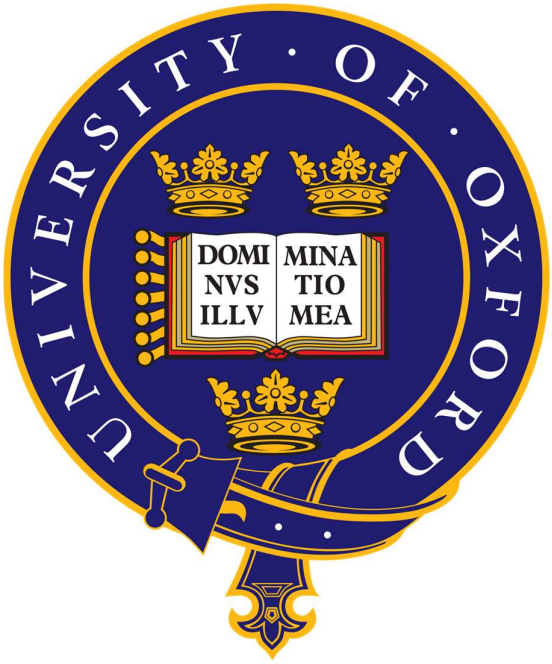
Temporal Prediction and the Mammalian Auditory System





Freddy Trinh

Department of Physiology, Anatomy and Genetics

Corpus Christi College

University of Oxford

This thesis is submitted for the degree of

Doctor of Philosophy

Hilary Term, 2022

To my family

Abstract

Neurons in the sensory systems are tuned to diverse but specific features of the sensory environment. Various hypotheses have been proposed to explain how this particular tuning may be the result of optimisation to statistical regularities of natural sensory stimuli under a given principle, such as efficient coding, sparse coding, and slowness. In this thesis, we investigate whether the auditory system is governed by temporal prediction. This principle states that the sensory features represented by neurons are optimised to predict immediate future input from recent past input. Experimental evidence indicates that neurons in the retina extract the most predictive information at a near-optimal efficiency level. Recent theoretical work has also demonstrated that temporal prediction can explain tuning across different sensory modalities. An artificial neural network optimised for temporal prediction of natural moving images can explain many of the receptive field features exhibited by neurons at different stages along the visual pathway. An auditory counterpart applied to spectrogram-like representations of natural sounds similarly produces receptive fields resembling those in the primary auditory cortex. In this thesis, we present work that further explores this explanatory principle for the auditory system. First, we show that an artificial neural network optimised for temporal prediction of raw waveforms of natural sounds produces tuning properties that are both qualitatively and quantitatively similar to those observed in the cochlea. Second, we show that a decision model based on auditory temporal prediction can predict the behaviour of mice in multiple auditory discrimination tasks. Finally, we present work that investigates whether a recurrent temporal prediction model can explain functional connectivity and neural responses to natural sounds in primary auditory cortex.

Together, these findings provide further support for the role of temporal prediction in accounting for diverse characteristics of the auditory system.

Acknowledgements

The scientific work presented in this thesis would not have been possible were it not for the help and support from several people. First of all, I want to express my deepest gratitude to Ben Willmore, Nicol Harper, and Andy King for their unwavering support and guidance both academically and personally, which were invaluable over the course of the project. I especially want to thank Nicol for always going over and beyond what is asked from a supervisor. His encouragement and advice were critical for my transition into theoretical neuroscience, which has been both challenging and highly rewarding. I also had the pleasure of collaborating with Chi Chen and Livia de Hoz, and want to thank them for expanding my scientific horizons. I am also deeply indebted to my past teachers and mentors, who have been great role models for me personally and as a scientist. I also want to thank the Medical Sciences Division of the University of Oxford and the Augustinus Foundation for their financial support. Furthermore, I want to thank past and present members of the King group for creating an excellent research environment and for teaching me so much.

I am thankful for all the friends I have made during my time in Oxford, who ensured that my time here truly provided an education in every sense of the word. I feel lucky to share so many fond memories of invigorating debates, stimulating conversations, and great times with so many wonderful people. They were a great source of comfort and laughter, and without them Oxford would not have felt like home. I also want to thank my friends outside of Oxford for their visits and all our shared adventures. I express my most heartfelt thanks to my parents and my brother for their endless love and tremendous support. And last, but by

no means least, I want to thank Nina. I am incredibly grateful for the time we have spent together, and I am excited to see what the future brings for us.

Contribution Statement

Chapter 2 and Chapter 4 were led by myself alone, supervised by Nicol Harper, Ben Willmore and Andrew King. The work in Chapter 3 is the result of a collaboration with Chi Chen and Livia de Hoz at Charité Medical University, Berlin. For this Chapter, I did all the modelling work. All behavioural experiments were done by Chi Chen. I made the figures together with Chi Chen. All participants contributed to data analysis. The work in Chapter 3 is available online as a BioRxiv preprint which all authors co-wrote and revised (Chen et al., 2021).

Table of Contents

Chapter 1. General Introduction 1

1.1 The Auditory System 2
1.1.1 The Auditory Pathway 3
1.1.2 The Cochlea 5
1.1.3 The Auditory Brainstem 8
1.1.4 Midbrain 10
1.1.5 Thalamus 10
1.1.6 Primary Auditory Cortex 10
1.2 Encoding Models and Receptive Fields in the Auditory System 13
1.3 Normative Models 16
1.3.1 Efficient Coding 18
1.3.2 Predictive Coding 18
1.3.3 Sparse Coding 19
1.3.4 Slowness and Slow Feature Analysis 20
1.3.5 Temporal Prediction 20
1.4 Decision Making 23
1.5 Overview of Chapters 24

Chapter 2. Temporal Prediction and the Cochlea 27

2.1 Introduction 27
2.2 Results 30
2.2.1 Cochlear Temporal Prediction Model 30
2.2.2 The Effects of Noise Injection to the Training Set 38
2.2.3 Cochlear Autoencoder Model 43
2.3 Discussion 49
2.3.1 Comparison to Other Models 50
2.3.2 Arguments for Temporal Prediction in the Cochlea 51
2.3.3 Weaknesses and Future Development of the Cochlear Temporal Prediction Model 52
2.4 Methods 54
2.4.1 Auditory Dataset 54
2.4.2 Artificial Neural Network Models 55
2.4.3 IR Analysis 57
2.4.4 Computational Implementation 57

Chapter 3. Temporal Prediction and Auditory Behaviour 59

3.1 Introduction 59
3.2 Results 62

3.2.1 Behavioural Responses to Frequency Modulated Sweeps 65
3.2.2 Behaviour Is Consistent with Temporal Prediction 69
3.2.3 Hierarchy of Stimulus Dimensions Relevant for Behaviour 72
3.2.4 Non-Preferred Stimulus Dimensions 77
3.2.5 Plasticity of Stimulus Dimensions 80
3.2.6 Perception of Envelope Features 85
3.2.7 Summary and Quantitative Comparison of Models 88
3.3 Discussion 91
3.3.1 Hierarchical Organisation of Acoustic Dimension Processing 92
3.3.2 Temporal Prediction and the Hierarchy of Stimulus Dimensions 93
3.3.3 Neurophysiological Considerations of Dimension Perception 97
3.4 Materials and Methods 99
3.4.1 Experiments 99
3.4.1.1 Animals 99
3.4.1.2 Apparatus: the Audiobox 100
3.4.1.3 Sounds 101
3.4.1.4 Discrimination Task 103
3.4.1.5 Multidimensional Generalisation Measurement 104
3.4.1.6 Analysis of Performance in the Audiobox 105
3.4.1.7 Statistical Analysis 106
3.4.2 Models 106
3.4.2.1 Introduction 106
3.4.2.2 The Cochleagram Model 109
3.4.2.3 The Normative Encoding Models 109
3.4.2.4 Frequency Scrambled Encoding Model 116
3.4.2.5 Measuring Encoding Model Receptive Fields 116
3.4.2.6 Decision Models 117
3.4.3 Modelling Implementation 121

Chapter 4. Temporal Prediction and Primary Auditory Cortex 123

4.1 Introduction 123
4.2 Results 126
4.2.1 Receptive Fields and Frequency Tuning 127
4.2.2 Functional Connectivity 133
4.2.3 Predicting A1 Neuron Responses to Natural Sounds 139
4.3 Discussion 144
4.3.1 Receptive Fields and Frequency Tuning 144
4.3.2 Functional Connectivity 145
4.3.3 Normative Encoding Models 147
4.3.4 Model Limitations and Future Investigations 149
4.4 Methods 151
4.4.1 The Cochleagram and Normative Models 151
4.4.2 Determining Model Spectro-Temporal Receptive Fields and Best Frequencies 151
4.4.3 Functional Connectivity 152

4.4.3.1 A1 Data....152
4.4.3.2 Functional Connectivity Analysis....153
4.4.4 A1 Neuron Modelling....155
4.4.4.1 Natural Sounds and Neural Recordings....155
4.4.4.2 Neuron Models....156
4.4.5 Modelling Implementation....158

Chapter 5. Discussion....161

5.1 Summary of Results....161
5.2 Temporal Prediction as a Normative Principle....162
5.3 Arguments Supporting Temporal Prediction in the Auditory System....164
5.4 Implementation of Temporal Prediction in the Auditory System....167
5.5 Model Limitations and Future Investigations....170
5.6 Experimental Predictions and Model Testing....175
5.7 Conclusion....177

Bibliography....181

Chapter 1. General Introduction

The brain is an exceptionally complex organ, which performs a plethora of vital biological tasks and functions. A crucial function is to obtain information from the surrounding environment, process and evaluate it, and finally use this information to enable the organism to act and adapt appropriately. Sensory systems, such as the auditory system, provide the first stage of this chain of processing. It is well known that neurons in the sensory systems are tuned to very diverse, but highly specific, features of the sensory environment. We also know that these patterns of tuning are conserved across different animals, and to some extent across species, indicating that they are significant for the normal functioning of sensory systems. However, we do not yet have a good understanding of how or why these particular tuning patterns have been selected by the brain.

The ultimate goal of any organism is to maximise fitness and reproductive success, and we would therefore expect that evolutionary pressures play an important role in shaping neuronal tuning patterns. However, there are many possible theories of what influence evolution may have on neural coding. For example, the brain may be optimised to minimise its overall number of neurons or to minimise energy use; other theories such as redundancy reduction will be discussed below (Barlow, 1961). Temporal prediction is another possible principle, which hypothesises that the sensory systems encode features of the most recent stimulus history that are most predictive of its immediate future.

In this thesis, I will investigate temporal prediction as a guiding principle to explain the physiological and anatomical features of different stages of the auditory system. First of all, I will focus on the cochlea, where auditory transduction occurs. Then, I will demonstrate how a recurrent temporal prediction model of the auditory cortex is able to decode and predict behaviour of mice performing an auditory discrimination task. Finally, I will investigate whether temporal prediction can explain patterns of functional connectivity in the primary auditory cortex and the responses of cortical neurons to natural sounds.

1.1 The Auditory System

Sensation is a critical feature of all biological life as a means to obtain information about the surrounding environment. In order to operate in a given habitat, information must be gathered and then processed to perform appropriate actions. Hearing is the sensory modality that allows for the sensation of oscillatory pressure changes through a medium such as atmospheric air. The ability to hear is shared amongst most animals, but comes in many different manifestations. In marine animals, where pressure waves propagate through a liquid medium, the details of the hearing mechanism differ from land animals, while the functional principle remains the same. Hearing in vertebrates most likely developed as a way to detect, localise and identify sources of vibrations (Fay and Popper, 2000). In some animals, hearing has evolved into a sense responsible for intraspecies communication, and for humans, auditory processing is crucial for advanced processes such as speech and the enjoyment of music.

1.1.1 The Auditory Pathway

The head and outer ear are the first structures to modify auditory stimuli by altering the amplitude and phase spectra of incoming sound, and this filtering is important for sound localisation. The sound wave is then funnelled through the ear canal, where the pressure wave pushes and pulls on the tympanic membrane, also known as the ear drum. The ear drum transmits its vibrations to the ossicles, an intricate, interlocked system of delicate bones that concentrates the vibrations felt by the ear drum onto the cochlea by vibrating the oval window of the inner ear. This overcomes the acoustic impedance when traversing from air to the liquid medium presented by the cochlea. The cochlea separates sound into its constituent frequency components, while also converting it from a mechanical to an electrochemical signal.

The information from the cochlea is then sent along the auditory nerve to the auditory brainstem. Here, after processing takes place in the cochlear nucleus, the information from both ears converges, and is used to derive more abstract information, such as the location of the sound source by comparing any intensity and timing differences of the sound from each ear. From the auditory brainstem, the information is then sent onwards to the auditory midbrain and then to the thalamus and finally to the auditory cortex (Fig. 1.1).

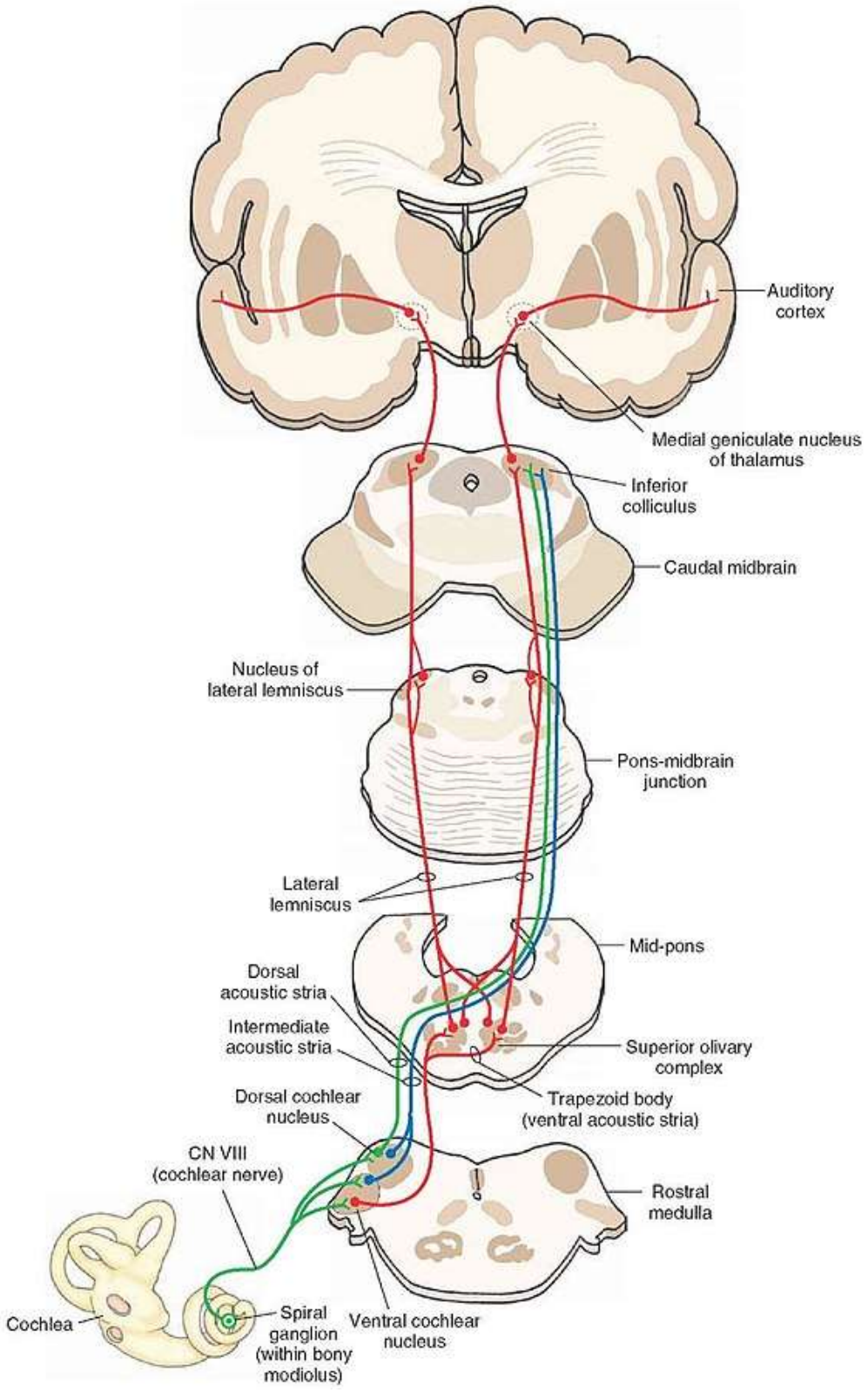


Figure 1.1. Overview of the human auditory system. The ascending pathway going from one ear to the auditory cortex. Figure from commons.wikimedia.org/wiki/File:The_Auditory_Pathway.jpg.

1.1.2 The Cochlea

The cochlea converts mechanical energy of sound waves into electrochemical signals used by the nervous system. It consists of a fluid-filled tube, bisected along its length by the basilar membrane (Fig. 1.2). The features of the basilar membrane vary longitudinally; at the basal end, the basilar membrane is thick, narrow and stiff, and gets progressively thinner, wider and more flexible towards the apical end where it is at its widest, thinnest and most flexible (Von Békésy, 1960; Liu et al., 2015; Raufer et al., 2019). Consequently, this creates a gradient that in concert with the fluid-filled compartments forms the passive cochlear mechanical basis for frequency tuning. The constituent frequency components of incoming sound waves are instantiated as travelling waves down the basilar membrane, which maximally displace the basilar membrane at frequency-tuned sites going from high-frequency tuned basal part to the low-frequency tuned apical end. The frequency preferred at each point on the basilar membrane depends approximately exponentially on the distance along the membrane from apical end (Greenwood, 1961, 1990, 1996; Liberman, 1982; Robles and Ruggero, 2001).

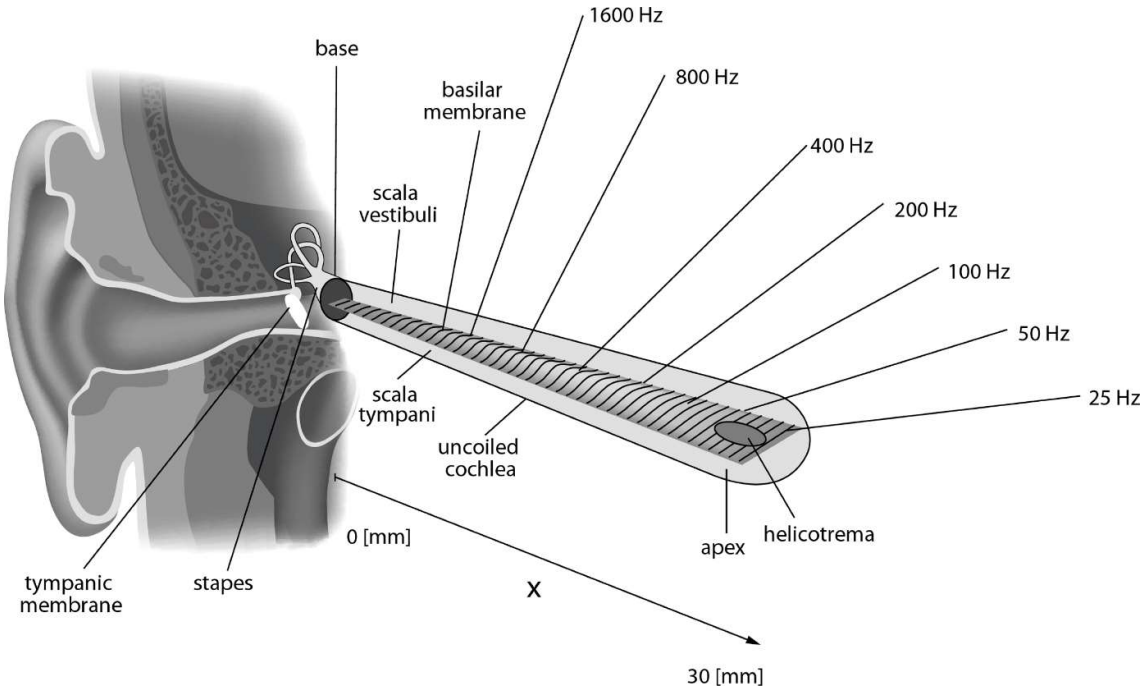


Figure 1.2. Overview of the cochlea and its frequency tuning. This figure illustrates the cochlea uncoiled, showing the basilar membrane and its frequency tuning. Figure from commons.wikimedia.org/wiki/File:Uncoiled_cochlea_with_basilar_membrane.png.

