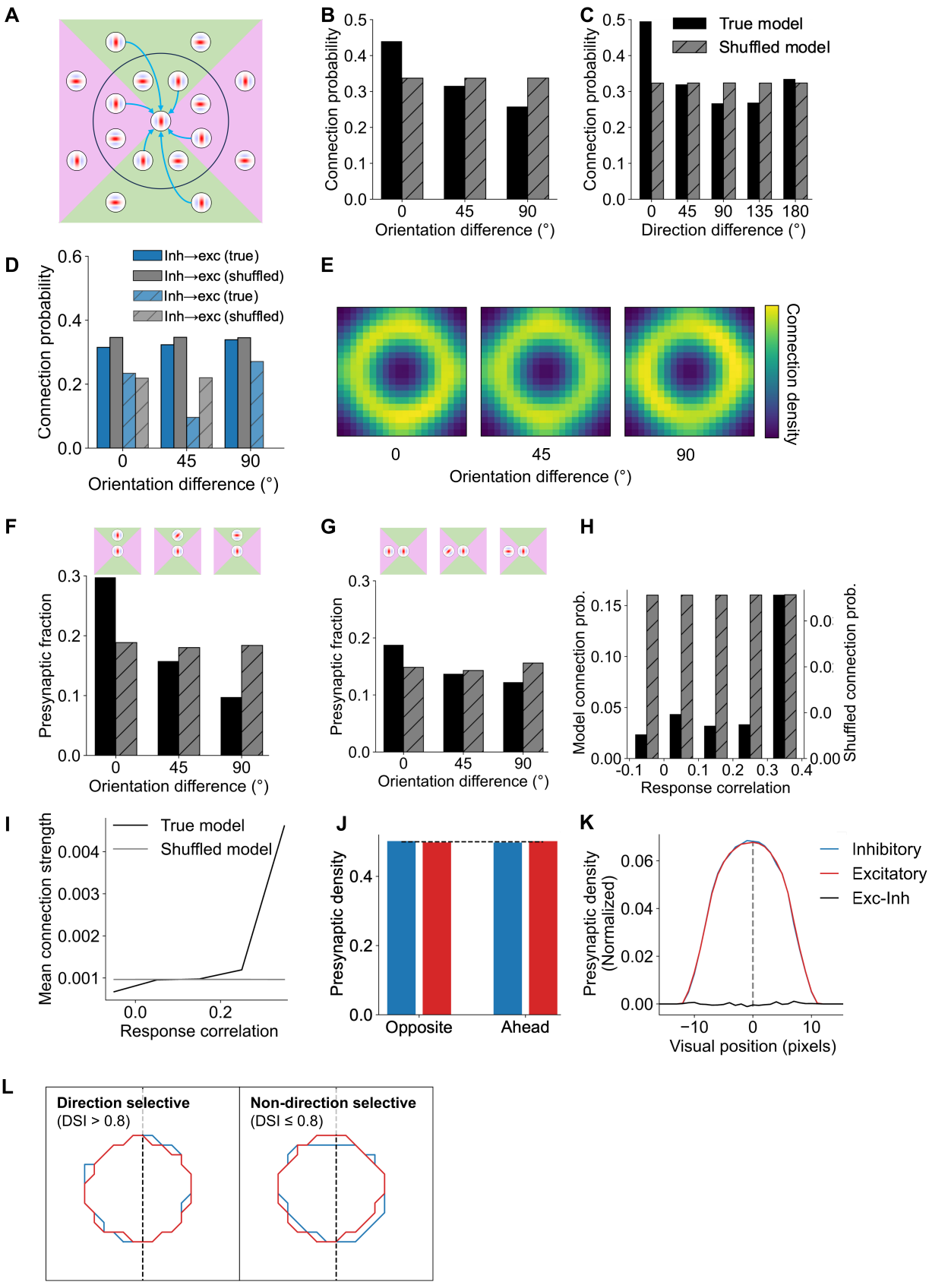


Figure S2.5: Functional connectivity resembling mouse V1 in the model is abolished when connectivity is measured with the recurrent weights randomly shuffled, Related to Figures 2.2 and 2.3.

(A) Schematic of local functional connectivity in mouse V1.
(B, C) Short-range connection probability as a function of the difference in orientation and direction tuning among model units.
(D) As in B, but for inhibitory-to-inhibitory and inhibitory-to-excitatory connections in the model.
(E-G) Long-range connection probability as a function of difference in orientation preferences for receptive fields located in co-axial (F) and co-orthogonal (G) locations relative to the receptive field of the post-synaptic unit. Heatmap (E) shows the shuffled connection probability over visual space across differences in orientation tuning for model units. Heatmap is smoothed for display purposes with a Gaussian filter (σ=2 pixels).
(H, I) Response correlation for model units as a function of connection probability (H) as well as the input connection strength (I).
(J) After shuffling, model unit presynaptic density for excitatory and inhibitory cells is equal in both halves of visual space. Dashed line represents equal density (0.5).
(K) Profile of model unit presynaptic density across horizontal visual space for excitatory and inhibitory inputs. Profiles smoothed with a 5-pixel moving average.
(L) Pooled density contours across all excitatory (red) and inhibitory (blue) model units for direction- and non-direction-selective post-synaptic excitatory units, showing no overall differences between them after shuffling.

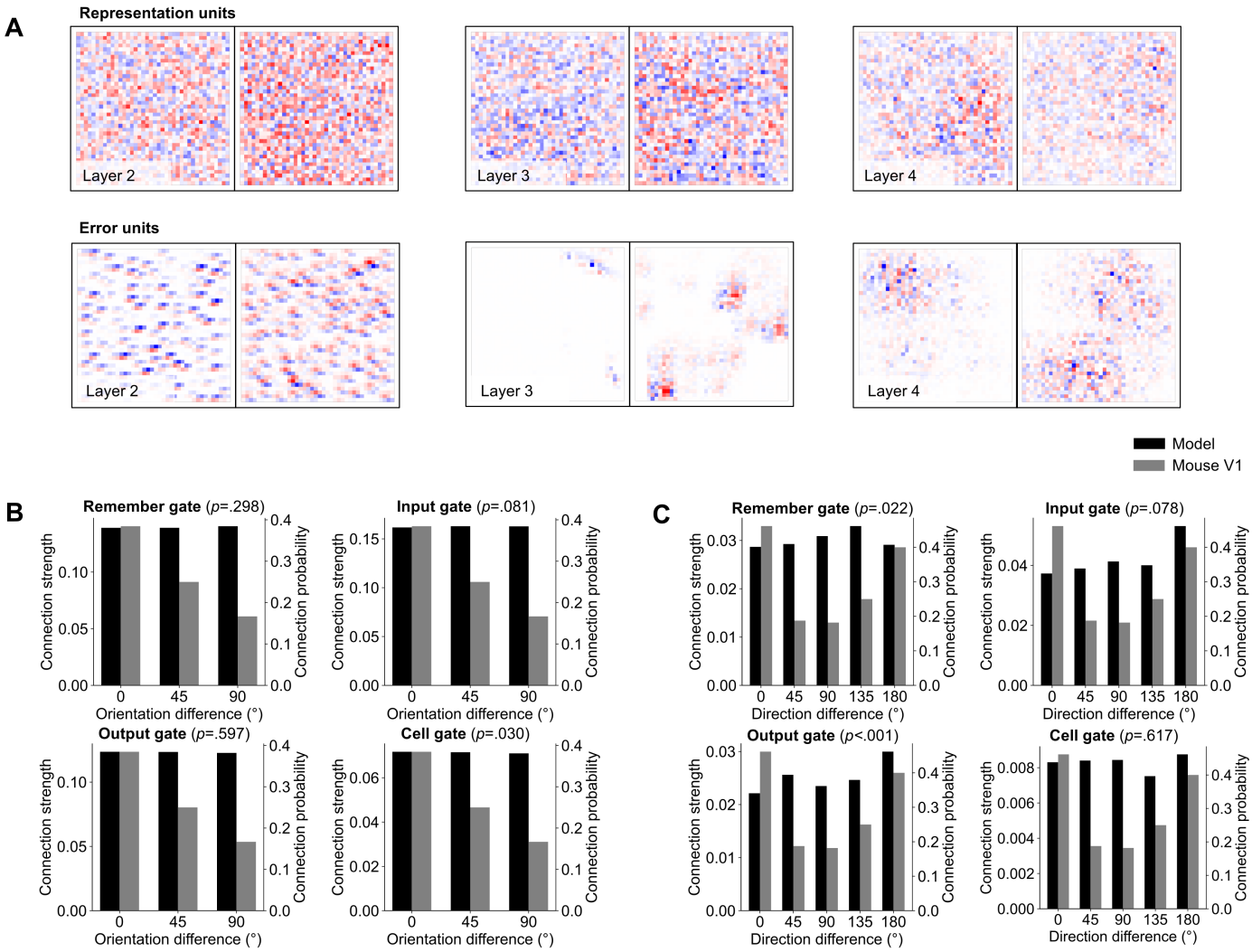


Figure S2.6: PredNet does not recapitulate short-range orientation- or direction-dependent connectivity, Related to Figure 2.5.

(A) Optimal stimuli for exemplar units in PredNet estimated using gradient ascent. While there is clear structure present in the error units, this structure and any spatial localization are markedly absent in the recurrently-connected representation units.

(B) PredNet does not capture well the finding that units with similar orientation tuning are more likely to connect, with little variation in connection strength as a function of orientation difference.

(C) PredNet shows more variation in connection strength as a function of the difference in preferred stimulus direction across its recurrent weights, but these properties do not vary in the same way as in V1, with the characteristic 'U'-shape absent.

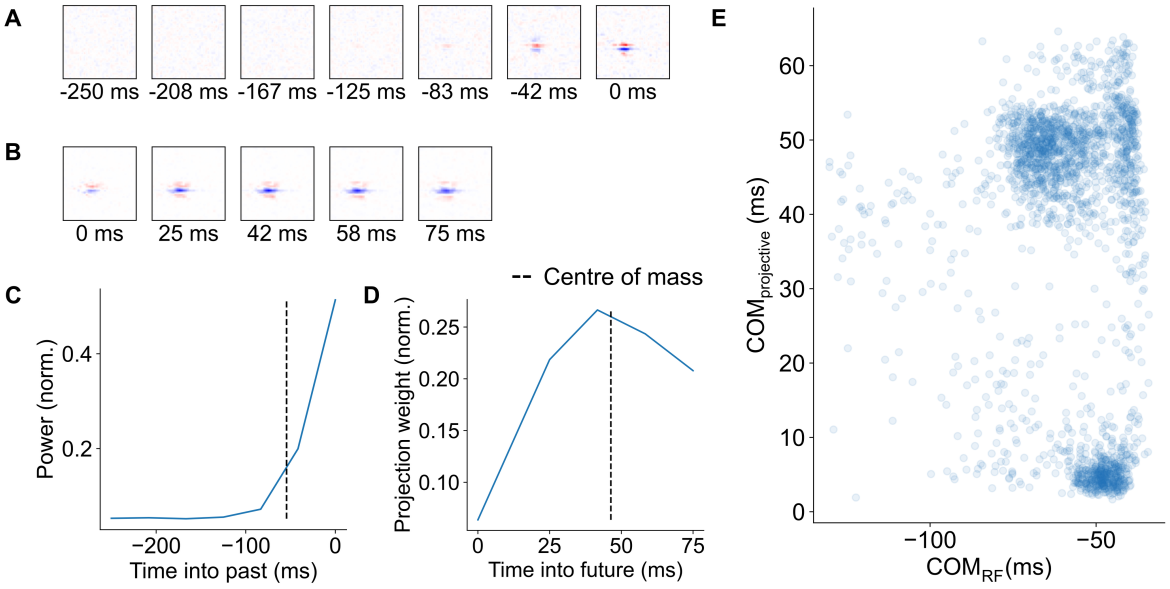


Figure S2.7: Units that integrate information further into the past project further into the future in the span-predicting temporal prediction model, Related to Figure 2.6.

(A) Example spatiotemporal receptive field for a single model unit showing a decay in power into the past.

(B) The 'projective' receptive field for the same model unit in A, showing this unit's linear output weights at each future prediction time step.

(C, D) Temporal power of the receptive field into the past (C) and of the projection weights into the future (D) for the example shown in A-B. The dashed line indicates the centre of mass for each curve.

(E) Scatter plot for the centre of mass of the temporal power of the receptive fields (COM_{RF}) versus the centre of mass of the temporal power of the corresponding projective fields ($COM_{projection}$). Each point is a hidden unit.

3

Hierarchical temporal prediction as a model of the mammalian dorsal visual pathway

Introduction

Classical theories of vision posit a series of hierarchically organized processing stages that gradually extract higher-level visual features.[171,172] In this way, the brain is thought to decompose the patterns of light which impinge on the retina into meaningful representations to be further processed in downstream areas, and ultimately, to guide behaviour. A key question, then, is how to understand these computations and whether an underlying principle of organization can explain how the resulting representations change across the visual pathway.[71,72]

As described in chapter two, one promising principle is that of temporal prediction, which argues that the sensory brain is optimized to represent stimulus features that are predictive of the immediate future.[71,89,90] This is likely to be useful for extracting underlying variables in the input, eliminating unnecessary information based on the behavioural relevance of stimuli,[173] and creating a representation that is useful for guiding future action given sensory and motor delays.[71,89,90] Temporal prediction can be applied in a hierarchical manner to capture spatiotemporal receptive field properties across multiple stages of the visual pathway.[71] Importantly, as an unsupervised principle,

temporal prediction does not rely on hand-labelled examples as with many deep learning models of the visual system.[174,175] Together, these features make temporal prediction a promising candidate for modelling neural processing in sensory systems.

Previous work has demonstrated that the tuning properties of neurons along the visual pathway can be reproduced by applying temporal prediction in a hierarchical manner,[71,90] while the work in chapter two of this thesis demonstrated that the functional specificity of connections within the primary visual cortex (V1) can be recapitulated by applying temporal prediction to a locally-recurrent network.[89] However, these models have so far neglected the substantial role of long-range feedback connectivity from higher-order visual areas, which has been implicated in a range of visual processing functions, imparting both modulatory and driver roles on downstream neural activity.[127,176] Among other functions, feedback from higher visual areas is believed to mediate extra-classical receptive field effects such as surround suppression,[177–180] maintain and gate working memory representations,[181–183] and to convey predictions that may drive learning via prediction errors.[99] Accordingly, incorporating these elements into normative models, including temporal prediction, is likely to be an important step in better capturing the structural and functional properties of the visual pathway.

Here, we demonstrate that a hierarchical recurrent temporal prediction model trained on movies of dynamic natural scenes better captures the tuning properties of visual neurons along the mammalian visual pathway than when implementing hierarchy in a feedforward model or when implementing local recurrency alone. The architectural addition of inter-areal recurrency and the network's optimization for temporal prediction jointly improved the model's fit to the tuning properties of visual cortical neurons measured across different studies. Moreover, feedback connectivity significantly improved the network's memory capacity and frame prediction performance, while recapitulating the dependence of feedback on pattern motion sensitivity and surround

suppression in the responses of V1 neurons.[177–180,184] Finally, compared with alternative normative models, the hierarchical recurrent temporal prediction model was most closely aligned with the stimulus representations in different areas of the visual cortex. Together, these results provide evidence that inter-areal feedback across the visual hierarchy may also be optimized for temporal prediction.

Results

The hierarchical recurrent model improves temporal prediction performance

The hierarchical recurrent model consisted of a recurrent neural network instantiating a hierarchical model of temporal prediction, where each group of units was trained to predict its lower-order inputs. Thus, the first group (G_1) was trained for next-frame prediction of video clips of dynamic natural scenes while the second and third groups (G_2 and G_3) were trained to predict the future activity of groups G_1 and G_2, respectively (Figure 3.1A). Within the recurrent layer of the network, internal connectivity ensured that each group received feedforward input from the previous group, feedback input from the subsequent group and local recurrent input from itself. We conceptualized each group as modelling increasingly deep regions of the dorsal visual pathway, with groups G_1, G_2, and G_3 corresponding to primary visual cortex (V1), secondary visual cortex (V2) and middle temporal area (MT), respectively (Figure 3.1B). However, the model structure and training data are not *per se* tailored to any specific mammalian species. Hence, while we will largely compare our model to macaque or other primate data, where such data are not available, we will compare to mouse data.

We first analysed the temporal prediction performance of the hierarchical recurrent model and compared this to several baseline models. Specifically, we compared the hierarchical recurrent temporal prediction model to a single-layer recurrent temporal prediction model without the addition of hierarchical groups, a purely feedforward

model which omitted both local and feedback recurrency, and a baseline comparison where we compared the predictions to simply copying the input frame. The copy frame comparison ensured that each model was truly acting as a generative model to predict the next frame, rather than merely reproducing its inputs as a trivial solution. For next frame prediction (Figure 3.1C), we calculated the mean squared error between the true and predicted next frame of the stimulus. Conversely, for 10 frame prediction (Figure 3.1D), we measured the mean squared error across 10 frames generated in an autoregressive manner. This involved repeatedly predicting the next frame using the model's own preceding predictions, rather than the true frame inputs. For next frame prediction, we assessed the models' performance both in-distribution (the held-out test set used for training; Figure 3.1E) and out-of-distribution (a novel dataset to that used during training; Figure 3.1F) to assess the generality of these findings. In the multi-frame case, the input consisted of oriented moving bar stimuli where there was an unambiguous 'true future' given the preceding input (Figures 3.1G, H).

Across all conditions, we found that the hierarchical model performed best with the overall lowest mean squared error for temporal prediction (paired t-test, all $p<.001$) (Figures 3.1E-G). Thus, the addition of hierarchy in the form of additional groups improved prediction performance compared with both a single-layer recurrent model and a feedforward model trained for temporal prediction. However, as the hierarchical and vanilla RNN were matched in terms of the number of parameters for the first group, but not the total number of parameters, this effect could partly have been explained by the large parameter count of the hierarchical network recurrent network. Finally, all three models exceeded the baseline performance for copying the preceding frame (paired t-test, all $p<.001$).

We also varied the beta parameter, which controls the influence of higher groups on the hierarchical recurrent model's loss function during training, to assess its impact on temporal prediction performance (Figure 3.1H). A higher beta value indicates a greater

weighting applied to higher groups, with β=0.5 indicating an equal weighting. Although higher beta values impaired next-frame prediction performance across the in- and out-of-distribution cases (Figure S3.1), we found that a small beta improved multi-frame prediction on the moving bar stimuli set (β=0 vs β=0.1: paired t-test, $t(71)=3.20$, $p=.002$). Thus, the addition of hierarchy to the loss function could improve the model's capacity to generalize to multi-frame prediction across novel stimuli.

Given the importance of hierarchy in the model for temporal prediction, we next asked whether this might in part be supported by increasing the memory capacity of the model, mediated by the network's feedback connectivity. In the visual cortex, feedback from higher visual areas to V1 is selectively recruited during working memory tasks,[182] while disruption of corticocortical feedback to higher visual areas has been shown to impair performance in working memory-dependent behaviour.[183] Thus, we reasoned that feedback in the hierarchical recurrent temporal prediction model might similarly support the network's memory capacity.

To test this hypothesis, we trained a linear decoder to predict the identity of hand-written MNIST digits from the network's hidden state. To probe how well the model's hidden state could maintain the digit representations, the decoder was trained on the model's hidden state after n frames of Gaussian noise input (Figure 3.1I). For each n-back decoder, we then calculated the decoder's accuracy on the MNIST test set for the model with and without feedback. Ablating feedback significantly reduced the decoder accuracy from 3-back onwards (z-test, $z=6.54$, $p<.001$), implying that feedback improved the model's capacity to maintain representations over longer timescales (Figure 3.1J). To test whether this improvement was related to the model's training or whether any form of feedback could account for the difference, we assessed the 5-back accuracy when shuffling an increasing proportion of the model's feedback weights (Figure 3.1K). As the proportion of shuffled weights increased, the accuracy decreased until it was non-significantly different from the model where feedback was ablated entirely (z-test,

z=0.494, p=.621). Thus, randomized feedback was ineffective, and feedback weights required the structure imposed by training to maintain network representations.

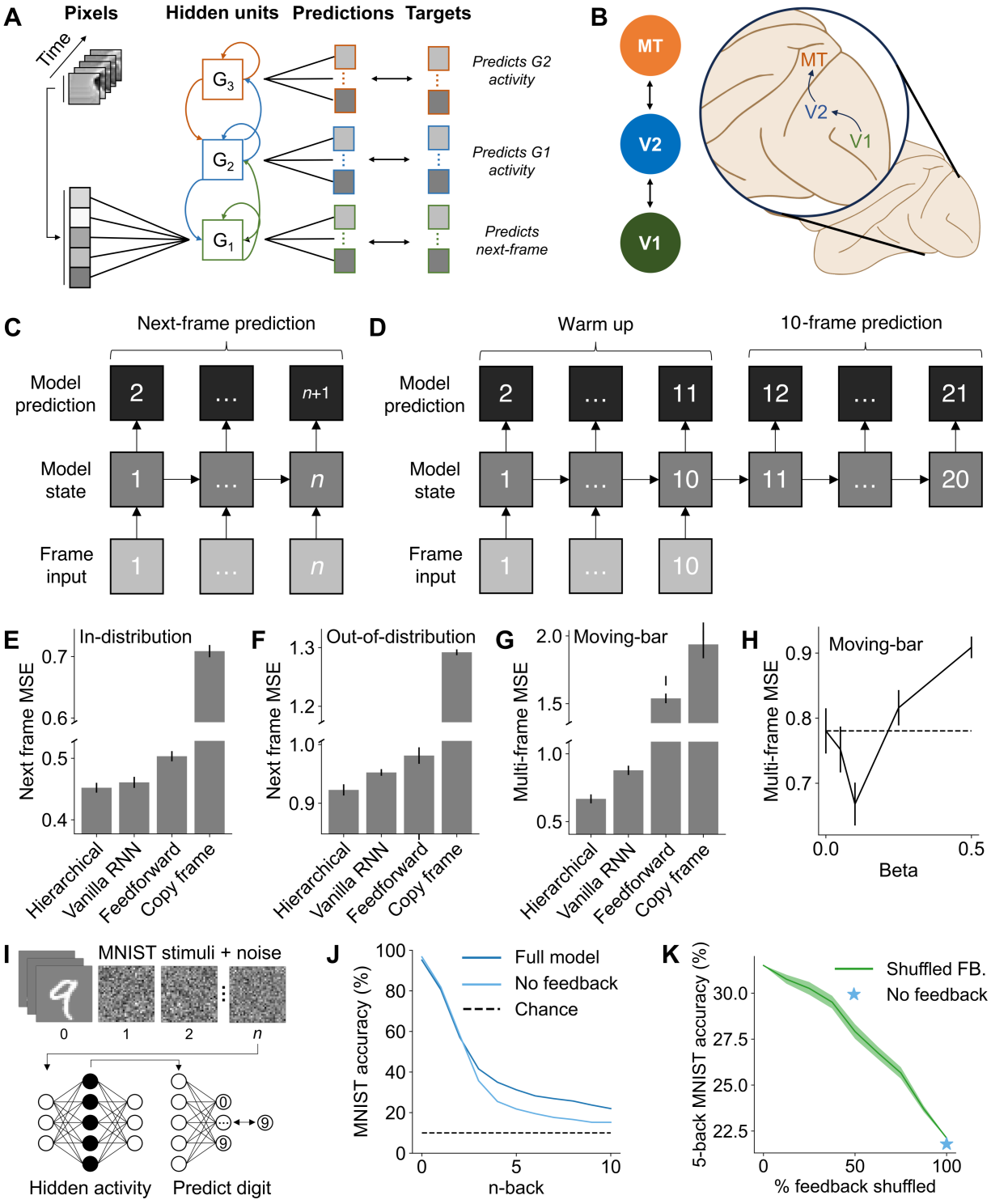


Figure 3.1. The hierarchical recurrent temporal prediction model shows improved frame prediction performance.

(A) Schematic of the temporal prediction model.
(B) A subset of the dorsal visual pathway, with regions V1, V2 and MT highlighted, which we conceptualized as corresponding to groups G_1, G_2 and G_3. Macaque graphic via Macauley Smith Breault (https://zenodo.org/records/3926117).
(C) The next frame prediction procedure, where the preceding frame and hidden state are inputted into the model at each timestep.
(D) The multi-frame prediction procedure, where the model was run autoregressively such that the model predicted multiple frames, relying on its own predictions and hidden state alone.
(E-G) Prediction performance across comparison models for (E) in-distribution, (F) out-of-distribution and (G) multi-frame prediction of a moving bar. In all cases, the hierarchical model performed best, exceeding the performance of the comparison models.
(H) Multi-frame prediction performance for a moving bar across differing beta values. The lowest mean squared error was found for a beta value of 0.1, implying that a hierarchical instantiation of the network can improve prediction performance.
(I) Schematic of the n-back procedure. MNIST digits are used as input to the model followed by *n* frames of Gaussian noise. The hidden state is then linearly mapped to predict the digit identity to calculate an accuracy score, repeating the same procedure for all *n* 0-10.
(J) MNIST accuracy declines more quickly when feedback is ablated, suggesting that model feedback increases the memory capacity of the network.
(K) MNIST accuracy declines as the percentage of shuffled feedback weights increases, implying that the increase in memory capacity is not purely a feature of the model's architecture, but depends on the structure of the feedback connectivity weights imposed by training.
See also Figure S3.1.

Hierarchical recurrent temporal prediction captures tuning properties across the visual hierarchy

We next investigated how model unit response properties varied across the model's groups relative to different stages of the visual system. To make these comparisons, we estimated the model units' receptive fields using the response-weighted average of each unit to Gaussian random noise. Units in the first group generally had a clear Gabor-like structure with alternating excitatory and inhibitory regions akin to the receptive fields of V1 simple cells (Figure 3.2Ai). In contrast, units in higher-order groups generally displayed little spatial structure (Figures 3.2Aii, iii).[129]

To further characterize the receptive field properties of model units, we recorded their responses to full-field sinusoidal gratings of varying temporal frequency, spatial frequency, and orientation. The preferred stimulus was defined as the drifting grating that maximally stimulated each unit, which we used to classify model units as simple-cell- or complex-cell-like in their responses.[128,129] In visual cortex, the activity of simple

cells is heavily modulated by a drifting grating, whereas complex-cell responses are phase invariant and only minimally modulated by the moving stimulus. We found model units that exhibited both kinds of response, including simple-cell- (Figure 3.2Bi) and complex-cell-like (Figures 3.2Bi, Bii) responses.

We quantified these responses using the modulation ratio – the ratio of their amplitudes to that of the average response – with a larger modulation ratio indicating a more simple-cell-like response (Figure 3.2C). G_1 units (mean modulation ratio=0.53) had a bimodal-like distribution similar to that of macaque V1[185] (mean=0.79), indicating two subpopulations of units (Figure 3.2Ci). Higher groups (G_2 mean=0.48, G_3 mean=0.45) had a more unimodal-like distribution, which tended to monotonically decline with increasing modulation ratio, similarly to the true distribution found in the secondary visual cortex (V2)[186] (mean=0.63; Figures 3.2Cii, Ciii). This was reflected in the average modulation ratio across groups (Figure 3.2Ciii inset), which declined for higher model groups, indicating a larger proportion of complex-cell-like responses. Finally, we calculated the joint distributions of the modulation ratio with the model unit's orientation selectivity (Figure S3.2A), as well as how well each unit was modelled as a Gabor (Figure S3.2B). As expected, and in tandem with the biology, simple-cell-like model units in G1 were more orientation selective (mean orientation selectivity index, OSI=0.77) and had a higher r^2 for their Gabor fits (mean=0.54) than complex-cell-like units (mean OSI=0.84; t-test, $t(181)=-3.84$, $p<.001$; mean Gabor fit=0.60; t-test, $t(251)=-4.23$, $p<.001$).

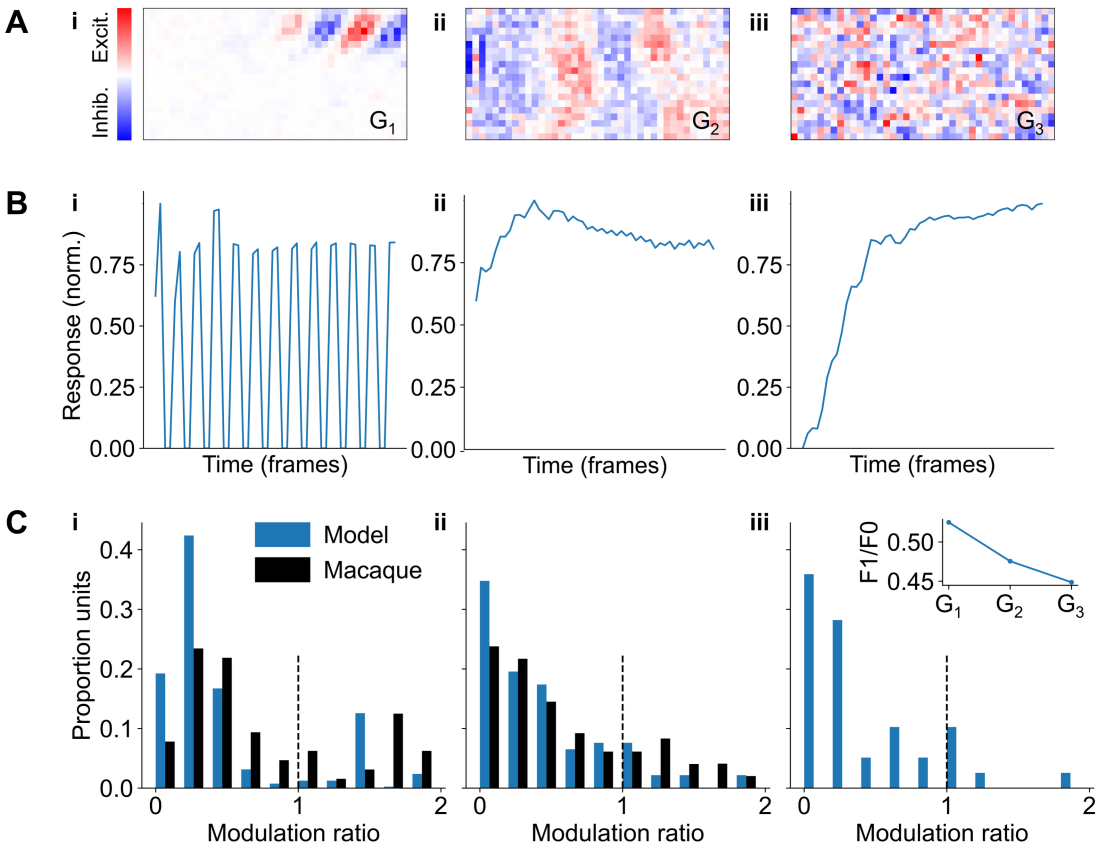


Figure 3.2. Hierarchical recurrent temporal prediction captures modulation tuning properties across the visual hierarchy.
(A) Exemplar model unit receptive fields from group G_1 (i), G_2 (ii) and G_3 (iii).
(B) Normalized response to the preferred grating stimulus for the same model units as in (A).
(C) Distribution of modulation ratio values across model groups G_1 (i), G_2 (ii) and G_3 (iii) (blue bars). Comparison data for G_1 from macaque V1[185] and for G_2 from macaque V2[186] (black bars). Inset shows the average modulation value for each model group.
See also Figures S3.2 and S3.3.

The distribution of spatial and temporal frequency tuning properties also varied systematically across the model. For each unit, we measured the tuning curve for that unit's response as a function of temporal and spatial frequency, with the preferred frequency taken as the tuning curve's peak. Units were generally well tuned, with tuning curves spanning a wide range of temporal and spatial frequencies (Figure 3.3A). The second group G_2 had a lower average preferred spatial frequency than the first group G_1 (t-test, $t(79.4)=3.91$, $p<.001$; Figure 3.3B). The same trend has been found in macaque visual cortex where the average preferred spatial frequency for units recorded in V1[187] was greater than in higher-order areas V2[188] and MT[189] (Figure 3.3C). However, whereas

the preferred spatial frequency monotonically declined for the macaque visual system, there was a U-shaped profile for the model, where the average preferred spatial frequency increased from G_2 to G_3. In contrast, the mean preferred temporal frequency increased across the model's groups similarly to the macaque visual system, with higher model groups and higher-order visual areas both showing greater selectivity to higher temporal frequency stimuli than the first model group (G_1 vs. G_2, t-test, $t(87.7)=-6.07$, $p<.001$; G_1 vs. G_3, t-test, $t(31.7)=-7.32$, $p<.001$) and macaque V1[189], respectively (Figure 3.3D, E).

Finally, we analysed the local connectivity of group one units and found that the network replicated the results previously described for the single-layer recurrent temporal prediction model in chapter two (Figure S3.3).[89] Specifically, we found that short-range connections were most likely for pairs of orientation selective G_1 units when they shared a similar orientation preference (Figure S3.3A) and most likely for pairs of direction selective G_1 units when they shared similar or opposite preferred directions of motion (Figure S3.3B), as is found in mouse V1.[120] Finally, for longer-range connections we found that the connection probability was greatest between units with a similar orientation preference whose receptive fields were aligned co-axially in visual space, again as is found in mouse V1[118] (Figures S3.3C-E).

Thus, overall, the model qualitatively matched the receptive field structure, distribution of modulation ratio and the change in preferred spatial and temporal frequencies across the macaque visual hierarchy.

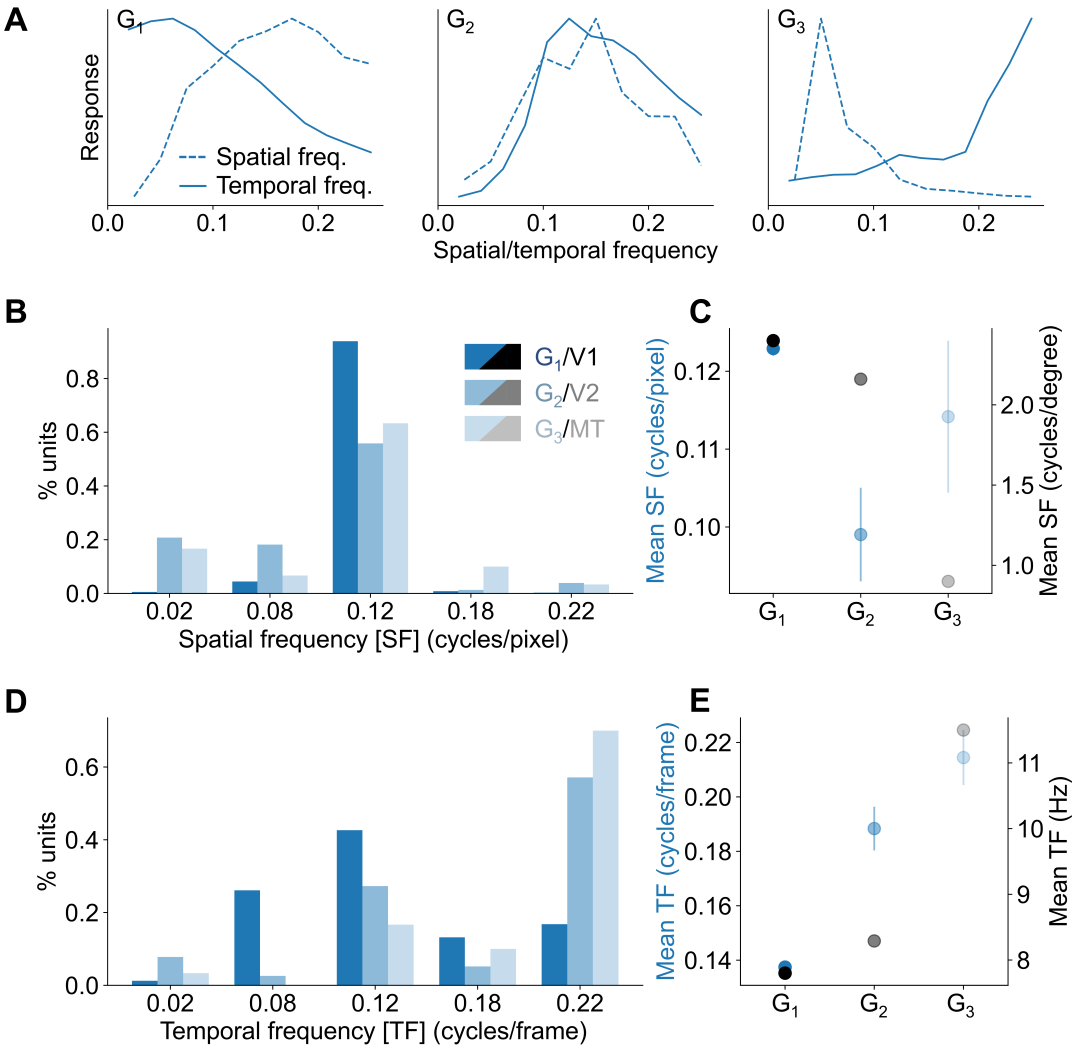


Figure 3.3. Hierarchical recurrent temporal prediction captures spatial and temporal frequency tuning properties across the visual hierarchy.

(A) Exemplar spatial and temporal frequency tuning curves for three units from G_1, G_2 and G_3.

(B) Distribution of preferred spatial frequencies for model units across each group.

(C) Mean preferred spatial frequency for each group compared with macaque V1,[187] V2,[188] and MT.[189]

(D) Distribution of preferred temporal frequencies for model units across each group.

(E) Mean preferred temporal frequency for each group compared with macaque V1,[189] V2,[188] and MT.[189]

Note that, during training, units in higher layers tended to progressively 'die' off – that is, they became inactive and non-responsive to stimuli. Thus, the larger errors bars for higher groups in (C) and (E) could largely be explained by this smaller sample size. In particular, groups 1-3

consisted of 800, 97 and 42 active units each.

Model units mirror the visual system's hierarchy of 2D motion sensitivity

Although direction selectivity is present in V1, neurons in this area are generally unable to represent the complex motion of two-dimensional moving surfaces or patterns.[190] This sensitivity to two-dimensional motion can be probed using plaid stimuli which consist of two overlaid full-field gratings moving in different directions (Figure 3.4Ai). Perceptually, the direction of motion corresponds to the average of the direction of the two components. Accordingly, representing this plaid pattern motion requires integrating the motion signals from the individual grating components. Notably, neurons across the visual system are differentially sensitive to the two-dimensional motion of the plaid versus its individual component gratings (Figure 3.4Ai). For example, in macaque V1, neurons generally respond to the motion of individual components making up the plaid pattern, whereas a proportion of MT neurons represent the overall motion of the plaid pattern.

Individual model units included examples of both kinds of response to plaid stimuli. For component-selective units, the tuning curve displays two peaks corresponding to the direction of motion of components of a plaid (Figure 3.4Bi), resulting in a cross-like contour plot when the direction of motion of each component is varied independently (Figure 3.4Ci). In contrast, pattern-selective units responded with a single peak corresponding to the overall direction of motion of the plaid stimulus (Figure 3.4Bii), resulting in a single response region in the contour plot (Figure 3.4Cii).

The plaid pattern index quantifies this preference, based on the difference in correlation between the unit's true response and an ideal component response versus an ideal plaid pattern response (Figure 3.4Aii). Thus, a higher plaid pattern index indicates greater plaid selectivity. The distribution of plaid pattern index values across model groups recapitulated the hierarchy found in the macaque visual system (V1,[191] V2,[188] and

MT[191]), with the plaid selectivity monotonically increasing from lower to higher groups (G_1 mean=-3.60, G_2 mean=-1.56, G_3 mean=-1.12; one-way ANOVA, $F(2, 931)=61.6$, $p<.001$) (Figure 3.4D). Indeed, no significant difference was found between the means of each group and the corresponding region of the macaque visual cortex (t-test, all $p>0.153$), indicating that progressively greater selectivity for the plaid stimulus compared with the individual grating components was present across both the hierarchical recurrent temporal prediction model and the macaque visual system.

To test whether model units were specifically selective to the direction of plaid motion – as opposed to the plaid's orientation, for example – we measured the plaid pattern direction selectivity of each pattern-selective model unit (where the plaid pattern index is greater than 0). As expected, the majority (73.2%) of pattern-selective units were also pattern-direction-selective, defined as having a direction selectivity index (DSI) greater than 0.3[89] (mean DSI=0.65). Thus, these units were specifically tuned to the direction of pattern motion.

Finally, we investigated the role of feedback in supporting pattern-like responses in model units. Although pattern direction selectivity is more weakly represented in V1, pattern-like motion responses are not entirely absent, which may result in part from feedback to V1 from higher cortical areas. Indeed, suppressing feedback from the middle suprasylvian gyrus has been shown to reduce pattern-like responses in early visual cortex of the cat.[184] In line with these experimental data, abolishing feedback in the model resulted in a significantly lower mean plaid pattern index in G_1 (t-test, $t(1519)=5.33$, $p<.001$; Figure 3.4E), which at the level of individual model units resulted in a change from a plaid-like response (Figure 3.4Fi) to a component-like response (Figure 3.4Fii).

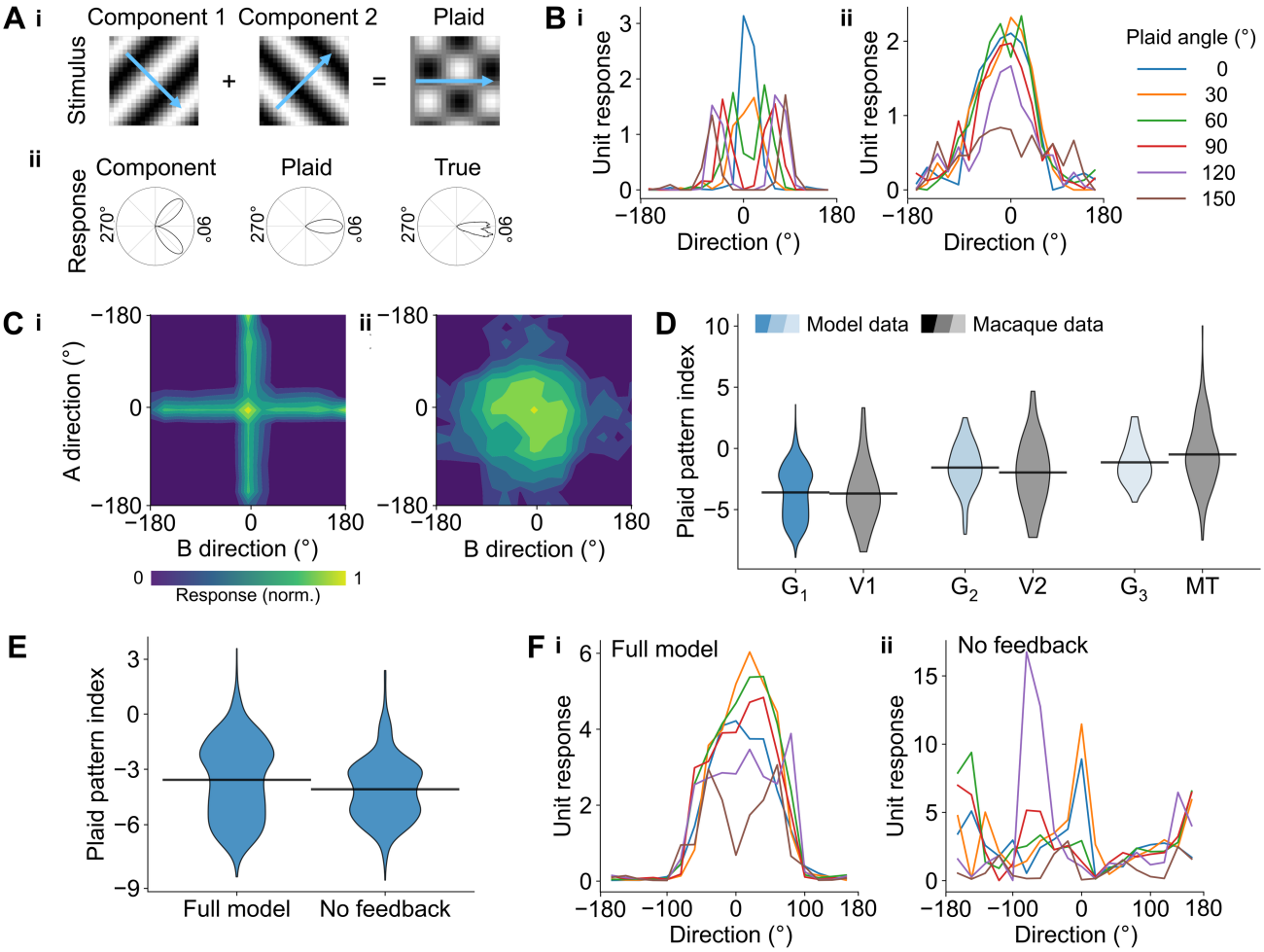


Figure 3.4. Model units mirror the visual system's hierarchy of 2D motion sensitivity.
(A) Construction of plaid stimuli through the additive combination of two grating stimuli (i). The plaid pattern index is computed by measuring the Fisher-transformed difference in the correlation between the true response of the model unit or visual neuron and an idealized component or plaid response (ii).
(B) Example direction tuning curves to plaid stimuli with different plaid angles (the angular separation between the two gratings) for component- (i) and pattern-selective (ii) units.
(C) Example contour plots for component (i) and pattern-selective (ii) units where the angle of each component is independently varied.
(D) Plaid pattern index across the three model groups, G_1, G_2 and G_3, and in macaque V1,[192] V2[188], and MT.[192] A larger plaid pattern index indicates greater response correlation to the theoretical plaid response.
(E) The mean plaid pattern index for G_1 units is reduced when feedback in the model is abolished.
(F) Example direction tuning curves for plaid stimuli across different plaid angles for the same model unit with (i) and without (ii) feedback, demonstrating a change from a pattern-like to a

component-like response when feedback is abolished.

Model units recapitulate feedback-dependent surround suppression

We next investigated the functional role of the model's feedback connectivity with respect to surround suppression. Surround suppression is a hallmark nonlinearity in the responses of V1 neurons and occurs where the neural response is inhibited by stimuli extending beyond the neuron's classical receptive field.[178,179] In the context of predictive coding, surround suppression has been interpreted as a consequence of the statistics of the animal's natural environment.[99] As contours are generally continuous in the environment, it is argued that short discontinuous contours violate the visual system's internal model and produce a larger response via feedback from higher areas, leading to the surround suppression effect. In support of this interpretation, reducing or ablating feedback from higher visual areas to V1 reduces the magnitude of surround suppression.[178,179] Although the temporal prediction model is not trained with explicit error propagation between groups, as in predictive coding models, we asked whether ablating feedback would similarly reduce the surround suppression effect.

In line with the experimental data, surround suppression was significantly reduced when model feedback was abolished. The average G_1 unit activity for the largest stimulus size was 24.8% smaller for the full model compared with the model where feedback was abolished (Figure 3.5Ai; t-test, $t(1432)=-4.48$, $p<.001$), which is comparable to mouse V1 data[179] (Figure 3.5Aii, 33.6% smaller response) though less than that for macaque V1 data[178] (Figure 3.5Aiii, 52% smaller response). This effect was also confirmed by considering the suppression index, which describes the extent to which each unit or neuron is suppressed with increasing stimulus size. The average suppression index was significantly greater for the full model (median=26.2) compared with the model without feedback (median=0.170; Mann-Whitney U test, $U=513498$, $p<.001$), indicating greater surround suppression with the full model (Figure 3.5Bi). This effect was comparable for

both mouse V1 (Figure 3.5Bii) and macaque V1 (Figure 3.5Biii), though the reduction in surround suppression was more modest in the biology.

We further analysed surround suppression in the model with respect to the receptive field properties with and without feedback. Specifically, we analysed the model's G_1 unit responses to stimuli within each unit's classical receptive field (defined as the maximally exciting stimulus size) as well as to stimuli in the proximal receptive field (defined as the ring of visual space area beyond the unit's classical receptive field). In line with data from marmoset V1,[177] the average G_1 model unit's classical receptive field size was significantly larger when feedback was abolished (paired t-test, $t(790)=-29.9$, $p<.001$; Figure 3.5C). We next analysed the average response of G_1 model units for stimuli in each unit's classical receptive field as well as stimuli which extended to include both the classical and proximal receptive fields. As for marmoset V1, the normalized response was significantly smaller for stimuli within each unit's classical receptive field when feedback was abolished (one-sample t-test, $t(790)=-48.3$, $p<.001$; Figure 3.5D). However, unlike in marmoset V1, we found that model G_1 units' average response was suppressed for stimuli spanning the classical and proximal receptive field when feedback was abolished, whereas it increased in marmoset V1 (Figure S3.4).

To better understand the circuit-level basis of surround suppression, we further investigated the structure of feedback weights in the model. Overall, feedback in the model was on average stronger for the inhibitory over the excitatory subpopulations (Figure S3.5A; $t(76)=-3.2$; $p=.001$), suggesting that feedback might impart an overall inhibitory effect on downstream model units. In mouse V1, the impact of feedback from higher visual areas is complex although – in line with these results – there is evidence that feedback has a larger impact in terms of the induced EPSCs among downstream interneuron versus pyramidal cell populations. Given these results, we hypothesised that ablating feedback connections to inhibitory but not excitatory downstream model units would abolish the surround suppression effect. In line with this hypothesis, we found

that surround suppression was significantly reduced when ablating feedback to downstream inhibitory units (median suppression index = 1.6) but not when ablating feedback to downstream excitatory units (median suppression index = 15.2; Mann-Whitney U test, U=396494, p<.001; Figure S3.5B). Thus, these results argue for a mechanism whereby surround suppression is maintained via excitation of lower-order inhibitory units. However, as surround suppression was more fully abolished when feedback to both populations was ablated, the surround suppression effect was not entirely dependent on feedback to inhibitory units alone.

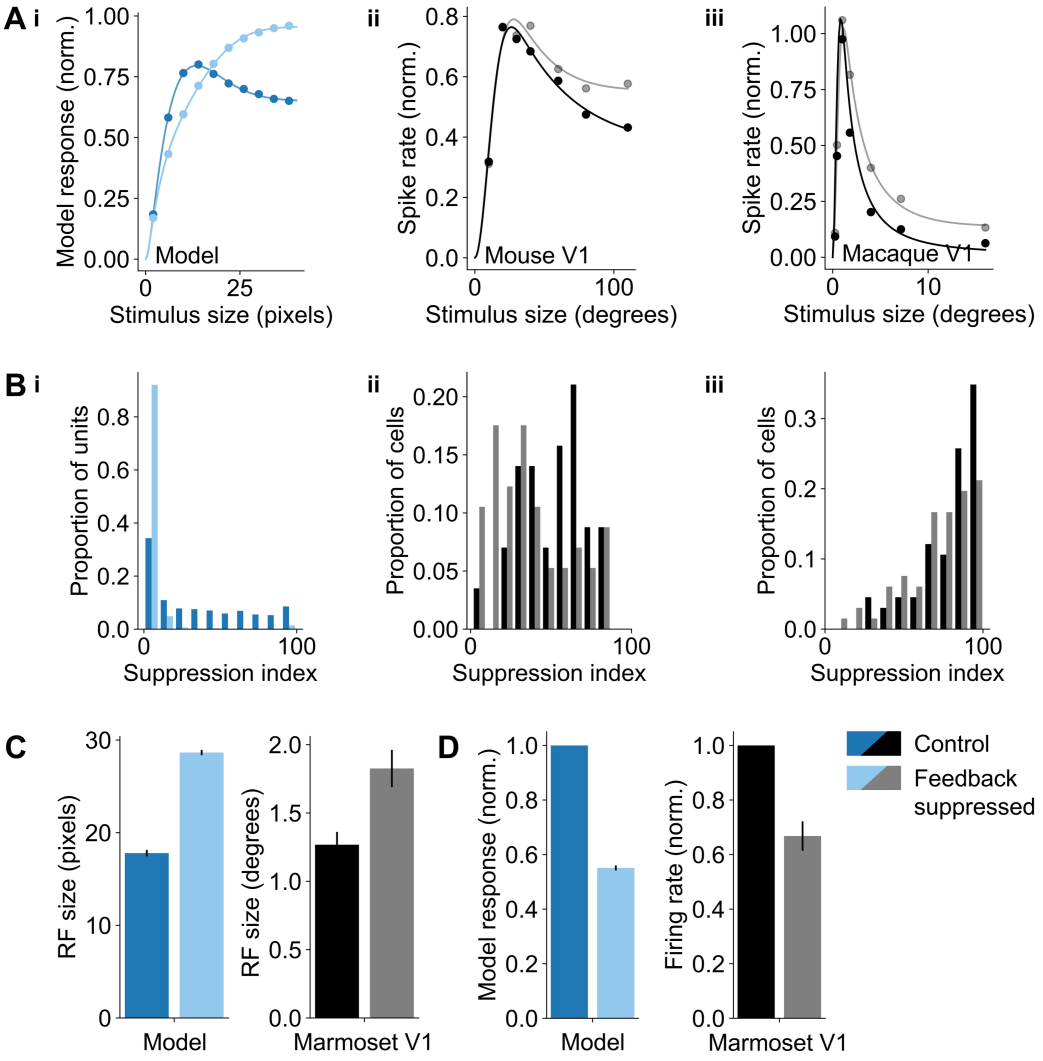


Figure 3.5. The role of feedback connectivity in surround suppression.
(A) Mean response across model G_1 units (i), mouse V1 neurons[179] (ii) and macaque V1 neurons[178] (iii) as a function of stimulus size with and without feedback for model units or with and without inactivation of higher visual areas (mouse) or V2 (macaque).

(B) Distribution of suppression index values for model G_1 units (i), mouse V1 neurons[179] (ii) and macaque V1 neurons[178] (iii) as a function of stimulus size with and without feedback for model units or with and without inactivation of higher visual areas (mouse) or V2 (macaque). A higher index indicates greater surround suppression.
(C) Receptive field size is larger when higher-order feedback is suppressed for both model G_1 units (left) and marmoset V1[177] (right).
(D) Model responses in the classical receptive field are suppressed when higher-order feedback is suppressed for both model G_1 units (left) and marmoset V1[177] (right).
See also Figures S3.4, S3.5.

The hierarchical recurrent temporal prediction model best captures cortical stimulus representations compared with alternative models

We next compared how well different normative models could recapitulate the response properties across the visual pathway in terms of the modulation ratio, plaid motion selectivity and the presence of surround suppression. To probe the impact of recurrency while maintaining a hierarchical representation, we compared the model to a purely feedforward hierarchical temporal prediction network. Unlike for the recurrent network, where temporal information was maintained via recurrency, we explicitly passed in the preceding four frames of stimulus history with each frame for the feedforward network in order to include a temporal component. Similarly, to assess the role of the training objective, we compared the hierarchical recurrent temporal prediction model with a hierarchical recurrent autoencoder. Thus, the hierarchical recurrent autoencoder was identical to the hierarchical recurrent temporal prediction model in all aspects, except that it was trained to reproduce its input frame rather than predicting the subsequent frame. For each model, we quantified the difference between the experimental and model distributions of each measure using the Kolmogorov-Smirnov distance, where a lower value indicates that the two distributions are more similar.

The distribution of modulation ratio values in the hierarchical recurrent temporal prediction model was closer to those measured for macaque V1 and V2 compared with the alternative comparison models (Figure 3.6Ai, ii). The autoencoder exhibited weak bimodality among the first group of units, but unlike macaque V2, the responses were

skewed towards a modulation ratio greater than one in the second group. In contrast, the feedforward temporal prediction model produced only simple-cell-like units in the first group, with complex-cell-like responses emerging only in the second group. Thus, recurrency was required for the emergence of simple- and complex-cell-like units within the first group.

Considering plaid motion sensitivity, we found a monotonic increase in the mean plaid pattern index across groups in both the feedforward and hierarchical recurrent temporal prediction models, which is consistent with the hierarchical organization of plaid pattern motion sensitivity found in the macaque visual system (Figures 3.6Bi, Bii). However, for the feedforward model, the population mean across groups was significantly lower than the experimental data (t-test, $t(1790)=28.1$, $p<.001$), with very few pattern-like responses. In contrast, in the autoencoder model, there was relatively little variation in pattern-selectivity over component-selectivity as a function of model group, although the mean pattern index was overall greater than that found in macaque V1 or V2 (t-test, $p<.022$; Figure 3.6Biii). Thus, the emergence of pattern-selectivity across the visual pathway was best captured by the hierarchical recurrent temporal prediction model.

For surround suppression, we first compared the mean response as a function of stimulus size with and without feedback for the hierarchical recurrent temporal prediction and autoencoder models (Figure 3.6Ci). For the full models with feedback, surround suppression was greater for the autoencoder model than for the temporal prediction model, although the response for the autoencoder was very weak once feedback was abolished. In contrast, for the feedforward temporal prediction model, the mean tuning curve showed no evidence of surround suppression for any group, though the mean response increased with model depth (Figure 3.6Cii). These results were further confirmed by comparing the distribution of surround suppression across all models (Figure 3.6Ciii). Although the bimodal distribution of the hierarchical recurrent temporal prediction model was qualitatively distinct from the distribution of these

values in mouse V1, the mean suppression index (mean=55.1) was closer to the V1 data (mean=37.2), and the Kolmogorov-Smirnov distance was much smaller than for the other models (Figure 3.6Civ). The mean suppression index was larger for the autoencoder model (mean=84.8), where fully suppressed units were overrepresented, whereas only a few units showed any degree of surround suppression in the feedforward model (mean=7.3). Thus, the emergence of robust surround suppression required recurrent interactions in the model, and emerged only weakly in the feedforward model.

To further relate the networks' learned representations to the visual system, we quantified the representational similarity of each model to different regions of the mouse visual cortex. Specifically, for each model we computed the response similarity matrices – the correlation of model or neural activity across pairs of stimuli – before measuring the distance between these similarity matrices (Figure S3.6). In this way, representational similarity gives a population-level measure of model-brain alignment. For all representational similarity analyses, the reported values are for the highest performing group or layer across each network (Figures 3.6D, E).