The organ of Corti is the sensory tissue of the cochlea (Fig. 1.3). It lies above the basilar membrane and below the tectorial membrane. The organ of Corti runs along the basilar membrane and contains inner and outer hair cells, which respond to vibrations of the basilar membrane by alternating depolarisation and hyperpolarisation. This is the transduction process that transforms the mechanical soundwave into an electrochemical signal that the brain is able to process. More exactly, the tectorial membrane forms a cover just above the rows of inner and outer hair cells. The vibrations of the basilar membrane cause the organ of Corti to slide back and forth relative to the tectorial membrane. This causes deflection of the many hair-like stereocilia of the hair cells, which cause depolarisation of hair cells through mechanical opening of potassium channels on the stereocilia via small fibres (tip links)

between their tips. The outer hair cells respond mechanically to changes in their membrane potential by rapidly alternating between an elongated and shortened state, which inputs additional motor action onto the basilar membrane (Brownell et al., 1985). This attribute, called electromotility, is believed to be important for cochlear amplification and is also responsible for the non-linear properties of the cochlea (Gillespie and Müller, 2009; Hudspeth, 2014). The inner hair cells are predominately responsible for transmitting the auditory information extracted by the cochlea to the brain. The inner hair cells synapse onto auditory nerve fibres in the 8th cranial nerve, which propagate this information to the cochlear nucleus in the brainstem and from there onwards to higher brain areas.

Adaptation mechanisms at the level of the hair cell modulate the response properties of the cochlea (Eatock, 2000). Fast negative feedback decreases the permeability of the potassium channels through a calcium-mediated mechanism at the stereocilia, while a slower mechanism of adaptation occurs through the myosin-dependent reconfiguration of transduction channel insertion points in the stereocilia, which decreases tension on the tip links and therefore also decreases the probability of open transduction channels. Adaptation also occurs at the synapse between the inner hair and the auditory nerve fibers (Zilany et al., 2009). Synaptic adaptation can be due to neurotransmitter vesicle depletion (Moser and Beutner, 2000) or postsynaptic desensitisation (Raman et al., 1994). These adaptation mechanisms may allow the cochlea to adjust its dynamic range to better encode the time varying distribution of sound intensity in the environment (Wen et al., 2012).

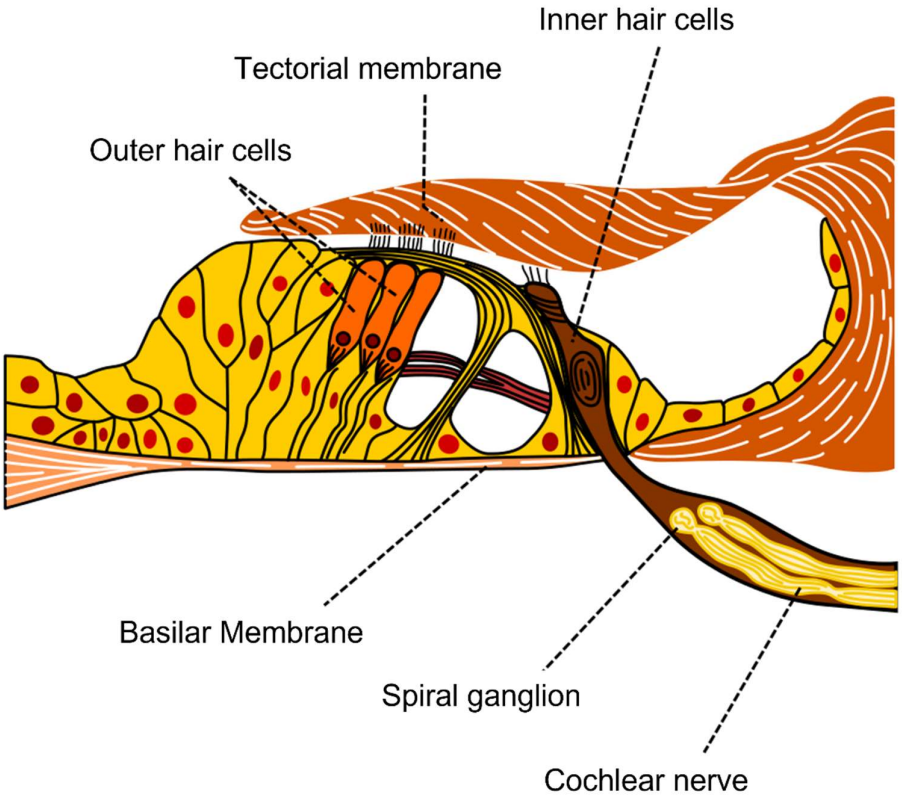


Figure 1.3. A cross-section of the organ of Corti. Figure adapted from https://commons.wikimedia.org/wiki/File:Organ_of_corti.svg.

1.1.3 The Auditory Brainstem

The auditory nerve fibres send information from the cochlea onwards to the brainstem. The auditory brainstem is a complex network that consists of the cochlear nucleus (CN), the superior olivary complex (SOC), nuclei of the lateral lemniscus (NLL) and the inferior colliculus (IC). The CN is the first brainstem structure to receive sensory input. The tonotopic organisation determined by the cochlea is maintained in the CN (Osen, 1970; Bourk et al., 1981; Ryugo and Parks, 2003). The auditory nerve innervates a wide range of different neurons in the CN with distinct firing patterns and morphology. This allows the CN to divide

the incoming stream of information into separate streams that can perform a range of computations and extract different information from the auditory nerve activity. From the CN the sensory information is then relayed through the trapezoid body to both the contra- and ipsilateral SOC. The CN also innervates other structures in the brainstem and midbrain, primarily contralaterally, including the nuclei of the lateral lemniscus and the IC.

The SOC consists of a group of brainstem nuclei, where especially the medial and lateral superior olive (MSO and LSO), and the medial nucleus of the trapezoid body (MNTB) have drawn the most interest experimentally as these nuclei are important for sound localisation. MSO cell activity is driven by sound arriving at each ear, LSO neurons are mainly stimulated by ipsilateral sound and inhibited by contralateral stimuli, and neurons in the MNTB are activated by sound received by the contralateral ear (Goldberg and Brown, 1968; Guinan et al., 1972; Yin, 2002).

Identifying the spatial position of where a particular sound is emitted from plays an important role for orientation, auditory attention and for sound segregation of different sources in the environment. The brain performs sound localisation by using different cues from the incoming sound, such as monaural cues derived from the direction-specific spectral transformation by the pinna and head (Slattery and Middlebrooks, 1994; Ito et al., 2020), which are primarily important for determining sound-source elevation and binaural cues, such as interaural time- and level differences, which are important for determining sound source lateralisation (Rayleigh, 1907).

1.1.4 Midbrain

The inferior colliculus (IC) in the midbrain integrates information from many brain structures and is the first auditory structure that processes both horizontal and vertical information about the spatial location of the incoming sound from each ear (Chase and Young, 2008). The IC receives ascending projections from the CN and SOC (Komune et al., 2015), but also top-down connections from the auditory cortex (Bajo and King, 2012). While the IC is involved in sound localisation (Davis et al., 2003), further processing takes place to generate a two-dimensional map of auditory space in the superior colliculus, where multi-sensory information is integrated (King and Hutchings, 1987).

1.1.5 Thalamus

The output of the IC projects to the medial geniculate body (MGB) of the thalamus and to the contralateral IC. The MGB plays a role in complex processes, such as auditory learning (Gilad et al., 2020), auditory fear conditioning (Weinberger, 2011) and multi-sensory processing (Lohse et al., 2021). It acts as a thalamic relay station and passes on information from the IC to the auditory cortex.

1.1.6 Primary Auditory Cortex

The auditory cortex forms the neural substrate for numerous tasks such as perception, learning, speech, and language (King et al., 2018). It has been proposed that the neocortex is a uniform structure composed of repeats of a canonical neural circuit, which is suggested to be a widely useful computational unit (Douglas et al., 1989; Douglas and Martin, 2004).

Although the idea of a canonical cortical circuit has captured the fascination of neuroscientists, it remains contentious (Miller, 2016). In sensory cortices, thalamic input is mainly concentrated to layer 4, which in turn synapse onto layer 2/3 and from there onto layer 5, which, along with layer 6, forms the descending output neurons for the cortical circuit (Harris and Mrsic-Flogel, 2013). Approximately 80% of neurons in the cerebral cortex are excitatory, while 20% are inhibitory (Wonders and Anderson, 2006; Marín, 2012).

The primary auditory cortex (A1) in mammals is defined as the first cortical structure that receives auditory input from subcortical auditory structures, although the MGB also projects to other auditory cortical areas, including another primary-like region, the anterior auditory field (AAF). The auditory cortex shares many similarities with the laminar organisation found in other sensory modalities. However, A1 differs in certain ways. A1 receives pre-combined binaural information from subcortical nuclei, unlike in the visual and somatosensory modalities, where information from both sides of the body largely converges for the first time at the cortical level. A1 also differs in its neural composition compared to its visual and somatosensory counterparts. In contrast to primary visual and somatosensory cortex, spiny stellate cells, a predominant target of thalamic input, are remarkably absent in layer 4 of A1. Instead, small pyramidal neurons are more numerous in layer 3 and 4, and they predominantly receive thalamic input (Smith and Populin, 2001).

The main thalamic input to A1 innervates both cortical layers 3 and 4 (Winer, 1992; Huang and Winer, 2000), while being largely contained to layer 4 in the visual and barrel cortex. In addition to this ‘lemniscal’ pathway, there is also a ‘non-lemniscal’ auditory

pathway which projects largely to layers 1, 3, and 6 of primary auditory cortex (Winer, 1992; Huang and Winer, 2000).

In addition to its inputs from the MGB, A1 is also innervated by the contralateral auditory cortex through the corpus callosum (Bamiou et al., 2007). Layer 3 neurons in A1 send projections to the contralateral A1 through the corpus callosum (Imig and Brugge, 1978), while the axons of the same type of neurons in visual and somatosensory cortex mostly tend to remain in ipsilateral cortex (Innocenti, 1980; Manzoni et al., 1980). In A1, layer 5a neurons sends axons to layer 2/3 neurons locally and more distally to ipsi- and contralateral cortex, and also to striatum (Games and Winer, 1988; Winer and Prieto, 2001; Yuan et al., 2011). Layer 5b neurons project to widespread subcortical auditory structures; the MGB in the thalamus, inferior colliculus in the midbrain and SOC and CN in the brainstem, but also lateral amygdala and striatum (Winer, 2006; Bajo and King, 2012; Asokan et al., 2018). In contrast, layer 6 neurons project more specifically to the thalamus (Linden & Schreiner 2003, Williamson & Polley 2019).

The primary auditory cortical areas inherit the tonotopic organisation established in the cochlea, which is ordered in contours that share best frequencies orthogonal to the tonotopic axis (Merzenich et al., 1975). This is valid globally in A1, but locally A1 neurons are quite heterogenous in terms of their frequency tuning (Bandyopadhyay et al., 2010; Gaucher et al., 2020). Bands of binaural response-specific neurons have also been reported running parallel to the tonotopic axis (Imig and Adrián, 1977; Middlebrooks et al., 1980), though it is likely that these properties are arranged in clusters (Panniello et al., 2018).

A1 neurons encode a wide range of auditory features, such as frequency, spatial location, temporal modulation, intensity (Chambers et al., 2014), frequency sweep and direction (Heil et al., 1992b), spectral bandwidth and amplitude modulation (Sloas et al., 2016), and more abstract features of sound such as echoes and ambient noise (Chechik and Nelken, 2012). The tuning characteristics of A1 neurons are more complicated than those of subcortical structures, and while a single stimulus dimension can describe IC neuron activity, multiple stimulus dimensions are needed for A1 neurons (Sharpee et al., 2004; Atencio et al., 2012). The tuning of A1 neurons has been suggested to be a consequence of A1 neurons extracting the most predictive features of sound (Singer et al., 2018). Furthermore, A1 is involved in complex cognitive auditory tasks, such as decision making, auditory learning, identifying auditory objects and deriving meaning from sounds (King et al., 2018).

1.2 Encoding Models and Receptive Fields in the Auditory System

Neural encoding models describe the relationship between a sensory stimulus and the activity of a neuron. They typically map a stimulus to the moment-by-moment response of a neuron or a population of neurons. Such models can be used to understand the tuning of neuronal activity for particular stimulus features, and to predict neural activity to novel stimuli. In sensory neuroscience, encoding models are closely associated with the notion of a receptive field. Encoding models typically simplify the transformation of sensory information by the brain. This means that they may not capture the full complexity of real neural processing,

especially for the responses of neurons further along the auditory pathway. Nevertheless, encoding models are able to provide insights into how the brain is organised functionally.

An example of a simple encoding model is the linear receptive field, which is the linear mapping from the stimulus to the neuronal response. For auditory neurons, a linear mapping can be estimated directly between the raw sound pressure waveform and neuronal responses, giving a receptive field which is a one-dimensional function of time lag. In 1969, Goblick and Pfeiffer established that, for a single cochlear nerve fibre, this function resembles a gammatone filter (Goblick and Pfeiffer, 1969). A gammatone filter is a sinusoid at a specific frequency that is ramped on and off using the gamma function. When a sound signal is convolved with a gammatone filter, components with frequencies similar to the frequency of the gammatone will be retained, and more distant frequencies will be filtered out.

The receptive fields for auditory neurons beyond the auditory nerve are traditionally characterised using spectro-temporal receptive field (STRF) models. The STRF of a neuron is the linear mapping to activity of the neuron from the spectrogram of the stimulus, rather than from the raw waveform. This describes how a neuron interacts with the stimulus based on the spectral content in a time-dependent manner. For auditory cortical neurons, STRFs have been widely used to characterise the firing patterns in response to natural sounds (Aertsen et al., 1981).

The STRF also forms the linear stage of linear-nonlinear (LN) models, where the spectrogram of the incoming sound is convolved with the STRF and then fed into a static nonlinear transformation, such as a sigmoid (Fig. 1.4). The nonlinear stage of these models is

usually necessary for accurate neuronal modelling. Without the nonlinear stage, linear models tend to predict that, under some circumstances, neurons will produce spike rates lower than zero. Additionally, a nonlinear output stage has been shown to improve prediction of neural responses in the auditory cortex (Machens et al., 2004; Atencio et al., 2008; Rabinowitz et al., 2011). The activity of IC neurons is fairly linear, which makes the LN model suitable for this auditory structure (Atencio et al., 2012).

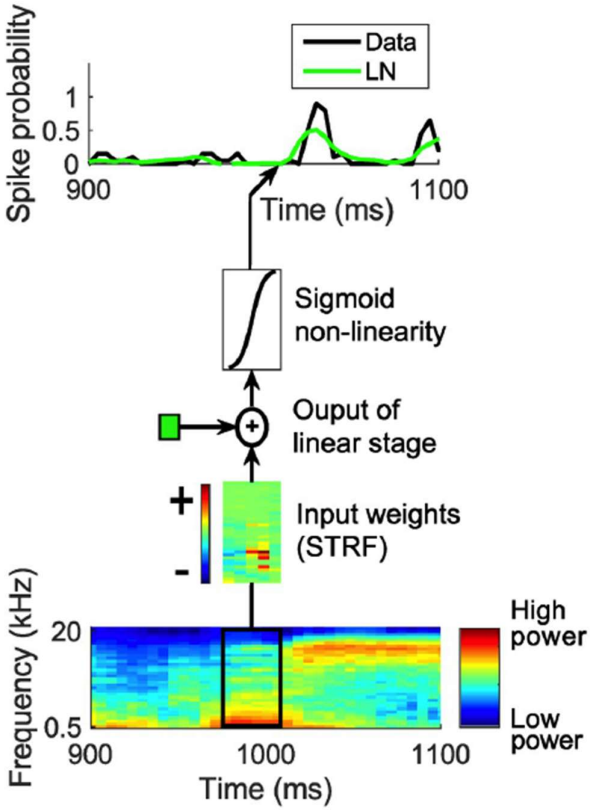


Figure 1.4. The linear-nonlinear (LN) model. The LN model maps sound to neural activity. The LN model consists of a linear stage, the spectro-temporal receptive field (STRF), followed by a nonlinearity. Figure adapted from (Rahman et al., 2019).

An important caveat about STRF modelling is that, even when output nonlinearities are included, these models represent a simplification of neuronal responses. As a result, no

STRF model will fully capture the complex responses of cortical neurons; also, any given model may fail to describe responses under all circumstances. For example, it has been shown that STRF models derived from neuronal responses to synthetic and natural sounds do not match (Theunissen et al., 2000), indicating that these models cannot necessarily generalise across different stimulus conditions.

More sophisticated analysis methods have been used to characterise the nonlinear responses of auditory cortical neurons. For example, maximally informative dimensions (MID) is a technique that identifies multiple spectro-temporal stimulus features which interact to influence the response of auditory cortical neurons (Sharpee et al., 2004; Atencio and Sharpee, 2017). It is believed that A1 neurons have composite receptive fields that are the result of an integration of multiple receptive fields from presynaptic neurons, which has been supported by studies on songbirds (Kozlov and Gentner, 2016). This insight has inspired network receptive field models that use neural networks to create encoding models consisting of an ensemble of LN model units that are tuned to various features of the spectro-temporal profile of sound, which then feed into a final unit (Harper et al., 2016). Network receptive fields have been found to predict auditory cortical neuron responses better than LN models.

1.3 Normative Models

The encoding models described in Section 1.2 above are methods for understanding the function of neuronal systems; how they process and represent sensory information. In contrast, normative models in neuroscience are a way of understanding why neuronal

systems are the way they are. A normative theory makes a hypothesis about the goal of a particular system, and represents this hypothesis using a model with an objective function (also called cost or utility function), which mathematically describes the problem the system has to solve. By optimising the objective function, the normative model instantiates the hypothesis as a set of concrete predictions about how the neuronal system should behave. If these predictions match the behaviour of the real neuronal system, this constitutes evidence that the hypothesis may accurately represent the goal of the real neuronal system.

Normative theories are derived from a first set of principles, and in sensory neuroscience this includes considerations based on the natural sensory environment. Sensory neurons are believed to adapt to the statistical properties of sensory stimuli from their surroundings using both evolutionary and developmental processes (Simoncelli and Olshausen, 2001). The properties of sensory systems are thought to be guided by three factors: 1) the actions that the organism must carry out, 2) the constraints on the sensory system and its neurons, such as biological limitations, and 3) the natural habitat where the organism lives (Simoncelli and Olshausen, 2001). Multiple normative theories have been proposed to explain what tasks sensory systems are optimised for. The theories are not necessarily mutually exclusive, for example one brain pathway may be governed by one normative approach and another by a different approach, or perhaps by some hybrid of two or more normative theories (Marblestone et al., 2016).

1.3.1 Efficient Coding

In 1954, Attneave proposed that the objective of the visual system is to encode natural visual stimuli as efficiently as possible (Attneave, 1954), and in 1961, Barlow suggested that the spikes evoked by sensory neurons formed an efficient code of the surrounding sensory world, where the brain removes redundant information between the neurons and uses the minimum number of spikes needed for this representation (Barlow, 1961). A vast amount of possible sensory input needs to be encoded, constrained by the limitations imposed by the neural hardware and its metabolic energy requirements. This encoding is done efficiently by employing prior knowledge about the regularities of the sensory environment, enabling the activity of different neurons to be more independent. There have been studies showing that the auditory system might operate under efficient coding (Lewicki, 2002; Smith and Lewicki, 2006). In the efficient coding framework, the entire stimulus is represented, and no sensory information is discarded.

1.3.2 Predictive Coding

Predictive coding can be thought of as an extension to efficient coding in the sense that it also represents the entire stimulus while minimising statistical redundancy (Friston, 2003; Huang and Rao, 2011). Natural stimuli contain temporal and spatial regularities that are highly redundant. Predictive coding posits that only non-predictable information and the residual errors in prediction are transmitted, and provides a principle for removal of redundant information from sensory stimuli. Given the hierarchical structure of the sensory systems, predictive coding suggests that each level merges information from lower and higher levels. Predictions of activity in lower levels are communicated from higher levels in a top-down

manner; there this activity is subtracted from the activity at that lower level, and the non-predictable residual errors are transmitted up to higher levels in a bottom-up manner. This framework has been applied especially to the visual system and can explain receptive fields in the primary visual cortex (V1) using a form of Hebbian learning rule (Rao and Ballard, 1999). Rao and Ballard (1999) found that prediction error units were tuned to oriented edges similar to simple cells in V1, while other units displayed end-stopping. The centre-surround receptive fields found in bipolar and retinal ganglion cells in the retina have also been explained using this principle (Srinivasan et al., 1982; Field, 1987; Ruderman and Bialek, 1994; Dong and Atick, 1995; Wessberg et al., 2000).

1.3.3 Sparse Coding

Sparse coding is also related to the efficient coding principle. The objective of sparse coding is to accurately represent the entire incoming input while having only a small subset of neurons in a population active for any given input. This has a number of advantages, such as using less energy and representing complex information in a manner that is easy to decode. This would also increase memory capacity as sensory information will be represented more explicitly and with no overlapping activity between neurons encoding different sensory information (Olshausen and Field, 2004). Olshausen and Field showed that this principle can capture the receptive fields found in simple cells in the primary visual cortex. This was achieved by finding a set of basis functions to represent natural images, where the coefficient values for the basis functions were as statistically independent as possible over the entire span of the natural images represented. The coefficients were given a sparse prior such that

it minimised the amount of non-zero coefficients to represent the natural images (Olshausen and Field, 1996).

1.3.4 Slowness and Slow Feature Analysis

The underlying sources of sensory information, such as the trajectories of flying birds, tend to change on a slower time-scale compared to unimportant noise in corresponding neural signals, such as the speckling in the air from atmospheric dust. Therefore, the slowness principle posits that the important sensory information is encoded in neural signals that vary slowly over time (Földiák, 1991). This is related to slow feature analysis where encoding models are optimised to maximise slowness by minimising the averaged squared derivative of the output signal over time, while also retaining information about the stimulus (Wiskott and Sejnowski, 2002). Slow feature analysis have shown to be able to explain various receptive field properties of A1 neurons (Carlin and Elhilali, 2013).

1.3.5 Temporal Prediction

The temporal prediction framework proposes that sensory systems are tuned to information that carries predictive value about the immediate future of a sensory stimulus based on its recent past. This idea is different from the efficient, sparse and predictive coding paradigms, as temporal prediction does not aim to represent all the incoming sensory information with high fidelity, but instead predominately encodes the information that is predictive of the immediate future (Bialek et al., 2001).

In this thesis, we expand on previous studies on temporal prediction (Singer et al., 2018, 2019). We use a normative approach that investigates if the auditory system is optimised for temporal prediction. Our work is related to the predictive information framework, which argues that a sensory system selectively encodes predictive features of the stimuli (Bialek et al., 2001; Palmer et al., 2015; Salisbury and Palmer, 2016; Heeger, 2017; Chalk et al., 2018). The idea takes inspiration from information theory, and states that the sensory system must be optimised to carry information from past input that is predictive of future sensory input (Bialek et al., 2001).

There are a number of arguments for temporal prediction as a useful principle in neural information processing. One reason stems from the fact that sensory information is necessary to guide actions. However, sensory information is delayed due to sensory transduction by the peripheral sensory system, followed by neuronal transmission and processing, and muscle activation (Maunsell and Gibson, 1992; Schmolesky et al., 1998; DiCarlo and Maunsell, 2005; Camalier et al., 2012). Yet the brain needs to know the state of the environment at the time of performing actions. Hence, it is argued that the only components of incoming sensory data that are useful are those that are predictive (Bialek et al., 2001). As an example, temporal prediction could overcome the delay in neural processing in order to catch a moving object by estimating its future position (Nijhawan, 1994).

Another reason is that optimisation for temporal prediction may encourage representations that better reflect underlying causes (Bialek et al., 2001), as efficiently predicting the future from the past requires building models of the worlds that capture its essential dynamics. Furthermore, sensory input that says nothing about the future is unlikely

to be useful for survival, and optimisation for temporal prediction causes the sensory systems to disregard this superfluous information and only encode the predictive information, reducing the amount of information the brain needs to process. Hence, temporal prediction argues that the most useful components of incoming sensory data are those components that are predictive (Bialek et al., 2001), and these components might be preferentially encoded in the nervous system.

Experiments suggests that the larval salamander retina might be optimised for temporal prediction (Palmer et al., 2015; Salisbury and Palmer, 2016). Each recorded retinal cell participates in a 5-cell group that encodes maximum predictive information, which suggests that predictive computation happens in the early stages of sensory processing. Theoretical studies also suggest that sensory cortex utilises temporal prediction as an encoding principle. When artificial neural networks were optimised for a simplified form of temporal prediction based on moving images, units were produced with receptive fields resembling those found for neurons in primary visual cortex (Singer et al., 2018) (Fig. 1.5). Furthermore, when networks were optimised to predict the future of sound spectrograms from their past for a dataset of natural sounds, the resulting receptive fields of the hidden units resembled those of real neurons recorded in auditory cortex (Singer et al., 2018).

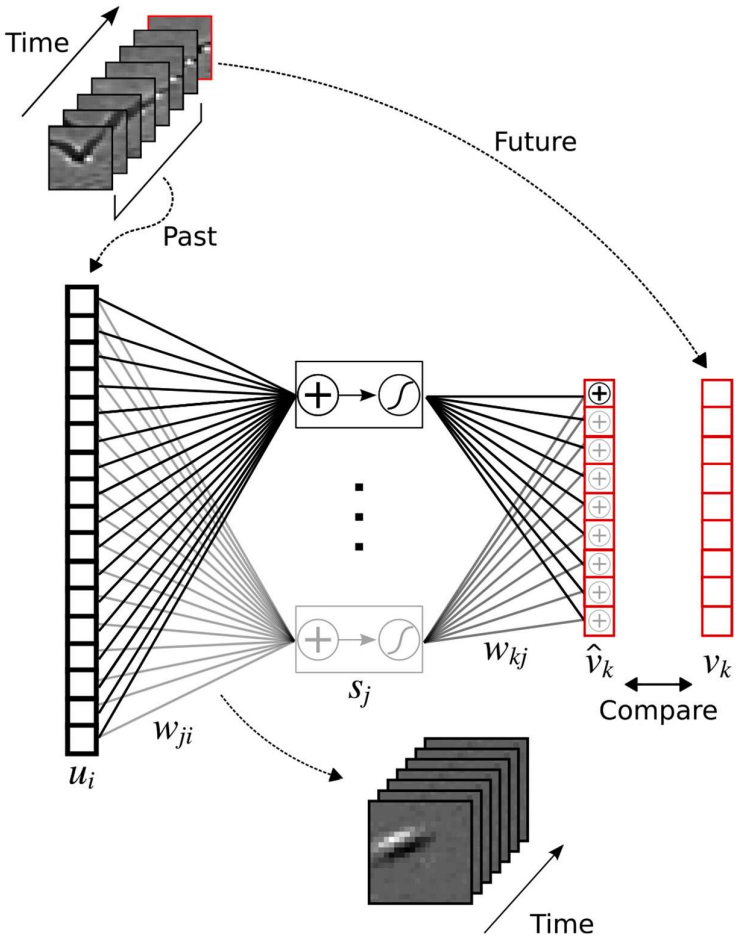


Figure 1.5. A temporal prediction model trained on movie clips. The feedforward model is optimised to predict the upcoming future movie-sample from the most recent past of the movie. Figure adapted from (Singer et al., 2018).

1.4 Decision Making

Sensory processing provides information from the environment to an animal, which is used to generate suitable actions for survival. Decision making can be described as a four-stage process (Doya, 2008). First, information about the present state of the environment and animal is gathered. Second, possible choice and action are evaluated and assigned a reward

or punishment value. Third, an action is chosen on the basis of step one and two, which ideally optimises the value of the chosen action. Finally, the possible and chosen actions are re-evaluated based on the outcome of the decision made.

A wide range of decision models have traditionally been used in neuroscience, e.g. accumulator models where evidence is integrated over a period of time, such as the drift diffusion model (Ratcliff, 1978) and the leaky competing accumulator (Usher and McClelland, 2001). Simpler models have also been used for decision-making tasks, where the decision variables are mapped onto the independent variables in a logistic regression, and the dependent variable represents the choice as a probability value. This approach has been used for rhesus monkeys trained on a visual decision task (Lau and Glimcher, 2005), and for a two-armed bandit task using rats (Kim et al., 2009).

1.5 Overview of Chapters

In chapter 2, we show that artificial neural networks optimised for temporal prediction of raw sound waves can explain the tuning properties in the cochlea, as the hidden units of the model exhibit similar response properties to those of inner hair cells and auditory nerve fibres that have been found experimentally.

In chapter 3, we demonstrate how a recurrent temporal prediction model can predict the behaviour of mice performing an auditory discrimination task using a wide variety of auditory stimulus features as decision parameters.

In chapter 4, we show that a recurrent temporal prediction model can explain local functional connectivity in A1, and investigate whether the temporal prediction principle can predict A1 neural responses to natural sounds.

In the 5th and final chapter, we will summarise our results and give our concluding remarks. We will discuss temporal prediction as a framework and suggest future investigations and experimental predictions of our work.

Chapter 2. Temporal Prediction and the Cochlea

2.1 Introduction

Neurons in the sensory systems are tuned to diverse but specific features of the sensory environment. Various principles have been proposed which explain how this specific tuning may be the result of adaptation to the statistical regularities of natural sensory stimuli, e.g. efficient coding (Barlow, 1959), sparse coding (Olshausen and Field, 1996, 1997; Hateren and Ruderman, 1998; Carlson et al., 2012) and slowness (Hyvärinen et al., 2004; Carlin and Elhilali, 2013). We suggest that sensory systems, including the peripheral auditory system, are governed by temporal prediction (Bialek et al., 2001; Palmer et al., 2015). Performing actions takes time and sensory input can only guide those actions accurately to the extent that it informs us about the state of the environment at the time they take place. Optimisation for efficient temporal prediction may also encourage representations that better reflect underlying causes and eliminate irrelevant information (Bialek et al., 2001). The most useful components of incoming sensory data are maybe those components that are predictive (Bialek et al., 2001). We argue that it is these components that might be preferentially encoded in the nervous system.

There is evidence for temporal prediction in other sensory systems and processing levels. Notably, the temporal prediction principle has been tested in the larval salamander retina (Palmer et al., 2015; Salisbury and Palmer, 2016). Recordings from the retina showed that every recorded cell participates in a 5-cell group that encodes near maximal predictive

information, and therefore that predictive computation occurs in the early stages of sensory processing. Theoretical work has also found that artificial neural networks optimised for a simplified form of temporal prediction applied to movies of natural scenes produces model units with receptive fields resembling those found for neurons in primary visual cortex (Singer et al., 2018). Furthermore, when the networks were optimised to instead predict the future of sound spectrograms from their past for a dataset of natural sounds, the resulting receptive fields of the hidden units resembled those of real neurons recorded in auditory cortex (Singer et al., 2018). In this work, the processing of the cochlea was fixed and assumed to be approximated by a cochleagram, which is a spectrogram-like process. Furthermore, a hierarchical temporal prediction model, where each subsequent layer is optimised to predict the activity of its input layer, can explain aspects of sensory tuning along the visual pathway, from retina, to primary visual cortex, to the middle temporal visual area of cortex (Singer et al., 2019).

In this chapter, we investigate whether the temporal prediction principle can explain processing by peripheral auditory structures. The peripheral auditory system consists of the outer, middle and inner ear, and the cochlea is situated in the inner part. The cochlea is often thought of as a frequency analyser and consists of a fluid-filled tube bisected length-ways by the basilar membrane. The basal regions of the membrane are maximally displaced by high frequency sound components, and the frequency of maximal displacement progressively decreases towards the less stiff apex. Thus, the preferred frequency increases approximately exponentially from apex to base (Liberman, 1982), and more length is dedicated to low frequencies than high. Vibration of a particular region of the basilar membrane depolarises inner hair cells located there, which send signals via the auditory nerve to the cochlear

nucleus in the brainstem. The impulse response (IR) of a point on the basilar membrane or the corresponding auditory nerve fibres can be approximated by a gammatone filter. In the time domain, gammatone filters are sinusoidal functions encapsulated within an asymmetric temporal envelope. This envelope has a fast onset and a slower more gradual offset (Goblick and Pfeiffer, 1969). They are spectrally tuned, with a filter bandwidth that increases with preferred sound frequency.

The representations of sounds in the cochlea and auditory nerve have previously been investigated by models optimised to efficiently encode sound (Lewicki, 2002; Smith and Lewicki, 2006). These models produced units with IRs that were similar to cochlear filters in many respects, showing sinusoidal ringing within a limited envelope, and a range of frequency tuning across units. However, the envelopes of the IRs in Lewicki (2002) were symmetrical in form in contrast to the asymmetric temporal envelopes of the biology. The model of Smith and Lewicki (2006) produced appropriately asymmetric envelopes, but the inference process of the model was also not very biologically plausible as it did not respect causality requiring an iterative process that requires all time points comprising a long waveform to be available simultaneously.

We asked if the temporal prediction principle could capture cochlear features and address some of these shortcomings. Our initial approach made relatively few assumptions about the components of the cochlea and their interactions, and we simply modelled the cochlea as an artificial neural network composed of elementary interacting nonlinear components. We then examined the tuning properties of the units of the networks.

2.2 Results

2.2.1 Cochlear Temporal Prediction Model

To examine whether optimisation for temporal prediction could explain cochlear tuning properties we trained a simple neural network model to predict the immediate future of sound waveforms from their recent past. The model was a feedforward neural network with one layer of hidden units (Fig. 2.1) (see Methods for more details). The input layer consisted of 128 units fully connected to 64 hidden units, and the hidden units were fully connected to 16 output units. The input was the immediate past waveform, and the activity of each hidden unit was a weighted sum of the input followed by a nonlinear transform by the scaled hyperbolic tangent function. The output weights of the network provided the linear prediction of the upcoming future from the activity of the hidden units.

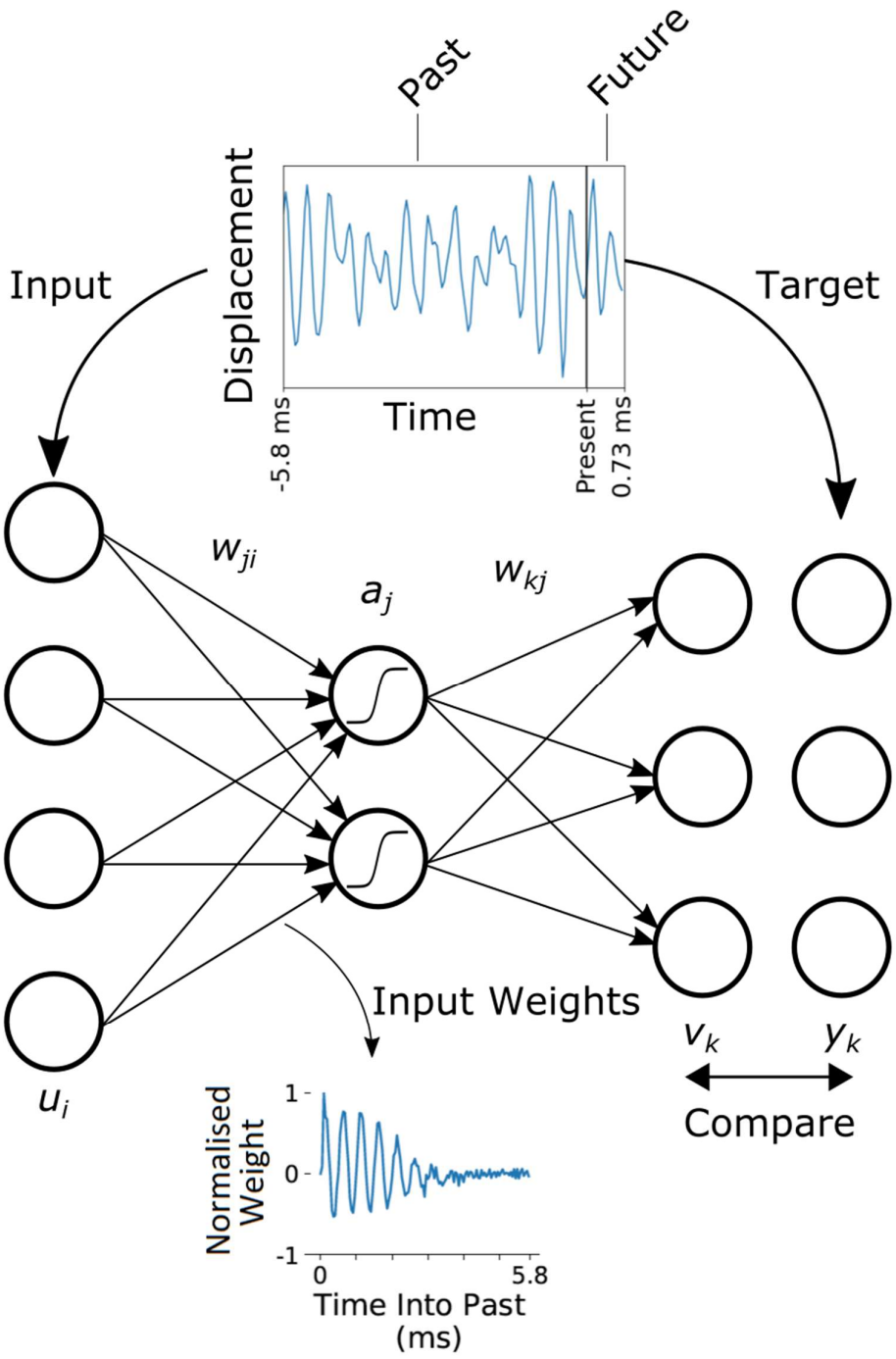


Figure 2.1. Schematic of the temporal prediction model of the cochlea. Raw sound waves were provided as inputs to a feedforward neural network, and optimised to perform temporal prediction by predicting the immediate future of sound waveform based on its most recent past. Each input unit represents a particular sample (time point) in the past 5.8 ms, while each output unit represents a particular sample in the future 0.73 ms. Here, the weights, w_{ji} and w_{kj}, and the biases, b_j, and b_k, were

optimised to predict the future, v_k, given the past, u_i, for a large dataset of natural sounds. The upper inset figure shows an example sound snippet representing the past, u_i, and the future, y_k, which served as input and target for the temporal prediction objective. An example of the tuning properties learned by the model is shown in the lower inset figure. The incoming weights, w_{ji}, to hidden unit j can be interpreted as the IR of this hidden unit.

To train the model we used a training set that consisted of sound recordings representative of the natural auditory environment. This dataset included birdsong, animal vocalisations, environmental sounds, human speech by children and adults, and baby vocalisations. The stimuli were linearly filtered to mimic the transformation performed by the middle ear. We then added weak Gaussian noise to the input, which may act as regularisation.

Regularisation of a network prevents overfitting, and improves the generalisation properties, and can be implemented through the addition of input noise (Sietsma and Dow, 1991). Another form of regularisation is Tikhonov regularisation (Bishop, 1995). In regression, if we want to estimate the mapping from a known matrix *X* to a vector *y* by finding a vector weight *w* such that $y = Xw$, we can minimise the ordinary least squares $\|Xw - y\|^2$, where $\|\ \ \|$ denotes the L2 norm. In this case, Tikhonov regularisation adds a regularisation term $\|\Gamma w\|^2$, and thus in this case minimises $\|Xw - y\|^2 + \|\Gamma w\|^2$, where Γ is the Tikhonov matrix. L2 weight regularisation, a common form of regularisation, uses a regularisation term of $\lambda\|w\|^2$, and is a type of Tikhonov regularisation. Small levels of input noise have shown to be equivalent to Tikhonov regularisation under certain assumptions (Bishop, 1995).

The model parameters were trained to predict the waveform over a brief timespan into the future (16 samples = 0.73 ms) from the present, given the recent past (128 samples = 5.8 ms) of the waveform immediately before the present. We assumed that the code used by the cochlea can be approximately linearly decoded to predict the future input. This is consistent with the observation that decoding stimuli from neural responses is often performed surprisingly well by a linear transformation (Eliasmith and Anderson, 1991). We interpreted the pattern of input weights, w_{ji}, to each hidden unit, j, as being analogous to the IR of a specific site along the basilar membrane.

The trained temporal prediction model mostly produced substantial IRs for the hidden units (51 out of 64) (Fig. 2.2). Hidden units with IR power below 1% of the average IR power were deemed non-active and removed from further analysis.