In terms of brain regions, we considered mouse V1 as well as two early regions of the ventral (rostrolateral visual cortex) and dorsal (lateromedial visual cortex) visual streams. In the primate visual system, the dorsal- and ventral-streams are characterized by their relative selectivity to motion and form respectively.[14] The mouse visual system similarly features a dorsal- and ventral-like stream[20,193] and while the distinction between these streams is less well characterized, they feature distinct motion tuning properties as in the primate visual system.[194] Accordingly, these areas provide an opportunity to investigate how the motion tuning of the temporal prediction model relates to its model-brain alignment.

For each of these cortical regions, the representational similarity of the hierarchical recurrent temporal prediction model exceeded that of both the feedforward model (paired t-test, all p<.023) and the autoencoder model (paired t-test, all p<.011; Figure

3.6D). We also compared the full model against an untrained variant and again found that the hierarchical recurrent temporal prediction model had higher similarity scores across all cortical areas, with significant differences in V1 (paired t-test, $t(13)=4.01$, $p=.001$) and rostrolateral visual cortex (paired t-test, $t(12)=7.31$, $p<.001$), but not in lateromedial visual cortex (paired t-test, $t(9)=2.26$, $p=.050$; Figure 3.6D). Given the greater motion sensitivity of the dorsal stream, we hypothesized that the hierarchical recurrent temporal prediction model would perform better as a model of a dorsal region (rostrolateral visual cortex) versus a ventral region (lateromedial visual cortex) relative to a model optimized on static images (Figure 3.6E). Accordingly, we compared the hierarchical recurrent temporal prediction model to a model trained for object recognition – namely, VGG-19.[145] In support of this hypothesis, we found no difference in the representational similarity score between the temporal prediction model and VGG-19 across V1 (paired t-test, $t(13)=-0.70$, $p=.499$) and lateromedial visual cortex (paired t-test, $t(9)=-0.41$, $p=.694$). In contrast, the temporal prediction model performed significantly better than VGG-19 for the dorsal rostrolateral visual cortex (paired t-test, $t(12)=2.60$, $p=.023$).

Thus, overall, we found that the distribution of the tuning parameters and model-brain alignment was best captured by the hierarchical recurrent temporal prediction model.

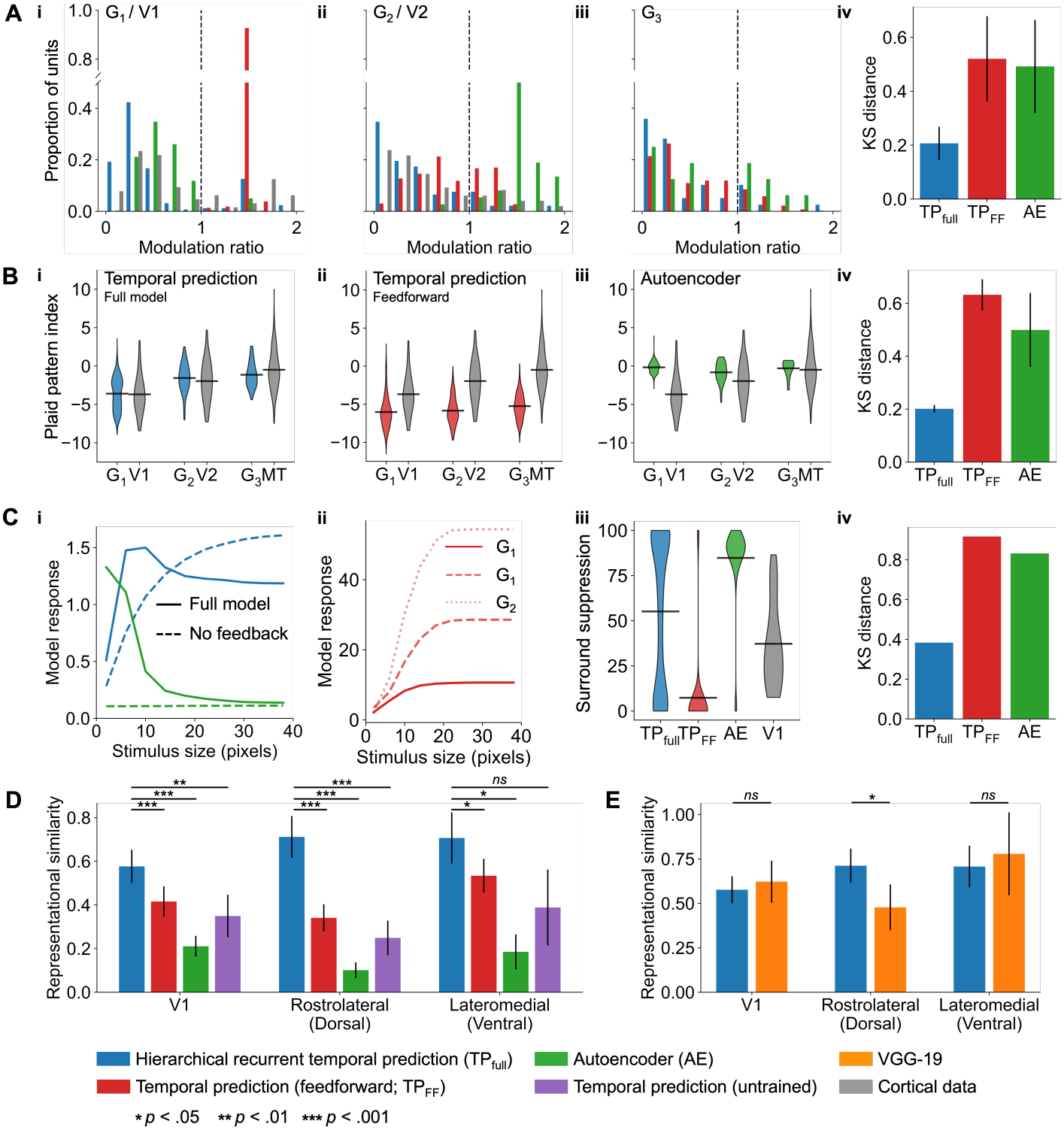


Figure 3.6. Comparison of tuning properties across the models.
(A) Distribution of modulation ratio values across models units and experimental data (i: G_1 vs macaque V1, ii: G_2 vs macaque V2, iii: G_3) and the Kolmogorov-Smirnov (KS) distance from the experimental data (iv).
(B) Distribution of plaid pattern index values in macaque V1, V2 and MT and across groups for the recurrent temporal prediction model (i), feedforward temporal prediction model (ii),

autoencoder model (iii), and the KS distance between these models and macaque data (iv). **(C)** Surround suppression among G_1 units for the hierarchical recurrent temporal prediction and autoencoder models with and without feedback ablated (i), for each group of the feedforward temporal prediction model (ii), the distribution of surround suppression index values for each model and mouse V1 (iii) and the KS distance between each model and mouse V1 (iv). **(D-E)** Representational similarity of the recurrent temporal prediction model relative to the alternative normative models with mouse V1, rostrolateral and lateromedial visual cortex. See also Figure S3.6.

Discussion

The principle of temporal prediction argues that the sensory brain is optimized to represent the immediate future based on the recent past. In line with this principle, we hypothesized that a hierarchical recurrent model optimized for temporal prediction should recapitulate the functional organization of visual cortex. With little fine tuning to the model, the network exhibited tuning properties akin to those found across several visual cortical areas, including not only the distribution of simple- and complex-like responses found in macaque V1 and V2, but also feedback-dependent surround suppression and the emergence of global motion sensitivity across those visual regions leading to area MT. Compared with alternative normative models, the response properties of visual cortical neurons were best captured by the hierarchical recurrent temporal prediction model, which similarly exhibited the closest alignment with the stimulus representations found in mouse visual cortex. Together, these results provide evidence for temporal prediction as an organizational principle across the visual cortex.

Relation to biology

As a normative model, the hierarchical recurrent temporal prediction model represents a fairly abstract representation of the visual brain,[63,72] rather than focusing on low-level mechanistic details. Nevertheless, there is a careful balance to be made between incorporating sufficient details to make meaningful comparisons possible, while avoiding potentially redundant features that could obscure the role of the more general normative principle of interest. In the case of the current temporal prediction model, the network's units obey Dale's law – projecting exclusively excitatory or inhibitory

connections – and incorporate both local recurrent and feedback connectivity. Indeed, these fundamental properties of neural circuits are generally omitted from normative models of sensory processing, particularly those implemented as deep object recognition networks.[112,175]

However, as point-like neurons, the model's units are fairly abstract representations of true biological neurons. That is, they lack distinct neuronal compartments, and their activity is rate based and lacks spiking behaviour. Nevertheless, these neuronal properties are fully compatible with temporal prediction as a principle, and implementing them in other studies has reproduced the action potential firing patterns and membrane time constants of neurons.[147,195] Finally, the model is trained via backpropagation which is generally considered to be biologically implausible.[196] However, the model itself is agnostic about the learning rule – certain elements of temporal prediction could be genetically hard-wired or arise via competitive activity-dependent mechanisms during development – and novel more biologically-plausible algorithms could, in principle, be applied to the current network.[110,160]

Comparison with alternative normative models

In the current study, we compared the hierarchical recurrent temporal prediction model to several alternative models to better understand how each model's features related to their capacity to capture different elements of the visual system.

The feedforward temporal prediction model incorporated both a hierarchical organization and the temporal prediction objective – that is, each higher group predicted the activity of its lower order inputs. However, in contrast to the full temporal prediction model, the feedforward model did not include local recurrency or feedback connectivity. Notably, plaid motion selectivity and surround suppression were considerably weaker than in the recurrent model, and the feedforward model performed worse in terms of its representational similarity to the visual cortex. Accordingly, the architectural addition of local and long-range recurrency had a large impact in improving the model's fit to the

visual system. Similarly, in terms of the model objective, the hierarchical recurrent autoencoder model performed worse on the response similarity measure and did not match the visual system's response properties as well as the temporal prediction model. Thus, both the temporal prediction objective, as well as the model's architecture, were important in producing a brain-like model.

Several other studies have also modelled the visual system by applying unsupervised learning principles to dynamic spatiotemporal inputs in a hierarchical manner. In particular, predictive coding has been influential as another hierarchical normative framework which has been applied to sensory processing. Predictive coding argues that brain is optimized to reduce statistical redundancies by passing forward only the residual prediction errors that cannot be "explained away" by the brain's internal model.[126] Predictive coding is inherently hierarchical in that it is organized around a series of stacked prediction modules, which – like temporal prediction – is consistent with a hierarchical view of cortical organization. The original model by Rao and Ballard[99] was able to account for non-linear effects in V1, such as surround suppression, though it was only trained on static images. In contrast, a more recent variant on predictive coding – PredNet[146] – has also been trained on dynamic natural movies. While PredNet could account for some properties such as surround suppression,[197] it does not generally recapitulate low-level features of the visual cortex, such as the Gabor-like receptive fields of V1,[89] nor has it been shown to capture the hierarchical organization of the visual system's tuning properties.

A number of studies have also investigated how well dual-stream networks can model both the dorsal and ventral streams when trained on dynamic moving visual inputs. Bakhtiari et al.[198] demonstrated how a dual-stream network trained for contrastive temporal prediction can develop dorsal stream-like representations that recapitulate some elements of motion processing, such as random dot kinematogram selectivity. Although more abstract in its form than the hierarchical recurrent temporal prediction

model, their contrastive model underscores the value of temporal prediction as an unsupervised principle for modelling sensory areas. However, unlike in this study, their contrastive objective is trained to predict the latent state of the future frame rather than the pixel-wise future frame itself. Considering the successes of contrastive methods elsewhere,[16] it will be instructive to explore this form of temporal prediction in future work. Nevertheless, their model – as is common with other contrastive networks – relies on a pretrained ResNet backbone and therefore is not truly unsupervised in the same manner as the models described in this study.

Finally, other studies employing a dual-stream network, such as Cadieu and Olshausen,[199] have described how optimizing a dual-pathway network – in their case to decompose inputs into amplitude and phase information – can produce ventral- and dorsal-like representations. After training, the optimized network produces simple- and complex-cell-like responses, as well as units with various kinds of form and motion invariance. Although their model lacks recurrency and is limited in the range of physiological properties that were investigated relative to the current study, it illustrates the utility of such dual-stream models, which could be integrated into the temporal prediction framework in future.

In conclusion, we have shown that a hierarchical recurrent network optimized for temporal prediction replicates many aspects of the neuronal response properties and functional organization of different areas within the mammalian visual cortex. These findings add to the growing evidence that temporal prediction can account for information processing across the sensory hierarchy.

Methods

Model dataset

The model was trained using a dataset drawn from around 2.5 hours of wildlife videos down sampled to 25-30 frames per second. Videos were pre-processed by converting them to grayscale, resizing each clip to a width and height of 600 pixels using bilinear interpolation and applying a bandpass filter. Each video was then cropped into 1200 spatially overlapping clips of 20x40 pixels at 40 contiguous frames each. Finally, each video clip was normalized by subtracting its mean and dividing by its standard deviation. This produced a training, test and validation dataset of 1000k, 200k and 200k clips, respectively.

The hierarchical recurrent temporal prediction model

The model is implemented as a recurrent network consisting of a series of hierarchically organized groups of units (Figure 3.1A). During training, each group in the model is optimized to predict the future value of its inputs – specifically, the future video frame for group one (G_1) units or the future internal state of the lower-order inputs for each group $n > 1$. Internally, the model is implemented as a single recurrent layer, with the hierarchy defined by 1) the hierarchy of prediction whereby higher-order areas are trained to predict the future value of their lower-order inputs and 2) by the restricted internal connectivity such that each group receive inputs from and projects to its immediate higher- and lower-order groups only. This internal hidden state is then mapped to the predicted future state by a linear output layer, with the difference between the true and expected future state minimized during training via backpropagation. Finally, an L1 regularization penalty in the cost function ensures that only weights that contribute to network performance are retained.

More formally, the model receives as input a tensor U of shape $T \times I$ where t=1 to 40 timesteps and i=1 to 800 pixels as a flattened 20×40 pixel image. The network itself

consists of 3 groups g=1 to 3, where each group g comprised j=1 to 800 units. The activity of each hidden unit h^g_{jt} is defined as:

$$a^g_{jt} = b^g_j + \sum_{j'=1}^{J^g}(r^{g,g}_{jj'} \cdot h^g_{j't-1}) + \begin{cases} \sum_{i=1}^{I}(w_{ji} \cdot u_{it-1}) + \sum_{j'=1}^{J^2}(r^{1,2}_{jj'} \cdot h^2_{j't-1}) & g = 1 \\ \sum_{j'=1}^{J^{g-1}}(r^{g,g-1}_{jj'} \cdot h^{g-1}_{j't-1}) + \sum_{i=1}^{I^{g+1}}(r^{g,g+1}_{jj'} \cdot h^{g+1}_{j't-1}) & 1 < g < G \\ \sum_{j'=1}^{J^{g-1}}(r^{g,g-1}_{jj'} \cdot h^{g-1}_{j't-1}) & g = G \end{cases}$$

$$h^g_{jt} = \mathrm{ReLU}(a^g_{jt})$$

where b^g_j is the bias of group g unit j, w_{ji} is the input weight between pixel i and group one unit j, $u_{it\text{-}1}$ is the activity of pixel i at the previous timestep, $r^{g,g}_{jj'}$ is the recurrent weight between presynaptic unit j' in group g to postsynaptic unit j in group g, and $h^g_{j't\text{-}1}$ is the activity of presynaptic unit j' in group g.

In addition, to enforce Dale's Law whereby units have exclusively excitatory or inhibitory connections, each recurrent weight $r^{g,g}_{jj'}$ was clamped after each forward pass as:

$$r^{g,g}_{jj'} = \begin{cases} +|r^{g,g}_{jj'}| & \text{if excitatory} \\ -|r^{g,g}_{jj'}| & \text{if inhibitory} \end{cases}$$

where 20% of units in each group were set as inhibitory (based on the approximate percentage of cortical inhibitory interneurons found in the brain[164,165]), with the additional constraint that inhibitory units could only project locally.

Finally, the network's internal hidden state was mapped to the output predictions as y^g_{kt} for group g, unit k at time t, which was defined as:

$$y^g_{kt} = b^g_k + \sum_{j=1}^{J^g} m^{g,g}_{kj} \cdot h^g_{jt}$$

where b^g_k is the bias for unit k of group g and $m^{g,g}_{kj}$ is the weight between hidden unit h^g_{jt} and output unit y^g_{kt}.

The trainable parameters in the network were optimized by minimizing the loss function:

$$\text{Loss} = \quad \cdot \text{L1} + (1 \qquad) \cdot E^1 + \sum_{g=2}^{G} \quad \cdot E^g$$

where *L1* is sum of the absolute value of all weights in the network, λ is a weighting parameter that describes the degree of regularization, and β is a weighting parameter that determines the relative contribution of the prediction error between the first and higher-order groups. Finally, E^g is the mean squared error between the true and predicted future value of the lower-order group *g-1*:

$$E^g = \frac{1}{NTK^g} \sum_{n=1}^{N} \sum_{t=4}^{T-1} \sum_{k=1}^{K^g} \begin{cases} (y_{knt}^1 - u_{knt+1})^2 & g = 1 \\ (y_{knt}^g - h_{knt+1}^{g-1})^2 & g > 1 \end{cases}$$

Where, for group one, $y_{knt}^1 - u_{knt+1}$ is the difference between the predicted and true future pixel value, and, for higher order groups, $y_{knt}^g - h_{knt+1}^{g-1}$ is the difference between the predicted and true future value of the lower order group. Finally, E^g is averaged across all clips *N*, timesteps *T* and pixels *I*. Note that the subscript *n*, which indicates clip number, was left off from all previous equations for brevity.

Comparison models

Three comparison models were developed in addition to the hierarchical temporal prediction model as described above.

1. Feedforward temporal prediction model – this model was trained to maximize the same loss function as the hierarchical temporal prediction model but using a strictly feedforward architecture. Unlike for the other recurrent models, at each time point, the network was provided with five frames of input – the current frame, and the preceding four frames – in order to include a temporal component. The training and architectural details were adapted from previous work,[71] though only three stacks were used in this chapter, and adapted the

kernel sizes as detailed below to equate the receptive field sizes in the hierarchical recurrent models:

Table 3.1. Architecture for the feedforward temporal prediction model.

Stack	Input size	Hidden layer size	Kernel size
0	1x40x800	800x36x1	5x800
1	800x36x1	800x32x1	5x1
2	800x32x1	800x28x1	5x1

2. Hierarchical recurrent autoencoder – the same architecture as the hierarchical recurrent temporal prediction network but with the loss function altered to predict the activity of the current rather than future lower-order group, with an additional regularization term, weighted by the λ_α hyperparameter, based on the summed activity in the network to ensure the sparsity of the network's internal representations.

$$\text{Loss} = \lambda \cdot \text{L1} + \lambda_\alpha \cdot \sum_{g=1}^{G} \sum_{j=1}^{J^g} h_{jt}^g + (1-\beta) \cdot E^1 + \sum_{g=2}^{G} \beta \cdot E^g$$

$$E^g = \frac{1}{NTK^g} \sum_{n=1}^{N} \sum_{t=4}^{T} \sum_{k=1}^{K^g} \begin{cases} (y_{knt}^1 - u_{knt})^2 & g = 1 \\ (y_{knt}^g - h_{knt}^{g-1})^2 & g > 1 \end{cases}$$

This network similarly included three groups of 800 units each as for the recurrent hierarchical temporal prediction model.

3. Untrained hierarchical recurrent temporal prediction model – the same architecture as described above, but with no training and hence randomly initialized weights.

Finally, one other pre-trained model developed by another group was also included to compare its capacity to capture the neural representation of naturalistic visual stimuli:

1. VGG-19[145] – a deep convolutional network trained on the ImageNet database to classify images into one of 1000 categories. This network is purely feedforward and does not incorporate any form of recurrency.

Implementation

The temporal prediction model and its variants were all implemented in PyTorch, with gradient descent performed using the ADAM optimizer set at a learning rate of $\alpha = 10^{-4}$. The hyperparameter λ was set as 10^{-6}, as this value was close to the global minimum which minimized the mean squared error for next frame prediction on the validation set while producing the most biologically realistic receptive fields. The hyperparameter β was set at 0.1. Note that for Figure 3.1C, D, where a large number of model variants were investigated, for computational efficiency, these models were trained on a smaller 20x20 pixel dataset,[89] rather than the 20x40 pixel dataset used throughout the rest of the study. All networks were trained for a maximum of 4000 epochs or until the validation loss converged on a steady state.

Data analysis

Receptive field estimation. Receptive fields were estimated using the response-weighted average, by taking the weighted response of each unit in the network to 20,000 frames of random Gaussian noise.

Unit tuning characteristics. To assess each unit's tuning properties, we measured its response to sinusoidal gratings varying in temporal frequency, spatial frequency and orientation for 40 frames. Each unit's preferred temporal frequency, spatial frequency and orientation was taken as the parameter combination that maximized the unit's mean response across frames.

Modulation ratio. The modulation ratio was computed as $F=F_1/F_0$, where F_0 is the mean response of the neuron to its preferred stimulus and F_1 is the amplitude of the fitted sinusoid to the neuron's response to its preferred stimulus.

Orientation and direction selectivity. Orientation selectivity was quantified via the orientation selectivity index (OSI):

$$OSI = \frac{R^{\mathrm{Or}}_{\mathrm{pref}} - R^{\mathrm{Or}}_{\mathrm{orth}}}{R^{\mathrm{Or}}_{\mathrm{pref}} + R^{\mathrm{Or}}_{\mathrm{orth}}}$$

where $R^{\mathrm{Or}}_{\mathrm{pref}}$ and $R^{\mathrm{Or}}_{\mathrm{orth}}$ are the unit responses at the preferred and orthogonal orientations for that unit. Similarly, direction selectivity was quantified via the direction selectivity index (DSI):

$$DSI = \frac{R^{\mathrm{Dir}}_{\mathrm{pref}} - R^{\mathrm{Dir}}_{\mathrm{opp}}}{R^{\mathrm{Dir}}_{\mathrm{pref}} + R^{\mathrm{Dir}}_{\mathrm{opp}}}$$

where $R^{\mathrm{Dir}}_{\mathrm{pref}}$ and $R^{\mathrm{Dir}}_{\mathrm{opp}}$ are the unit responses at the preferred and opposite directions for that unit.

Plaid responses. Plaid stimuli were produced by summing two drifting gratings, each at half amplitude (0.5). The gratings differed by a 'plaid angle' – the angular separation of each plaid component ($|\alpha - \beta|$) around the overall direction of the plaid stimulus ($\frac{\alpha+\beta}{2}$). To identify the degree to which units responded to the direction of the individual plaid components versus the overall direction of the plaid stimulus, we computed the partial correlation of the unit's response to the plaid stimulus with an idealized response assuming the unit were entirely component or plaid selective.[192] Thus, the plaid correlation r_p was defined as the correlation between the true plaid response and the predicted plaid response, controlling for the predicted component response. Similarly, the component correlation r_c was taken as the correlation between the true component response and the predicted component response, controlling for the predicted plaid response. To compare across units, we computed the Z-transforms Z_p and Z_c of r_p and r_c using Fisher's Z-transformation. This process was repeated across component separation angles of 60, 90, 120 and 150 degrees. The plaid pattern index was then taken as the difference between the average Z_p and Z_c values across all component separation angles.

A value greater than 0 indicates greater selectivity for the plaid stimulus over its individual components, whereas a value less than 0 indicates greater selectivity to the individual components over the composite plaid stimulus.

Surround suppression. For each G_1 model unit, we first fitted a Gabor function to each unit's response-weighted average to extract the receptive field centre. We then presented a drifting grating stimulus determined by the unit's preferred orientation, spatial and temporal frequency, with a mask applied to localize the grating stimulus to the unit's receptive field centre. We recorded the response of that unit as the diameter of the mask was increased up to 40 pixels. We calculate the surround suppression index based on this surround suppression tuning curve as:

$$\text{surround suppression} = 100 \cdot \frac{r_{max} - r_{large}}{r_{max}}$$

where r_{max} is the unit's maximum response across all grating diameters and r_{large} is the unit's response to the large diameter stimulus.

Representational similarity analysis

Neural data were taken from the Allen Brain Institute's Neuropixels Visual Coding dataset of *in vivo* electrophysiological recordings of the mouse brain.[200] All recordings were pre-processed and spike-sorted by the Allen Institute. For these analyses, only data from the natural movie presentations were included ("Natural Movie One" and "Natural Movies Three", 150 seconds total). All data were from wildtype mice with recordings from V1 (n=15 mice), rostrolateral visual cortex (n=14 mice) and lateromedial visual cortex (n=11 mice) used. In addition, only those units whose noise power to signal power ratio did not exceed 60 were included.[166]

To calculate the representational similarity between each model and the neural data, we first computed the representational similarity matrix (RSM).[198,201] For this, we binned the responses of the model or real neurons into 5-frame chunks and took the mean of each bin such that each model's output was represented as an $M \times N$ matrix with M rows

of stimuli and N columns of units. The RSM was calculated as the correlation between each pair of columns in the matrix (the response to a given stimulus m), resulting in an $M \times M$ RSM. The representational similarity was then calculated by taking the distance between the flattened lower triangle of the RSMs of each model and the neural data using Spearman's correlation coefficient.[202]

To compare across mice, this similarity measure was normalized by dividing by the noise ceiling for that recording, where a value ≥ 1 indicates that model similarity is as high as could be achieved once accounting for random noise.[198] Thus, the noise ceiling describes the theoretical maximum similarity that could be achieved by even a perfect model given the inherent noise present in the neural system. It was computed by randomly dividing the neural dataset into two halves of units and by taking the similarity of the RSMs for each half of units. This shuffling process was then repeated for 100 iterations, with the noise ceiling taken as the mean similarity value across these 100 shuffle iterations.

Supplemental information

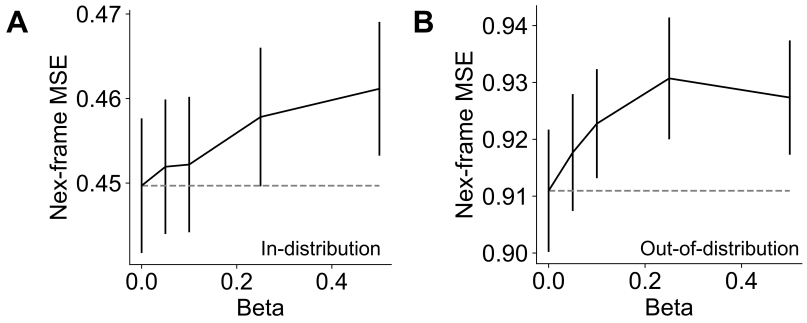


Figure S3.1. Next-frame prediction performance as a function of the beta hyperparameter, Related to Figure 3.1.
Next-frame prediction error increased as a function of beta for both the in-distribution **(A)** and out-of-distribution datasets **(B)**.

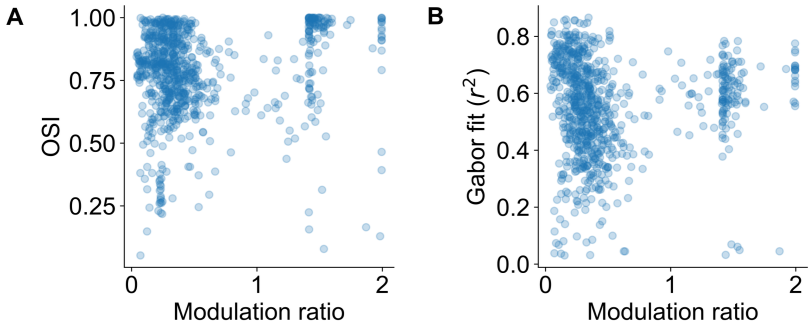


Figure S3.2. Joint distribution of modulation ratio with the orientation selectivity index (OSI) and Gabor fit r^2 for the first model group, Related to Figure 3.2.
The simple-cell-like units (modulation ratio > 1) had higher mean orientation selectivity (**A**) and were better fitted by the Gabor function (**B**) than the complex-cell-like units.