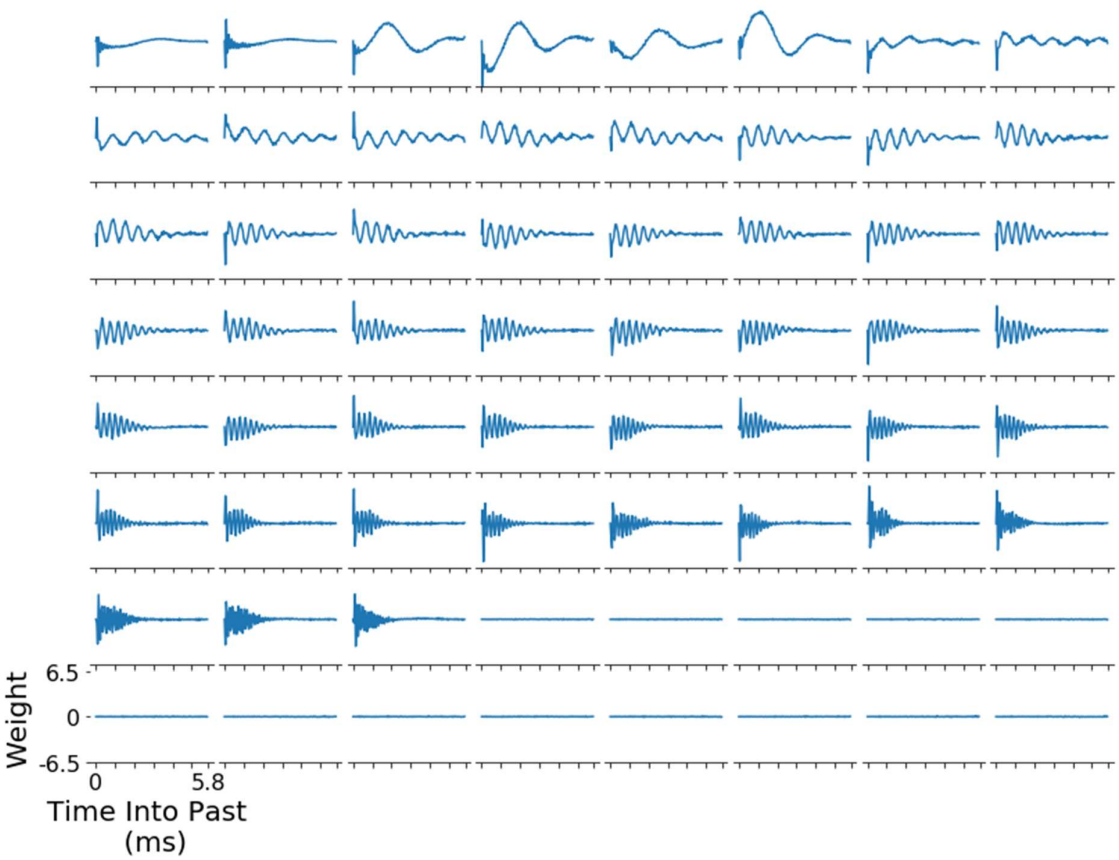


Figure 2.2. All IRs for the temporal prediction model. The IRs of all 64 hidden units of the model sorted by centre frequency, which was determined by its peak in frequency space. The IR for hidden unit *j* is the weights, w_{ji}, from all input units, *i*, representing the past to that hidden unit. If the power of the IR failed to exceed a threshold of 1% of the average model IR power, it was removed from further analysis (last 13 hidden units shown).

We compared the IRs of the trained model (Fig. 2.3A) with experimental data taken from auditory nerve fibres of anaesthetised cats (de Boer, 1978; Carney and Yin, 1988; Lewicki, 2002) (Fig. 2.3B), and found a qualitative resemblance. They showed the characteristic sinusoidal ringing elicited by the basilar membrane with a quick onset followed by a more prolonged offset. The model IRs were temporally tuned by preferentially encoding

the most recent part of the auditory stimulus, similar to cochlear filters. Most model IRs replicated the auditory nerve fibre IRs, although the onset of the sinusoidal ringing was generally faster for the model IRs than in the real basilar membrane response (Fig. 2.2). Unlike biological IRs, some model IRs also showed a strong preference for the most recent stimulus sample, as seen by the click-like rise.

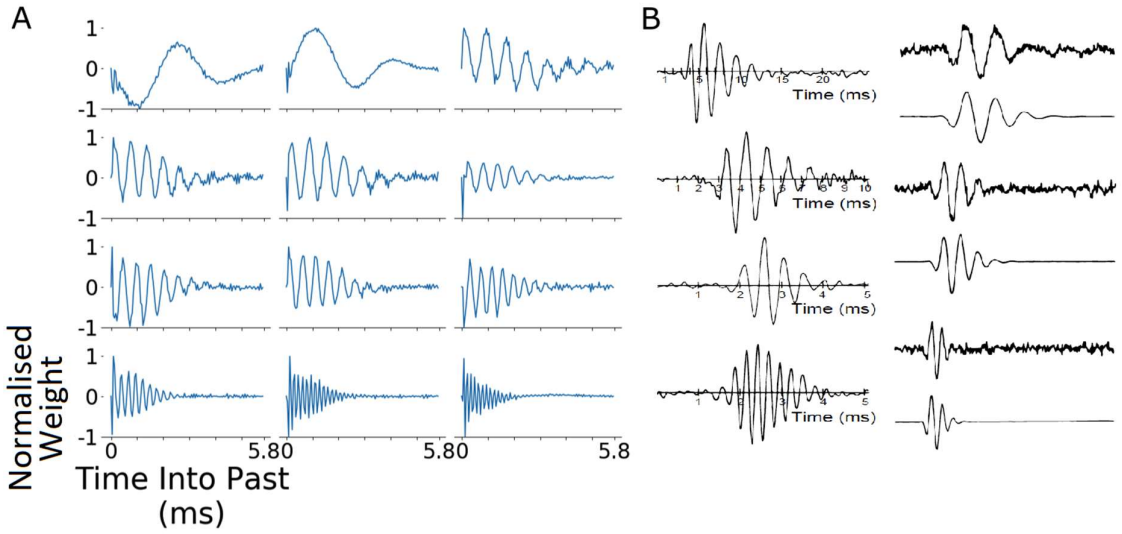


Figure 2.3. IRs of 12 example hidden units of the temporal prediction model and auditory nerve fibres. The hidden unit IRs resembled the IRs shown by the auditory nerve fibres. **(A)** IRs of the network model. The IR for each hidden unit was normalised to its respective maximum absolute value. **(B)** IRs from seven different auditory nerve fibres derived from anaesthetised cats using reverse correlation (de Boer, 1978; Carney and Yin, 1988; Lewicki, 2002).

We determined the centre frequency of each hidden unit as the peak of its IR in frequency space. The spectral tuning of the hidden units was sharply located in frequency space, and spanned from the lowest frequency bin of around 175 Hz up to around 7000 Hz (Fig. 2.4A). We fit the centre frequency (f) as a function of the units' rank order with an exponential equation that describes the tonotopic cochlear map in mammals (Greenwood,

1961, 1990, 1996; Robles and Ruggero, 2001) (Fig 4B), given by $f = A(10^{Cz} + B)$, where *z* is the distance from the cochlea apex scaled from 0 to 1, and *A*, *B* and *C* are fitting parameters. We do not have an absolute measure of distance from the cochlea apex in our model, but we take each unit, ordered by centre frequency, to represent an equal-sized adjacent space along the basilar membrane. Thus, together the units represent a contiguous section somewhere along the basilar membrane. Hence, we take z = 0 at the position of the lowest-frequency model unit (not the apex), and z = 1 to be the position of the highest-frequency model unit (not the base), with each unit in between assigned an value of z proportional to its rank order r between these limits, that is $z = (r-1)/(r_{max}-1)$, where r_{max} = 51 is the number of units with substantial impulse responses. The exponential function fit better than a linear fit (F-test, $F(49,48) = 320.05$, $p = 1.11\times10^{-16}$). All together the IRs spanned a wide range of centre frequencies and the model hidden layer comprised a filter-bank resembling the basilar membrane of the cochlea, and approximately captured the smooth, near-exponential dependence of tuning frequency of basilar membrane position exhibited by the cochlea (Liberman, 1982). The bandwidth of the IRs also increased as the centre frequency increased (Pearson's correlation coefficient, denoted as $\rho = 0.816$, $p = 7.01\times10^{-12}$), as seen for auditory filters (Fig. 2.4C) (Moore, 1986). The sharpness of each IR was estimated by dividing the centre frequency of the IR by its bandwidth. Physiological data suggests that the filter sharpness increases with increasing centre frequency ($\rho = 0.816$, $p = 2.43\times10^{-24}$) (Evans, 1975). The filter sharpness of the hidden units of our model increased steadily with centre frequency ($\rho = 0.44$, $p = 0.00248$), which matched experimental findings fairly well up until around 7000 Hz (Fig. 2.4D). The temporal envelope of an IR is the temporal span that captures 95% of the IR's power (Lewicki, 2002). We found that this decreased with increasing centre frequency ($\rho = -0.86$, $p = 7.592\times10^{-14}$) (Fig. 2.4E).

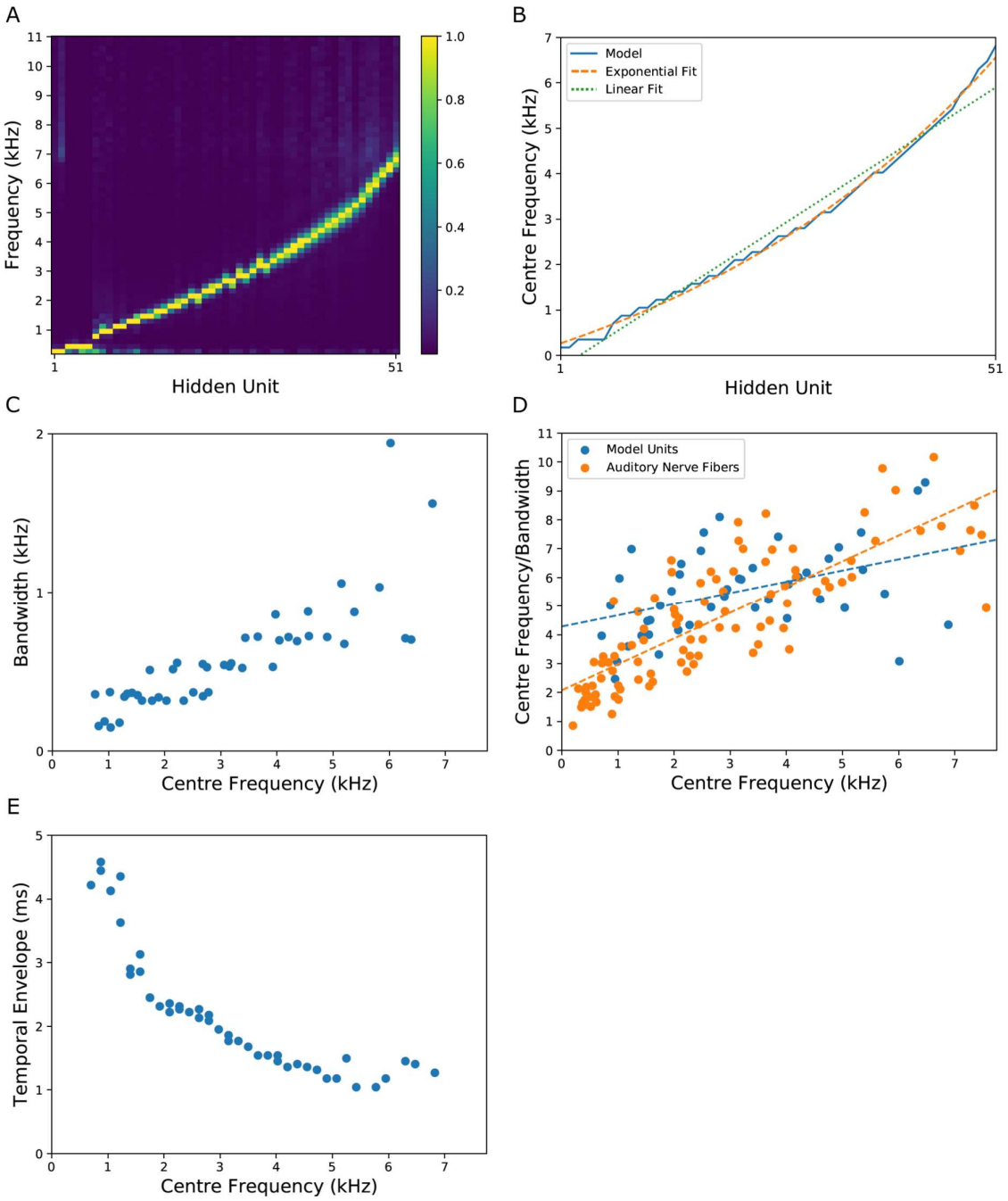


Figure 2.4. IR characteristics of the temporal prediction model. (A) The normalised power spectrum of the IRs of the hidden units of the feedforward model. The colour-bar represents the normalised power. If the power of the IR failed to exceed a threshold of 1% of the average model IR power, it was removed from analysis (13 out of 64 hidden units). **(B)** The centre frequency of each model unit as a function of hidden unit rank. The hidden units are interpreted to span a contiguous section of the basilar membrane in tonotopic order, with each unit evenly spaced along this section. The model

centre frequencies are fitted with an exponential equation that has been used to describe the relationship between distance along the basilar membrane and centre frequencies (Greenwood, 1961, 1990, 1996; Robles and Ruggero, 2001). The best linear fit of the model centre frequencies is also shown. **(C)** The IR bandwidth of each hidden unit plotted against centre frequency. The IR bandwidth was estimated as a 10 dB decrease on either side of the spectral peak of the IR. **(D)** The filter sharpness of each hidden unit plotted against centre frequency. The filter sharpness was determined by dividing centre frequency by IR bandwidth. Blue dots indicate data points taken from the temporal prediction model. Orange dots indicate data points taken from cat auditory nerve fibres (Evans, 1975). The best linear fit of both the model and experimental data are shown as dashed lines in the respective colour. **(E)** The temporal envelope of each hidden unit plotted against centre frequency. The temporal envelope of each IR was determined by the shortest time window possible that captured 95% of the power of the IR.

2.2.2 The Effects of Noise Injection to the Training Set

We found that the addition of weak Gaussian noise to the input was necessary for the model to learn cochlea-like IRs (Fig. 2.3). An absence of noise in the training set produced a representation not found in the cochlea or auditory nerve fibres (Fig. 2.5). The IRs did not reproduce the gammatone IRs found in the cochlea, and did not exhibit the gradual offset. Instead, the IRs displayed sequences of staggered onset and offset with some IRs containing an additional underlying low frequency oscillation (Fig. 2.5B). This may be due to overfitting. Most IRs were not sharply tuned in frequency space and the centre frequencies of the IRs were separated onto opposite ends of the frequency space with virtually no tuning to the mid-range frequencies (Fig. 2.5A).

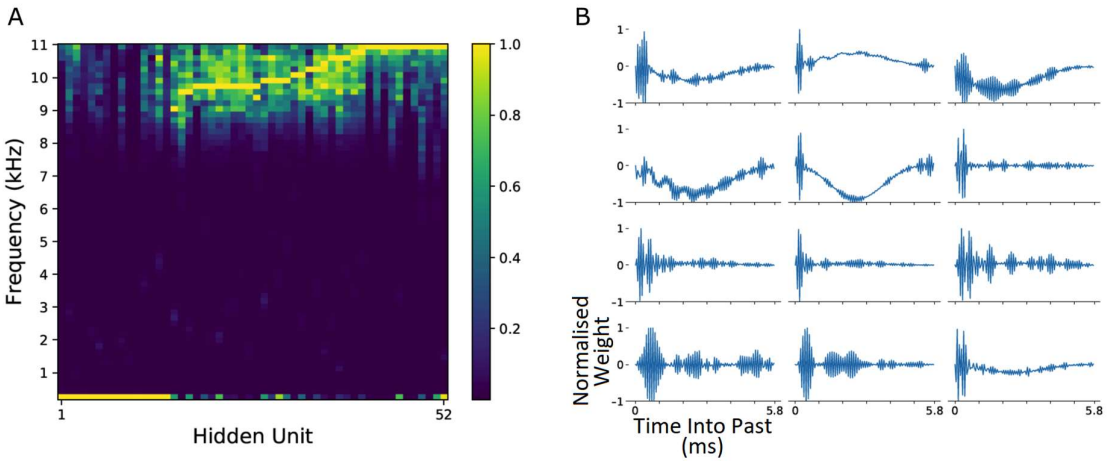


Figure 2.5. IR characteristics of the temporal prediction model trained on noiseless sound. (A) The normalised power spectrum of the IR of the hidden units of the feedforward model. The colour-bar represents the normalised power. If the power of the IR failed to exceed a threshold of 1% of the average model IR power, it was removed from analysis (12 out of 64 hidden units). **(B)** IRs of the network model. The IR for each hidden unit was normalised to its respective maximum absolute value.

We further examined the effect of noise on the training set of the temporal prediction model. When Gaussian noise was added to the future targets only, while keeping the past input clean, there was a deterioration of the biological similarity of the model (Fig. 2.6C and 2.6D), as with the noiseless case. Without input noise, the IRs displayed a more jagged structure, which may be due to overfitting. However, noise on both input and target resulted in a similar representation to the model trained on noisy input (Fig. 2.6A and 2.6B), although there were fewer active units (34 active hidden units) and a subset of these lacked sharp frequency tuning. We concluded that noise added to the input was the crucial factor in order to explain cochlear tuning properties.

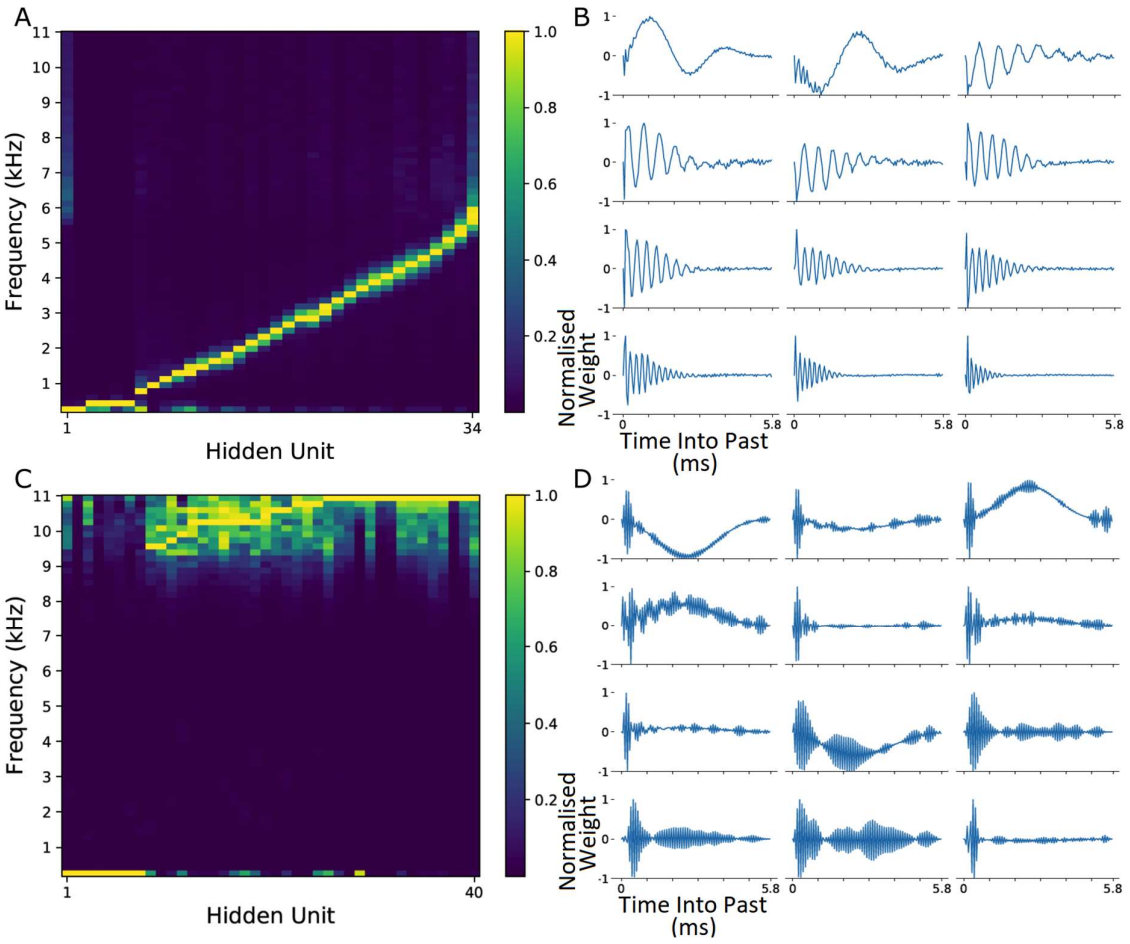


Figure 2.6. IR characteristics of the temporal prediction model trained on different noise schemes. **(A, B)** Noisy input and target. **(C, D)** Noisy target. **(A, C)** The normalised power spectrum of the IR of the hidden units of the feedforward model. The colour bar represents the normalised power. If the power of the IR failed to exceed a threshold of 1% of the average model IR power, it was removed from analysis (30 and 24 hidden units respectively). **(B, D)** IRs of the network model. The IR for each hidden unit was normalised to its respective maximum absolute value.

To determine the effect of input noise on the prediction performance of our network, we explored the effect of applying noise to the input at a range of different signal-to-noise ratios (SNRs) (Fig. 2.7). Across a range from the 3 to 15 dB SNR, hidden units exhibited cochlear-like tuning (Fig. 2.7C and D). The similarity between the hidden units and biological

tuning deteriorated for noise levels outside this SNR range (Fig. 2.7A, B and E). We found that the noise level previously presented in this chapter was the one that best mimics the cochlear physiology overall (9.1 dB SNR). Ideally, we might expect that this same noise level should enable the temporal prediction model to best predict future input when evaluated on a held-out noiseless test set. However, we found, in fact, that the no-noise case provided the best predictions (Fig. 2.7F).