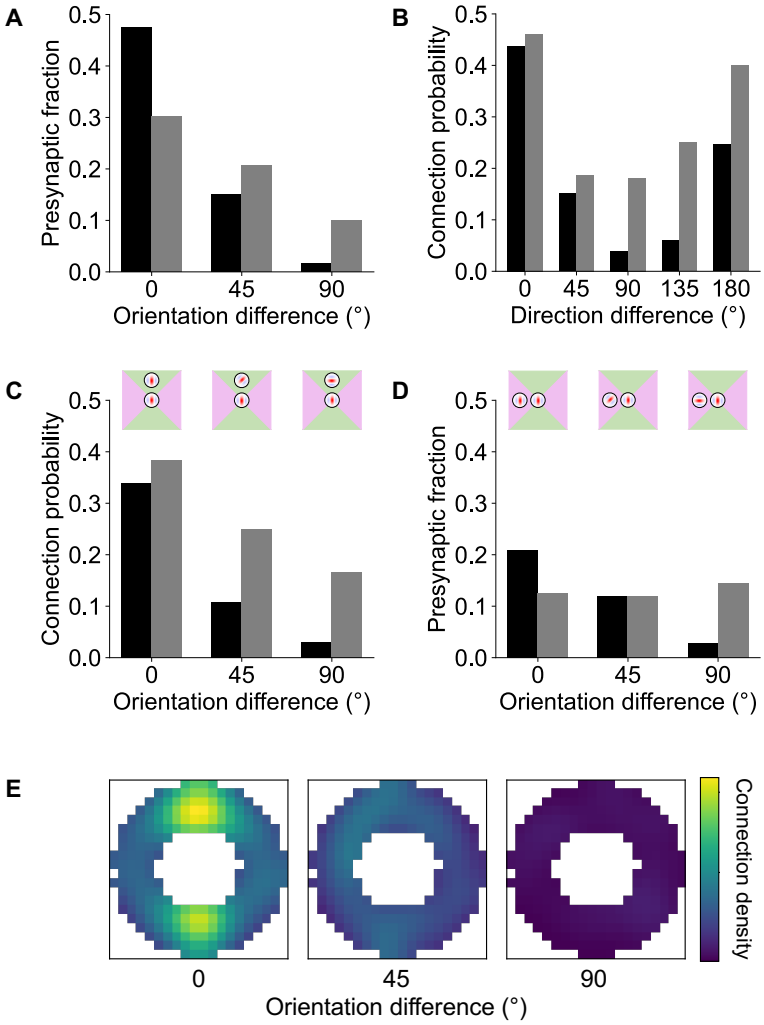


Figure S3.3. Functional connectivity of group one model units, Related to Figure 3.2.
A-B) Short-range connections between model units are more prevalent for excitatory units that have similar orientation tuning (A) and direction-tuned units that have similar or opposite preferred directions of motion (B), as is also the case in V1.[120]
(C-E) In both the model and V1,[118] long-range connection probability is higher for units with similar orientation preferences when their receptive fields are located in co-axial (C) than in co-orthogonal (D) locations. Heatmap (E) shows the normalized connection probability over visual space across differences in orientation tuning for model units.
For detailed methods, see chapter two.

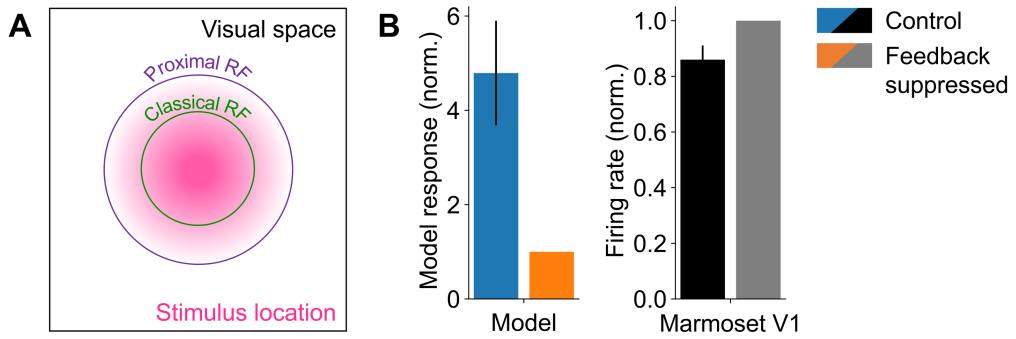
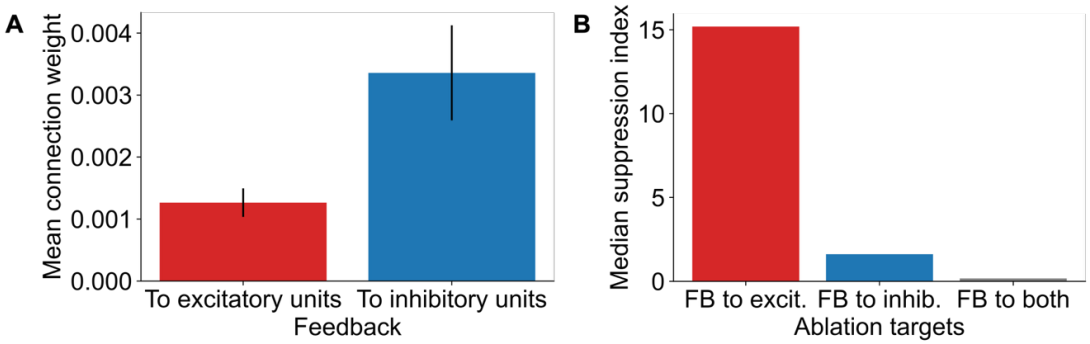


Figure S3.4. Model and V1 response to stimuli in the classical and proximal receptive field, Related to Figure 3.5.

(A) Diagram of the classical and proximal receptive field for an exemplar neuron.

(B) The full model's average response rate is higher for stimuli in the classical and proximal receptive field compared with the model when feedback is suppressed. This contrasts with the data from marmoset V1, where the opposite trend is observed, such that the average neural response for stimuli in the classical and proximal receptive field is greater when feedback is suppressed.[177]



Supplemental Figure 3.5. Role of excitatory and inhibitory feedback in surround suppression, Related to Figure 3.5.

(A) Mean feedback weights from each G2 unit to excitatory and inhibitory G1 units.

(B) Median surround suppression index did not significantly differ when ablating feedback to excitatory versus inhibitory subpopulations.

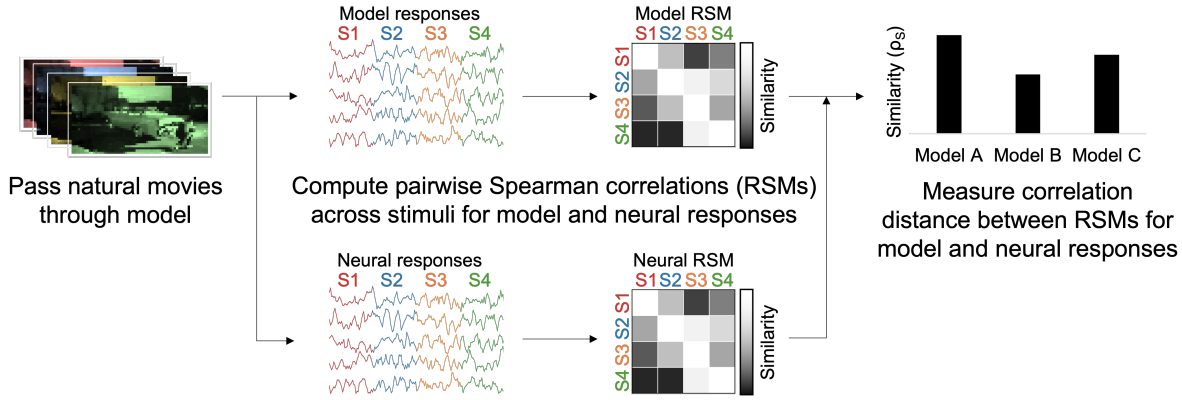


Figure S3.6. Schematic of the representational similarity analysis procedure, Related to Figure 3.6.

Representational similarity was computed by first calculating the representational similarity matrices (RSMs) for the responses of model units and mouse visual neurons to the natural movie stimuli. The RSMs consisted of the pairwise similarity of model units or visual neurons to each set of stimuli. Finally, the representational similarity was computed as the distance between these RSMs, calculated as the Spearman correlation between the lower triangle of each RSM.

4

Exploring the impacts of modelling choices on predicting neural responses across the auditory pathway

Introduction

Why does the brain function as it does? Within cognitive and systems neuroscience, normative modelling provides one approach to answering this question, by providing a conceptual framework for testing computational theories about neural function.[63,72,73] These models are often described as 'top-down' in the sense that they characterize the target system in terms of a higher-level goal or function, and thus provide an optimization perspective to answer 'why' questions about the brain.[63,72,203]

A particular appeal of normative models is their capacity to explain a wide variety of experimental data via a few fundamental computational principles. The findings in chapter two this thesis, for example, demonstrated that models optimized for temporal prediction can account for many aspects of functional connectivity in V1, while chapter three showed how temporal prediction can account for feedback-dependent response properties among neurons across the dorsal visual pathway. These results imply that much of the complexity of neural sensory systems could be an emergent phenomenon as a result of optimizing for some comparatively simpler normative principle. In this way,

these models provide one mechanism for tackling the complexity of the brain by identifying the fundamental computational processes which drive the brain's organization.

Normative network models have proven particularly influential as models of sensory cortex, where they are among the state of the art for predicting neural responses.[73,204] Beyond their raw predictive power, an implicit appeal of these models is the potential insight they offer into the computations underlying sensory processing. However, the exact relation between model learning objectives and the supposed computational function of the neural system is not always clear. For example, in the context of modelling the auditory and visual systems, existing large scale comparisons have demonstrated that many different models often perform roughly equivalently in predicting neural responses across varied learning objectives, model architectures and training datasets.[16,204] Moreover, existing approaches often compare off-the-shelf models co-opted from an engineering context without considering the impact of these hyperparameter choices.[72] Together, these two considerations make it difficult to conclude how any particular variable relates to the model's characteristics and, in turn, how these relate to different hypotheses about brain function. In other words, the presence of multiple confounding variables undermines the stated goal of explaining neural function at a normative level.

To better disentangle these variables, we systematically assessed the impact of different hyperparameters on network representations as well as the capacity to predict the responses of auditory neurons to natural sounds. In this chapter, we therefore move from the visual to the auditory system to provide a comparative approach and to test the generality of these normative investigations across different sensory modalities. In particular, we analysed neural responses from across the auditory pathway – using recordings from the inferior colliculus (IC) as well as primary (A1) and non-primary (posterior ectosylvian gyrus [PEG]) auditory cortex of ferrets. This approach, following

on from the multi-region investigation of the dorsal visual pathway in chapter three, provides an opportunity to relate model characteristics across different depths of the sensory hierarchy.

Overall, the model objective and model architecture had the largest impacts on neural predictivity, while the training dataset had a comparatively weaker effect. Notably, networks with more brain-like receptive fields were better at predicting neural responses as well as those models which were better able to generalize to novel downstream objectives. Thus, hyperparameter choices which encouraged the models to learn more general representations tended to better predict neural responses, implying that more general basis functions may better model the auditory pathway. These results emphasize the importance of considering hyperparameter choices carefully when investigating normative models and provide novel insights into how these variables impact the networks' learned representations.

Results

Normative model hyperparameters

We trained a large number of normative models by varying the training dataset, model objective and model architecture (Figure 4.1Ai). For each model, we then analysed the network's properties across a number of dimensions, including the network's learned representations, receptive field properties and capacity to generalize to novel downstream tasks (Figure 4.1Aii). Finally, for each model, we assessed the neural prediction performance by fitting a linear-nonlinear mapping between the model's hidden activity and auditory neural responses to natural sound stimuli (Figure 4.1Aiii). Specifically, we evaluated how well each model could predict neural responses to natural sounds across the ferret IC, primary auditory cortex (A1 and AAF) and higher auditory cortex (PEG, including the PPF and PSF subfields; Figure 4.1B). For primary auditory cortex, we made use of two separate datasets, $A1_2$ and $A1_3$, which provided an internal

replication by using two distinct sets of experimental recordings for the same brain area. By systematically varying the models' hyperparameters, we were able to dissect their individual contributions while minimizing the effect of unintended confounds.

We used three different datasets for model training: recordings of zebra finch vocalizations, human speech by male and female speakers, and environmental field recordings in a variety of natural locations (Figure 4.1C). There is ample evidence from both experimental and modelling work for the role of stimulus statistics in shaping learned representations.[205–207] For example, the differing spectral content which is relatively biased towards lower frequencies for the zebra finch and speech datasets, and towards higher frequencies for the environmental dataset (Figures 4.1D, E), could lead to preferential specialization at these frequency ranges in the trained models. On the other hand, the naturalistic and varied datasets chosen might provide sufficient coverage across the relevant regions of spectrotemporal space to avoid any biases on aggregate.

We similarly varied the model training task across five different objectives – chosen to represent plausible neural computations that could underlie different functions of the auditory system (Figure 4.1F).[90,208–211] Specifically, the classifier networks were trained in a supervised manner to predict the true class for a given input, while the remaining four model objectives were trained in an unsupervised (temporal prediction) or self-supervised (denoising, dereverberation, source-separation) manner, in the sense that they did not require manually labelled targets.

Finally, we varied the model architecture by comparing feedforward networks with either a shallow (single hidden layer) or deeper (three hidden layer) architecture. By varying the depth of the model architecture, we sought to assess whether additional non-linearities would improve the networks' capacity to model auditory neural activity. In particular, there is some evidence that deeper networks are better able to model activity in deeper areas along the sensory pathway.[111] However, this hierarchical relationship has not always held empirically.[212,213] Similarly, there have been recent arguments made for the so-called

'shallow brain hypothesis' which argues that the structure of the brain is not necessarily consistent with a strictly deep or hierarchical perspective.[214] In the case of the deeper networks, we modelled neural responses using activity from the deepest hidden layer such that the total receptive field size was equated across both architectures (200 ms), to ensure that only the number of non-linear processing stages varied. For each model, we also varied the L1 regularization strength during training, which controls the extent of the wiring constraint among model units. Unless otherwise noted, we used the lambda value which minimized the model's loss on the held-out validation set.

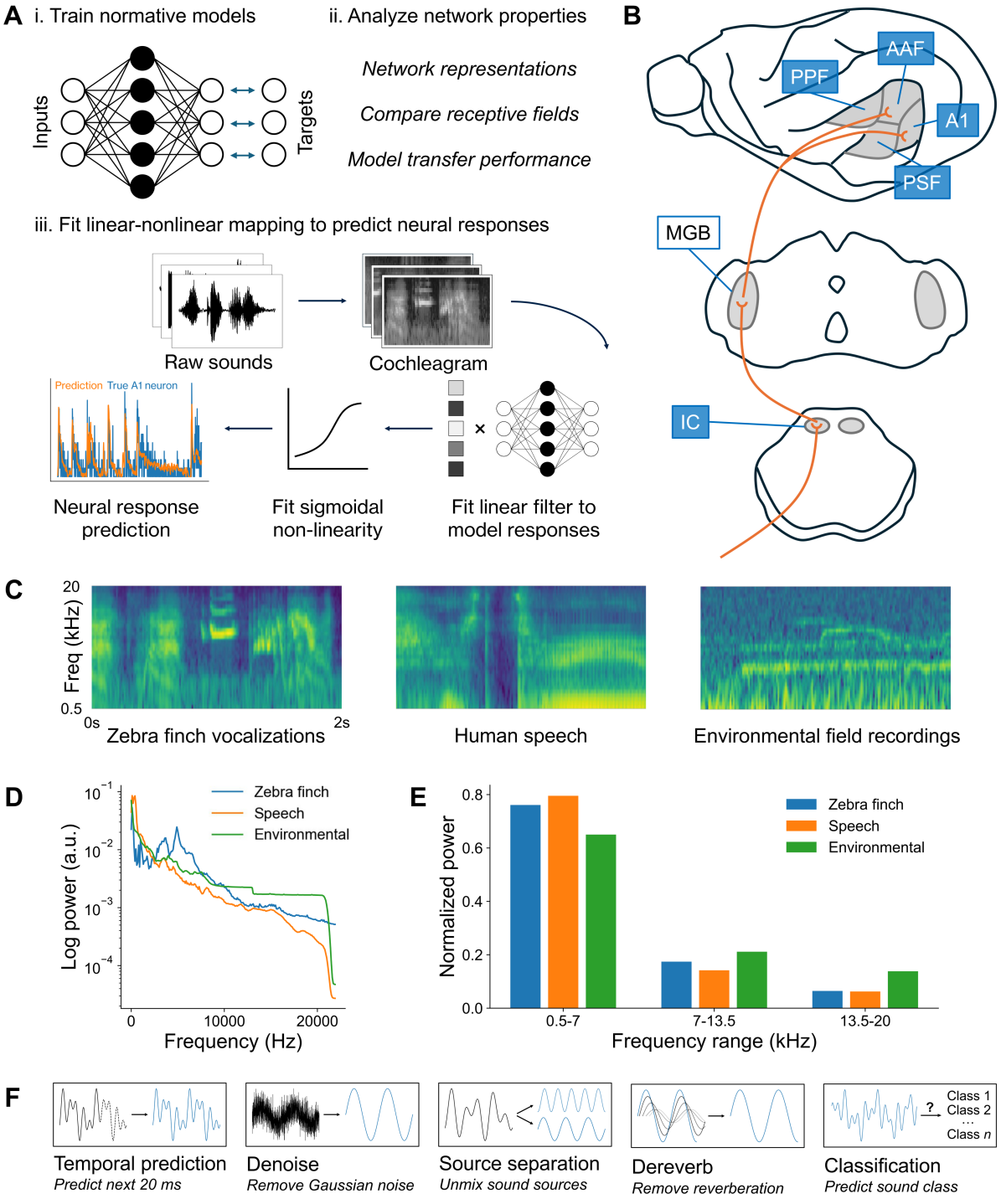


Figure 4.1. Overview of the normative model hyperparameters explored in this chapter.
(A) Schematic of the model training and fitting process. Each model was trained independently for the given combination of training dataset, model objective and network architecture, before each model's properties were analysed and a linear-nonlinear mapping learned to predict neural responses across the auditory pathway.
(B) Diagram of a subset of the ferret auditory pathway. The highlighted regions (IC, primary auditory cortex [A1, AAF], and higher auditory cortex [PSF, PPF]) represent those areas used for neural prediction.

(C) Three example cochleagram inputs to the models from the zebra finch, human speech and environmental datasets demonstrating the differing spectrotemporal content across datasets.
(D) Log power spectral density across each dataset.
(E) Normalized power at three different frequency bands demonstrating the differing spectral power distributions across training datasets. The environmental dataset has higher power at higher frequencies (7-20 kHz) compared with the zebra finch vocalization and human speech datasets.
(F) Diagram of the model objectives used for model training: temporal prediction, denoising, source separation, dereverberation and supervised sound classification.

Model objective dominates the shallow networks' learned representations

We first analysed the results of the shallow feedforward models. Overall, prediction performance varied substantially across the different brain areas (Figure 4.2Ai), with the mean CC_{norm} for each model varying from 0.36 to 0.43 in IC, from 0.21 to 0.29 in $A1_2$, from 0.20 to 0.32 in $A1_3$ and from 0.16 to 0.22 in PEG. Averaged across all models, there was a significant main effect (one-way ANOVA, $F(3, 56)=208$, $p<.001$) of brain area on neural prediction performance (IC mean = 0.40, $A1_2$ mean = 0.25, $A1_3$ mean = 0.28, PEG mean = 0.19), with a monotonic decrease from lower to higher regions of the auditory pathway, as has been previously reported.[215]

Notably, we equated the total recording time (the neural recording dataset size) used to fit the model-brain mappings to the smallest dataset ($A1_2$), to ensure that the different neural recording datasets were comparable (see Methods). Indeed, the total recording time had a large effect on the resulting neural prediction performance measure, increasing the final performance by as much as 50% (Figure 4.1Aii). Moreover, extrapolating the scaling curve of neural recording dataset size versus neural prediction performance, these results imply that a significantly larger recording dataset would be required before further improvements in neural prediction performance fully saturated (Figure 4.2B). Thus, in all cases, the given CC_{norm} values should be considered as relative measures with respect to the different models, rather than the ceiling performance that would be expected with a larger, unconstrained neural recording dataset size.

Networks generally, but did not always, perform above the baseline for an untrained model (Figures 4.2C-F). Accordingly, neural prediction performance showed a definite benefit from training, though some proportion of this performance could be attributed to fitting the model-brain mapping, rather than training for the normative objective alone. Overall, the temporal prediction and dereverberation networks tended to perform best as model objectives, though the relative performance of each model objective varied across brain areas (Figures 4.2C-F). For example, while the denoising networks performed among the best for IC, they were worse at predicting neural responses in PEG. To quantify these changes along the auditory pathway, we measured the rank correlation in neural prediction performance across model objectives between each pair of brain areas (Figure 4.2G; see Methods). The strength of this rank correlation varied substantially across the different pairs of brain areas (correlation range: 0.44-0.87) and was correlated with the rank anatomical distance between these brain areas (Pearson's r, r=.89, p=.018; Figure 4.2H). More specifically, the relative performance of each model objective was most consistent for the same brain region ($A1_2$ and $A1_3$) and least consistent between the most anatomically 'distant' brain areas (IC and PEG). Thus, the performance of different model objectives at predicting neural responses varied consistently across the auditory pathway. In contrast, when repeating this analysis over the training dataset rather than model objective, there was no clear relation between neural prediction performance and dataset across brain areas (Pearson's r, r=-.55, p=.257).

To further probe the relationship between the training dataset and model objective, we next analysed how the models' learned representations varied according to these two factors. We first calculated the representational similarity for each pair of models, before averaging the similarity scores according to whether the pair of models shared the same model objective or training dataset. The average representational similarity was significantly greater for pairs of models which shared the same model objective (t-test,

$t(21)=6.5$, $p<.001$), while there was no significant difference in representational similarity across models which shared the same training dataset versus those that differed in their training dataset (t-test, $t(33)=-0.69$, $p=.498$; Figure 4.2I). Thus, on average, the training objective remained the dominant factor in determining the model's learned representations, implying that models converged on similar representations regardless of the training dataset.

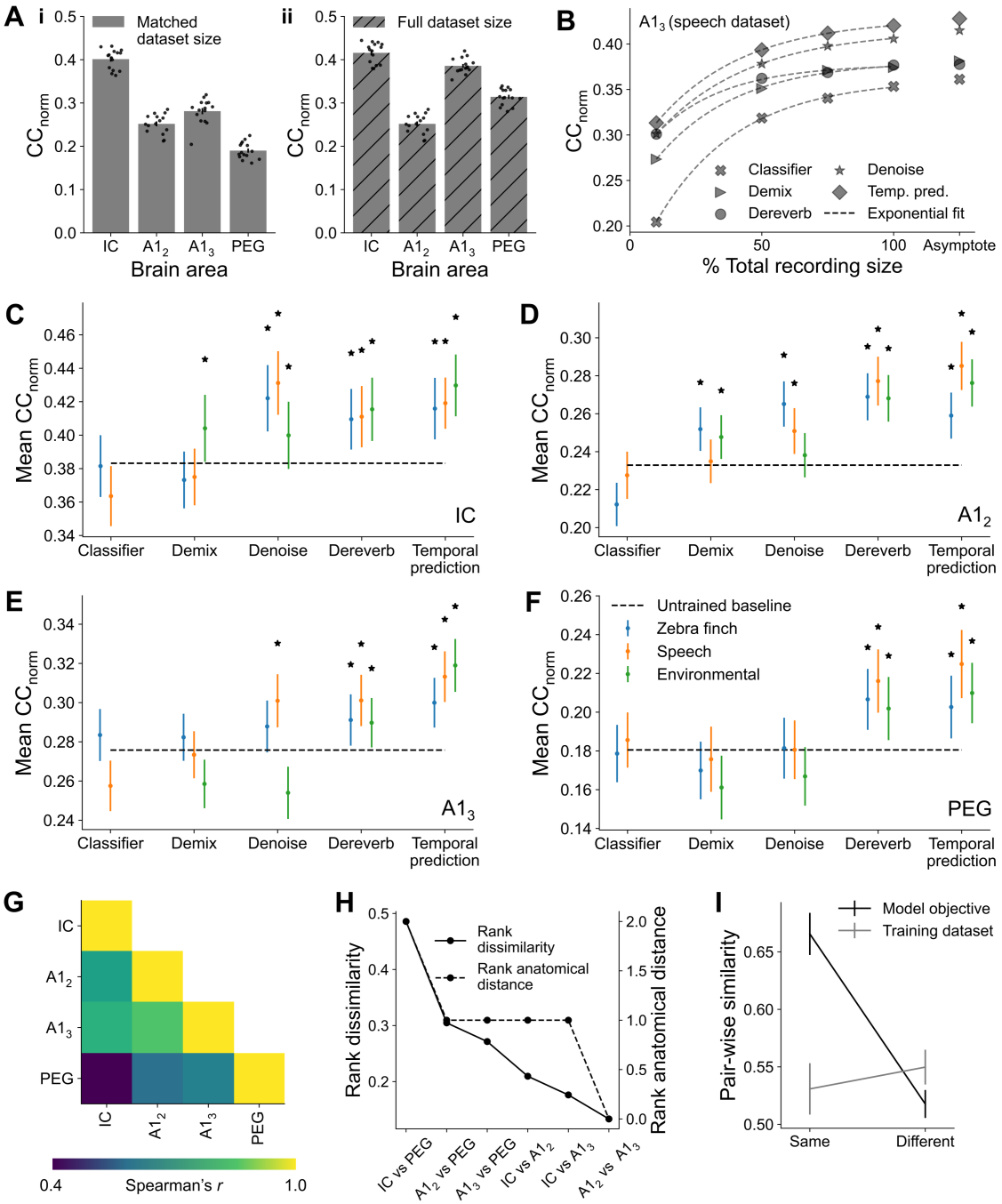


Figure 4.2. Model objective dominates the shallow networks' learned representations.
(A) Neural prediction performance monotonically decreased across higher areas of the auditory pathway. As the CC_{norm} score depended to a large extent on the size of the neural recording dataset, we equalized the length of the neural recordings across brain areas (i) rather than using the full dataset size (ii). As we equalised the size of all the neural datasets to $A1_2$, the full recording for $A1_2$ is the same as the matched recording. Error bars indicate s.e.m.
(B) Scaling curve for CC_{norm} as a function of the neural recording dataset size for different model objectives trained on the speech dataset for $A1_3$. The asymptote gives the ceiling for neural performance, extrapolated from an exponential fit ($y=ae^{bx}+c$), implying that a significantly larger dataset would be required before increases in the CC_{norm} score saturated.

(C-F) Neural prediction performance across shallow networks for each brain area, dataset and model objective. Asterisks indicate significant improvement over the untrained baseline (p<.05, uncorrected). While the dereverberation and temporal prediction networks showed a consistent benefit of training across brain areas, this effect was more marginal for the denoising, demixing and classifier networks which were in many cases non-significantly better than the untrained baseline. Error bars indicate s.e.m.
(G) The rank correlation (Spearman's *r*) of neural prediction performance by model objective between each pair of neural recording datasets.
(H) The same data as (G) ordered by dissimilarity (1-*r*) and plotted against the rank anatomical distance between pairs of brain regions (see Methods). More anatomically distant regions are less consistent in which objectives perform best.
(I) Pair-wise representational similarity between networks is greater for models trained on the same versus different objective, while no significant difference in pair-wise representational similarity was found for pairs trained using the same versus different dataset. Error bars indicate s.e.m.

Influence of model receptive fields on neural prediction

We next asked how the models' learned representations might relate to their capacity to predict neural responses.

In particular, we probed whether networks whose learned basis functions were more biologically realistic as receptive fields would be better at predicting the responses of auditory neurons. For each model unit's receptive field, we measured the maximum Pearson correlation coefficient (CC_{max}) between that unit's receptive field and the set of true spectrotemporal receptive fields (STRFs) in the given brain area. We then took the mean of these maximum correlation coefficient values over all model units as a measure of that network's receptive field similarity to the brain. For this analysis, we included all models across all L1 regularization values.

Overall, we found a significant correlation between neural prediction performance and receptive field similarity across all brain areas (Pearson's *r*, all $p \leq .002$; Figures 4.3A-C), such that networks with more neural-like receptive fields tended to better predict the responses of neurons in those brain areas. The temporal prediction models consistently produced the most 'brain-like' STRFs (trained on the environmental dataset for IC and $A1_2$, and when trained on the speech dataset for $A1_3$ and PEG), while the speech classifier was consistently the least 'brain-like' across all regions. The strength of this effect was

lowest for PEG (Pearson's r, r=.44, p=.002), reflecting a weaker correspondence between linear STRFs and neural predictivity in this region of secondary auditory cortex. However, when assessed using the full PEG neural recording dataset – with higher average CC_{norm} – we found that the strength of this effect was comparable to the other brain areas, implying the weaker relationship was primarily driven by overall lower neural prediction performance (Figure S4.1A).

Finally, we considered the role of the model training dataset on receptive field similarity. In line with the network representational similarity analyses, there was no significant effect of model dataset on average receptive field similarity (two-way ANOVA, $F(2,132)=1.72$, p=.183; Figure 4.3F), though there was a main effect of brain area (two-way ANOVA, $F(3,132)=88.3$, $p<.001$) with the receptive field similarity greatest for the two A1 datasets ($A1_2$ and $A1_3$). Thus, the input dataset statistics did not bias the learned model receptive fields in terms of their overall similarity to the true neural STRFs.