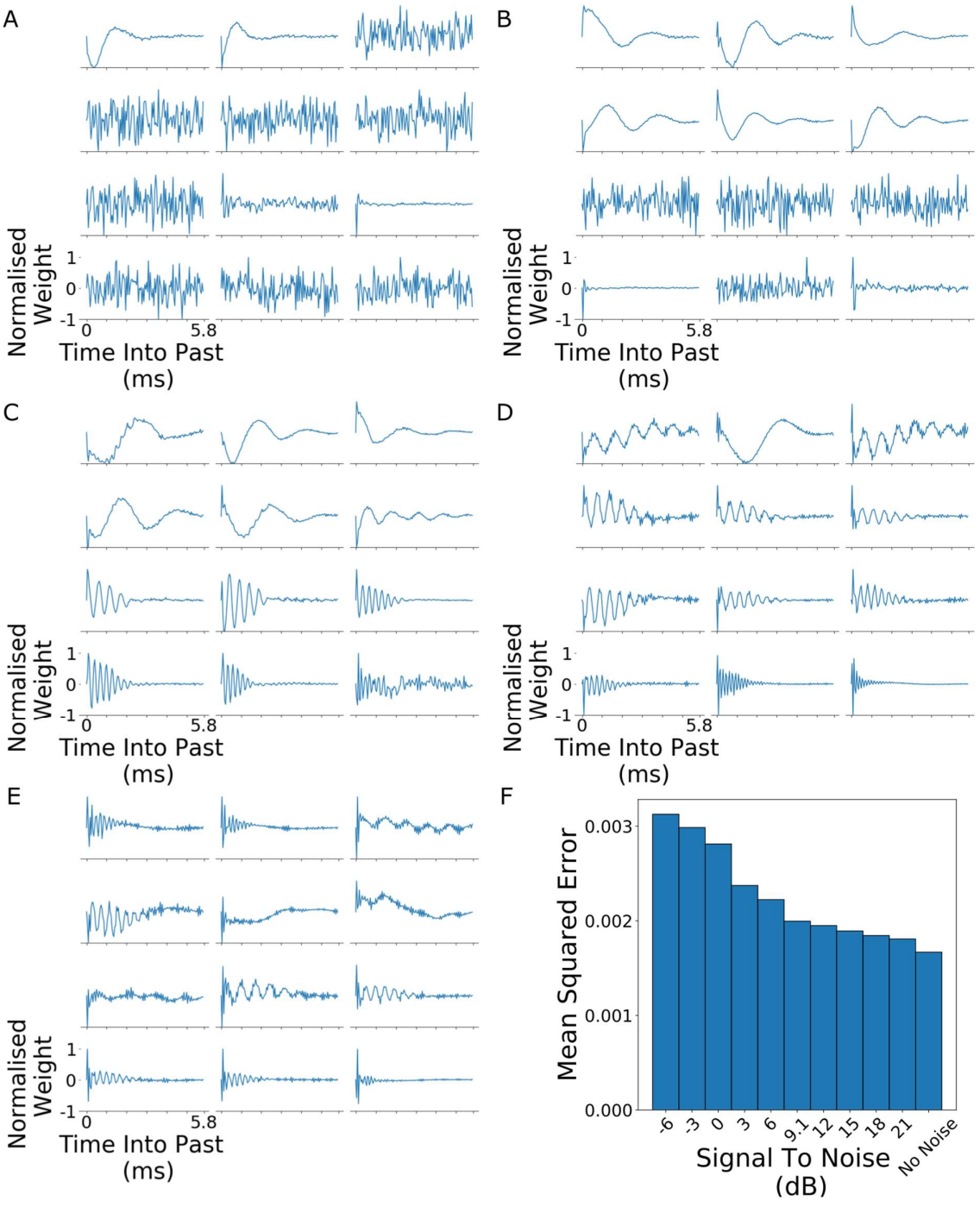


Figure 2.7. IRs of temporal prediction models trained on various levels of noise added to the input. **(A-E)** IRs of models trained on a signal-to-noise ratio (SNR) of -6, 0, 3, 15, 21 dB, respectively. The IR for each hidden unit was normalised to its respective maximum absolute value. **(F)** Prediction performance of temporal prediction models trained with various levels of noise added to the input. In

each case, the models were tested on a held-out test set consisting of natural sounds without added noise.

2.2.3 Cochlear Autoencoder Model

We also asked whether temporal prediction itself was crucial for our model to learn cochlea-like tuning. It is possible that a network that merely compresses auditory stimuli is enough to produce cochlea-like tuning, and that temporal prediction *per se* is not required. Two previous models based on the principles of efficient coding have had some success in reproducing aspects of cochlear filters (Lewicki, 2002; Smith and Lewicki, 2006), although they have some limitations. Efficient coding seeks to use coding resources efficiently to represent all the incoming sensory input, and can be viewed as a form of compression of the input (Chalk et al., 2018). This is in contrast with temporal prediction, where incoming sensory input is only represented to the degree that it is predictive of future input. To investigate how much shared characteristics between the temporal prediction model and cochlear features can be explained as a consequence of compression rather than prediction, we trained an autoencoder neural network. The autoencoder model is identical to the temporal prediction model apart from the output layer, which recreates the input (Fig 2.8), and was trained to recreate the input from a compressed encoding on noiseless and noisy sound (Fig. 2.9 and 2.10 respectively).

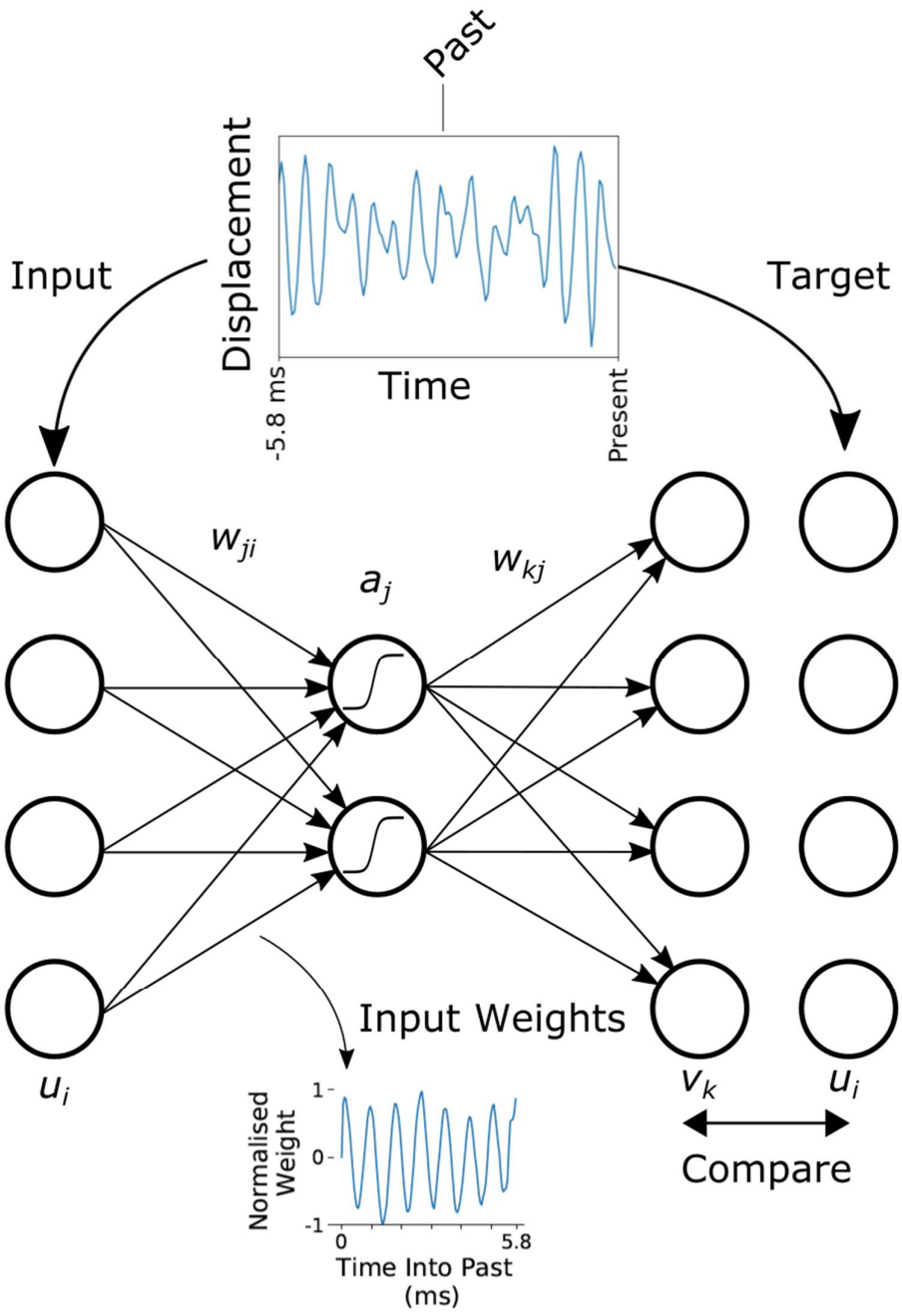


Figure 2.8. Schematic of the autoencoder model of the cochlea. Raw sound waves were provided as inputs to a feedforward neural network, and optimised to recreate its input. Here, the weights, w_{ji} and w_{kj}, and the biases, b_j and b_k, were optimised to recreate its input, v_k, given the identical input, u_i, for a large dataset of natural sounds. The upper inset figure shows an example sound snippet representing the past, u_i, which served as both input and target for the autoencoder objective. Each input unit represents a single sample of the past 5.8 ms, while each output unit represent a sample of

the model recreation of the same 5.8 ms of the stimulus. An example of the tuning properties learned by the model is shown in the lower inset figure. The incoming weights, w_{ji}, to hidden unit j can be interpreted as the IR of this hidden unit.

In both cases, the autoencoders were unable to capture the same range in frequency tuning and also did not display sharp frequency tuning apart from a small subset of hidden units (Fig. 2.9A and 2.10A). A few gammatone-like IRs were displayed by the noisy autoencoder, but without the temporal asymmetry demonstrated by biological cochlear filters, by being more strongly driven by recent stimuli (Fig. 2.10B). For noiseless sounds, we observed a small subset of units that exhibit oscillatory sinusoid-like IRs, but without strongly resembling gammatone filters (Fig 2.9B). The bandwidth of the IRs for the noiseless autoencoder did positively correlate with center frequency ($\rho = 0.326$, $p = 0.0086$) (Fig 2.9C), but not for the noisy autoencoder ($\rho = 0.19$, $p = 0.132$) (2.10C). The filter sharpness overshot the biological data across the frequency range (Fig 2.9D and 2.10D), although both the noiseless and noisy autoencoder did exhibit a positive correlation with center frequency ($\rho = 0.45$, $p = 1.9\times10^{-4}$) and ($\rho = 0.6$, $p = 1.5\times10^{-7}$) respectively. The temporal envelope was largely uncorrelated with the centre frequency for both the noiseless ($\rho = 0.182$, $p = 0.15$) and noisy autoencoder ($\rho = -0.044$, $p = 0.731$), which is unlike cochlear filters (Fig. 2.9E and 2.10E).

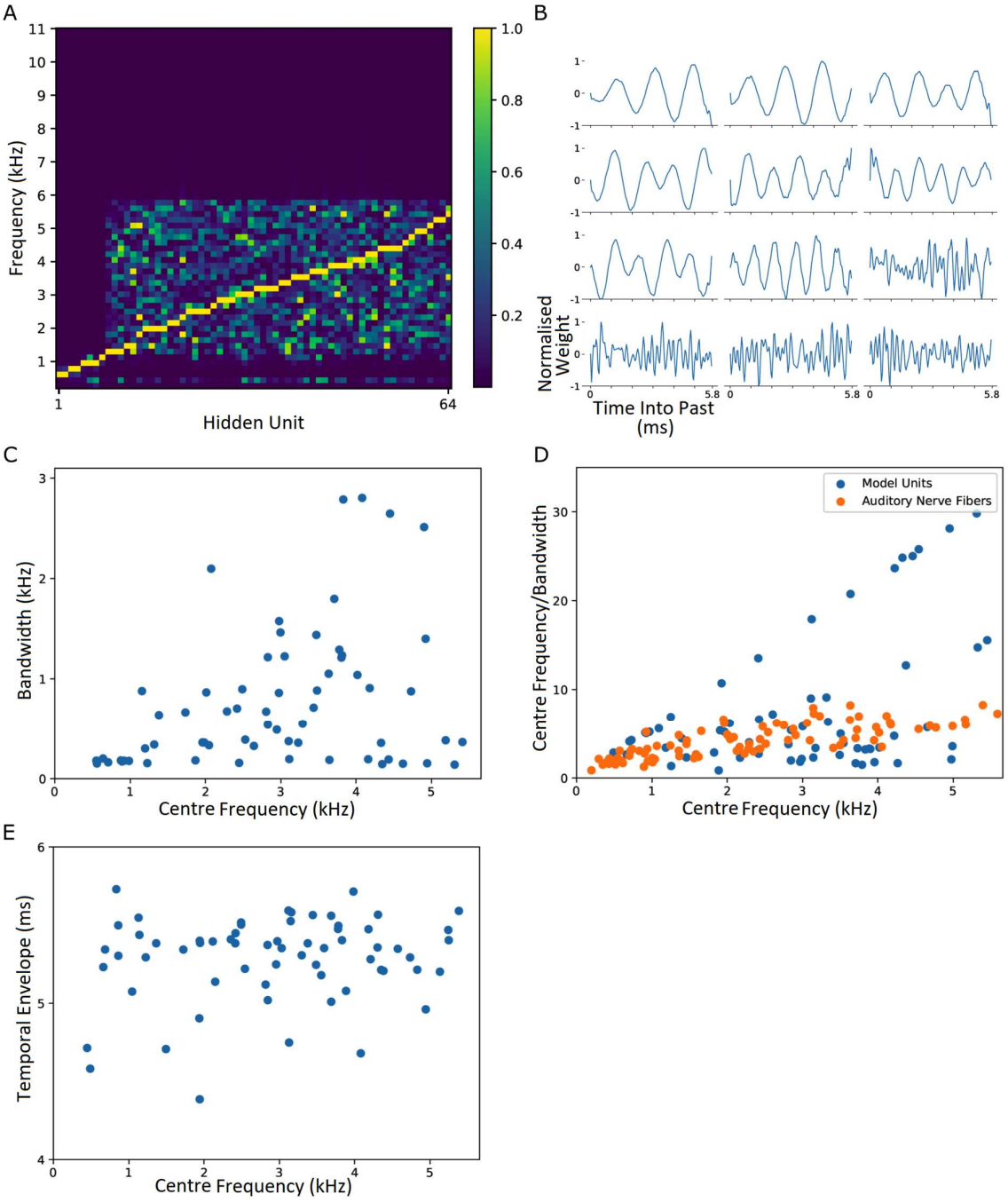


Figure 2.9. Autoencoder model trained on noiseless sound. (A) The normalised power spectrum of the IR of the hidden units of the feedforward model. The colour bar represents the normalised power. **(B)** IRs of the network model. The IR for each hidden unit was normalised to its respective maximum absolute value. **(C)** The IR bandwidth of each hidden unit plotted against centre frequency. The IR bandwidth was estimated as a 10 dB decrease on either side of the spectral peak of the IR. **(D)** The

filter sharpness of each hidden unit plotted against centre frequency. The filter sharpness was determined by dividing centre frequency by IR bandwidth. Blue dots indicate data points taken from the temporal prediction model. Orange dots indicate data points taken from cat auditory nerve fibres (Evans, 1975). **(E)** The temporal envelope of each hidden unit plotted against centre frequency. The temporal envelope of each IR was determined by the shortest time window possible that captured 95% of the power of the IR.

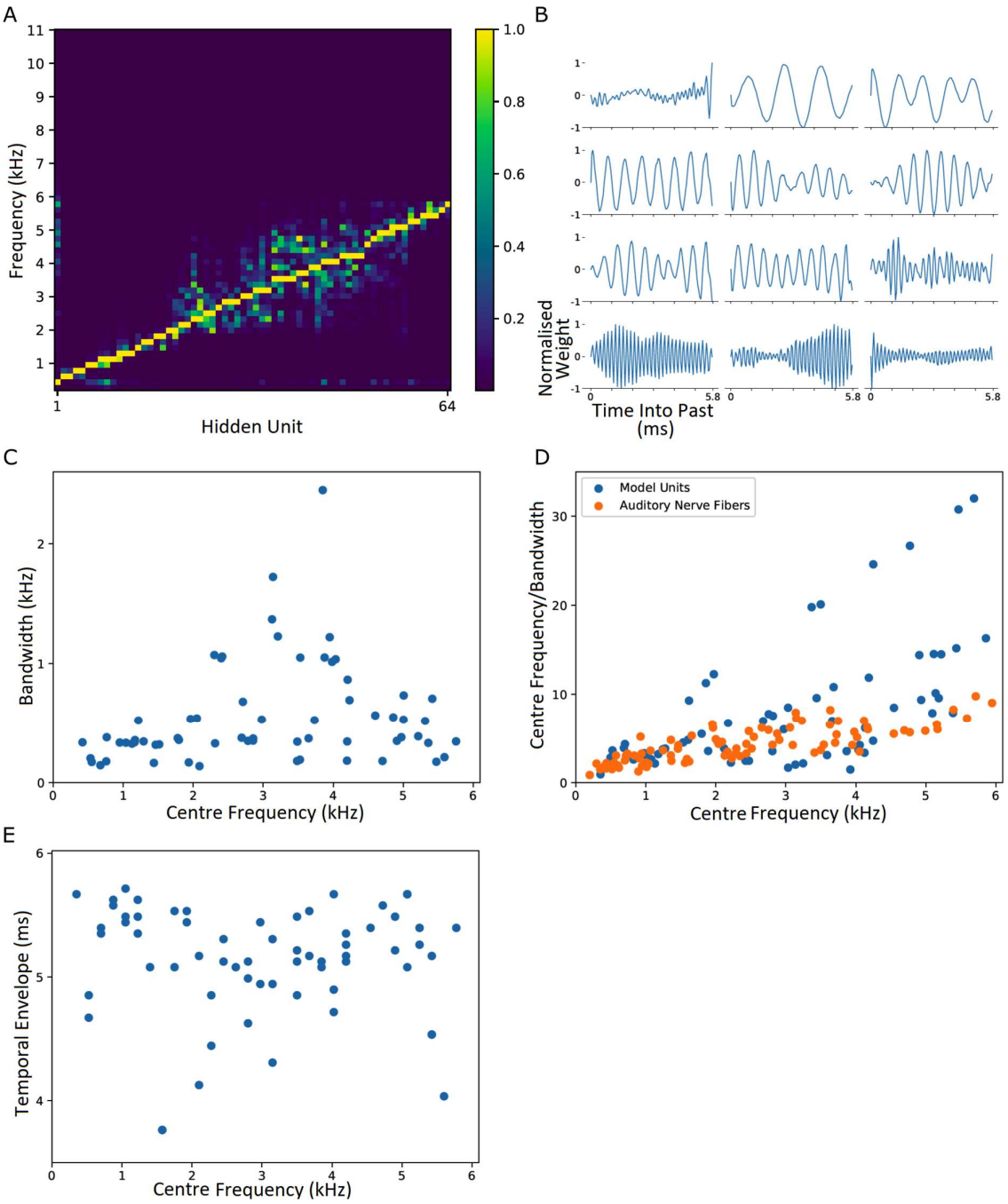


Figure 2.10. Autoencoder model trained on noisy sound. (A) The normalised power spectrum of the IR of the hidden units of the feedforward model. The colour bar represents the normalised power. **(B)** IRs of the network model. The IR for each hidden unit was normalised to its respective maximum absolute value. **(C)** The IR bandwidth of each hidden unit plotted against centre frequency. The IR bandwidth was estimated as a 10 dB decrease on either side of the spectral peak of the IR. **(D)** The

filter sharpness of each hidden unit plotted against centre frequency. The filter sharpness was determined by dividing centre frequency by IR bandwidth. Blue dots indicate data points taken from the temporal prediction model. Orange dots indicate data points taken from cat auditory nerve fibres (Evans, 1975). **(E)** The temporal envelope of each hidden unit plotted against centre frequency. The temporal envelope of each IR was determined by the shortest time window possible that captured 95% of the power of the IR.