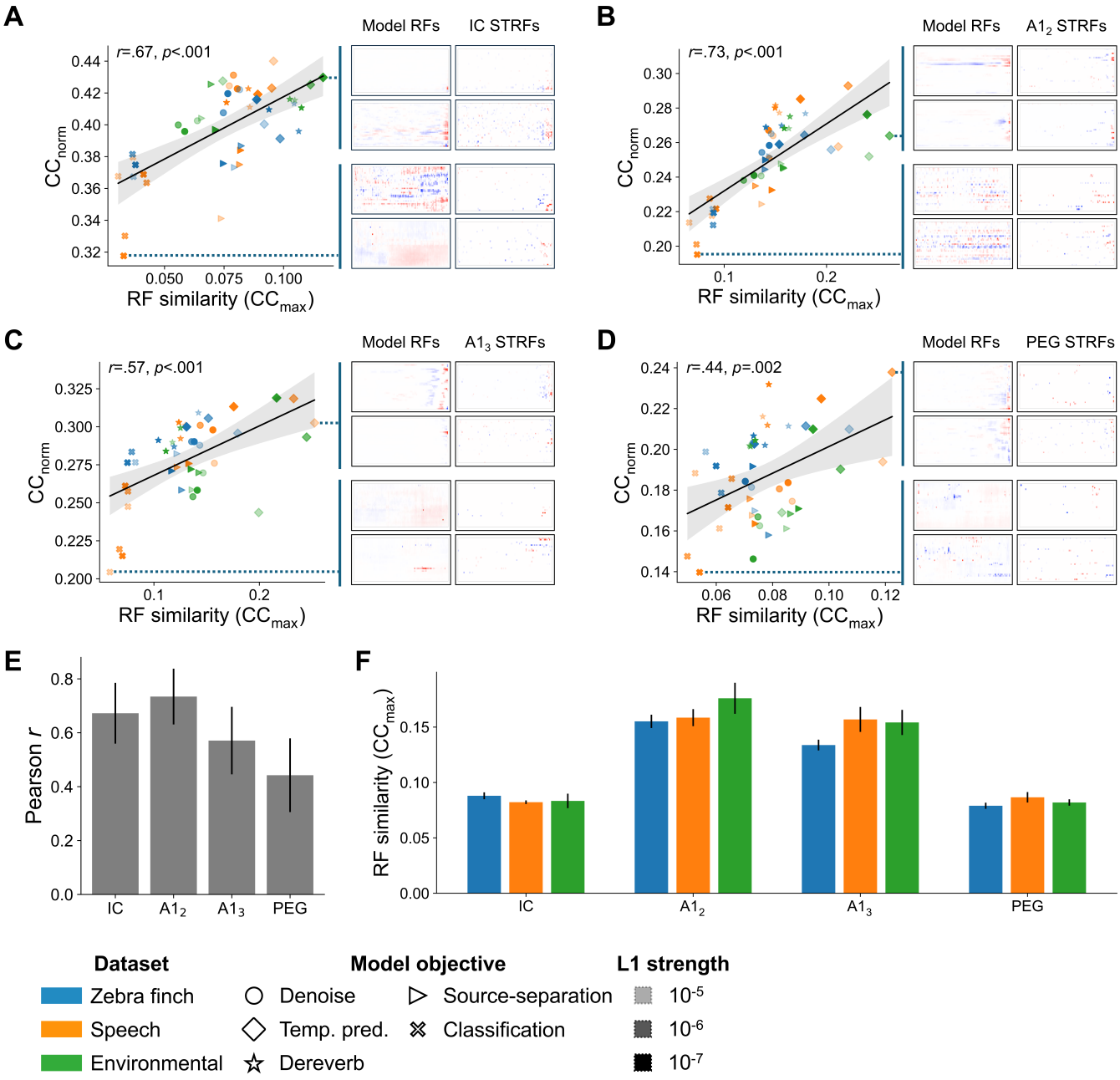


Figure 4.3. Receptive field similarity explains shallow network performance.
(A-D) Neural prediction performance (CC_{norm}) versus receptive field similarity (CC_{max}) for (A) IC, (B) $A1_2$, (C) $A1_3$ and (D) PEG. Networks with more similar receptive fields generally performed better at predicting neural responses.
(E) The Pearson correlation coefficient for the results in (A-D), demonstrating that this effect is consistent across brain areas. Error bars indicate s.e.m.
(F) Aggregate receptive field similarity across brain areas and datasets. There was no main effect of dataset, although the speech and environmental datasets tended towards higher scores on average. For this analysis, we excluded the classifier models to ensure that model objectives were equated across all training datasets. Error bars indicate s.e.m.
See also Figure S4.1.

Influence of model generalizability on neural prediction

Beyond the receptive field properties of each network model, we reasoned that there may be a more fundamental organizational principle linking the networks' learned representations to their capacity to predict neural responses.

In particular, we asked whether networks which learn more general representations might be better able to predict neural responses along the auditory pathway. The single objective networks described so far are constrained in a way which contrasts with the general nature of the auditory system. Indeed, the auditory system subserves numerous concurrent functions which cannot be reduced to a single objective. Accordingly, networks which are highly task specialized might struggle to predict neural responses by processing their auditory inputs in an overly narrow, task-specific manner.

We assessed network generalizability by measuring each network's transfer performance. For each model, we fixed its hidden representations while re-training the linear output layer for each objective, taking the generalization score as the average performance rank across all objectives (Figure 4.4A). Thus, a network with a lower transfer rank was better able to co-opt its hidden representations to novel downstream objectives and therefore is more generalizable. As for Figure 4.3, we again included models for all L1 regularization values for these analyses.

Overall, the temporal prediction networks trained on the speech and zebra finch datasets had the lowest transfer rank scores, indicating the greatest capacity to generalize across novel tasks, followed by the denoising and dereverberation networks trained on the zebra finch dataset. In contrast, the supervised classification networks consistently had the highest transfer rank scores, indicating their poor performance at generalizing to downstream objectives.

In support of our initial hypothesis, we found a significant relationship between transfer performance and neural prediction performance across the IC and A1 datasets

(Pearson's r, all $p<.001$; Figures 4.4B, C, D). However, the relationship was much weaker for PEG and did not reach significance (Pearson's r, $r=-.26$, $p=.081$; Figures 4.4E and F). Nevertheless, as for receptive field similarity, this weaker relationship could be explained by the overall lower neural prediction performance for PEG, as we found a significant effect when performing this analysis using the full PEG neural recording dataset (Figure S4.1B).

Notably, when grouping the results by training dataset, we found that the relationship between transfer generalizability and neural prediction performance was much weaker for the environmental dataset (Figure S4.2). This weaker effect could partly, but not entirely, be explained by the absence of classifier models for the environmental dataset. However, even when excluding the classifier models from all datasets, we found that this relationship remained consistent for the speech and zebra finch datasets (speech: mean $r=-.35$, zebra finch: mean $r=-.62$), while the transfer effect was actually in the *opposite* direction for the environmental dataset on average (mean $r=.21$). Thus, for network generalizability, we found that the training dataset could have a meaningful impact on the relationship between the networks' learned representations and neural prediction performance.

Mechanistically, more generalizable networks tended to have more 'average' representations with respect to the other networks (Figure 4.4G). Specifically, for each network, we calculated its representational similarity matrix (RSM), before taking the average across all RSMs. There was a significant relationship between transfer rank – and therefore a network's capacity to generalize to novel tasks – and each network's distance to the average RSM (Pearson's r, $r=-.53$, $p<.001$). Thus, generalizability was a direct consequence of being linearly closer to the average network representation.[216]

If more generalizable networks were better at predicting neural responses, we reasoned that training across multiple objectives could encourage more general network representations and thereby improve neural prediction performance, as has been

previously reported.[111] Based on the top performing model objectives, we trained a single network simultaneously to predict its future inputs (temporal prediction) as well as to produce an anechoic copy of its reverberant inputs (dereverberation). Thus, the network shared a hidden representation but learned two sets of linear weights corresponding to each objective. There was a small advantage in neural prediction performance for the multi-task model across all regions except for IC (Figures S4.3A). However, reflecting this small effect size, there was no significant main effect of model objective when comparing the multi-task versus temporal prediction networks (two-way ANOVA, $F(1,17)= 0.39$, $p=.582$), though there was a significant improvement when comparing the multi-task networks with the dereverberation networks (two-way ANOVA, $F(1,17)=11.3$, $p=.004$). The multi-task advantage was also variable across training datasets (Figure S4.3B). For the zebra finch and speech training datasets, there was a 3.6% and 5.0% advantage respectively for the multi-task network, while for the environmental dataset, the multi-task network performed 1.3% worse than the average neural prediction performance for the single task networks. This dataset-specific effect was similarly reflected in generalization scores, where there was a lower transfer rank (implying greater generalizability) for the temporal prediction compared with the multi-task network for the environmental and speech datasets but not for the zebra finch dataset (Figure S4.3C).

Thus, we found only weak evidence for a multi-task advantage in terms of network generalizability and neural prediction performance.

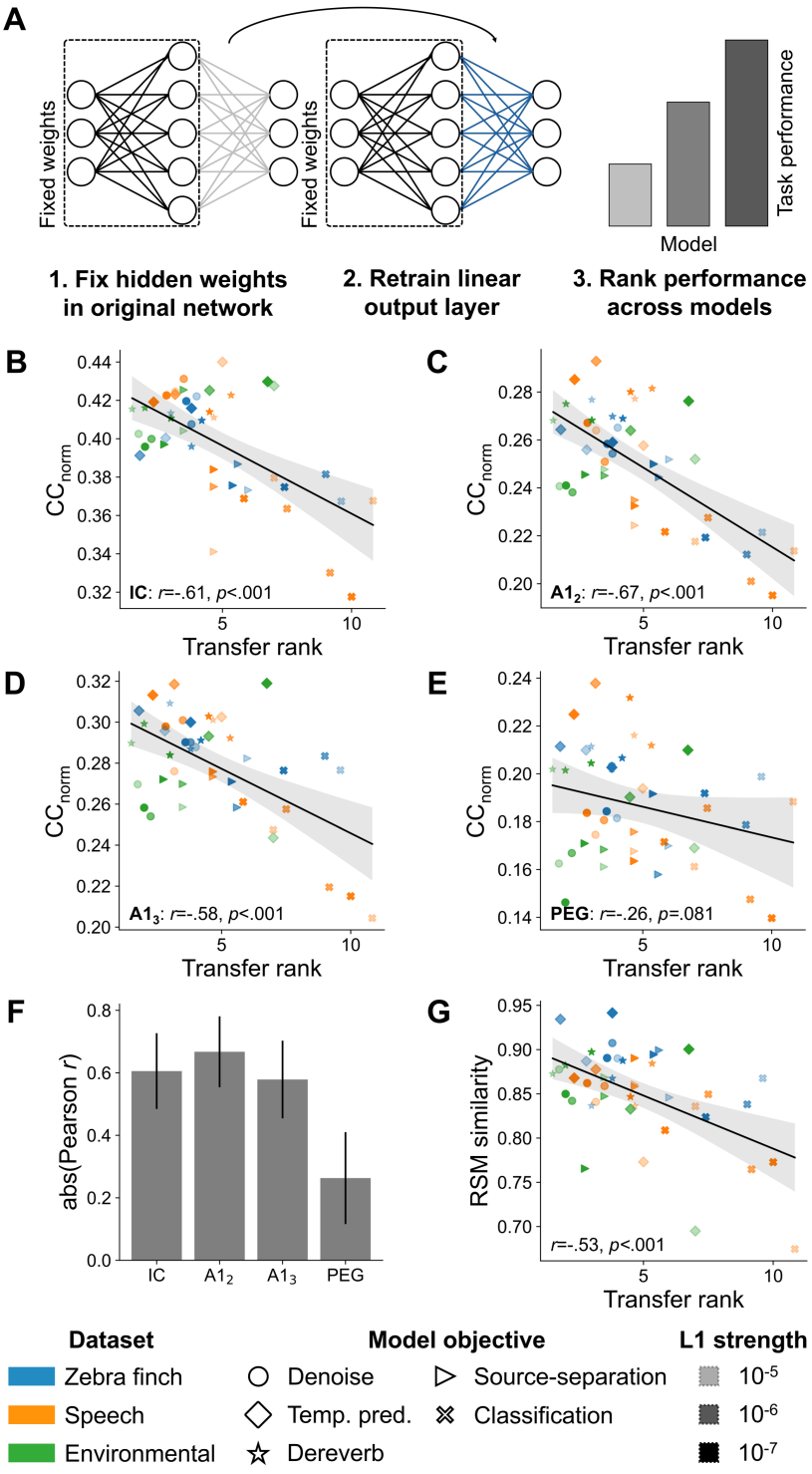


Figure 4.4. Network generalizability predicts neural prediction performance.
(A) Schematic of the transfer generalization procedure. We first fixed the hidden weights before retraining the linear mapping – from a random initialization – on a novel objective. Finally,

we took the rank for the transfer performance across all networks.
(B-E) Neural prediction performance (CC_{norm}) versus generalizability (transfer rank) for (B) IC, (C) $A1_2$, (D) $A1_3$ and (E) PEG. Networks with greater capacity to generalize to downstream objectives tended to perform better at predicting neural responses, although this effect was non-significant for PEG.
(F) The Pearson correlation coefficient for the results in (B-E) show this effect is highly consistent across both brain areas and datasets, with the exception of PEG. Error bars indicate s.e.m.
(G) Transfer rank predicts the representational similarity to the average representational similarity matrix across all networks.
See also Figures S4.1, S4.2 and S4.3.

Trade-off between generalizability and task-specialization impacts deep network neural prediction performance

Finally, to better understand the role of the network's architecture and the number of non-linearities, we compared the performance of the shallow versus deep models across their task performance, capacity to generalize across novel objectives, and neural prediction performance.

Task performance was uniformly greater for the deeper architecture, demonstrating the functional benefit of additional non-linear processing stages (Figure 4.5A). This effect was consistent for both within-distribution (one-sample t-test, $t(11)=4.13$, $p=.002$; the held-out test set) and out-of-distribution datasets (one-sample t-test, $t(11)=3.60$, $p=.004$). In the out-of-distribution case, we assessed the network's performance on each dataset not used for training – for example, the speech and environmental datasets for a model trained on the zebra finch dataset – and again found a greater performance benefit for the deep networks. These results indicate that the improvement in task performance for the deep networks could not be attributed to overfitting via the larger parameter count and instead reflected a true improvement across datasets.

Given this improvement in task performance for the deep networks, we next asked whether this would also be reflected in superior generalization performance relative to the shallow networks. To that end, we measured the pairwise transfer performance for each objective- and training dataset-matched shallow and deep network. The transfer rank score was calculated as before, and in this case for two networks was bounded

between zero and one, with a score of zero indicating superior transfer performance across all objectives. Contrary to their improved task performance, the deep networks were almost uniformly worse at generalizing to novel objectives (paired t-test, $t(14)=5.42$, $p<.001$; Figure 4.5B). This inferior transfer performance was not due to an absence of the generalization effect for the deep networks *per se* as, when looking at the deep feedforward models alone, more generalizable networks were better able to predict neural responses across all regions as for the shallow networks (Pearson r, all $p\leq.014$; Figure S4.4). Accordingly, deep networks demonstrated an implicit trade-off between task specialization and task generalization, reflected in improved task performance at the expense of generalizing to novel objectives.

In line with the results for the shallow networks, this impaired generalization performance was reflected in poorer neural prediction performance relative to the shallow networks (paired t-test, all $p\leq0.031$; Figures 4.5C and S4.5). Thus, contrary to expectations, deep models performed worse at predicting neural responses which may

have resulted from poorer task generalization performance, reflecting greater task specialization in the deep models.

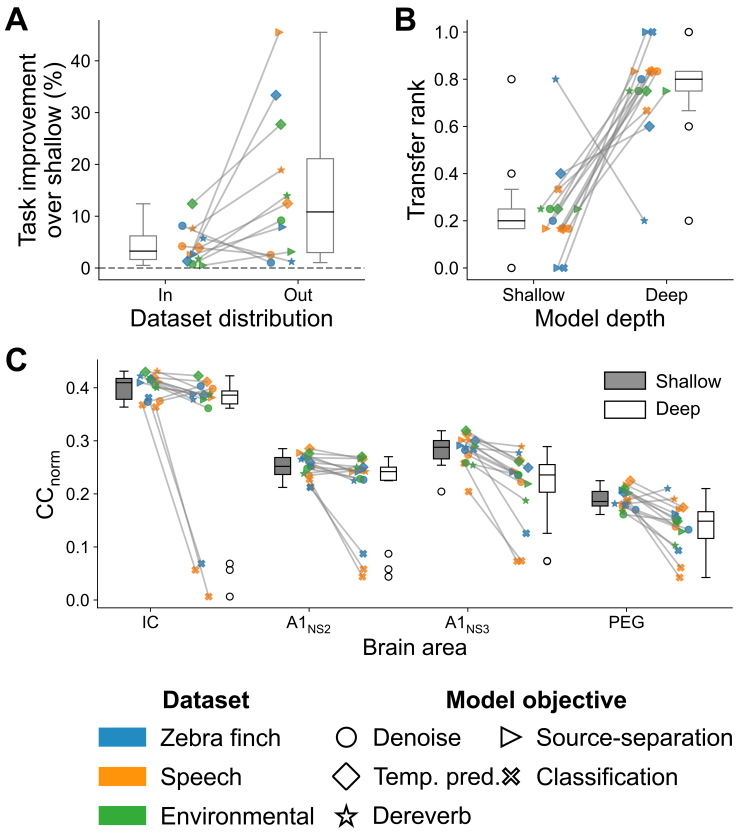


Figure 4.5. Deep networks demonstrate inferior neural prediction performance despite improved task performance.
(A) The deep networks show improved task performance, and this relative performance gain over the shallow network was proportionally greater for the out-of-distribution case. Data points for this and subsequent plots have been jittered along the *x*-axis for visualization purposes.
(B) The deep networks had higher transfer ranks compared with shallow networks, indicating poorer transfer generalization capacity.
(C) Across all brain areas, the deep networks are worse at predicting auditory neural responses, which may have resulted from poorer generalizability as a result of increased task-specialization.
See also Figures S4.4 and S4.5.

Discussion

By systematically varying the training dataset, model objective and model architecture, we probed how hyperparameter choices impacted neural prediction performance across

the auditory pathway. The training objective and model architecture had the largest impacts, whereas the training dataset had comparatively little impact on neural prediction performance or network representational similarity. Notably, networks with receptive fields more similar to the biology, and networks which learned more general representations, tended to better predict neural responses in these areas. Surprisingly, deep networks were worse at predicting neural responses than their shallow equivalents, despite their greater task performance. These results highlight an implicit trade-off between task specialization and task-generalizability and show how this can impact the capacity of a model to predict neural responses. Overall, this work provides insights into how modelling choices impact neural prediction performance and offers a novel framework in terms of transfer generalization for future investigations into normative modelling of sensory systems.

Role of training

What can normative modelling tell us about the auditory system? A key premise of this approach is that some relation exists between the normative training objective and the modelled system of interest. However, the results of the current study present a comparatively mixed picture.

Most models performed above the untrained baseline – particularly so for the deep networks, indicating that training for task-specific objectives can produce learned representations which improve neural prediction performance. Equally, anatomically closer regions were more similar in which model objectives best predicted neural responses in those regions, indicating that different task objectives produced representations consistent with those found in different brain areas. However, the improvement over the untrained baseline was often modest – and in some cases (for example, the classifier models) training *reduced* neural prediction performance. Moreover, the deep networks – though uniformly better in their task performance –

were worse at predicting neural responses. Accordingly, the relationship between task-optimization and neural prediction performance was complex and not always clear-cut.

Notably, training objectives which produced more general representations tended to better predict neural responses. We found that the temporal prediction objective was the most generalizable, and that this generalizability was a consequence of its learned representations being closest to the average similarity matrix over the set of models tested (Figure 4.5G). Why might this be the case? For one, successful temporal prediction requires extracting invariant features in the input data. For example, temporal prediction may necessitate disentangling non-predictive features – including random noise – from predictive features. In this way, temporal prediction has inherent conceptual links with other normative objectives such as denoising. Accordingly, these more abstract features learned by the temporal prediction objective may also be more useful for downstream tasks, hence producing more generalizable representations.

Role of brain area

All models were – on average – worse at predicting the responses of neurons in higher regions of the auditory system relative to lower-level regions. These results are in line with previous findings and reflect the additional non-linear computations imposed by subsequent brain regions as information ascends from the subcortical to primary cortical and non-primary cortical brain areas.[215]

Conversely, looking across different brain regions in this study, we found that the relationships between different model properties – for example, their receptive field similarity and capacity to generalize to novel tasks – and the capacity to predict neural responses were highly consistent across the auditory pathway. Indeed, while there was a somewhat weaker effect for PEG, we found that this could be largely attributed to the overall lower neural prediction performance of this brain area. Thus, these model-brain relationships were robust and were not limited to a particular brain area or neural

recording dataset. In particular, these results imply that generalizability may usefully explain how well a model is able to predict neural responses across the auditory pathway.

Relation to prior work

Compared with the large body of work on the visual system, comparatively few studies have considered normative modelling of the auditory pathway. Of those, there is a predominant focus on human functional magnetic resonance imaging data[111,217–219] and, more specifically, on human speech encoding in auditory cortex.[111,219–221] In line with this chapter, these studies consistently demonstrate that many different models can successfully capture auditory cortical responses, and that model training can – but does not always – improve the fit to the neural data.[111,219,221] Conversely, far fewer studies have considered how model hyperparameters can impact neural prediction performance. Where the model objective has been varied, there is some evidence that self-supervised learning objectives are superior at predicting auditory cortical responses[220,221] – a result in alignment with the performance of the temporal prediction model versus supervised classifier models in this study. Nevertheless, the specific roles of training dataset and model architecture have not been systematically varied across these studies.[111,220,221]

In contrast, far fewer studies have probed the capacity to predict single-cell responses, particularly considering subcortical regions. In terms of single-neuron encoding models of the auditory system, these have primarily been fit directly to neural data[85,215,222,223] and therefore comprise a complementary but parallel research program to the normative models described in this chapter. Like the present study, normative models trained on a variety of tasks – including sparse autoencoder and temporal prediction objectives – have quantitatively and qualitatively reproduced the properties of auditory neurons across the cortex and midbrain.[90,224] However, these studies did not compare how well these models could predict auditory neural responses to natural sounds. Accordingly, this study represents a relatively novel approach by considering the effects of multiple modelling

hyperparameters on predicting neural responses across the auditory pathway at the level of single-unit recordings.

Limitations and future work

A key limitation of this current study is the lack of matched stimuli across neural datasets. Matched stimuli would facilitate cross-area comparisons but, due to the availability of experimental datasets, this was not possible for IC and $A1_2$. Nevertheless, all efforts were made to equate the neural recordings for each brain area. The total neural recording dataset size was matched (the total duration of stimuli used for fitting the model-brain mapping) while the test set was chosen to include qualitatively similar stimuli across neural datasets (including ferret vocal calls, human speech, etc.). Accordingly, we do not believe that matched neural recording datasets would radically alter this chapter's conclusions.

This study is also necessarily limited by the set of chosen hyperparameters. Due to the exhaustive nature of the model hyperparameter search, where we trained every combination of the chosen hyperparameters, adding in additional variants quickly increases the total number of models needed and makes training computationally expensive. Nevertheless, this study was restricted to two comparatively simple architectures compared within the wider space of network architectures used in modern machine learning. In particular, future work might consider including recurrency, which might better model the dynamical aspects of auditory neural responses.

Finally, to a large extent, this chapter rests on the assumed validity of the neural prediction performance measure. Here, this is operationalized as the normalized correlation coefficient (CC_{norm}) between the true and predicted single-unit responses of auditory neurons.[169,170] However, linear predictivity is one among many possible similarity metrics including reverse linear predictivity,[225] representational similarity analysis,[202] and centred kernel alignment.[226] Notably, the choice of metric has been shown to influence the resulting conclusions, making comparisons between studies

employing different metrics fraught with difficulties.[213] Accordingly, future work engaging with alternative metrics would strengthen these findings with regards to neural predictivity.

In conclusion, this study demonstrates how modelling choices across training dataset, model architecture and model objective systematically impact neural prediction performance across the auditory pathway. These findings suggest that generalizability may be a useful framework for normative modelling, which could provide improved insights into the underlying computations of the sensory brain.

Methods

Datasets

Three different datasets with varied natural statistics were used for model training – namely: human speech, zebra finch birdsong and environmental field recordings. Human speech was taken from a previously published dataset,[227] consisting of short sentences spoken by English speakers, with a balance of male and female speakers. Zebra finch birdsong was taken from a previously published dataset,[228] consisting of isolated vocal recordings from male zebra finches. Finally, environmental recordings were sourced from https://freesound.org and consisted of recordings from a variety of natural locations incorporating weather, distant animal vocalizations as well as non-specific percussive sounds. The specific users and sources taken for the environmental dataset were: *Mørke Mose* (https://freesound.org/s/459912, *secondharmonicgeneration* (https://freesound.org/s/519519), *frederic.font* (https://freesound.org/s/516565), *Philip_Goddard* (https://freesound.org/s/668382), *csengeri* https://freesound.org/s/235447, and *klankbeeld* (https://freesound.org/s/561541).

Prior to further processing, all sound files were mixed down to a single audio channel, if they were recorded in stereo, before normalizing the resulting *wav* file. Finally, we

processed the sound waveforms into cochleagrams to approximate the auditory nerve fibre representation of sound. More specifically, we used the spec-power model which is based on a mel-scaled spectrogram representation to which a logarithmic compression nonlinearity is applied.[167] The cochleagrams were binned at a temporal resolution of 2 ms, with 32 frequency channels which ranged from 0.5-20 kHz,[167] and clipped into one second (500 bin) segments. Each cochleagram clip was then individually normalized. All datasets were 30600 clips in size, equivalent to 8.5 hours of audio in length (7.65 hours for the training set and 0.85 hours for the validation set).

Model architectures

We considered two architectures – a shallow and deep convolutional neural network,[85] which were designed to equate the total receptive field size of the final output units, while varying the network depth and total number of non-linearities present. Thus, all networks had a final temporal receptive field size of 200 ms.

For the shallow network, the model consisted of a single 2D convolutional layer followed by a ReLU non-linearity and a linear transposed convolution layer to map the hidden state to the output predictions. As the kernel size along the frequency dimension equalled the number of frequency channels in the cochleagram input, this could be equivalently understood as a form of 1D convolution. That is, the model was trained to learn a series of filters of 100 timesteps in 'length' and 32 frequency channels in 'height' such that convolving each individual filter over the 500 x 32 input cochleagram produced a one-dimensional vector of 401 timesteps in length. For the deep network, the model consisted of 3 convolutional layers, similarly using the ReLU non-linearity and the linear transposed convolutional output layer. For both networks, the number of output channels n depended on the objective (for example, the temporal prediction model predicted a span of 10 timesteps into the future such that n=10).

The network architectures for both models are described below:

Table 4.2. Architecture for the shallow convolutional network.

Layer	Layer type	Kernel size	Input channels	Output channels
1	2D convolution	100x32	1	200
	ReLU			
2	2D transposed convolution	1x32	200	*n*

Table 4.3. Architecture for the deep convolutional network.

Layer	Layer type	Kernel size	Input channels	Output channels
1	2D convolution	34x12	1	25
	ReLU			
2	2D convolution	34x12	25	50
	ReLU			
3	2D convolution	34x12	50	200
	ReLU			
4	2D transposed convolution	1x32	200	*n*

Model objectives

We trained the networks on several supervised, self-supervised and unsupervised learning objectives based on a number of ecologically plausible auditory tasks.

The denoising network was trained to produce the 'clean', noiseless cochleagram from noisy cochleagram inputs produced by adding Gaussian noise into the input waveform at a signal-to-noise ratio of 5 dB.

The source-separation network was designed to produce two cochleagrams corresponding to the two mixed input sources. The input sources were produced by summing the waveforms of two sound clips each selected randomly from the chosen natural sound dataset. In the case of the environmental dataset, we ensured that at least

one clip included distinct 'auditory objects' (e.g., animal calls), to ensure that the superimposed clips did not consist of two ambient soundscapes.

The dereverberation network was trained to produce a 'dry' cochleagram from the reverberant cochleagram input. To produce reverberant copies of the input sounds, we used the 'pyroomacoustics' package to simulate a 'shoebox' room. The room dimensions were 1.5x1.5x15 m, with a virtual microphone positioned centrally 3.75 m from the back at a height of 0.15 m. The room's energy absorption was calibrated to a limestone wall, resulting in an RT_{60} value of 2s.

The temporal prediction network was trained to predict the upcoming 20 ms based on the preceding 200 ms of input. For this network no further transformations were applied to the original datasets.

Finally, for the classifier networks, we produced a set of target labels for each dataset and trained the network to predict the relevant label for each input. For the speech dataset, recordings were labelled according to both speaker gender (male or female) and speaker ID (a separate label for each individual speaker). Similarly, for the zebra finch dataset, we labelled each clip according to the corresponding bird ID. For the environmental dataset, as there was no obvious dimension for labelling, we did not produce a classifier for this set of recordings.

The denoising, source-separation, dereverberation and temporal prediction networks were trained to minimize the mean-squared error between the predicted and target cochleagrams. Conversely, the classifier networks were trained to minimize the cross-entropy between the predicted and target logits. Finally, for each network, we included an L1 regularization term over all weights in the network to encourage the development of sparse filters and to minimize the risk of overfitting to the training dataset. The L1 regularization term was weighted by the lambda hyperparameter controlling the strength of regularization which was chosen to minimize the loss (mean-squared error or cross-

entropy) on the held-out validation set. Thus, we used the model with the optimal lambda value unless otherwise noted (Figures 4.4, 4.5, S4.1, and S4.2). In general, we found no consistent relation between L1 regularization strength or model performance and neural prediction performance.

Representational similarity analysis (RSA)

We used RSA as a means of producing an aggregate score for how similarly different network models represented their sensory inputs. More specifically, RSA measures the distance between the representational similarity (correlation) matrices for each pair of models.[198,201]

Representational similarity matrices (RSM) were computed by first binning the responses to natural stimuli into 200 ms chunks and taking the mean response across time. Thus, each model's output was represented as an $M \times N$ matrix with M rows of stimuli and N columns of units. The RSM was then calculated as the correlation between each pair of columns in the matrix (the response to a given stimulus m), resulting in an $M \times M$ RSM. Representational similarity was then calculated by taking the distance between the flattened lower triangle of the RSMs of each model and the neural data using Spearman's correlation coefficient.[202]

Neural response fitting

All neural recordings were taken from existing datasets, performed on passively listening ferrets (*Mustela putorius furo*), across IC, A1 and PEG. Neural data for A1 and PEG were taken from previously published datasets.[85,167,229,230] To disambiguate the two A1 datasets, we term these $A1_2$ and $A1_3$, referring to the recordings from Lopez Espejo et al.[229] and Pennington et al.,[230] respectively. The IC dataset consisted of recordings from an unpublished dataset, but formed part of a series of experiments whose methods have been published elsewhere (NS1).[222]

In brief, electrophysiological recordings from central IC were performed in three anaesthetized ferrets under ketamine (5 mg/kg/h) and medetomidine (0.022 mg/kg/h)

using silicon probe electrodes. 20 different five second stimuli were presented using earphones for 20 repeats in a random order. Animal experiments were conducted under licence from the United Kingdom Home Office and with approval from the local institutional ethics committee.

The dataset characteristics are detailed below:

Table 4.3. Neural recording dataset characteristics.

	IC	$A1_2$	$A1_3$	PEG
Animal *n*	3	6	5	5
Single-unit *n*	32	186	211	111
Awake or anaesthetized	Anaesthetized	Awake	Awake	Awake

In all cases, data were pre-processed in the same way by binning spike times into 2 ms peristimulus time histograms (PSTHs), while auditory stimuli were similarly processed into cochleagram representations using the same method as the model training datasets. The 2 ms timestep (250 Hz Nyquist frequency) was chosen to model IC responses which can track amplitude modulations up to approximately 300 Hz,[231] while remaining close to the range for cortical responses which can track amplitude modulations up to approximately 100 Hz.[232] For each model, we produced a linear-nonlinear mapping from the model's hidden state to predict the neural response to natural sounds. We included all recorded units in both datasets whose noise-to-signal power ratio in response to the natural sounds was below 40.[166]

For each model, the neural fitting process consisted of first learning a linear mapping using Lasso regression before fitting a rectified sigmoidal non-linearity.[167] Thus, each biological neuron was predicted as a non-linear combination of the model's hidden units. The non-linearity $f(x)$ was defined as:

$$f(x) = \mathrm{ReLU}\left(\frac{a}{1 + e^{\frac{c-x}{b}}} + d\right)$$

where the parameters a, b, c and d were optimized to minimize the mean squared error between the true and predicted neural firing rate using the SciPy "curve_fit" function.[168] The L1 regularization strength (α) of the Lasso regression procedure was chosen via cross-validation from 20 values log-spaced between 10^{-1} and 10^{-9} to maximize the average CC_{norm} across each fold's validation set for the combined linear-nonlinear mapping. The reported values are finally taken as the performance on the held-out test set. The test set was chosen to be representative of the wider dataset, and across all datasets included a similar set of stimulus types including ferret vocalizations, human speech and outdoor field recordings.

Model objective rank consistency

To measure how consistently each model objective performed across brain areas, we took the Spearman correlation between the model objective's performance for each pair of neural recording datasets, averaging the final correlation value across the three training datasets. Finally, we compared these correlation values for each pair of brain areas to the rank anatomical distance between brain areas. For example, the $A1_2$/$A1_3$ dataset pair had a rank distance of 0, the IC/PEG dataset pair had rank distance of 2, while the remaining pairs had a rank distance of 1.

Receptive field mapping

For the shallow convolutional models, we took the receptive field as the filter input weights. For deep convolutional models, we estimated the receptive fields of the final layer's hidden units using the response-weighted-average. Specifically, we took the average response of each output unit in the model weighted by its response to 20,000 clips of Gaussian noise. For neural data, we estimated the STRFs as the linear coefficients from the neural response fitting procedure as described above. That is, for each neuron, we fitted a linear mapping between the input cochleagram and that cell's response, taking the STRF as the resulting linear filter.

Transfer performance

To assess the generality of the models' learned representations, we computed a 'generalization' score which describes how well each network can generalize to novel downstream objectives. Specifically, we fixed the hidden weights and retrained the linear output layer for the different networks and learning objectives. Then, for each downstream objective, we ranked the models' performance as the average loss across all minibatches on the test set, taking the generalization score as the average rank across objectives. Where two models' scores were non-significantly different, their ranks were tied for that task objective. A lower average rank thus corresponds to a model whose hidden representation can be better repurposed across different objectives.

Supplemental information

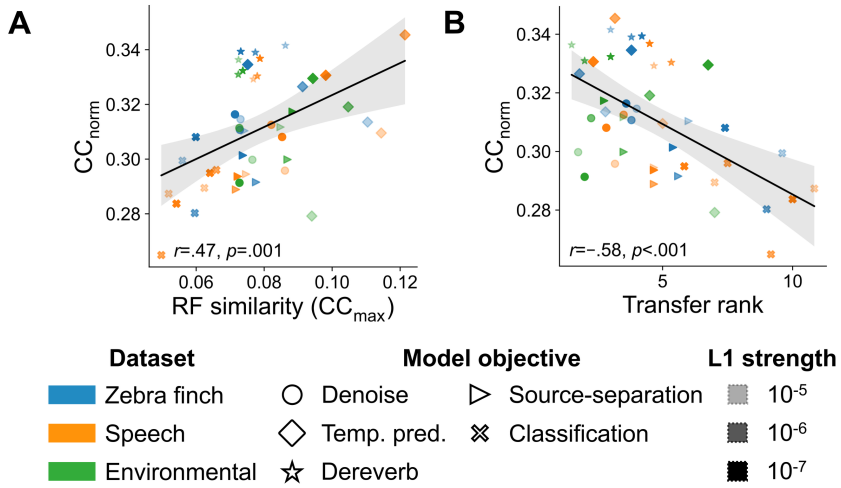


Figure S4.1. CC_{norm} as a function of receptive field similarity and transfer rank PEG using the full neural recording, Related to Figures 4.3 and 4.4.
(A) Neural prediction performance (CC_{norm}) versus receptive field similarity (CC_{max}) for PEG. Networks with more similar receptive fields performed better at predicting neural responses.
(B) Neural prediction performance (CC_{norm}) versus generalizability (transfer rank) for PEG. Networks with greater capacity to generalize to downstream objectives tended to perform better at predicting neural responses.
In both cases (A and B), there is a significant or strengthened relationship for PEG when using the full recording dataset, implying the weaker relationship for the matched dataset could be attributed to the lower overall neural prediction performance score.

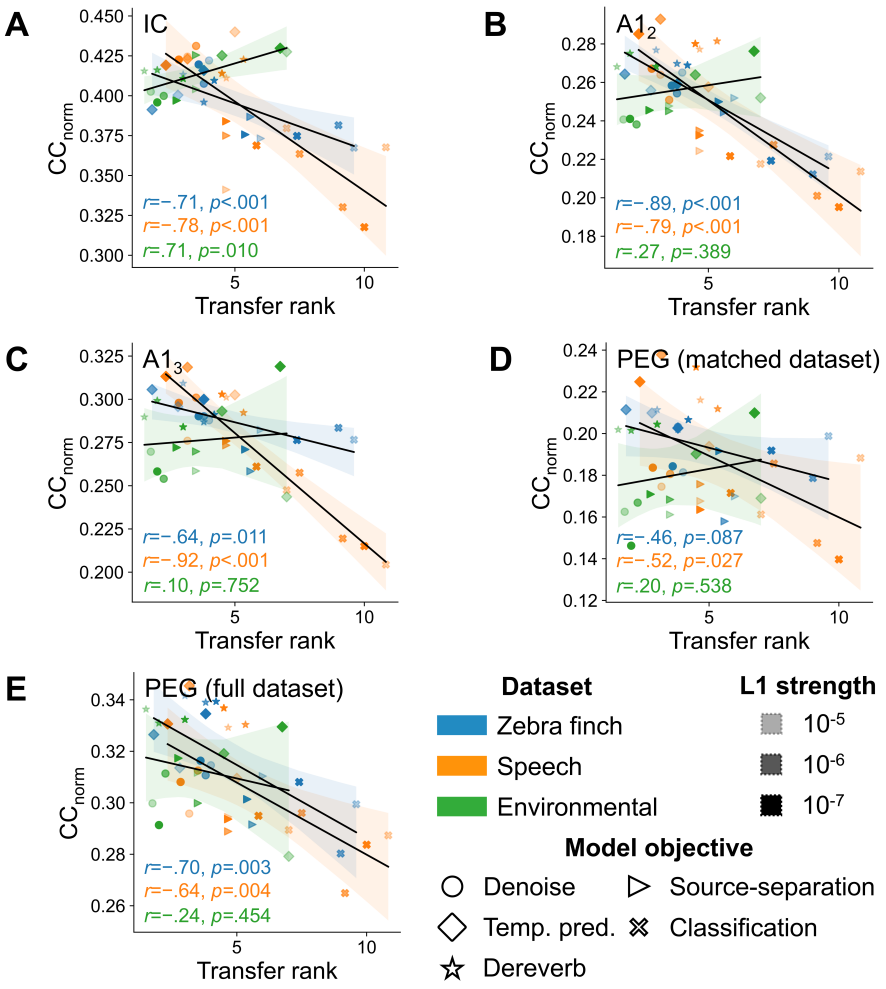


Figure S4.2. CC_{norm} as a function of transfer rank grouped by training dataset, Related to Figure 4.4.

Neural prediction performance (CC_{norm}) versus generalizability (transfer rank), broken down by training dataset for IC (A), $A1_2$ (B), $A1_3$ (C), PEG (matched dataset, D) and PEG (full dataset, E). There was a weaker effect with the environmental dataset – often in the opposite direction – implying this relationship may depend on the underlying training dataset.

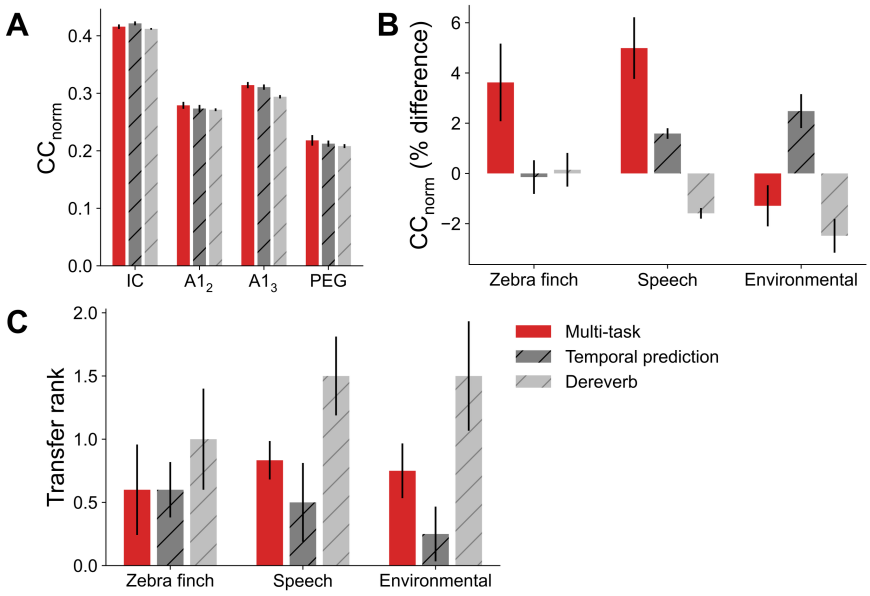


Figure S4.3. Weak evidence for an improvement in neural prediction performance for multi-task networks, Related to Figure 4.4.

(A) Neural prediction performance (CC_{norm}) across the multi-task, temporal prediction and dereverb networks averaged across training datasets. Error bars indicate s.e.m.

(B) Percentage change in CC_{norm} compared to the single task average (temporal prediction and dereverb) for each network, averaged across brain areas. Error bars indicate s.e.m.

(C) Transfer rank for the multi-task, temporal prediction and dereverb networks for each training dataset. Error bars indicate s.e.m.

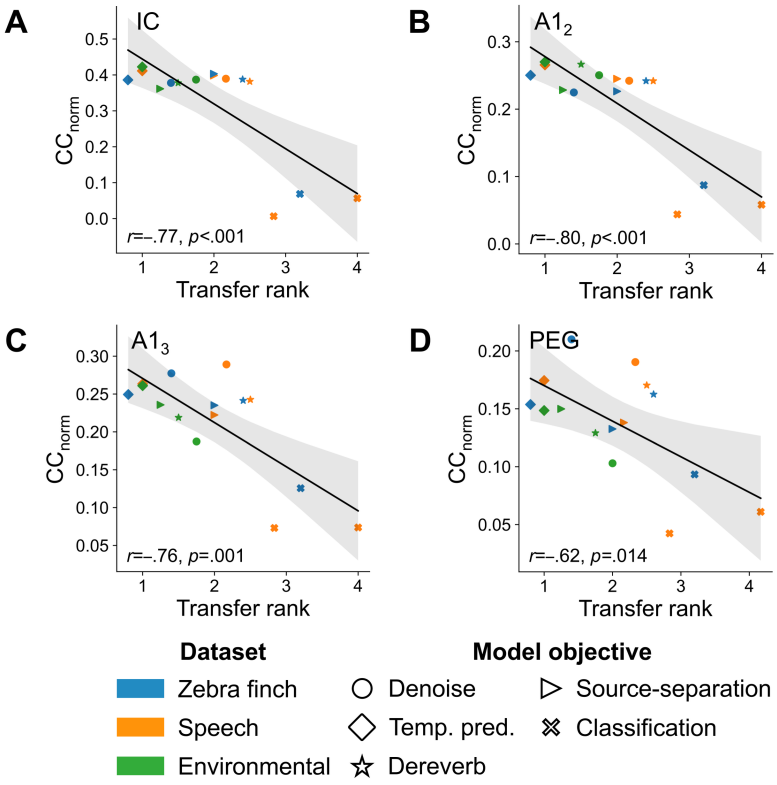


Figure S4.4. More generalizable deep networks better predict auditory neural responses, Related to Figure 4.5.

Networks with a lower transfer rank – which are better able to generalize across novel tasks – are better at predicting the responses of auditory neurons (A: IC, B: $A1_2$, C: $A1_3$, D: PEG). Unlike for the shallow networks, this effect was consistent across all brain areas.

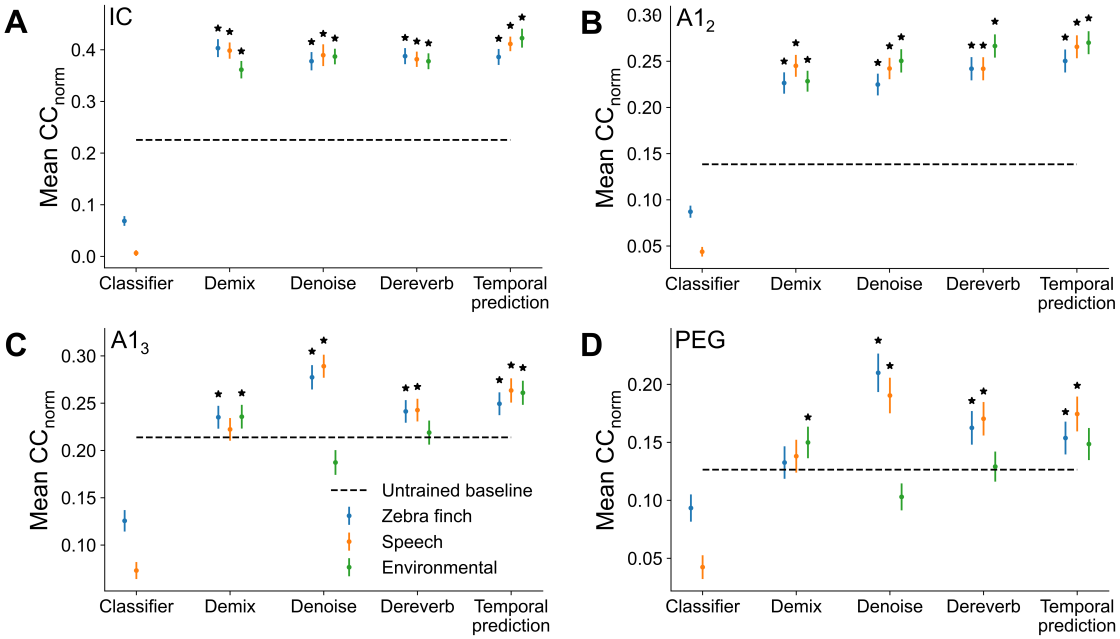


Figure S4.5. Deep network performance across all models, Related to Figure 4.5. Neural prediction performance across deep networks for each brain area, dataset and model objective for each brain area (A: IC, B: A1$_2$, C: A1$_3$, D: PEG). Asterisks indicate significant improvement over the untrained baseline ($p<.05$, uncorrected). Notably, there was a larger effect of training for the deep models – with a greater number of models showing a significant improvement in neural prediction performance over the untrained baseline compared with the shallow models. Error bars indicate s.e.m.

5

General discussion and conclusions

A major goal of this thesis has been to better understand the computational principles underlying the brain's sensory systems through the use of normative modelling approaches. To that end, I have investigated how task-optimized neural network models can explain the local connectivity patterns of primary visual cortex, the response properties of neurons across the dorsal visual pathway and neural activity across the auditory pathway. I will begin by summarizing the main findings of this thesis, before outlining directions for future work and discussing some pertinent questions raised by this thesis.

Summary of findings

In chapter two, I developed a recurrent model optimized for temporal prediction with the goal of capturing functional connectivity in mouse V1. The model successfully recapitulated the distribution of orientation- and direction-dependent connectivity as well as cell-type specific connectivity motifs across excitatory and inhibitory subpopulations. Indeed, we found that optimizing the model for temporal prediction improved the model's fit to mouse V1 compared with a model optimized for compression. Moreover, the better the model was able to predict its future inputs, the

better it captured V1-like connectivity patterns. These results imply that local connectivity in mouse V1 could reflect optimization for temporal prediction.

In chapter three, I extended the recurrent temporal prediction model into a hierarchical model of the visual cortex. While the single region model predicted the next frame in its inputs, the hierarchical model was organized as a series of groups where each group predicted the future value of its lower-order inputs. We hypothesized that higher groups would extract increasingly abstract properties in their sensory inputs, with the goal of capturing similar transformations in the visual cortex. In support of this hypothesis, the distribution of tuning properties across the model's groups mirrored those across increasingly deep regions of the dorsal visual pathway. We found that both the temporal prediction objective and the hierarchical recurrent architecture were important in modelling the response properties and neural activity of visual neurons. An important aim for any principle of neural function is that it can be applied in a generic manner across the length of the sensory hierarchy. Thus, the success of temporal prediction in modelling response properties across primary and non-primary visual cortex provides evidence for temporal prediction as a general normative principle of sensory systems.

Finally, in chapter four, I investigated temporal prediction within the wider landscape of normative modelling principles, moving from the visual to the auditory system. In particular, a key goal was to better characterize the contributions of different hyperparameter choices and how this impacts the capacity of network models to predict neural responses. By taking a comparative approach, using neural recordings across different auditory regions, I was able to investigate how these results varied as a function of depth along the auditory pathway. The training objective and model architecture had the largest impacts on the networks' learned representations and their capacity to predict neural responses. In particular, despite training across different datasets, networks trained on the same task tended to learn similar representations. Moreover, I found that networks which were better able to generalize to novel downstream tasks were better able

to predict auditory neural responses across both subcortical, primary and secondary cortical regions. Overall, this study highlighted the importance of different modelling choices when building neural network models to study the transformation of representations across sensory systems.

Future directions

This thesis has explored normative modelling of sensory systems, providing evidence for temporal prediction as an explanatory principle as well as providing some intuition as to why different normative principles might better predict neural activity. Going forwards, there are a number of open questions and new research directions to build upon this work.

Firstly, there are several extensions to temporal prediction as a normative principle which future work could address. In this thesis, temporal prediction has been operationalized by minimizing the mean squared error between the true and predicted future. This approach is easy to define and easily computable, in contrast to other formulations such as the information bottleneck.[91,105] From the perspective of maximum likelihood estimation, by minimizing the mean squared error, we can interpret the network's output as an estimate of the mean under the assumption of a Gaussian noise model – that is, the 'average' true future. Accordingly, a potential extension to these modelling efforts would be to estimate the *distribution* of possible futures – for example, by explicitly estimating the parameters of some distribution of future states. In this way, the variance of that distribution could also provide an estimate of the model's uncertainty in its predictions – akin to the idea of 'precision' in predictive coding,[126] which could forge further links between these two normative frameworks.

Secondly, going beyond the current auditory and visual models described in this thesis, an important challenge will be integrating these networks both into a cross-modal

framework and with the motor system. So far, this thesis has focused exclusively on the unimodal response properties of sensory regions. However, cross-modal sensory integration is widespread and occurs at a relatively early stage of the visual and auditory pathways. In the visual system, neural responses in the LGN can be modulated by auditory stimuli,[233] while neural activity in the IC can be modulated by visual inputs.[234] Accordingly, integrating visual and auditory stimuli into a multisensory framework would provide a useful way of investigating cross-modal interactions. Going beyond sensory systems, integrating the motor system with temporal prediction represents an outstanding challenge. Although the motor system has not been considered in this thesis, the behavioural output of perception is ultimately some form of motor command.[235–237] Typically, motor-learning problems are approached using some variant of reinforcement learning.[238,239] Accordingly, integrating temporal prediction with reinforcement learning represents one approach, for example using temporal prediction as a preprocessing stage to transform sensory inputs into more behaviourally useful representations for the downstream reinforcement learning model. Together, these two research extensions would help build out temporal prediction into a fuller and more complete model of sensory systems.

Finally, the role of biological realism in modelling represents a line of thought which could be investigated further. As noted in the preceding chapters, the networks described in this thesis are comparatively abstract as models of the brain, in the sense of using rate-based, point-like neurons trained with backpropagation. While increasing biological realism can be incorporated into these models, there is an implicit assumption that more biologically detailed models will better capture the relevant characteristics of the target system. However, it is largely untested whether more biologically realistic models are, for example, better able to predict neural responses. From the opposite perspective, it remains to be tested in greater depth whether including features from the biology could have certain advantages for engineering. For example, could a

multicompartment model improve next-frame prediction performance in a temporal prediction model, or are these biological details merely an outcome of the brain's physical constraints with limited functional benefit?[240] These efforts would represent a welcome extension to the modelling work in chapter five to further characterize how modelling choices impact network representations and neural prediction performance.

What can we gain from normative models?

Normative network models form the core of this thesis as an approach to studying the brain. However, their use as models of the brain is not uncontroversial. It is worth asking, then, what we can learn from these models and what might be the potential caveats.

The promise of normative modelling is that it provides a neural-like system which can be easily optimized and manipulated to test hypotheses about neural function. In particular, normative models are well suited to understanding how different constraints impact the resulting representations learned by models and exhibited in different brains regions. For example, in this thesis, I have investigated how variable input statistics across training datasets impact network characteristics (Figure 4.3), and how constraining the L1 regularization in the model can produce more retinal- or visual cortical-like receptive fields (Figure 2.5). More generally, normative modelling offers a means of precisely formulating hypotheses about neural computations. For example, in this thesis, by contrasting temporal prediction with alternative normative principles (Figures 2.4, 3.6, Chapter four), I have argued that many neural response properties can be explained as a result of optimizing for temporal prediction.

If these are the successes of normative modelling, what are the challenges? In particular, model interpretability and biological realism have both been levelled as key weaknesses of this modelling approach.[73]

In terms of interpretability, these networks are often referred to as 'black box' models in the sense that the network's underlying computations are relatively obscure. Although these models are constrained by their learning objective during training, the resulting solution which minimizes that objective is generally inscrutable.[241] Thus, even where the model is empirically successful – it captures some phenomena of interest, it brings us no closer to understanding how that solution is derived. Hence, the argument goes that, from the perspective of a scientist rather than an engineer, these networks do not provide an adequate framework for theory building.[73]

However, while the appeal to interpretability is not without merit, it would be premature to discard these models as a neuroscientific tool on these grounds alone. Firstly, as normative models, these networks may be intrinsically valuable as a research tool independent of a mechanistic understanding of their function. In this way, normative modelling is perhaps better understood as a form of empirical investigation.[63,73] Experiments are conducted based on a given hypothesis and the resulting networks are analysed with this hypothesis in mind. Even where the underlying network is not understood completely, valuable insights can still be gained by relating how different constraints can produce different representations in the model.[240] More generally, viewing these networks as entirely inscrutable may be overly pessimistic. Indeed, equivalent complaints could be levelled against systems neuroscience itself in the study of biological neural networks.[242] Moreover, to argue that we have no insight into the internal functioning of these models is incorrect. In this thesis, 'virtual physiology' and ablation experiments have helped produce clear – albeit incomplete – insights into potential network mechanisms. Indeed, interpretable machine learning is a rapidly emerging field, with the toolkit to analyse and better understand these models ever improving.[243]

Setting aside questions of interpretability, the biological realism of these networks has also come under question. Indeed, the idealized point-like units commonly used in these

networks are a far cry from the complexity of biological neurons, while the algorithm used to train these networks – backpropagation – is widely considered to be biologically implausible.[74] On the one hand, incorporating more biologically realistic units – including spiking behaviour[244] and multiple compartments[245] – as well as local learning rules is certainly possible.[196,246] On the other hand, viewing these models as highly abstract may be a feature as much as it is a flaw. By abstracting away from the specifics of neural physiology, these models can – in theory – help reveal the fundamental principles of neural computation.

Thus, on both these fronts – interpretability and biological realism – there are strong grounds for the continued use of normative network models as an explanatory tool in systems neuroscience.

Comparing normative models

Once trained, how can we adjudicate a model's success? Typically, this is approached by comparing different models to some set of experimentally informed criteria. That is, success is judged according to how similar a model is to the brain and how well it performs relative to other models optimized for alternative normative principles.

However, as with all modelling, this approach must reckon with the problem of underdetermination.[247] Firstly, we can never exhaust the potential space of models to rule out that some untested model does not better explain the data. More generally, even among the subset of models we *are* evaluating, there is the possibility that multiple models could equally well describe the phenomena of interest. These concerns are not merely academic. In particular, recent large-scale modelling competitions such as BrainScore,[204] which ranks models according to their neural prediction performance, demonstrate the latter objection well. In the leaderboard rankings, many models using qualitatively different architectures and trained on a range of datasets perform roughly

equivalently. If many models can predict neural responses equally well, what does any individual model actually tell us about the brain?

To resolve this question, one approach would be to expand the space of model-brain comparisons. In the case of BrainScore, models are compared on a functional basis – that is, on their capacity to predict neural responses. However, an alternative approach is to compare models on a structural basis – that is, to what extent do these models recapitulate known connectivity motifs found in the brain. This approach was explored in chapter two of this thesis, where we investigated the emergence of V1-like connectivity across a variety of normative models. Expanding on this research in the future could facilitate a broader comparison of these models and help answer to what extent the structural basis (and therefore the underlying computational mechanisms) of these models mirrors that of the brain.

More generally, as explored extensively in chapter four, there is a tendency to compare across models without considering the source of advantage, which is usually unknown. In other words, models are often directly compared despite differing in their training dataset, network architecture or model objective.[111,112] Accordingly, where the goal is to relate modelling choices to brain function, the presence of confounding variables severely limits the possibility of making strong claims about neural computations.

Finally, even when comparisons are well controlled, there is a tendency to treat model development as a winner-takes-all-process, disregarding the potential merits of lower-ranked models entirely. This approach is flawed because it neglects the possibility that different models might fall along a wider continuum as a result of more fundamental links between models. Indeed, some modelling work is moving in this direction, uncovering the relationship between ideas like efficient coding, prediction and excitation-inhibition balance in artificial and biological networks.[91,248] Similarly, parametric approaches – producing a continuum of models – can tackle this problem

by directly relating hyperparameter choices to the network's learned representations,[89] such as explored in chapters two and four in this thesis.

What conclusions can we take from this discussion? Firstly, normative modelling is arguably most successful when making targeted hypotheses looking at specific phenomena for a given neural system – for example, looking at local connectivity as in chapter two of this thesis. Secondly, the explanatory power of normative modelling is strengthened when comparisons are made by avoiding potential confounds – an area which I explore in detail in chapter four. Overall, this thesis has aimed to keep these ideals in mind, with the goal of moving towards a more integrative outlook to understand how temporal prediction can be used to explain the brain's sensory systems.