2.3 Discussion

We optimised a feedforward network with one hidden layer to predict the immediate future of sound waveforms given their recent past. The model was able to capture several features of the cochlea. The hidden units of the model mimicked a filter-bank and showed a wide range of frequency tuning. The density of frequency-tuned hidden units was inversely proportional to its centre frequency, with an approximately logarithmic distribution similar to that found in the cochlea (Liberman, 1982). The IRs of the hidden units had gammatone-like temporal profiles and resembled the ones recorded from cat auditory nerve fibres using electrophysiology (Fig. 2.2 and 2.3). More specifically, the IRs exhibited ringing with a rapid onset followed by a slow, gradual offset, and were temporally asymmetric by weighting the most recent past higher in order to predict the future waveform. The bandwidth and the filter sharpness of the IRs also increased as a function of centre frequency, while the temporal envelope decreased in a fashion consistent with the biology (Fig. 2.4) (Evans, 1975; Moore, 1986; Lewicki, 2002).

The addition of Gaussian noise to the input proved to be essential for the model to replicate cochlear features (Fig. 2.5 and 2.6). The addition of noise to a training set is often used as a type of data augmentation in the field of machine learning (Goodfellow et al., 2016). From a statistical point of view, the addition of noise to a dataset works as regularisation and helps with generalisation of the model by preventing overfitting (Sietsma and Dow, 1991). It is equivalent to Tikhonov or L2 regularisation when the noise amplitude is small (Bishop, 1995, 2006). However, while noise at about 3-15 dB SNR did enable the reproduction of cochlea-like features, it did not improve generalisation of prediction on a held-out test set.

An autoencoding, compressive network trained to recreate the input did not emulate cochlear features in the same way, which suggested that the cochlear features produced by temporal prediction are not solely due to compression (Fig. 2.9 and 2.10). Hence, our results suggest that the features of the cochlea may be optimised to encode the most predictive features of the natural soundscape, not just features that enable efficient compression.

2.3.1 Comparison to Other Models

Sensory coding has previously been approached using various other theoretical principles, one being efficient coding (Barlow, 1959). Efficient coding postulates that sensory systems are optimised to represent natural stimuli by minimising redundant activity between transmitting neurons (Lewicki, 2002). Cochlear models optimised to efficiently represent auditory stimuli produced filters that resemble the IR of cochlear auditory nerve fibres (Lewicki, 2002; Smith and Lewicki, 2006). However, the efficient coding approach used by Lewicki (2002) did not capture the asymmetry of biological IRs. The encoding algorithm of

Smith and Lewicki (2006) is also biologically implausible as it produces temporally-sparse spikes, each with a scalar value, rather than binary spikes as in the auditory nerve or a continuous real number as with the inner hair cell membrane potential. Also, this encoding algorithm does not place its spikes sequentially over time, but rather unrealistically treats time as a dimension one can move back and forth over (Smith and Lewicki, 2006). Our model has addressed some of these shortcomings by using temporal prediction as a governing paradigm. The IRs from the temporal prediction model capture the temporal asymmetry of cochlear filters. The model is also more biologically plausible and can be applied in real-time, without the need for a causal iterative settling process to produce the model output.

2.3.2 Arguments for Temporal Prediction in the Cochlea

There are several arguments as to why sensory structures, such as the cochlea, may be governed by temporal prediction. First of all, efficient temporal prediction may induce encoding of more 'underlying' features behind the stimuli (Bialek et al., 2001). Second, sensory stimuli guide future action, rendering the future sensory environment very relevant; optimisation for temporal prediction enables accurate and fast future action. Even a millisecond faster action enabled by prediction can make the difference in a survival context. Third, a sensory representation optimised for temporal prediction will disregard superfluous information of the input stimulus that is not predictive of the future. This can provide an initial filter to restrict the information the brain has to deal with, as gathering and transmission of unnecessary information is energetically costly (Marzen and DeDeo, 2017). Fourth, temporal prediction, as we instantiate it, may be more biologically realistic as it can be implemented in real-time, and does not require an iterative settling process as found in other studies (Smith

and Lewicki, 2006). Finally, temporal prediction is able to account for the temporal and spectral components of receptive fields found in the sensory system across different modalities (Palmer et al., 2015; Singer et al., 2018, 2019). Previous models have not been able to capture the temporal asymmetries of these receptive fields very well (Lewicki, 2002).

2.3.3 Weaknesses and Future Development of the Cochlear Temporal Prediction Model

The temporal prediction model captures some important cochlear features, but still has a few limitations. The rise on the model IRs is somewhat faster than in the biology, and some show an initial click not seen in the experimental data. We are currently trying to address this by investigating different combinations of time windows into the past and future, delays between the two, and building models that also predict the variance of the future rather than just the mean. The dissimilarities between the model and the biology, such as the click-like onset seen in some model units, might be due to factors unaccounted for by our model in its current form. It may be due to technical limitations such as the size of our sound dataset. It could be due to restrictions imposed onto a biological system by physical dynamics, for example the basilar membrane is a physical system perhaps better captured by optimising a mechanical model than a simple filter-bank. Or there may be important biological processes we have not incorporated, such as adaptation processes at the hair cell level (Eatock, 2000). Our model is currently quite simple, and a more explicit model that includes these considerations might produce IRs that are closer to cochlear filters.

The frequency tuning shown by the model ranges from around 175 to 7000 Hz and does not fully capture the range of humans hearing, which can reach an upper boundary of 20000 Hz. Reproducing cochlear features by training a neural network to optimise temporal prediction of waveforms was helped by the application of both a high- and low-pass filter to the dataset with cut-off frequencies of 500 Hz and 5512.5 Hz, respectively, in addition to a middle-ear filter. The application of the low-pass filter improved the model in terms of producing cochlear features, but limits the frequency range elicited. We have yet to explore the cut-off frequency of this low-pass filter as a hyperparameter, but suspect that the frequency range of the model can be increased while maintaining its biological fidelity by increasing the low-pass frequency threshold.

There are notable biological complications in the cochlea, such as the outer hair cells, which are also arranged in rows along its length, and respond to changes in their membrane potential due to basilar membrane movement by motor action that feeds back to the membrane. At least partly due to this, the cochlea exhibits some nonlinear behaviours. For example, basilar membrane displacement increases compressively with preferred-frequency sound amplitude (Rhode, 1971; Ruggero et al., 1997), and two-tone suppression resulting from the simultaneous presentation of two tones with nearby frequencies nonlinearly attenuates basilar membrane responses (Suga et al., 1963; Frishkopf, 1964). A future avenue would be to extend the current feedforward network architecture to one that more explicitly represents the basilar membrane, the outer and inner hair cells, and the auditory nerve as distinct layers that are appropriately connected, for example recurrently in the case of the outer hair cells and basilar membrane.

2.4 Methods

2.4.1 Auditory Dataset

We used an auditory dataset composed of natural sounds: birdsong and other animal calls, environmental sounds, and human speech by children and adults as well as baby vocalisations. Sound recordings of birdsong and mammal and insect calls were drawn from the Macauley Library at the Cornell Lab of Ornithology[1] (https://www.macaulaylibrary.org/) and the British Library Sounds Environment & Nature Collection (https://sounds.bl.uk/Environment. We also used a corpus of ferret vocalisations recorded in our laboratory and an additional recording from https://freesound.org/people/J.Zazvurek/sounds/155115/. Recordings of adult human speech were taken from http://databases.forensic-voice-comparison.net/ (Morrison et al., 2012). We also used some human speech and environmental sounds, such as snapping twigs, which were recorded in our laboratory in an anechoic chamber. Recordings of children speaking were from the CHILDES database (MacWhinney, 2000; Beckman and Edwards, 2010) from TalkBank (talkbank.org). Baby vocalisations were from the OxVoc database (Parsons et al., 2014).

Each recording was resampled to 22050 Hz and convolved using a filter that mimics the transformation of raw sound by the middle ear. This filter was provided by the Python

[1] Specifically, we use the following recordings from the Macauley Library: ML3429, ML3891, ML53181, ML55348, ML55350, ML59291, ML61494, ML62970, ML67895, ML70539, ML72995, ML80731, ML100723, ML100797, ML100857, ML111595, ML116302, ML116303, ML126289, ML129248, ML130909, ML132511, ML132526, ML136504, Ml164696, ML171372, ML191165, ML191178, ML191290, ML197064, ML199078, ML201215, ML205481, ML205774, ML207517, ML206448, ML207181, ML210645, ML213055, ML213356, ML217313, ML217490, ML220049, ML233387, ML516692, ML527184 and ML527292.

package “Brian Hears” (Fontaine et al., 2011). To lessen edge effects, we also applied a high- and low-pass filter (5th-order Butterworth) with cut-off frequencies of 500 Hz and 5512.5 Hz respectively. The high-pass filter was applied as the time window of the network input limited the capacity to capture frequencies with a period below the time window, and the high-pass filter provided a smooth fall-off in power near this cut-off. The low-pass filter was justified as high frequency components in the sound dataset contained relatively little power, and also because auditory nerve fibres do not phase lock at high sound wave frequencies. We exclusively selected sound snippets with a root mean square value above a threshold (0.105) to avoid silent sound clips as training input. Each individual sound snippet was then normalised to its Euclidian norm. Finally, we added Gaussian noise to the input of the sound snippets.

For the feedforward neural network, the auditory recordings were divided into N small audio snippets. Each snippet *n* was $I+K$ = 144 time steps (samples) long, where the first I = 128 (5.80 ms) acted as the input data u_{in}, indexed by i = 1 to I, and the last K = 16 (0.73 ms) as the target values y_{kn}, indexed by k = 1 to K.

2.4.2 Artificial Neural Network Models

The feedforward network has one hidden layer, and is fully connected between all layers. It contains i = 1 to I = 128 input units, one for each input sample, and j = 1 to J = 64 hidden units, and k = 1 to K = 16 output units, one for each output sample, to train for temporal prediction. The autoencoder model was trained to recreate its input and has K = I = 128 input and output units.

The activity of each hidden unit is the linear weighting from all input units followed by an activation function. The activation x_{jn} is a linear mapping from the input:

$$x_{jn} = b_j + \sum_{i=1}^{I} w_{ji} u_{in}$$

where b_j is the bias value for hidden unit j, w_{ji} the weight from input unit i to hidden unit j, and u_{in} the activity of input unit i for training example n.

The scaled hyperbolic tangent is used as the activation function for the hidden units:

$$\mathrm{g}(x) = \tanh(\alpha x)\beta$$

where α and β are scaling factors for the input and output respectively, and are set to 2/3 and 1.7159 respectively (LeCun et al., 1998).

a_{jn} is the activity of hidden unit j for training example n and is given by applying the activation function to the activation:

$$a_{jn} = \mathrm{g}(x_{jn})$$

The prediction v_{kn} for output unit k for training example n will then be:

$$v_{kn} = b_k + \sum_{j=1}^{J} w_{kj} a_{jn}$$

where w_{kj} is the weight from hidden unit j to output unit k, and b_k the bias for output unit k.

The trainable parameters w_{ji}, b_j, w_{kj} and b_k are optimised to perform temporal prediction, or to recreate the input for the autoencoder model, by using the objective function:

$$E = \frac{1}{NK}\sum_{n=1}^{N}\sum_{k=1}^{K}(v_{kn} - y_{kn})^2$$

where N is the total amount of training examples, and y_{kn} is the target value for output unit k for training example n, which is the future sound snippet or the identical input, u_{in}, for the temporal prediction model and autoencoder model respectively.

The training was performed using stochastic gradient descent on mini-batches of 100 training examples and the update function Adaptive Moment Estimation (Kingma and Ba, 2015).

2.4.3 IR Analysis

Hidden unit IRs with a power value of less than 1% of the mean IR power for all hidden units in a model were removed from analysis. The IR characteristics were found by determining the spectrum using the fast Fourier transform. The centre frequency was defined as the frequency bin containing the spectral peak for each IR. The bandwidth was interpreted as the frequency span within a 10 dB drop on either side of the centre frequency of the IR. Only units with a drop on both sides of the spectral peak were included in the IR characteristic analysis. The temporal envelope duration of each IR was interpreted as being the shortest temporal span that covers 95% of the total IR power.

2.4.4 Computational Implementation

Custom code was written in Python for data handling, modelling and data analysis, and run on NVIDIA GeForce GTX-1080 GPUs.

Chapter 3. Temporal Prediction and Auditory Behaviour

3.1 Introduction

In the previous chapter, we showed that the peripheral auditory system may be consistent with temporal prediction of sound. The encoded sound information from the peripheral auditory system travels up the auditory pathway where it is further processed, and then used for decision making and performing appropriate actions. In this chapter, we investigated if the application of temporal prediction models to the outputs of the auditory periphery provides a representation that explains aspects of sound categorisation by mice for multiple auditory discrimination tasks.

Animals, including humans, learn to discriminate between the diverse sensory stimuli in their environments and to generalise their behaviours to new instances of those stimulus types. For example, rain can sound dramatically different as it falls on different surfaces and yet we have no problem classifying it as rain and discriminating it from speech. What is the basis of this 'carving nature at its joints' (Plato, 1925), by which our brains separate the world into objects which fall into different classes e.g., footsteps, motorbikes, human voices, the sound of rain? Do we discriminate two stimuli to be of different classes, or generalise them into the same class, because of isolated physical properties such as frequency or colour, or because of their characteristics in some more abstract multi-dimensional space? Is this classification determined by the relative sensitivity of sensory neurons to different stimulus

dimensions? And is this a reflection of the contribution of these dimensions to the information in natural stimuli?

Most of our knowledge of perception is derived from tests along a single stimulus dimension. For example, psychophysics and electrophysiology studies have shown that the visual system can discriminate or categorise along dimensions of colour, contrast, or orientation. Similarly, the auditory system is able to discriminate or categorise sounds along dimensions of frequency (Rohrbaugh et al., 1971; Fay, 1974; Ehret, 1975; Syka et al., 1996; Kurt and Ehret, 2010), intensity (Terman, 1970; Rohrbaugh et al., 1971; Syka et al., 1996; Kobrina et al., 2018), frequency modulation direction (Wetzel et al., 1998; Ohl et al., 2001; Mercado et al., 2005; Rybalko et al., 2006) or amplitude modulation frequency (Schulze and Scheich, 1999; Kurt and Ehret, 2010). Yet discrimination along a single dimension can be influenced by stimulus features outside that dimension, as has been demonstrated in auditory processing (Nabelek and Hirsh, 1969; Heffner and Masterton, 1980; Florentine, 1983; Schouten, 1985; Gaese et al., 2006). In the visual modality, dimension integration has been studied by training animals, such as pigeons, to use dimension combinations (Vyazovska, 2021), and by asking human subjects to identify pop-up stimuli of different level of complexity (Ahissar and Hochstein, 2004); nevertheless, little is known about how multiple dimensions within any given modality are naturally integrated when classifying environmental stimuli. To understand the interactions between stimulus dimensions, as well as the hierarchy of importance of the dimensions, we focused on the auditory domain. Mice were trained to perform 2-dimensional sound discriminations before testing untrained generalisation. For example, the mice were trained to classify a high-frequency upward-sweep as safe and a low-frequency downward-sweep as unsafe. These two sounds differed, therefore, in both

frequency and the direction of the modulation. The mice were then tested for generalisation by examining how they classified novel stimuli that varied independently in both dimensions. For example, we determined if they judged a low-frequency upward-sweep to be safe, as the direction of modulation would suggest, or unsafe, as the frequency would suggest. By measuring how mice perceived and acted upon these changes we aimed to understand which dimensions tend to dominate in the discrimination of sounds (e.g. frequency or sweep direction).

Furthermore, we also investigated whether there might be general principles of sound classification that could explain the observed behavioural patterns of discrimination and classification. We hypothesised that the degree to which a dimension influences discrimination is a consequence of the degree of representation of that dimension in the neural population activity, and that this in turn reflects the degree of representation in the diverse array of natural sounds experienced by an animal over its lifetime or evolutionary history. There are several possible principled frameworks that have been posited to govern this representation (Salisbury and Palmer, 2016; Chalk et al., 2018). One hypothesis is that our brain seeks to efficiently compress the incoming input, removing redundant information (Attneave, 1954; Barlow, 1959; Olshausen and Field, 1996, 1997; Rao and Ballard, 1999). Another hypothesis is that our brain seeks to predict future sensory input (Bialek et al., 2001; Salisbury and Palmer, 2016; Chalk et al., 2018; Singer et al., 2018, 2019).

Here, we hypothesised that the similarity or difference of novel sounds to the safe and unsafe sounds will be judged in this compressed or predictive space. To test these general

hypotheses, we optimised network models for compression or prediction and subsequently trained a simple decision model to discriminate the safe and unsafe sounds based on the representation by the network models. We then compared the capacity of the different network models to explain the responses of the mice to the novel sounds.

3.2 Results

Mice were trained to discriminate between two frequency-modulated sounds, with one sound being 'safe', indicating water would be delivered at a spout upon a nose-poke, and the other being 'unsafe', indicating an aversive air puff would be delivered instead of water. The two sounds differed in two out of four dimensions: frequency range, sweep direction, modulation rate, and duration (Fig. 3.1A). For example, for sounds differing in frequency range and sweep direction, the safe sound could be a 9-18 kHz upward-sweep (Fig. 3.1B, blue frame), and the unsafe sound, a 6-3 kHz downward-sweep (Fig. 3.1B, red frame). Once the animals had learned to discriminate between the safe and the unsafe sounds, we sparsely introduced novel sounds (always safe) that varied in one or both of the two dimensions. Examining whether the mice responded to these sounds as 'safe' or 'unsafe' allowed us to explore how they generalise what they had learned. For example, a novel sound could have the frequency range of the safe sound but the direction of the unsafe sound, forcing the animal to make a choice whether to treat the new sound as similar to the safe or to the unsafe (Fig. 3.1B, other sounds in matrix). Measuring generalisation gave us access to the mouse's perception of the two sound dimensions at play and to establish whether decisions were based on just one of the dimensions or on both dimensions (Fig. 3.1C).

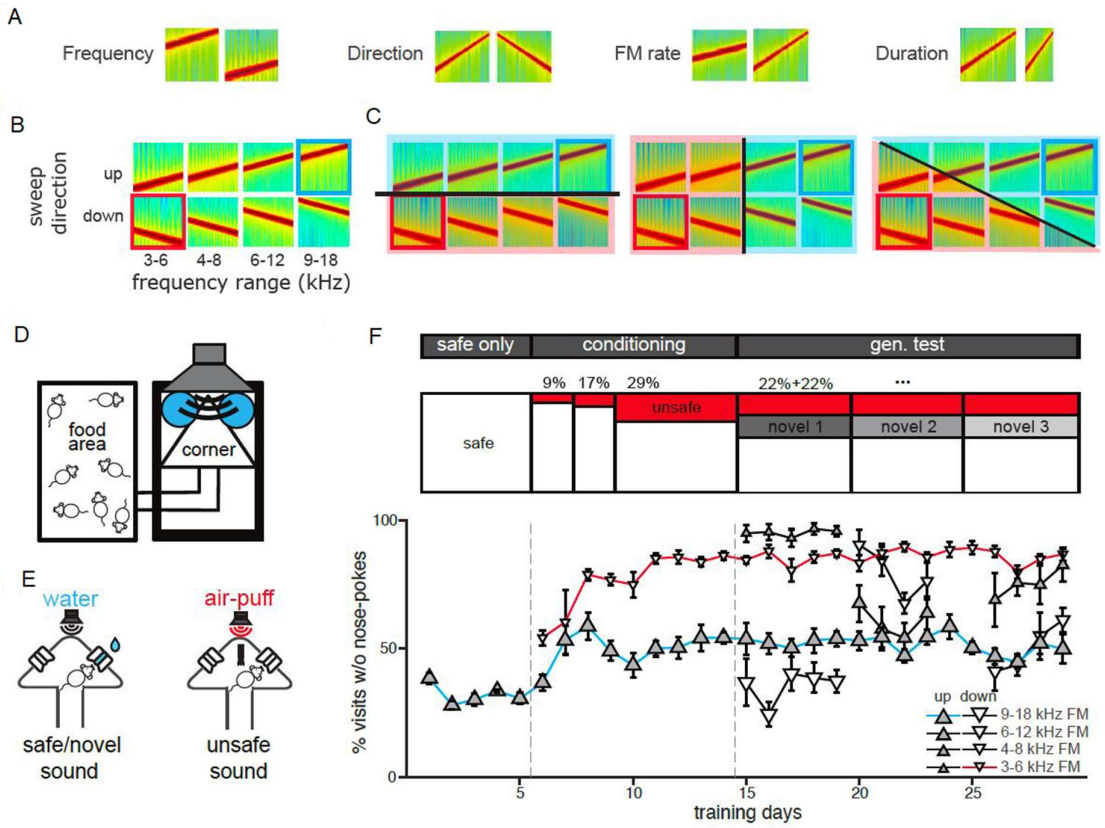


Figure 3.1. Discrimination and generalisation of frequency modulated sweeps (FMs) differing in two dimensions. (A) 4 pairs of spectrograms of FM sweeps differing in frequency range, direction, rate or duration. **(B)** Spectrograms of stimulus sets varying along frequency range and sweep direction. The safe sound was a 9-18 kHz upward FM sweep (blue rectangle) and the unsafe sound was a 6-3 kHz downward FM sweep (red rectangle). **(C)** Possible generalisation gradients: generalisation occurring along sweep direction (left), frequency range (middle) or both dimensions (right). **(D)** Schematic representation of the Audiobox. **(E)** Schema of a safe/novel (left) and unsafe (right) visit. Nose-poking was followed by access to water (safe, left) or an air-puff (conditioning, right). **(F) Top**: Experimental design. The task consists of three phases: the safe only phase with 100% safe visits (white), the conditioning phase with increasing probability of unsafe visits (red), and the generalisation testing phase in which novel sounds were introduced (grey). **Bottom**: Mean daily response for mice trained with the combination of frequency range and sweep directions shown in **(B)** Response was expressed

as the fraction of visits without nose-pokes for different types of stimuli. Error bars represent standard error.