6

References

1. Arshavsky, V.Y., Lamb, T.D., and Pugh Jr, E.N. (2002). G proteins and phototransduction. Annu. Rev. Physiol. *64*, 153–187. https://doi.org/10.1146/annurev.physiol.64.082701.102229.

2. Baden, T., Berens, P., Franke, K., Román Rosón, M., Bethge, M., and Euler, T. (2016). The functional diversity of retinal ganglion cells in the mouse. Nature *529*, 345–350. https://doi.org/10.1038/nature16468.

3. Solomon, S.G. (2021). Retinal ganglion cells and the magnocellular, parvocellular, and koniocellular subcortical visual pathways from the eye to the brain. In Handbook of Clinical Neurology (Elsevier), pp. 31–50.

4. Kim, U.S., Mahroo, O.A., Mollon, J.D., and Yu-Wai-Man, P. (2021). Retinal ganglion cells—diversity of cell types and clinical relevance. Front. Neurol. *12*, 661938. https://doi.org/10.3389/fneur.2021.661938.

5. Nassi, J.J., and Callaway, E.M. (2009). Parallel processing strategies of the primate visual system. Nat. Rev. Neurosci. *10*, 360–372. https://doi.org/10.1038/nrn2619.

6. Livingstone, M., and Hubel, D. (1988). Segregation of form, color, movement, and depth: anatomy, physiology, and perception. Science *240*, 740–749. https://doi.org/10.1126/science.3283936.

7. Van Essen, D.C., Anderson, C.H., and Felleman, D.J. (1992). Information processing in the primate visual system: An integrated systems perspective. Science *255*, 419–423. https://doi.org/10.1126/science.1734518.

8. Wong-Riley, M. (1979). Changes in the visual system of monocularly sutured or enucleated cats demonstrable with cytochrome oxidase histochemistry. Brain Res. *171*, 11–28. https://doi.org/10.1016/0006-8993(79)90728-5.

9. Matsubara, J.A., and Boyd, J.D. (2002). Relationship of LGN afferents and cortical efferents to cytochrome oxidase blobs. In The Cat Primary Visual Cortex (Elsevier), pp. 221–258.

10. Federer, F., Ichida, J.M., Jeffs, J., Schiessl, I., McLoughlin, N., and Angelucci, A. (2009). Four projection streams from primate V1 to the cytochrome oxidase stripes of V2. J. Neurosci. *29*, 15455–15471. https://doi.org/10.1523/JNEUROSCI.1648-09.2009.

11. Yabuta, N.H., and Callaway, E.M. (1998). Functional streams and local connections of layer 4C neurons in primary visual cortex of the macaque monkey. J. Neurosci. *18*, 9489–9499. https://doi.org/10.1523/JNEUROSCI.18-22-09489.1998.

12. Hendry, S.H., and Reid, R.C. (2000). The koniocellular pathway in primate vision. Annu. Rev. Neurosci. *23*, 127–153. https://doi.org/10.1146/annurev.neuro.23.1.127.

13. Livingstone, M.S., and Hubel, D.H. (1983). Specificity of cortico-cortical connections in monkey visual system. Nature *304*, 531–534. https://doi.org/10.1038/304531a0.

14. Mishkin, M., Ungerleider, L.G., and Macko, K.A. (1983). Object vision and spatial vision: two cortical pathways. Trends Neurosci. *6*, 414–417. https://doi.org/10.1016/0166-2236(83)90190-X.

15. Born, R.T., and Bradley, D.C. (2005). Structure and function of visual area MT. Annu Rev Neurosci *28*, 157–189. https://doi.org/10.1146/annurev.neuro.26.041002.131052.

16. Nayebi, A., Kong, N.C., Zhuang, C., Gardner, J.L., Norcia, A.M., and Yamins, D.L. (2023). Mouse visual cortex as a limited resource system that self-learns an ecologically-general representation. PLOS Comput. Biol. *19*, e1011506. https://doi.org/10.1371/journal.pcbi.1011506.

17. Seabrook, T.A., Burbridge, T.J., Crair, M.C., and Huberman, A.D. (2017). Architecture, function, and assembly of the mouse visual system. Annu. Rev. Neurosci. *40*, 499–538. https://doi.org/10.1146/annurev-neuro-071714-033842.

18. Glickfeld, L.L., and Olsen, S.R. (2017). Higher-order areas of the mouse visual cortex. Annu. Rev. Vis. Sci. *3*, 251–273. https://doi.org/10.1146/annurev-vision-102016-061331.

19. Murakami, T., Matsui, T., and Ohki, K. (2017). Functional segregation and development of mouse higher visual areas. J. Neurosci. *37*, 9424–9437. https://doi.org/10.1523/JNEUROSCI.0731-17.2017.

20. Wang, Q., Sporns, O., and Burkhalter, A. (2012). Network analysis of corticocortical connections reveals ventral and dorsal processing streams in mouse visual cortex. J. Neurosci. *32*, 4386–4399. https://doi.org/10.1523/JNEUROSCI.6063-11.2012.

21. Han, X., and Bonin, V. (2024). Higher-order cortical and thalamic pathways shape visual processing streams in the mouse cortex. Curr. Biol. *34*, 5671–5684. https://doi.org/10.1016/j.cub.2024.10.048.

22. Saleem, A.B. (2020). Two stream hypothesis of visual processing for navigation in mouse. Curr. Opin. Neurobiol. *64*, 70–78. https://doi.org/10.1016/j.conb.2020.03.009.

23. Rowley, D.P., and Sedigh-Sarvestani, M. (2024). Structural and functional evidence supports re-defining mouse higher order visual areas into a single area V2. bioRxiv, 2024–10. https://doi.org/10.1101/2024.10.10.617533.

24. Barnstedt, O., Keating, P., Weissenberger, Y., King, A.J., and Dahmen, J.C. (2015). Functional microarchitecture of the mouse dorsal inferior colliculus revealed through in vivo two-photon calcium imaging. J. Neurosci. *35*, 10927–10939. https://doi.org/10.1523/JNEUROSCI.0103-15.2015.

25. Bajo, V.M., Nodal, F.R., Bizley, J.K., Moore, D.R., and King, A.J. (2007). The ferret auditory cortex: Descending projections to the inferior colliculus. Cereb. Cortex *17*, 475–491. https://doi.org/10.1093/cercor/bhj164.

26. Blackwell, J.M., Lesicko, A.M., Rao, W., De Biasi, M., and Geffen, M.N. (2020). Auditory cortex shapes sound responses in the inferior colliculus. Elife *9*, e51890. https://doi.org/10.7554/eLife.51890.

27. Xiong, X.R., Liang, F., Zingg, B., Ji, X., Ibrahim, L.A., Tao, H.W., and Zhang, L.I. (2015). Auditory cortex controls sound-driven innate defense behaviour through corticofugal projections to inferior colliculus. Nat. Commun. *6*, 7224. https://doi.org/10.1038/ncomms8224.

28. Druga, R., and Syka, J. (1984). Projections from auditory structures to the superior colliculus in the rat. Neurosci. Lett. *45*, 247–252. https://doi.org/10.1016/0304-3940(84)90234-9.

29. Calford, M., and Aitkin, L. (1983). Ascending projections to the medial geniculate body of the cat: evidence for multiple, parallel auditory pathways through thalamus. J. Neurosci. *3*, 2365. https://doi.org/10.1523/JNEUROSCI.03-11-02365.1983.

30. Huffman, R.F., and Henson, O.W. (1990). The descending auditory pathway and acousticomotor systems: connections with the inferior colliculus. Brain Res. Rev. *15*, 295–323. https://doi.org/10.1016/0165-0173(90)90005-9.

31. Vasquez-Lopez, S.A., Weissenberger, Y., Lohse, M., Keating, P., King, A.J., and Dahmen, J.C. (2017). Thalamic input to auditory cortex is locally heterogeneous but globally tonotopic. Elife *6*, e25141. https://doi.org/10.7554/eLife.25141.

32. Woods, D.L., Herron, T.J., Cate, A.D., Yund, E.W., Stecker, G.C., Rinne, T., and Kang, X. (2010). Functional properties of human auditory cortical fields. Front. Syst. Neurosci. *4*. https://doi.org/10.3389/fnsys.2010.00155.

33. Kaas, J.H., and Hackett, T.A. (2000). Subdivisions of auditory cortex and processing streams in primates. Proc. Natl. Acad. Sci. *97*, 11793–11799. https://doi.org/10.1073/pnas.97.22.11793.

34. Bizley, J.K., Nodal, F.R., Nelken, I., and King, A.J. (2005). Functional organization of ferret auditory cortex. Cereb. Cortex *15*, 1637–1653. https://doi.org/10.1093/cercor/bhi042.

35. Souffi, S., Nodal, F.R., Bajo, V.M., and Edeline, J.-M. (2021). When and how does the auditory cortex influence subcortical auditory structures? New insights about the roles of descending cortical projections. Front. Neurosci. *15*. https://doi.org/10.3389/fnins.2021.690223.

36. Lee, T.-Y., Weissenberger, Y., King, A.J., and Dahmen, J.C. (2024). Midbrain encodes sound detection behavior without auditory cortex. eLife *12*, RP89950. https://doi.org/10.7554/eLife.89950.

37. Escera, C. (2023). Contributions of the subcortical auditory system to predictive coding and the neural encoding of speech. Curr. Opin. Behav. Sci. *54*, 101324. https://doi.org/10.1016/j.cobeha.2023.101324.

38. King, A.J., Teki, S., and Willmore, B.D.B. (2018). Recent advances in understanding the auditory cortex. F1000Research *7*, F1000 Faculty Rev-1555. https://doi.org/10.12688/f1000research.15580.1.

39. Atencio, C.A., Sharpee, T.O., and Schreiner, C.E. (2012). Receptive field dimensionality increases from the auditory midbrain to cortex. J. Neurophysiol. *107*, 2594–2603. https://doi.org/10.1152/jn.01025.2011.

40. Atencio, C.A., and Sharpee, T.O. (2017). Multidimensional receptive field processing by cat primary auditory cortical neurons. Neuroscience *359*, 130–141. https://doi.org/10.1016/j.neuroscience.2017.07.003.

41. Homma, N.Y., Atencio, C.A., and Schreiner, C.E. (2021). Plasticity of multidimensional receptive fields in core rat auditory cortex directed by sound statistics. Neuroscience *467*, 150–170. https://doi.org/10.1016/j.neuroscience.2021.04.028.

42. David, S.V., Fritz, J.B., and Shamma, S.A. (2012). Task reward structure shapes rapid receptive field plasticity in auditory cortex. Proc. Natl. Acad. Sci. U. S. A. *109*, 2144–2149. https://doi.org/10.1073/pnas.1117717109.

43. Da Costa, S., van der Zwaag, W., Miller, L.M., Clarke, S., and Saenz, M. (2013). Tuning in to sound: frequency-selective attentional filter in human primary auditory cortex. J. Neurosci. *33*, 1858–1863. https://doi.org/10.1523/JNEUROSCI.4405-12.2013.

44. Kauramäki, J., Jääskeläinen, I.P., and Sams, M. (2007). Selective attention increases both gain and feature selectivity of the human auditory cortex. PLoS One *2*, e909. https://doi.org/10.1371/journal.pone.0000909.

45. Weinberger, N.M. (2010). Reconceptualizing the primary auditory cortex: learning, memory and specific plasticity. In The auditory cortex (Springer), pp. 465–491.

46. Mohn, J.L., Downer, J.D., O'Connor, K.N., Johnson, J.S., and Sutter, M.L. (2021). Choice-related activity and neural encoding in primary auditory cortex and lateral belt during feature-selective attention. J. Neurophysiol. *125*, 1920–1937. https://doi.org/10.1152/jn.00406.2020.

47. Francis, N.A., Winkowski, D.E., Sheikhattar, A., Armengol, K., Babadi, B., and Kanold, P.O. (2018). Small networks encode decision-making in primary auditory cortex. Neuron *97*, 885–897. https://doi.org/10.1016/j.neuron.2018.01.019.

48. Griffiths, T.D., and Warren, J.D. (2004). What is an auditory object? Nat. Rev. Neurosci. *5*, 887–892. https://doi.org/10.1038/nrn1538.

49. Ding, N., and Simon, J.Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. Proc. Natl. Acad. Sci. *109*, 11854–11859. https://doi.org/10.1073/pnas.1205381109.

50. Christison-Lagay, K.L., and Cohen, Y.E. (2018). The contribution of primary auditory cortex to auditory categorization in behaving monkeys. Front. Neurosci. *12*. https://doi.org/10.3389/fnins.2018.00601.

51. Heller, C.R., Hamersky, G.R., and David, S.V. (2024). Task-specific invariant representation in auditory cortex. Elife *12*, RP89936. https://doi.org/10.7554/elife.89936.2.

52. King, A.J., and Nelken, I. (2009). Unraveling the principles of auditory cortical processing: can we learn from the visual system? Nat. Neurosci. *12*, 698–701. https://doi.org/10.1038/nn.2308.

53. Poort, J., Wilmes, K.A., Blot, A., Chadwick, A., Sahani, M., Clopath, C., Mrsic-Flogel, T.D., Hofer, S.B., and Khan, A.G. (2022). Learning and attention increase visual response selectivity through distinct mechanisms. Neuron *110*, 686–697. https://doi.org/10.1016/j.neuron.2021.11.016.

54. Flossmann, T., and Rochefort, N.L. (2021). Spatial navigation signals in rodent visual cortex. Curr. Opin. Neurobiol. *67*, 163–173. https://doi.org/10.1016/j.conb.2020.11.004.

55. Petro, L., Paton, A., and Muckli, L. (2017). Contextual modulation of primary visual cortex by auditory signals. Philos. Trans. R. Soc. B Biol. Sci. *372*, 20160104. https://doi.org/10.1098/rstb.2016.0104.

56. Booth, M., and Rolls, E.T. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. Cereb. Cortex N. Y. NY 1991 *8*, 510–523. https://doi.org/10.1093/cercor/8.6.510.

57. Hung, C.P., Kreiman, G., Poggio, T., and DiCarlo, J.J. (2005). Fast readout of object identity from macaque inferior temporal cortex. Science *310*, 863–866. https://doi.org/10.1126/science.1117593.

58. Wallisch, P., and Movshon, J.A. (2019). Responses of neurons in macaque MT to unikinetic plaids. J. Neurophysiol. *122*, 1937–1945. https://doi.org/10.1152/jn.00486.2019.

59. Goldstein, R.E. (2018). Are theoretical results 'Results'? Elife *7*, e40018. https://doi.org/10.7554/eLife.40018.

60. Bialek, W. (2017). Perspectives on theory at the interface of physics and biology. Rep. Prog. Phys. *81*, 012601. https://doi.org/10.1088/1361-6633/aa995b.

61. Phillips, R. (2015). Theory in biology: Figure 1 or Figure 7? Trends Cell Biol. *25*, 723–729. https://doi.org/10.1016/j.tcb.2015.10.007.

62. Shou, W., Bergstrom, C.T., Chakraborty, A.K., and Skinner, F.K. (2015). Theory, models and biology. Elife *4*, e07158. https://doi.org/10.7554/eLife.07158.

63. Levenstein, D., Alvarez, V.A., Amarasingham, A., Azab, H., Chen, Z.S., Gerkin, R.C., Hasenstaub, A., Iyer, R., Jolivet, R.B., and Marzen, S. (2023). On the role of theory and modeling in neuroscience. J. Neurosci. *43*, 1074–1088. https://doi.org/10.1523/JNEUROSCI.1179-22.2022.

64. Stevenson, I.H., and Kording, K.P. (2011). How advances in neural recording affect data analysis. Nat. Neurosci. *14*, 139–142. https://doi.org/10.1038/nn.2731.

65. Callebaut, W. (2012). Scientific perspectivism: A philosopher of science's response to the challenge of big data biology. Stud. Hist. Philos. Sci. Part C Stud. Hist. Philos. Biol. Biomed. Sci. *43*, 69–80. https://doi.org/10.1016/j.shpsc.2011.10.007.

66. Chirimuuta, M. (2014). Minimal models and canonical neural computations: The distinctness of computational explanation in neuroscience. Synthese *191*, 127–153. https://doi.org/10.1007/s11229-013-0369-y.

67. Ferster, D., and Miller, K.D. (2000). Neural mechanisms of orientation selectivity in the visual cortex. Annu. Rev. Neurosci. *23*, 441–471. https://doi.org/10.1146/annurev.neuro.23.1.441.

68. Levy, A. (2016). The unity of neuroscience: a flat view. Synthese *193*, 3843–3863. https://doi.org/10.1007/s11229-016-1256-0.

69. Kording, K. (2007). Decision theory: what" should" the nervous system do? Science *318*, 606–610. https://doi.org/10.1126/science.1142998.

70. Benjamin, A.S., Zhang, L.-Q., Qiu, C., Stocker, A.A., and Kording, K.P. (2022). Efficient neural codes naturally emerge through gradient descent learning. Nat. Commun. *13*, 7972. https://doi.org/10.1038/s41467-022-35659-7.

71. Singer, Y., Taylor, L., Willmore, B.D., King, A.J., and Harper, N.S. (2023). Hierarchical temporal prediction captures motion processing along the visual pathway. eLife *12*, e52599. https://doi.org/10.7554/eLife.52599.

72. Kanwisher, N., Khosla, M., and Dobs, K. (2023). Using artificial neural networks to ask 'why' questions of minds and brains. Trends Neurosci. *46*, 240–254. https://doi.org/10.1016/j.tins.2022.12.008.

73. Cichy, R.M., and Kaiser, D. (2019). Deep neural networks as scientific models. Trends Cogn. Sci. *23*, 305–317. https://doi.org/10.1016/j.tics.2019.01.009.

74. Richards, B.A., Lillicrap, T.P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R.P., de Berker, A., Ganguli, S., et al. (2019). A deep learning framework for neuroscience. Nat. Neurosci. *22*, 1761–1770. https://doi.org/10.1038/s41593-019-0520-2.

75. Block, H.-D. (1962). The perceptron: A model for brain functioning. i. Rev. Mod. Phys. *34*, 123. https://doi.org/10.1103/RevModPhys.34.123.

76. Minsky, M., and Papert, S. (1969). An introduction to computational geometry. Camb. Tiass HIT *479*, 104. https://doi.org/10.1109/tit.1969.1054388.

77. Alom, Z., Taha, T.M., Yakopcic, C., Westberg, S., Sidike, P., and Nasrin, M.S. The history began from AlexNet: A comprehensive survey on deep learning approaches. https://doi.org/10.48550/arxiv.1803.01164.

78. Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2017). ImageNet classification with deep convolutional neural networks. Commun. ACM *60*, 84–90. https://doi.org/10.1145/3065386.

79. Sutton, R. (2019). The bitter lesson. Incomplete Ideas Blog *13*, 38.

80. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog *1*, 9.

81. Lim, B., and Zohren, S. (2021). Time-series forecasting with deep learning: a survey. Philos. Trans. R. Soc. A *379*, 20200209. https://doi.org/10.1098/rsta.2020.0209.

82. LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. nature *521*, 436–444. https://doi.org/10.1038/nature14539.

83. Kriegeskorte, N., and Douglas, P.K. (2019). Interpreting encoding and decoding models. Curr. Opin. Neurobiol. *55*, 167–179. https://doi.org/10.1016/j.conb.2019.04.002.

84. van Gerven, M.A.J. (2017). A primer on encoding models in sensory neuroscience. J. Math. Psychol. *76*, 172–183. https://doi.org/10.1016/j.jmp.2016.06.009.

85. Pennington, J.R., and David, S.V. (2023). A convolutional neural network provides a generalizable model of natural sound coding by neural populations in auditory cortex. PLOS Comput. Biol. *19*, e1011110. https://doi.org/10.1371/journal.pcbi.1011110.

86. O'Keefe, J., Burgess, N., Donnett, J.G., Jeffery, K.J., and Maguire, E.A. (1998). Place cells, navigational accuracy, and the human hippocampus. Philos. Trans. R. Soc. Lond. B. Biol. Sci. *353*, 1333–1340. https://doi.org/10.1098/rstb.1998.0287.

87. Banino, A., Barry, C., Uria, B., Blundell, C., Lillicrap, T., Mirowski, P., Pritzel, A., Chadwick, M.J., Degris, T., Modayil, J., et al. (2018). Vector-based navigation using grid-like representations in artificial agents. Nature *557*, 429–433. https://doi.org/10.1038/s41586-018-0102-6.

88. Cueva, C.J., and Wei, X.-X. (2018). Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. ArXiv Prepr. ArXiv180307770. https://doi.org/10.48550/arxiv.1803.07770.

89. Klavinskis-Whiting, S., Fristed, E., Singer, Y., Iacaruso, M.F., King, A.J., and Harper, N.S. Prediction of future input explains lateral connectivity in primary visual cortex. Curr. Biol. https://doi.org/10.1016/j.cub.2024.11.073.

90. Singer, Y., Teramoto, Y., Willmore, B.D., Schnupp, J.W., King, A.J., and Harper, N.S. (2018). Sensory cortex is optimized for prediction of future input. eLife *7*, e31557. https://doi.org/10.7554/eLife.31557.

91. Chalk, M., Marre, O., and Tkačik, G. (2018). Toward a unified theory of efficient, predictive, and sparse coding. Proc. Natl. Acad. Sci. *115*, 186–191. https://doi.org/10.1073/pnas.1711114115.

92. Manookin, M.B., and Rieke, F. (2023). Two sides of the same coin: Efficient and predictive neural coding. Annu. Rev. Vis. Sci. *9*, 293–311. https://doi.org/10.1146/annurev-vision-112122-020941.

93. Barlow, H.B. (1961). Possible principles underlying the transformation of sensory messages. Sens. Commun. *1*, 217–233. https://doi.org/10.7551/mitpress/9780262518420.003.0013.

94. MacKay, D.M. (1956). Towards an information-flow model of human behaviour. Br. J. Psychol. *47*, 30. https://doi.org/10.1111/j.2044-8295.1956.tb00559.x.

95. Srinivasan, M.V., Laughlin, S.B., and Dubs, A. (1982). Predictive coding: a fresh view of inhibition in the retina. Proc. R. Soc. Lond. B Biol. Sci. *216*, 427–459. https://doi.org/10.1098/rspb.1982.0085.

96. Lewicki, M.S. (2002). Efficient coding of natural sounds. Nat. Neurosci. *5*, 356–363. https://doi.org/10.1038/nn831.

97. Röth, K., Shao, S., and Gjorgjieva, J. (2021). Efficient population coding depends on stimulus convergence and source of noise. PLoS Comput. Biol. *17*, e1008897. https://doi.org/10.1371/journal.pcbi.1008897.

98. Olshausen, B.A., and Field, D.J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? Vision Res. *37*, 3311–3325. https://doi.org/10.1016/S0042-6989(97)00169-7.

99. Rao, R.P.N., and Ballard, D.H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci. *2*, 79–87. https://doi.org/10.1038/4580.

100. Karklin, Y., and Simoncelli, E. (2011). Efficient coding of natural images with a population of noisy linear-nonlinear neurons. Adv. Neural Inf. Process. Syst. *24*.

101. Salisbury, J.M., and Palmer, S.E. (2016). Optimal prediction in the retina and natural motion statistics. J. Stat. Phys. *162*, 1309–1323. https://doi.org/10.1007/s10955-015-1439-y.

102. Palmer, S.E., Marre, O., Berry, M.J., and Bialek, W. (2015). Predictive information in a sensory population. Proc. Natl. Acad. Sci. *112*, 6908–6913. https://doi.org/10.1073/pnas.1506855112.

103. Machens, C.K., Gollisch, T., Kolesnikova, O., and Herz, A.V.M. (2005). Testing the efficiency of sensory coding with optimal stimulus ensembles. Neuron *47*, 447–456. https://doi.org/10.1016/j.neuron.2005.06.015.

104. Lamberti, M., Tripathi, S., van Putten, M.J.A.M., Marzen, S., and le Feber, J. (2023). Prediction in cultured cortical neural networks. PNAS Nexus *2*, pgad188. https://doi.org/10.1093/pnasnexus/pgad188.

105. Bialek, W., Nemenman, I., and Tishby, N. (2001). Predictability, complexity, and learning. Neural Comput. *13*, 2409–2463. https://doi.org/10.1162/089976601753195969.

106. Creutzig, F., and Sprekeler, H. (2008). Predictive coding and the slowness principle: An information-theoretic approach. Neural Comput. *20*, 1026–1041. https://doi.org/10.1162/neco.2008.01-07-455.

107. Wiskott, L., and Sejnowski, T.J. (2002). Slow feature analysis: Unsupervised learning of invariances. Neural Comput. *14*, 715–770. https://doi.org/10.1162/089976602317318938.

108. Berkes, P., and Wiskott, L. (2005). Slow feature analysis yields a rich repertoire of complex cell properties. J. Vis. *5*, 9–9. https://doi.org/10.1167/5.6.9.

109. Johnson, P.A., Blom, T., van Gaal, S., Feuerriegel, D., Bode, S., and Hogendoorn, H. (2023). Position representations of moving objects align with real-time position in the early visual response. eLife *12*, e82424. https://doi.org/10.7554/eLife.82424.

110. Millidge, B., Tang, M., Osanlouy, M., Harper, N.S., and Bogacz, R. (2024). Predictive coding networks for temporal prediction. PLOS Comput. Biol. *20*, e1011183. https://doi.org/10.1371/journal.pcbi.1011183

111. Tuckute, G., Feather, J., Boebinger, D., and McDermott, J.H. (2023). Many but not all deep neural network audio models capture brain responses and exhibit

correspondence between model stages and brain regions. Plos Biol. *21*, e3002366. https://doi.org/10.1371/journal.pbio.3002366.

112. Yamins, D.L., and DiCarlo, J.J. (2016). Using goal-driven deep learning models to understand sensory cortex. Nat. Neurosci. *19*, 356–365. https://doi.org/10.1038/nn.4244.

113. Burbank, K.S. (2015). Mirrored STDP implements autoencoder learning in a network of spiking neurons. PLoS Comput. Biol. *11*, e1004566. https://doi.org/10.1371/journal.pcbi.1004566.

114. Bhand, M., Mudur, R., Suresh, B., Saxe, A., and Ng, A. (2011). Unsupervised learning models of primary cortical receptive fields and receptive field plasticity. Adv. Neural Inf. Process. Syst. *24*.

115. Oord, A. van den, Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. ArXiv Prepr. ArXiv180703748. https://doi.org/10.48550/arxiv.1807.03748.

116. Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. ArXiv Prepr. ArXiv190405862. https://doi.org/10.48550/arxiv.1904.05862.

117. Francl, A., and McDermott, J.H. (2022). Deep neural network models of sound localization reveal how perception is adapted to real-world environments. Nat. Hum. Behav. *6*, 111–133. https://doi.org/10.1038/s41562-021-01244-z.

118. Iacaruso, M.F., Gasler, I.T., and Hofer, S.B. (2017). Synaptic organization of visual space in primary visual cortex. Nature *547*, 449–452. https://doi.org/10.1038/nature23019.

119. Cossell, L., Iacaruso, M.F., Muir, D.R., Houlton, R., Sader, E.N., Ko, H., Hofer, S.B., and Mrsic-Flogel, T.D. (2015). Functional organization of excitatory synaptic strength in primary visual cortex. Nature *518*, 399–403. https://doi.org/10.1038/nature14182.

120. Ko, H., Hofer, S.B., Pichler, B., Buchanan, K.A., Sjöström, P.J., and Mrsic-Flogel, T.D. (2011). Functional specificity of local synaptic connections in neocortical networks. Nature *473*, 87–91. https://doi.org/10.1038/nature09880.

121. Ding, Z., Fahey, P.G., Papadopoulos, S., Wang, E., Celii, B., Papadopoulos, C., Kunin, A.B., Chang, A., Fu, J., Ding, Z., et al. (2023). Functional connectomics reveals general wiring rule in mouse visual cortex. Preprint at bioRxiv, https://doi.org/10.1101/2023.03.13.531369.

122. Rossi, L.F., Harris, K.D., and Carandini, M. (2020). Spatial connectivity matches direction selectivity in visual cortex. Nature *588*, 648–652. https://doi.org/10.1038/s41586-020-2894-4.

123. Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M.C., DiCarlo, J.J., and Yamins, D.L. (2021). Unsupervised neural network models of the ventral visual stream. Proc. Natl. Acad. Sci. *118*. https://doi.org/10.1073/pnas.2014196118.

124. O'Reilly, R.C., Wyatte, D., and Rohrlich, J. (2014). Learning through time in the thalamocortical loops. https://doi.org/10.48550/arXiv.1407.3432.

125. Chariker, L., Shapley, R., and Young, L.-S. (2016). Orientation selectivity from very sparse LGN inputs in a comprehensive model of macaque V1 cortex. J. Neurosci. *36*, 12368–12384. https://doi.org/10.1523/JNEUROSCI.2603-16.2016.