Each dimension pair was tested in a different group of mice. For all 4 dimensions, we chose a span of values known to elicit discriminative responses in neurons in the auditory system. The mice were trained in the Audiobox (Fig. 3.1D), an apparatus in which mice lived in groups for the duration of the experiment (several weeks) while performing the task *ad libitum*. The Audiobox consists of 2 main compartments: a home-cage where food was always available and a 'drinking chamber' where water was delivered upon a nose-poke. As the mouse entered the chamber, a given sound began to play repeatedly for the duration of the visit. In those visits in which the 'safe sound' was presented, mice had access to water (Fig. 3.1E, left), whereas when the 'unsafe sound' was played, nose-poking was followed by an aversive air puff and no access to water (Fig. 3.1E, right). The task, therefore, resembled a Go/No-go paradigm. During the generalisation phase, two novel sounds (appearing in 11% of the chamber visits each) were introduced, and changed every 2 to 4 days (Fig. 3.1F). These sounds behaved like 'safe sounds' in the sense that in the visits in which either of these sounds played, the door giving access to water opened upon a nose-poke (Fig. 3.1E, left. We measured a mouse's response to a given sound as its level of nose-poke avoidance, i.e. the percentage chance that it will not poke its nose in that visit (the 'response').

3.2.1 Behavioural Responses to Frequency Modulated Sweeps

The first group of mice was trained to discriminate sweeps that differed in frequency range and sweep direction (9 to 18 kHz upward vs. 6 to 3 kHz downward sweeps, n = 9; Fig. 3.1B). Each day, discrimination was measured as the difference between the response to the safe sound (Fig. 3.1F, blue line) and that to the unsafe sound (Fig. 3.1F, red line). Mice that showed stable discrimination performance for at least three consecutive days (see Methods), were included in the analysis and then tested with novel sounds varying along both dimensions (Fig. 3.1B). Mice tended to either avoid or approach the novel sounds, reflecting whether they perceived them as similar to the 'unsafe' or the 'safe' sound respectively (Fig. 3.1F, black lines).

The mean response for this first set of sounds (safe, unsafe and novel, Fig. 3.2A, top), what we term the 'generalisation pattern', is shown in both the form of a performance plot (middle) and colour-coded matrix (bottom). We found that whether a novel sound was treated more like the unsafe or safe sound, that is how it was generalised, depended mostly on its frequency (the average correlation coefficient between frequency and behaviour, ρ(freq), was -0.89) and not its sweep direction (ρ(dir) = 0.17). Regardless of the direction, mice approached novel sounds of high frequency ranges (6-12 kHz and 9-18 kHz), and avoided novel sounds of low frequency ranges (3-6 kHz and 4-8 kHz). A two-way ANCOVA revealed a significant main effect of frequency range ($F(1,44) = 117.73$ $p = 0$), and a weak effect of sweep direction ($F(1,44) = 6.85$, $p = 0.01$), but no frequency × direction interaction ($F(1,44) = 2.52$, $p = 0.12$).

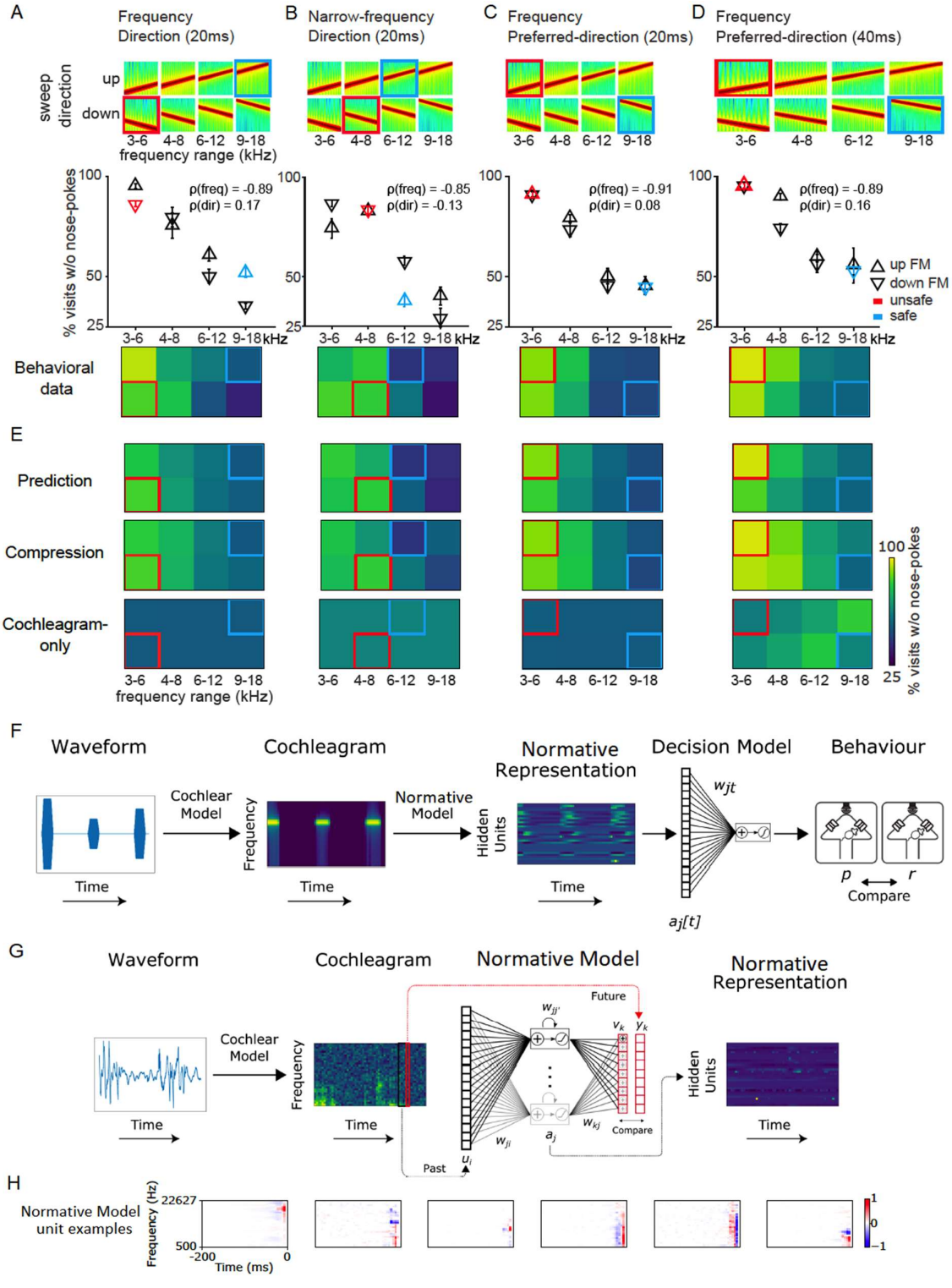


Figure 3.2. The frequency range of a FM sweep dominates over its direction. (A) Top: Spectrograms of the stimulus set used in the task. The animals were trained to discriminate the safe sound (blue

ringed) and the unsafe sound (red ringed) that differed in frequency range and sweep direction. They were then tested on novel stimuli (the other six sounds). **Middle:** Generalisation gradients showing the average response to each stimulus. Error bars represent standard error. **Bottom:** Display of the behavioural data as a heatmap. **(B)** Similar to **(A)** for mice trained to discriminate FM sweeps with partially overlapping frequency range. **(C)** Similar to **(A)** for mice trained to discriminate FMs with frequency-dependent preferred direction. Sound duration used for **(A-C)** was 20 ms, and the rate of frequency modulation was 50 octaves/s. **(D)** Similar to **(C)** for mice trained to discriminate FMs with duration of 40 ms and frequency modulation at 25 octave/s. **(E)** The responses of the models to the safe (blue ringed) and unsafe sounds (red ringed) on which they were trained, and to the novel sounds. Results shown for the prediction (top row), compression (middle row), and cochleagram-only models (bottom row). Columns correspond to the task and behavioural data above in **(A-D)**. **(F)** Schematic of the modelling of the behavioural responses. The model has three stages: 1) The cochlear model converts the sound waveform to a cochleagram, which is a spectro-temporal representation. 2) The normative encoding model processes the cochleagram into a new representation, which reflects the midbrain or cortex. The normative encoding model is pre-trained to predict or compress sounds, using diverse natural sounds for this pre-training. 3) The decision model maps from the normative encoding representation to the no-nose poke probability, which is the behavioural response. The decision model is trained separately for each normative encoding model and for each task to predict the safe and unsafe stimuli. Once trained, the decision model is tested using the novel sounds. (**G)** The normative encoding model used in the prediction model. This normative model embodies temporal prediction and is a recurrent neural network trained to predict the immediate future of natural sound cochleagrams from their recent past. **(H)** Example spectro-temporal receptive fields of prediction model units after being trained with natural sounds. Red – excitatory, blue – inhibitory.

To rule out the possibility that the dominant role of frequency resulted from a more salient difference along the frequency dimension, the frequency dimension was narrowed and mice were trained to discriminate sweeps of partially overlapping frequency ranges but opposite directions (6 to 12 kHz upward vs. 8 to 4 kHz downward sweeps; n = 7; Fig. 3.2B, top). The generalisation pattern revealed a similar pattern as before. Mice discriminated the two sounds well and, when exposed to novel sounds, responded mainly to their frequency range ignoring direction (the average correlation coefficient between frequency and behaviour, ρ(freq), was -0.85; between direction and behaviour, ρ(dir) = -0.13; Fig. 3.2B, bottom). A two-way ANCOVA revealed a significant main effect of frequency range (F(1,36) = 89.31, p = 0), but no effect of sweep direction (F(1,36) = 1.49, p = 0.23)) or frequency x direction interaction (F(1,36) = 1.62, p = 0.21).

Physiological studies have found a systematic representation of preferred sweep direction along the tonotopic axis at both the cortical and subcortical level, such that low-frequency neurons (with a characteristic frequency of < 8 kHz) prefer upward sweeps and high-frequency neurons prefer downward sweeps (Heil et al., 1992a; Hage et al., 2003; Zhang et al., 2003). To further confirm that the observed frequency dominance was not caused by the particular choice of frequency-direction combination, mice were trained with sweeps of two frequency ranges that were modulated in their preferred direction (9 to 18 kHz downward vs. 6 to 3 kHz upward sweeps; n = 9; Fig. 3.2C, top). Our results showed that the mice selectively attended to the frequency range of the novel sounds, and ignored sweep direction (the average correlation coefficient between frequency and behaviour, ρ(freq), is -0.91; between direction and behaviour, ρ(dir) = 0.08; Fig. 3.2C, bottom). A two-way ANCOVA for generalisation patterns revealed a significant main effect of frequency range (F(1,68) =

204.86, p = 0), but no effect of sweep direction (F(1,68) = 1.04, p = 0.31) or frequency x direction interaction (F(1,68) = 0.32, p = 0.57) .

Up to this point we had used sweeps of 20 ms duration, based on discriminable durations in mice (Ehret, 1992). Since previous studies have shown that sweep direction discrimination can be limited by the sweep duration (Nabelek and Hirsh, 1969; Schouten, 1985; Gaese et al., 2006), mice were then trained and tested with 40 ms long sweeps (9 to 18 kHz downward vs. 6 to 3 kHz upward sweeps; n = 7; Fig. 3.2D, top). We found that an increase in sweep duration did not change the generalisation pattern (the average correlation coefficient between frequency and behaviour, ρ(freq), is -0.89; between direction and behaviour, ρ(dir) = 0.16; Fig. 3.2D, bottom). A two-way ANCOVA revealed a significant main effect of frequency range (F(1,52) = 122.37, p = 0), but no effect of sweep direction (F(1,52) = 3.44, p = 0.07) or frequency x direction interaction (F(1,52) = 0.54, p = 0.47) . Taken together, these results (Fig. 3.2A-D) show that sound frequency is perceived as more relevant than sweep direction.

3.2.2 Behaviour Is Consistent with Temporal Prediction

We examined if some general principle could explain how the mice generalise over frequency and sweep direction, or any given set of dimensions. We hypothesised that the generalisation patterns reflect the relevant characteristics of the auditory environments that the animal evolved in and was reared in. However, this begs the question of what defines the relevant characteristics. We considered two alternative hypotheses, that those characteristics of

natural sound that are predictive of future input are important, or alternatively that those characteristics that allow a compressed representation are important.

To this end, we constructed three models: the cochleagram-only model, the prediction model and the compression model – and used them to predict the behavioural responses (Fig. 3.2E). The models consisted of several stages (Fig. 3.2F). The first stage of each model was a model of the cochlea (Rahman et al., 2020) that decomposed the sound waveform into frequency bands. For each task, the sounds were put through this simple model to produce a cochleagram for each sound, the activity in each frequency band over time resembling the activity in the auditory nerve or brainstem (Rahman et al., 2020).

The second stage consisted of an encoding model. Although abstract, this stage of the model embodies the neural processing of the cochlear activity that provides a representation of sound used for decision. Given the temporal resolution of the model (5ms time-bins) is approximately that of the primary auditory cortex (Rees and Møller, 1983; Singer et al., 2018), this is perhaps the most appropriate analogue, but auditory midbrain might also be appropriate. In the prediction model, we used a temporal prediction model (Singer et al., 2018) pre-trained to predict the future of natural sounds from their past. This embodies the hypothesis that the neural representation is optimised for prediction of future natural input from its past. In the compression model, we used an autoencoder (Goodfellow et al., 2016) pre-trained to compress natural sounds efficiently. This embodies the hypothesis that the neural representation is optimised for the efficient compression of the incoming natural sounds. Finally, the cochleagram-only model has no encoding model, instead applying the

decision model directly to the cochleagram representation. This embodies the hypothesis that the neural representation was not optimised to predict or compress the statistics of natural stimuli beyond the frequency decomposition of the cochlea (Lewicki, 2002; Smith and Lewicki, 2006).

More specifically, for the prediction model we used a temporal prediction model based on that in Singer et al. (2018), which has been shown to produce tuning properties of auditory cortical neurons, except that we developed a new version that included recurrent connections. To do this, we trained a recurrent neural network to predict the immediate future (5ms) of cochleagrams of natural sounds from the recent past (20ms), using a large corpus of natural sounds (Fig. 3.2G). The network had a single recurrent hidden layer (1600 hidden units) whose activity over time provides the representational space and was the input to the decision model. The network connection weights were constrained to be sparse. The trained network was able to learn representations that resemble spectro-temporal receptive fields (STRFs), as found in the auditory cortical neurons (Fig. 3.2H). The encoding model for the compression model was similar, but instead of predicting the future, it estimated its own input - the cochleagram in the recent past (20ms). The network weights were constrained to be sparse, as was its activity (making it a sparse recurrent autoencoder). The trained network also learned representations that had some STRF-like characteristics.

The final stage was the decision model, which was used to predict the mean response of the mice for each stimulus of each task (Fig. 3.2G). Here, for each task, the simple decision model was fit to map representations of the safe and unsafe stimuli from the encoding model

activity to the corresponding behavioural responses. We then tested what response the full model would estimate for the novel sounds using the fitted decision model. The decision model was a logistic regression, which mapped from the encoding model activity to a number between 0 and 100%, which was the model estimate of the animal response – chance of giving no-nose-pokes. The regression weights were constrained to be sparse - this drives any regression weights that are not useful to zero, finding a parsimonious mapping. The strength of this constraint for a given task was set as the strength that on-average best predicted the responses to the novel sounds on all the other tasks.

The prediction model qualitatively captured the behavioural data relatively well, which was also the case for the compression model (Fig. 3.2E, top and middle). However, the slight dependence of response on sweep direction that is evident in the behavioural data is more apparent for the prediction model than in the compression model. The cochleagram-only model failed to respond differently for the different auditory stimuli (Fig. 3.2E, bottom), indicating that for this model it is better to give a fixed output than to let the output depend on the cochleagram.

3.2.3 Hierarchy of Stimulus Dimensions Relevant for Behaviour

To test whether and how these results extended to other dimension combinations, we performed the same experiment using sounds that differed in the combination of modulation rate and direction, sweep duration and direction, or sweep duration and frequency-range.

Mice were trained to discriminate sounds that differed in modulation rate and direction, but had the same centre frequency of 8 kHz (100 octave/sec upward sweep as the safe sound vs. 50 octave/sec downward sweep as the unsafe sound; n = 8; Fig. 3.3A, top). We chose 50 and 100 octave/s because these rates are within the range (25-250 octaves/s) that elicit discriminable neural responses (Hage and Ehret, 2003). Generalisation patterns showed that mice responded to novel sounds according mainly to their sweep direction, and largely ignoring modulation rate (the average correlation coefficient between direction and behaviour, ρ(dir), was -0.83; between rate and behaviour, ρ(rate) = -0.33; Fig. 3.3A, bottom). A two-way ANCOVA analysis revealed a significant main effect of sweep direction (F(1,60 = 107.18, p = 0), a significant main effect of modulation rate (F(1,60) = 16.53, p = 0.0001), but no interaction (F(1,60)= 1.09, p = 0.30). This result further supports the idea that sound dimensions interact with each other such that a given dimension (e.g. direction) can have a stronger effect on behaviour than another (e.g. rate), but be ignored when presented with a third (e.g. frequency). This suggests a hierarchy of relevance of sound dimensions in sound discrimination.