126. Bastos, A.M., Usrey, W.M., Adams, R.A., Mangun, G.R., Fries, P., and Friston, K.J. (2012). Canonical microcircuits for predictive coding. Neuron *76*, 695–711. https://doi.org/10.1016/j.neuron.2012.10.038.

127. Muckli, L., and Petro, L.S. (2013). Network interactions: Non-geniculate input to V1. Curr. Opin. Neurobiol. *23*, 195–201. https://doi.org/10.1016/j.conb.2013.01.020.

128. Adelson, E.H., and Bergen, J.R. (1985). Spatiotemporal energy models for the perception of motion. Josa A *2*, 284–299. https://doi.org/10.1364/josaa.2.000284.

129. Rust, N.C., Schwartz, O., Movshon, J.A., and Simoncelli, E.P. (2005). Spatiotemporal elements of macaque V1 receptive fields. Neuron *46*, 945–956. https://doi.org/10.1016/j.neuron.2005.05.021.

130. Niell, C.M., and Stryker, M.P. (2008). Highly selective receptive fields in mouse visual cortex. J. Neurosci. *28*, 7520–7536. https://doi.org/10.1523/JNEUROSCI.0623-08.2008.

131. Durand, S., Iyer, R., Mizuseki, K., Vries, S. de, Mihalas, S., and Reid, R.C. (2016). A comparison of visual response properties in the lateral geniculate nucleus and primary visual cortex of awake and anesthetized mice. J. Neurosci. *36*, 12144–12156. https://doi.org/10.1523/JNEUROSCI.1741-16.2016.

132. Gavornik, J.P., and Bear, M.F. (2014). Learned spatiotemporal sequence recognition and prediction in primary visual cortex. Nat. Neurosci. *17*, 732–737. https://doi.org/10.1038/nn.3683.

133. Fiser, A., Mahringer, D., Oyibo, H.K., Petersen, A.V., Leinweber, M., and Keller, G.B. (2016). Experience-dependent spatial expectations in mouse visual cortex. Nat. Neurosci. *19*, 1658–1664. https://doi.org/10.1038/nn.4385.

134. Westerberg, J.A., Xiong, Y.S., Nejat, H., Sennesh, E., Durand, S., Cabasco, H., Belski, H., Gillis, R., Loeffler, H., Bawany, A., et al. (2024). Stimulus history, not expectation, drives sensory prediction errors in mammalian cortex. bioRxiv, 2024.10.02.616378. https://doi.org/10.1101/2024.10.02.616378.

135. Price, B.H., Jensen, C.M., Khoudary, A.A., and Gavornik, J.P. (2023). Expectation violations produce error signals in mouse V1. Cereb. Cortex *33*, 8803–8820. https://doi.org/10.1093/cercor/bhad163.

136. Jamali, S., Bagur, S., Brémont, E., Van Kerkoerle, T., Dehaene, S., and Bathellier, B. (2024). Parallel mechanisms signal a hierarchy of sequence structure violations in the auditory cortex. https://doi.org/10.7554/elife.102702.1.

137. Li, J., Liao, X., Zhang, J., Wang, M., Yang, N., Zhang, J., Lv, G., Li, H., Lu, J., Ding, R., et al. (2017). Primary auditory cortex is required for anticipatory motor response. Cereb. Cortex *27*, 3254–3271. https://doi.org/10.1093/cercor/bhx079.

138. Wang, M., Li, R., Li, J., Zhang, J., Chen, X., Zeng, S., and Liao, X. (2018). Frequency selectivity of echo responses in the mouse primary auditory cortex. Sci. Rep. *8*, 49. https://doi.org/10.1038/s41598-017-18465-w.

139. Musall, S., Haiss, F., Weber, B., and von der Behrens, W. (2017). Deviant processing in the primary somatosensory cortex. Cereb. Cortex *27*, 863–876. https://doi.org/10.1093/cercor/bhv283.

140. Auksztulewicz, R., Rajendran, V.G., Peng, F., Schnupp, J.W.H., and Harper, N.S. (2023). Omission responses in local field potentials in rat auditory cortex. BMC Biol. *21*, 130. https://doi.org/10.1186/s12915-023-01592-4.

141. Tang, M.F., Kheradpezhouh, E., Lee, C.C., Dickinson, J.E., Mattingley, J.B., and Arabzadeh, E. (2023). Expectation violations enhance neuronal encoding of sensory information in mouse primary visual cortex. Nat. Commun. *14*, 1196. https://doi.org/10.1038/s41467-023-36608-8.

142. Allen Brain Observatory (2019). Neuropixels Visual Coding.

143. Scholl, B., Wilson, D.E., Jaepel, J., and Fitzpatrick, D. (2019). Functional logic of layer 2/3 inhibitory connectivity in the ferret visual cortex. Neuron *104*, 451-457.e3. https://doi.org/10.1016/j.neuron.2019.08.004.

144. Harris, K.D., and Mrsic-Flogel, T.D. (2013). Cortical connectivity and sensory coding. Nature *503*, 51–58. https://doi.org/10.1038/nature12654.

145. Simonyan, K., and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. Preprint at arXiv, https://doi.org/10.48550/arXiv.1409.1556.

146. Lotter, W., Kreiman, G., and Cox, D. (2017). Deep predictive coding networks for video prediction and unsupervised learning. Preprint at arXiv, https://doi.org/10.48550/arXiv.1605.08104.

147. Taylor, L., Zenke, F., King, A.J., and Harper, N.S. (2024). Temporal prediction captures retinal spiking responses across animal species. Preprint, https://doi.org/10.1101/2024.03.26.586771.

148. Cohen, U., Chung, S., Lee, D.D., and Sompolinsky, H. (2020). Separability and geometry of object manifolds in deep neural networks. Nat. Commun. *11*, 746. https://doi.org/10.1038/s41467-020-14578-5.

149. Rousseeuw, P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

150. Di Lazzaro, V., Capone, F., Paolucci, M., Assenza, F., Brunelli, N., Ricci, L., and Florio, L. (2016). Canonical cortical circuits: current evidence and theoretical implications. Neurosci. Neuroeconomics, 1. https://doi.org/10.2147/NAN.S70816.

151. Nelson, S.B. (2002). Cortical microcircuits: diverse or canonical? Neuron *36*, 19–27. https://doi.org/10.1016/S0896-6273(02)00944-3.

152. Shepherd, G.M.G., and Yamawaki, N. (2021). Untangling the cortico-thalamo-cortical loop: cellular pieces of a knotty circuit puzzle. Nat. Rev. Neurosci. *22*, 389–406. https://doi.org/10.1038/s41583-021-00459-3.

153. Brown, S.P., and Hestrin, S. (2009). Cell-type identity: a key to unlocking the function of neocortical circuits. Curr. Opin. Neurobiol. *19*, 415–421. https://doi.org/10.1016/j.conb.2009.07.011.

154. Reid, R.C. (2012). From functional architecture to functional connectomics. Neuron *75*, 209–217. https://doi.org/10.1016/j.neuron.2012.06.031.

155. Ecke, G.A., Bruijns, S.A., Hoelscher, J., Mikulasch, F.A., Witschel, T., Arrenberg, A.B., and Mallot, H.A. (2019). Sparse coding predicts optic flow specificities of zebrafish pretectal neurons. Neural Comput. Appl., 1–10. https://doi.org/10.1007/s00521-019-04500-6.

156. Capparelli, F., Pawelzik, K., and Ernst, U. (2019). Constrained inference in sparse coding reproduces contextual effects and predicts laminar neural dynamics. PLOS Comput. Biol. *15*, e1007370. https://doi.org/10.1371/journal.pcbi.1007370.

157. Garrigues, P., and Olshausen, B. (2007). Learning horizontal connections in a sparse coding model of natural images. Adv. Neural Inf. Process. Syst. *20*.

158. Pachitariu, M., and Sahani, M. (2017). Visual motion computation in recurrent neural networks (Neuroscience) https://doi.org/10.1101/099101.

159. Iyer, R., and Mihalas, S. (2017). Cortical circuits implement optimal context integration. bioRxiv, 158360. https://doi.org/10.1101/158360.

160. Bengio, Y., Lee, D.-H., Bornschein, J., Mesnard, T., and Lin, Z. (2016). Towards biologically plausible deep learning. Preprint at arXiv. https://doi.org/10.48550/arxiv.1502.04156.

161. Ko, H., Cossell, L., Baragli, C., Antolik, J., Clopath, C., Hofer, S.B., and Mrsic-Flogel, T.D. (2013). The emergence of functional microcircuits in visual cortex. Nature *496*, 96–100. https://doi.org/10.1038/nature12015.

162. Ko, H., Mrsic-Flogel, T.D., and Hofer, S.B. (2014). Emergence of feature-specific connectivity in cortical microcircuits in the absence of visual experience. J. Neurosci. *34*, 9812–9816. https://doi.org/10.1523/JNEUROSCI.0875-14.2014.

163. Kiorpes, L. (2015). Visual development in primates: neural mechanisms and critical periods. Dev. Neurobiol. *75*, 1080–1090. https://doi.org/10.1002/dneu.22276.

164. Meyer, H.S., Schwarz, D., Wimmer, V.C., Schmitt, A.C., Kerr, J.N., Sakmann, B., and Helmstaedter, M. (2011). Inhibitory interneurons in a cortical column form hot zones of inhibition in layers 2 and 5A. Proc. Natl. Acad. Sci. *108*, 16807–16812. https://doi.org/10.1073/pnas.1113648108.

165. van Versendaal, D., and Levelt, C.N. (2016). Inhibitory interneurons in visual cortical plasticity. Cell. Mol. Life Sci. *73*, 3677–3691. https://doi.org/10.1007/s00018-016-2264-4.

166. Sahani, M., and Linden, J. (2002). How linear are auditory cortical responses? Adv. Neural Inf. Process. Syst. *15*.

167. Rahman, M., Willmore, B.D., King, A.J., and Harper, N.S. (2020). Simple transformations capture auditory input to cortex. Proc. Natl. Acad. Sci. *117*, 28442–28451. https://doi.org/10.1073/pnas.1922033117.

168. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2.

169. Schoppe, O., Harper, N.S., Willmore, B.D., King, A.J., and Schnupp, J.W. (2016). Measuring the performance of neural models. Front. Comput. Neurosci. *10*, 10. https://doi.org/10.3389/fncom.2016.00010.

170. Hsu, A., Borst, A., and Theunissen, F.E. (2004). Quantifying variability in neural responses and its application for the validation of model predictions. Netw. Comput. Neural Syst. *15*, 91–109. https://doi.org/10.1088/0954-898X_15_2_002.

171. Poggio, T. (1981). Marr's computational approach to vision. Trends Neurosci. *4*, 258–262. https://doi.org/10.1016/0166-2236(81)90081-3.

172. Herzog, M.H., and Clarke, A.M. (2014). Why vision is not both hierarchical and feedforward. Front. Comput. Neurosci. *8*, 135. https://doi.org/10.3389/fncom.2014.00135.

173. Bialek, W., Van Steveninck, R.R.D.R., and Tishby, N. (2006). Efficient representation as a design principle for neural coding and computation. In (IEEE), pp. 659–663. https://doi.org/10.1109/ISIT.2006.261867.

174. Kindel, W.F., Christensen, E.D., and Zylberberg, J. (2019). Using deep learning to probe the neural code for images in primary visual cortex. J. Vis. *19*, 29–29. https://doi.org/10.1167/19.4.29.

175. Yamins, D.L., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., and DiCarlo, J.J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc. Natl. Acad. Sci. *111*, 8619–8624. https://doi.org/10.1073/pnas.1403112111.

176. Briggs, F. (2020). Role of feedback connections in central visual processing. Annu. Rev. Vis. Sci. *6*, 313–334. https://doi.org/10.1146/annurev-vision-121219-081716.

177. Nurminen, L., Merlin, S., Bijanzadeh, M., Federer, F., and Angelucci, A. (2018). Top-down feedback controls spatial summation and response amplitude in primate visual cortex. Nat. Commun. *9*, 2281. https://doi.org/10.1038/s41467-018-04500-5.

178. Nassi, J.J., Lomber, S.G., and Born, R.T. (2013). Corticocortical feedback contributes to surround suppression in V1 of the alert primate. J. Neurosci. *33*, 8504–8517. https://doi.org/10.1523/JNEUROSCI.5124-12.2013.

179. Vangeneugden, J., van Beest, E.H., Cohen, M.X., Lorteije, J.A.M., Mukherjee, S., Kirchberger, L., Montijn, J.S., Thamizharasu, P., Camillo, D., Levelt, C.N., et al. (2019). Activity in lateral visual areas contributes to surround suppression in awake mouse V1. Curr. Biol. *29*, 4268-4275.e7. https://doi.org/10.1016/j.cub.2019.10.037.

180. Angelucci, A., Bijanzadeh, M., Nurminen, L., Federer, F., Merlin, S., and Bressloff, P.C. (2017). Circuits and mechanisms for surround modulation in visual cortex. Annu. Rev. Neurosci. *40*, 425–451. https://doi.org/10.1146/annurev-neuro-072116-031418.

181. Merrikhi, Y., Clark, K., Albarran, E., Parsa, M., Zirnsak, M., Moore, T., and Noudoost, B. (2017). Spatial working memory alters the efficacy of input to visual cortex. Nat. Commun. *8*, 15041. https://doi.org/10.1038/ncomms15041.

182. Lawrence, S.J., van Mourik, T., Kok, P., Koopmans, P.J., Norris, D.G., and de Lange, F.P. (2018). Laminar organization of working memory signals in human visual cortex. Curr. Biol. *28*, 3435–3440. https://doi.org/10.1016/j.cub.2018.08.043.

183. Voitov, I., and Mrsic-Flogel, T.D. (2022). Cortical feedback loops bind distributed representations of working memory. Nature *608*, 381–389. https://doi.org/10.1038/s41586-022-05014-3.

184. Schmidt, K.E., Lomber, S.G., Payne, B.R., and Galuske, R.A. (2011). Pattern motion representation in primary visual cortex is mediated by transcortical feedback. Neuroimage *54*, 474–484. https://doi.org/10.1016/j.neuroimage.2010.08.017.

185. Ringach, D.L., Shapley, R.M., and Hawken, M.J. (2002). Orientation selectivity in macaque V1: diversity and laminar dependence. J. Neurosci. *22*, 5639–5651. https://doi.org/10.1523/jneurosci.22-13-05639.2002.

186. Levitt, J.B., Kiper, D.C., and Movshon, J.A. (1994). Receptive fields and functional architecture of macaque V2. J. Neurophysiol. *71*, 2517–2542. https://doi.org/10.1152/jn.1994.71.6.2517.

187. Jones, H., Grieve, K., Wang, W., and Sillito, A. (2001). Surround suppression in primate V1. J. Neurophysiol. *86*, 2011–2028. https://doi.org/10.1152/jn.2001.86.4.2011.

188. Hu, J., Ma, H., Zhu, S., Li, P., Xu, H., Fang, Y., Chen, M., Han, C., Fang, C., and Cai, X. (2018). Visual motion processing in macaque V2. Cell Rep. *25*, 157–167. https://doi.org/10.1016/j.celrep.2018.09.014.

189. Bair, W., and Movshon, J.A. (2004). Adaptive temporal integration of motion in direction-selective neurons in macaque visual cortex. J. Neurosci. *24*, 7305–7323. https://doi.org/10.1523/JNEUROSCI.0554-04.2004.

190. Movshon, J.A. (1986). The analysis of moving patterns. Pattern Recognit. Mech., 117–151. https://doi.org/10.1007/978-3-662-09224-8_7.

191. Movshon, J.A., and Newsome, W.T. (1996). Visual response properties of striate cortical neurons projecting to area MT in macaque monkeys. J. Neurosci. *16*, 7733–7741. https://doi.org/10.1523/JNEUROSCI.16-23-07733.1996.

192. Rust, N.C., Mante, V., Simoncelli, E.P., and Movshon, J.A. (2006). How MT cells analyze the motion of visual patterns. Nat. Neurosci. *9*, 1421–1431. https://doi.org/10.1038/nn1786.

193. Wang, Q., Gao, E., and Burkhalter, A. (2011). Gateways of ventral and dorsal streams in mouse visual cortex. J. Neurosci. *31*, 1905–1918. https://doi.org/10.1523/JNEUROSCI.3488-10.2011.

194. Sit, K.K., and Goard, M.J. (2020). Distributed and retinotopically asymmetric processing of coherent motion in mouse visual cortex. Nat. Commun. *11*, 3565. https://doi.org/10.1038/s41467-020-17283-5.

195. Taylor, L., Zenke, F., King, A.J., and Harper, N.S. (2024). Temporal prediction captures key differences between spiking excitatory and inhibitory V1 neurons. bioRxiv, 2024–05. https://doi.org/10.1101/2024.05.12.593763.

196. Whittington, J.C., and Bogacz, R. (2019). Theories of error back-propagation in the brain. Trends Cogn. Sci. *23*, 235–250. https://doi.org/10.1016/j.tics.2018.12.005.

197. Lotter, W., Kreiman, G., and Cox, D. (2020). A neural network trained for prediction mimics diverse features of biological neurons and perception. Nat. Mach. Intell. *2*, 210–219. https://doi.org/10.1038/s42256-020-0170-9.

198. Bakhtiari, S., Mineault, P., Lillicrap, T., Pack, C., and Richards, B. (2021). The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. In Advances in Neural Information Processing Systems (Curran Associates, Inc.), pp. 25164–25178.

199. Cadieu, C.F., and Olshausen, B.A. (2012). Learning intermediate-level representations of form and motion from natural movies. Neural Comput. *24*, 827–866. https://doi.org/10.1162/NECO_a_00247.

200. Siegle, J.H., Jia, X., Durand, S., Gale, S., Bennett, C., Graddis, N., Heller, G., Ramirez, T.K., Choi, H., Luviano, J.A., et al. (2021). Survey of spiking in the mouse visual system reveals functional hierarchy. Nature, 1–7. https://doi.org/10.1038/s41586-020-03171-x.

201. Deitch, D., Rubin, A., and Ziv, Y. (2021). Representational drift in the mouse visual cortex. Curr. Biol. *31*, 4327-4339.e6. https://doi.org/10.1016/j.cub.2021.07.062.

202. Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., and Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. PLOS Comput. Biol. *10*, e1003553. https://doi.org/10.1371/journal.pcbi.1003553.

203. Saxe, A., Nelli, S., and Summerfield, C. (2021). If deep learning is the answer, what is the question? Nat. Rev. Neurosci. *22*, 55–67. https://doi.org/10.1038/s41583-020-00395-8.

204. Schrimpf, M., Kubilius, J., Hong, H., Majaj, N.J., Rajalingham, R., Issa, E.B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., et al. (2020). Brain-Score: Which artificial neural network for object recognition is most brain-like? https://doi.org/10.1101/407007.

205. Homma, N.Y., Hullett, P.W., Atencio, C.A., and Schreiner, C.E. (2020). Auditory cortical plasticity dependent on environmental noise statistics. Cell Rep. *30*, 4445–4458. https://doi.org/10.1016/j.celrep.2020.03.014.

206. Insanally, M.N., Köver, H., Kim, H., and Bao, S. (2009). Feature-dependent sensitive periods in the development of complex sound representation. J. Neurosci. *29*, 5456–5462. https://doi.org/10.1523/JNEUROSCI.5311-08.2009.

207. Pysanenko, K., Bureš, Z., Lindovský, J., and Syka, J. (2018). The effect of complex acoustic environment during early development on the responses of auditory cortex neurons in rats. Neuroscience *371*, 221–228. https://doi.org/10.1016/j.neuroscience.2017.11.049.

208. Ivanov, A.Z., King, A.J., Willmore, B.D., Walker, K.M., and Harper, N.S. (2022). Cortical adaptation to sound reverberation. Elife *11*, e75090. https://doi.org/10.7554/eLife.75090.

209. Willmore, B.D., and King, A.J. (2023). Adaptation in auditory processing. Physiol. Rev. *103*, 1025–1058. https://doi.org/10.1152/physrev.00011.2022.

210. Mesgarani, N., David, S.V., Fritz, J.B., and Shamma, S.A. (2014). Mechanisms of noise robust representation of speech in primary auditory cortex. Proc. Natl. Acad. Sci. *111*, 6792–6797. https://doi.org/10.1073/pnas.1318017111.

211. Fishman, Y.I., Kim, M., and Steinschneider, M. (2017). A crucial test of the population separation model of auditory stream segregation in macaque primary auditory cortex. J. Neurosci. *37*, 10645–10655. https://doi.org/10.1523/JNEUROSCI.0792-17.2017.

212. Farahat, A., and Vinck, M. (2025). Neural responses in early, but not late, visual cortex are well predicted by random-weight CNNs with sufficient model complexity. bioRxiv, 2025–02. https://doi.org/10.1101/2025.02.05.636721.

213. Soni, A., Srivastava, S., Khosla, M., and Kording, K.P. (2024). Conclusions about neural network to brain alignment are profoundly impacted by the similarity measure. bioRxiv, 2024–08. https://doi.org/10.1101/2024.08.07.607035.

214. Suzuki, M., Pennartz, C.M., and Aru, J. (2023). How deep is the brain? The shallow brain hypothesis. Nat. Rev. Neurosci. *24*, 778–791. https://doi.org/10.1038/s41583-023-00756-z.

215. Lohse, M., King, A.J., and Willmore, B.D. (2024). Subcortical origin of nonlinear sound encoding in auditory cortex. bioRxiv, 2024–02. https://doi.org/10.1101/2024.02.01.578402.

216. Dwivedi, K., and Roig, G. (2019). Representation similarity analysis for efficient task taxonomy & transfer learning. In, pp. 12387–12396. https://doi.org/10.48550/arXiv.1904.11740

217. Giordano, B.L., Esposito, M., Valente, G., and Formisano, E. (2023). Intermediate acoustic-to-semantic representations link behavioral and neural responses to natural sounds. Nat. Neurosci. *26*, 664–672. https://doi.org/10.1038/s41593-023-01285-9.

218. Güçlü, U., Thielen, J., Hanke, M., and Van Gerven, M. (2016). Brains on beats. Adv. Neural Inf. Process. Syst. *29*.

219. Kell, A.J., Yamins, D.L., Shook, E.N., Norman-Haignere, S.V., and McDermott, J.H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. Neuron *98*, 630–644. https://doi.org/10.1016/j.neuron.2018.03.044.

220. Vaidya, A.R., Jain, S., and Huth, A.G. (2022). Self-supervised models of audio effectively explain human cortical responses to speech. ArXiv Prepr. ArXiv220514252. https://doi.org/10.48550/arxiv.2205.14252.

221. Millet, J., Caucheteux, C., Boubenec, Y., Gramfort, A., Dunbar, E., Pallier, C., and King, J.-R. (2022). Toward a realistic model of speech processing in the brain with self-supervised learning. Adv. Neural Inf. Process. Syst. *35*, 33428–33443.

222. Harper, N.S., Schoppe, O., Willmore, B.D., Cui, Z., Schnupp, J.W., and King, A.J. (2016). Network receptive field modeling reveals extensive integration and multi-feature selectivity in auditory cortical neurons. PLoS Comput. Biol. *12*, e1005113. https://doi.org/10.1371/journal.pcbi.1005113.

223. Rahman, M., Willmore, B.D., King, A.J., and Harper, N.S. (2019). A dynamic network model of temporal receptive fields in primary auditory cortex. PLoS Comput. Biol. *15*, e1006618. https://doi.org/10.1371/journal.pcbi.1006618.

224. Zhang, Q., Hu, X., Hong, B., and Zhang, B. (2019). A hierarchical sparse coding model predicts acoustic feature encoding in both auditory midbrain and cortex. PLOS Comput. Biol. *15*, e1006766. https://doi.org/10.1371/journal.pcbi.1006766.

225. Higgins, I., Chang, L., Langston, V., Hassabis, D., Summerfield, C., Tsao, D., and Botvinick, M. (2021). Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. Nat. Commun. *12*, 6456. https://doi.org/10.1038/s41467-021-26751-5.

226. Williams, A.H., Kunz, E., Kornblith, S., and Linderman, S. (2021). Generalized shape metrics on neural representations. Adv. Neural Inf. Process. Syst. *34*, 4738–4750.

227. Valentini-Botinhao, C. (2017). Noisy speech database for training speech enhancement algorithms and tts models. Univ. Edinb. Sch. Inform. Cent. Speech Technol. Res. CSTR.

228. Medina, C.A., Vargas, E., Munger, S.J., and Miller, J.E. (2022). Vocal changes in a zebra finch model of Parkinson's disease characterized by alpha-synuclein overexpression in the song-dedicated anterior forebrain pathway. PLoS One *17*, e0265604. https://doi.org/10.1371/journal.pone.0265604.

229. Lopez Espejo, M., Schwartz, Z.P., and David, S.V. (2019). Spectral tuning of adaptation supports coding of sensory context in auditory cortex. https://doi.org/10.5281/zenodo.3445557.

230. Pennington, J., and David, S. (2023). Auditory cortex single unit population activity during natural sound presentation -- dataset. Version 1.0.0 (Zenodo). https://doi.org/10.5281/zenodo.7796574.

231. Vollmer, M., Snyder, R.L., Leake, P.A., Beitel, R.E., Moore, C.M., and Rebscher, S.J. (1999). Temporal properties of chronic cochlear electrical stimulation determine temporal resolution of neurons in cat inferior colliculus. J. Neurophysiol. *82*, 2883–2902. https://doi.org/10.1152/jn.1999.82.6.2883.

232. Schreiner, C.E., and Urbas, J.V. (1988). Representation of amplitude modulation in the auditory cortex of the cat. II. Comparison between cortical fields. Hear. Res. *32*, 49–63. https://doi.org/10.1016/0378-5955(88)90146-3.

233. Kimura, A. (2020). Cross-modal modulation of cell activity by sound in first-order visual thalamic nucleus. J. Comp. Neurol. *528*, 1917–1941. https://doi.org/10.1002/cne.24865.

234. Stitt, I., Galindo-Leon, E., Pieper, F., Hollensteiner, K.J., Engler, G., and Engel, A.K. (2015). Auditory and visual interactions between the superior and inferior colliculi in the ferret. Eur. J. Neurosci. *41*, 1311–1320. https://doi.org/10.1111/ejn.12847.

235. Rao, R.P.N. (2024). A sensory–motor theory of the neocortex. Nat. Neurosci. *27*, 1221–1235. https://doi.org/10.1038/s41593-024-01673-9.

236. Morillon, B., Hackett, T.A., Kajikawa, Y., and Schroeder, C.E. (2015). Predictive motor control of sensory dynamics in auditory active sensing. Curr. Opin. Neurobiol. *31*, 230–238. https://doi.org/10.1016/j.conb.2014.12.005.

237. Ahissar, E., and Assa, E. (2016). Perception as a closed-loop convergence process. eLife *5*, e12830. https://doi.org/10.7554/eLife.12830.

238. Izawa, J., and Shadmehr, R. (2011). Learning from sensory and reward prediction errors during motor adaptation. PLoS Comput. Biol. *7*, e1002012. https://doi.org/10.1371/journal.pcbi.1002012.

239. Tang, C., Abbatematteo, B., Hu, J., Chandra, R., Martín-Martín, R., and Stone, P. (2024). Deep reinforcement learning for robotics: A survey of real-world successes. Annu. Rev. Control Robot. Auton. Syst. *8*. https://doi.org/10.48550/arxiv.2408.03539.

240. Lindsay, G.W. (2021). Convolutional neural networks as a model of the visual system: Past, present, and future. J. Cogn. Neurosci. *33*, 2017–2031. https://doi.org/10.1162/jocn_a_01544.

241. Nanda, N., Chan, L., Lieberum, T., Smith, J., and Steinhardt, J. (2023). Progress measures for grokking via mechanistic interpretability. ArXiv Prepr. ArXiv230105217. https://doi.org/10.48550/arXiv.2301.05217.

242. Jonas, E., and Kording, K.P. (2017). Could a Neuroscientist Understand a Microprocessor? PLOS Comput. Biol. *13*, e1005268. https://doi.org/10.1371/journal.pcbi.1005268.

243. Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. Entropy *23*, 18. https://doi.org/10.3390/e23010018.

244. Tavanaei, A., Ghodrati, M., Kheradpisheh, S.R., Masquelier, T., and Maida, A. (2019). Deep learning in spiking neural networks. Neural Netw. *111*, 47–63. https://doi.org/10.1016/j.neunet.2018.12.002.

245. Guerguiev, J., Lillicrap, T.P., and Richards, B.A. (2017). Towards deep learning with segregated dendrites. Elife *6*, e22901. https://doi.org/10.7554/eLife.22901.

246. Whittington, J.C., and Bogacz, R. (2017). An approximation of the error backpropagation algorithm in a predictive coding network with local hebbian synaptic plasticity. Neural Comput. *29*, 1229–1262. https://doi.org/10.1162/NECO_a_00949.

247. Stanford, K. (2023). Underdetermination of scientific theory. In The Stanford Encyclopedia of Philosophy, E. N. Zalta and U. Nodelman, eds. (Metaphysics Research Lab, Stanford University).

248. Denève, S., and Machens, C.K. (2016). Efficient codes and balanced networks. Nat. Neurosci. *19*, 375–382. https://doi.org/10.1038/nn.4243.