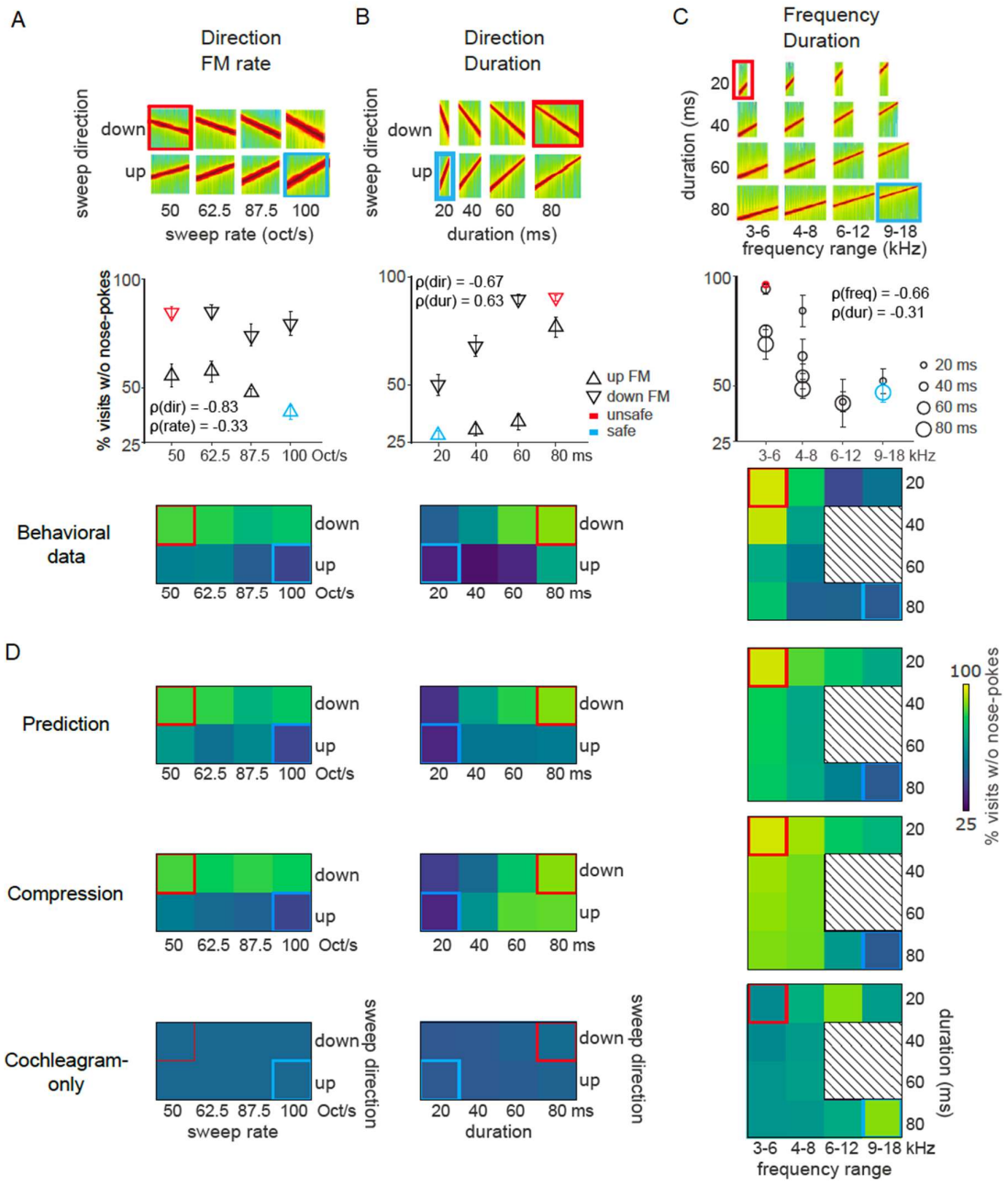


Figure 3.3. Bi-dimensional generalisation reveals hierarchical perception of FM sweep in mice. (A) Top: Spectrograms of the stimulus set used in the task. The safe sound (blue), is a 100 octave/s upward FM sweep, and the unsafe sound (red), is a 50 octave/s downward FM sweep. All FM sweeps were centred at 8 kHz. **Middle:** Generalisation gradients showing the average response to each stimulus. Error bars represent standard error. **Bottom:** Display of the same behavioural data as in **middle**, but as a colour image arranged as in **top**. **(B)** Similar to (**A)** for the combination of sweep direction and

duration dimensions. Average generalisation gradients (middle and bottom) revealed an effect on behaviour of both dimensions. **(C)** Similar to **(A)**, for the combination of frequency range and duration dimensions. Average generalisation gradients (middle and bottom) showed that both dimensions controlled the behaviour of the mice. **(D)** The responses of the prediction (top row), compression (middle row), and cochleagram-only models (bottom row). Columns are different tasks, corresponding to **(A-C).**

A new set of mice were then tested with sounds that combined sweep duration and direction (20 ms 4-16 kHz upward sweeps as safe vs 80 ms 16-4 kHz downward sweeps as unsafe; n = 10; Fig. 3.3B, top). For the sweep duration, we used 20 and 80 ms (Ehret, 1992). Responses changed monotonically as a function of changes in both sound dimensions (the average correlation coefficient between direction and behaviour, ρ(dir), was -0.67; between duration and behaviour, ρ(dur) = 0.63; Fig. 3.3B, bottom). It is worth noting that changing the duration of a sweep of fixed frequency range inevitably also alters the modulation rate. Here, the modulation rates for the safe and the unsafe sound were 100 and 25 Oct/s respectively, i.e. further apart than in the previous test, where they varied between 100 and 50 Oct/s. Since in the previous experiment the role of modulation rate in discrimination was weak, it is unlikely that mice based their choice on the velocity dimension instead of the duration dimension. A two-way ANCOVA revealed a significant main effect of sound duration (F(1,76) = 119.23, p = 0), a significant main effect of sweep direction (F(1,76) = 132.03, p = 0), but no duration x direction interaction (F(1,76) = 0.37, p = 0.55). This result indicates that both dimensions approximately equally contribute to sound discrimination.

We then investigated how mice generalised along the combined dimensions of frequency range and sweep duration (9 to 18 kHz 80 ms sweeps as safe vs. 3 to 6 kHz 20 ms sweeps as unsafe; $n = 8$; Fig. 3.3C, top). We found that both dimensions exerted strong control over the animals' behaviour at the lower frequency ranges (around the unsafe sound) while changes in either dimension were ignored near the safe sound (the average correlation coefficient between frequency and behaviour, ρ(freq), was -0.66; between duration and behaviour, ρ(dur) = -0.31; Fig. 3.3C, bottom). A two-way ANCOVA revealed a significant main effect of frequency range ($F(1,92) = 66.01$, $p = 0$), a significant main effect of duration ($F(1,92) = 14.13$, $p = 0.0003$) and a significant interaction between the two dimensions ($F(1,92) = 8.86$, $p = 0.004$). This suggests that the influence of duration on animals' behaviour is frequency dependent.

Finally, we examined the model predictions for these tasks (Fig. 3.3D). The prediction and compression models both qualitatively captured the behavioural data of direction/rate and direction/duration tasks fairly well (Fig. 3.3D, top and middle). For the frequency/duration task, the prediction model captured the behaviour better than the compression model, as the compression model was not as sensitive when the novel sound became more different to the unsafe sound in frequency interval or duration. The cochleagram-only model performed poorly, as for the frequency/direction tasks, but differentiated response probabilities were observed for stimulus set frequency/duration, although it failed to produce response probabilities similar to the behavioural data (Fig. 3.3D, bottom).

3.2.4 Non-Preferred Stimulus Dimensions

Up to this point we have tested the untrained response to sound stimuli varying in one or both dimensions with respect to either the safe or unsafe sounds. To further characterise the interaction dynamics between dimensions, we first forced mice to discriminate along a non-preferred dimension (e.g. sweep direction, when against frequency range) and then examined generalisation again. In particular, we assessed whether the trained dimension (e.g. sweep direction, non-preferred when combined with frequency) can take over behavioural control in detriment of the other dimension (frequency range, preferred).

A new group of mice was trained to discriminate sweep pairs of opposite direction but same frequency range in oredr to force direction discrimination (Fig. 3.4A, 5-10 kHz upward sweeps as safe vs. 10-5 kHz downward sweeps as unsafe, *n* = 8; Fig. 3.4A, top). We found that discrimination learning was slow, but the safe and unsafe sound were well discriminated after 50 trials of training (Fig. 3.4A, coloured symbols). As mentioned before, direction selectivity is tonotopically organised in the auditory system, such that the physiological preferred direction of neurons of low characteristic frequency is an upward sweep (Heil et al., 1992a; Zhang et al., 2003; Kuo and Wu, 2012). Therefore, we expected direction discrimination to be frequency-specific. Generalisation to novel sounds was later measured with sweeps of successively different frequency range and direction (8-4 kHz downward sweeps and 6-12 kHz upward sweeps). Learning did not generalise across direction at different frequency ranges, as revealed by strong avoidance to both tested sounds despite opposite sweep direction (Fig. 3.4A, black symbols). To strengthen the conditioned-valence of the downward direction, we then conditioned the mice to the already tested downward sweep sound (8-4 kHz downward

sweeps; Fig. 3.4B, coloured symbols) before testing again generalisation on the various novel sounds (Fig. 3.4B, black symbols). This conditioning did succeed in generating direction generalisation across wider frequency ranges (Fig. 3.4B, black symbols). The average correlation coefficient between direction and behaviour was -0.82, whereas the value between frequency and behaviour was 0.12 (Fig. 3.4B). However, generalisation across direction was still limited in frequency range, as reflected by the absence of categorical responses to the sweeps of frequency range beyond the trained ones (3-6 kHz upward and 18-9 kHz downward sweeps; Fig. 3.4B outer most black symbols). A two-way ANCOVA revealed a significant main effect of sweep direction ($F(1,60) = 74.2$, $p = 0$), but no effect of frequency range ($F(1,60) = 1.89$, $p = 0.17$) or interaction between the two dimensions ($F(1,60) = 0.53$, $p = 0.47$). Therefore, generalisation along the non-preferred dimension could be trained but remained specific in frequency, possibly because plasticity occurred only within a specific region of the tonotopic map.

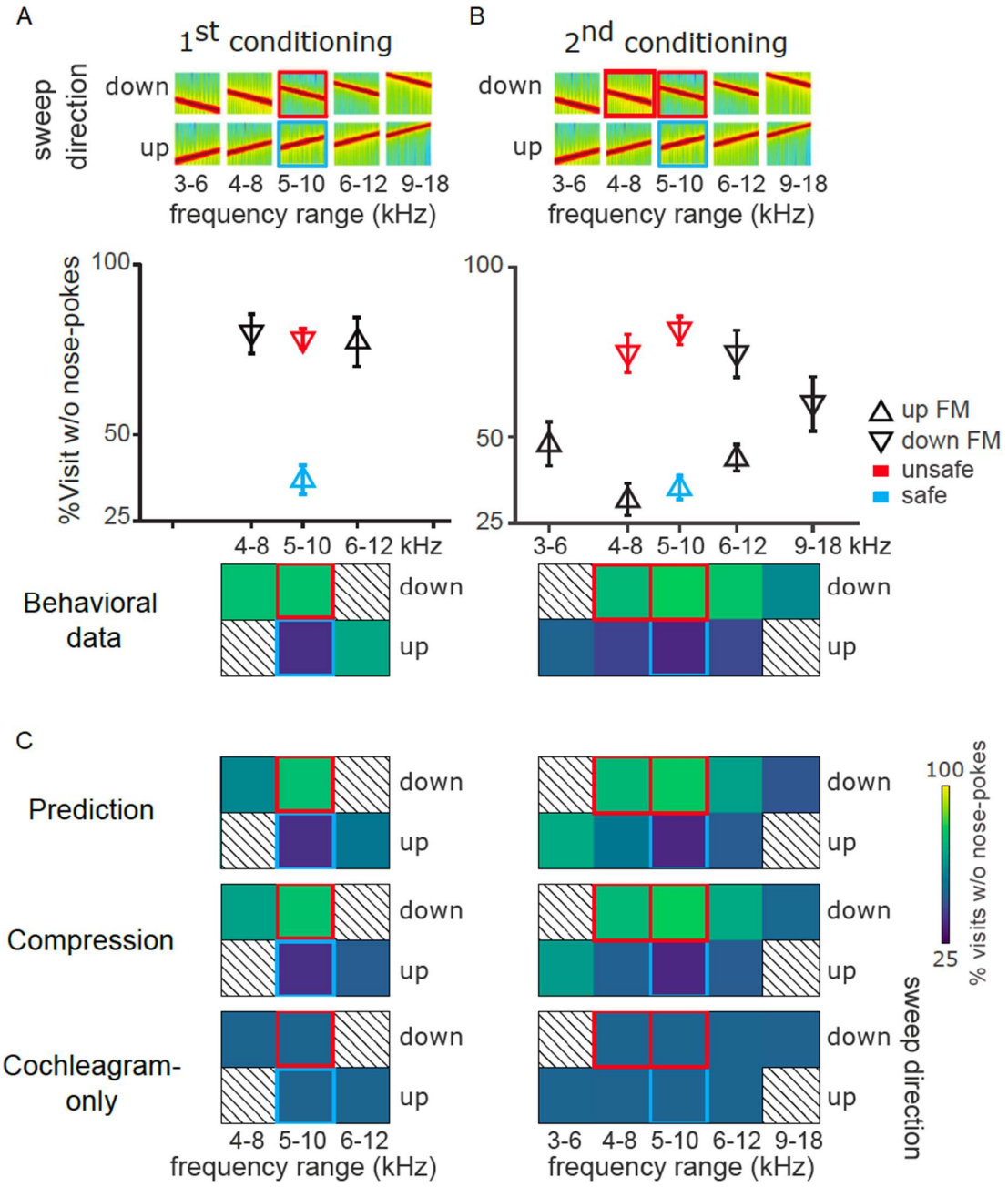


Figure 3.4. Generalisation along the non-preferred dimension was localised. (A) Top, Spectrograms of FM sweeps used in the task. The safe sound (blue) is a 5 to 10 kHz upward FM sweep, and the unsafe sound (red), a 10 to 5 kHz downward FM sweep. **(A) Middle:** Generalisation gradients showing percentage of visits without nose pokes for each sound stimuli. Error bar represents standard error. **(A) Bottom:** Display of generalisation gradients as a colour image. **(B)** Same as **(A)** for generalisation after the 2nd conditioning during which the already tested downward FM sound (8-4 kHz downward

FM sweep) was also conditioned. **(C)** The responses of the prediction (top row), compression (middle row), and cochleagram-only models (bottom row). Columns are different tasks, corresponding to (**A-B).**

We then examined the model predictions for the task. We trained the decision model for the generalisation test after the initial conditioning in the same manner as for the models previously described. For the second generalisation test, we initialised the decision model parameters with the ones learnt from the first learning phase to emulate the chronological order of the learning process, and then trained the decision model to predict the behavioural responses for the two unsafe stimuli and one safe stimulus. The result shows that both the prediction and compression model were able reproduce the lack of generalising to the test stimuli of flanking frequency ranges as observed in the behavioural experiment, whereas the cochleogram-only model could not (Fig. 3.4C).

3.2.5 Plasticity of Stimulus Dimensions

We then tested whether mice that had already shown a preferred dimension, could be trained to discriminate along the other non-preferred dimension. Here, animals were used that had been trained with the combination of frequency and sweep-direction (Fig. 3.2A, repeated in Fig. 3.5A left), where they showed a preference for the frequency dimension, while ignoring direction. These same animals were now given a second conditioning with the same safe sound, but a high-frequency downward-sweep as the unsafe sound to force them to discriminate along the direction dimension (Fig. 3.5A, middle). The second conditioning led to an increase in avoidance to the new unsafe sound (from 35% to 65%), but their

discrimination did not reach levels comparable to those of the first conditioning even after 4 days (Fig. 3.5A right; paired t-test for d', p = 0.002). A similar result was observed in mice previously trained with the combination direction-rate (Fig. 3.5B, left). When the upward 50 octaves/s was explicitly conditioned to force discrimination along the velocity dimension, we observed increased avoidance to the new unsafe sound (from 56% to 69%) but weaker discrimination than in the original conditioning (Fig. 3.5B, middle and right; paired t-test for d', p = 0.003).

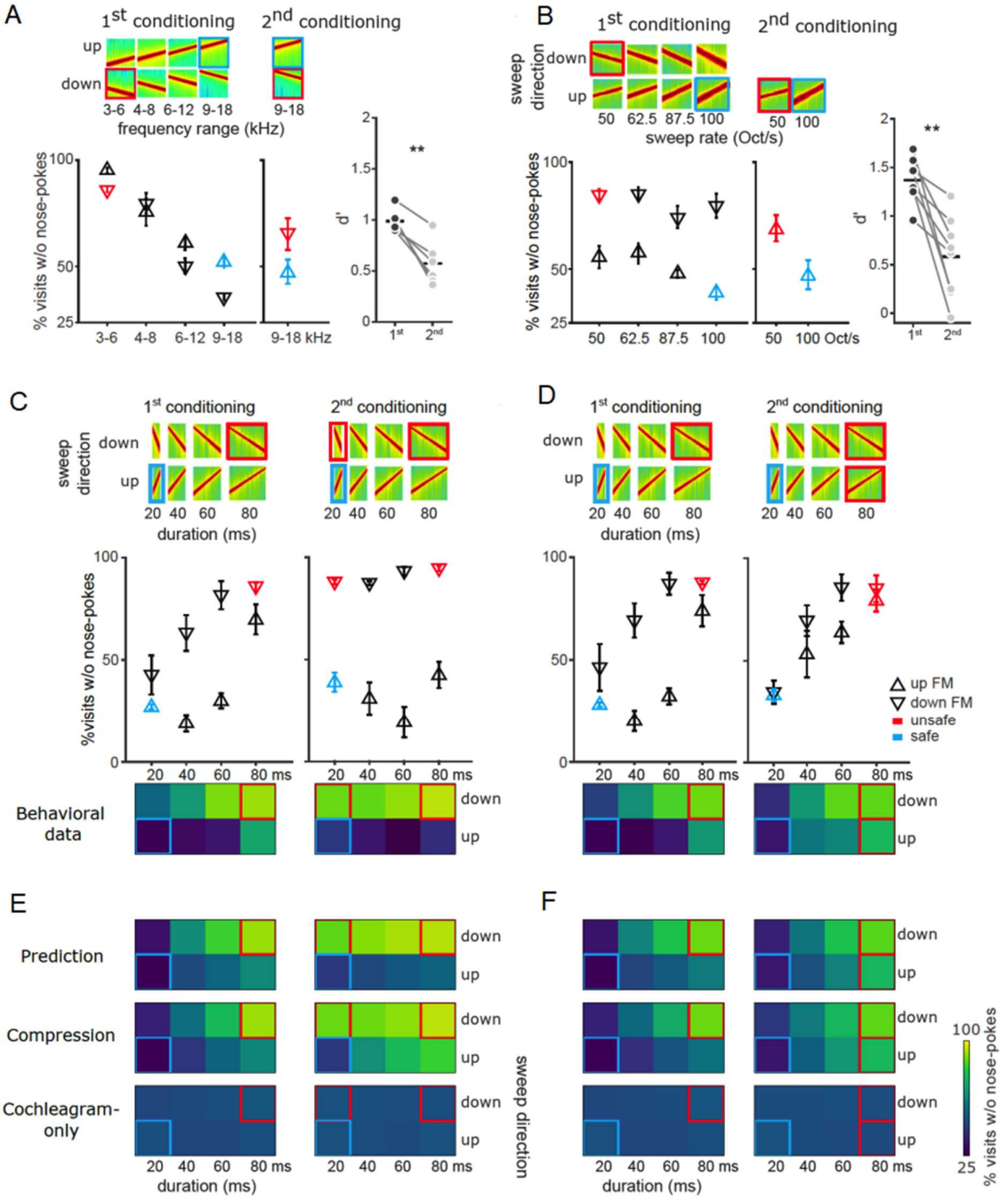


Figure 3.5. Behavioural control of sound dimensions is plastic and amenable to training. (A) Left, Generalisation gradients after initial training (1st conditioning) with safe (blue square) and unsafe (red) sounds combined differences in frequency range and sweep direction. **(A) Middle:** Generalisation gradient after subsequent conditioning to the non-dominant dimension, sweep direction (the 2nd conditioning, half the mice). **(A) Right:** d' for the last 50 unsafe visits of the 1st and the 2nd conditioning for individual mice. **(A)** Initially, mice selectively generalised along frequency range (left), and were

worse at direction discrimination (middle). **(B)** Same as **(A)** for mice initially trained with sweep direction and rate (left), and subsequently trained to discriminate the non-dominant dimension, rate (middle). **(C-D) Left:** Generalisation gradients after the 1st conditioning with direction and duration. Mice initially generalised along both direction and duration dimensions. **(C-D) Right:** Shift in generalisation pattern after the 2nd conditioning to discriminate either direction **(C)** or duration **(D)**. Responses to the safe and unsafe sounds are marked as blue and red respectively. **(E-F)** The responses of the prediction (top row), compression (middle row), and cochleagram-only models (bottom row). Columns are different tasks, corresponding to (**C-D).**

We next examined flexibility in a task in which both tested dimensions influenced behaviour. Mice that had been trained with the combination of direction-duration (Fig. 3.3B), were divided in two equally sized groups of comparable performances. The first half were subsequently trained with an additional unsafe sound (the second unsafe tone) with the same direction as the initial unsafe sound but the shorter duration of the safe sound (Fig. 3.5C, n = 5). Thus, the safe and the unsafe sounds differed now only in sweep direction. After this second conditioning, mice now avoided nose-poking in the presence of the added unsafe tone (Fig. 3.5C, right, coloured symbols; 88% of avoidance, d' = 1.45). Importantly, learning not only affected the newly unsafe sound, but was extended to novel sounds in between the safe and unsafe ones (Fig. 3.5C, right, black symbols). The discrimination gradient was now mainly along the direction dimension. A two-way ANCOVA for performance of the retrained mice revealed a significant main effect of sweep direction (F(1,36) = 264.73, p = 0), but no effect of duration (F(1,36) = 0.58, p = 0.45) or interaction between those two dimensions (F(1,36) = 0.62, p = 0.43).

The other half of the mice (Fig. 3.5D) were exposed to an additional unsafe tone that had the same duration as the first unsafe tone, but the direction of the safe sound. This resulted in safe and unsafe sounds differing only in sound duration (Fig. 3.5D, right, coloured symbols). Interestingly, even though the newly unsafe sound was already treated like an unsafe sound by the animals after the initial conditioning (from 74% to 79%; $d'_{2nd} = 1.24$), responses to novel sounds changed dramatically. Mice now mainly discriminated along the duration dimension (Fig. 3.5D, right, black symbols). A two-way ANCOVA revealed a significant main effect of sweep duration ($F(1,28) = 51.28$, $p = 0$), a weaker but significant effect of direction ($F(1,28) = 6.27$, $p = 0.02$), but no interaction between those two dimensions ($F(3,24) = 0.65$, $p = 0.43$). Thus, changes in task conditions can shift the behavioural pattern from one where both dimensions are used in discrimination to one in which one dimension is ignored. In summary, mice could be trained to flexibly use a single dimension as the behaviourally relevant one, however, discrimination along the non-preferred dimension remained more difficult and did not generalise beyond narrow ranges.

As before, we then examined the model predictions for the task with two unsafe sounds. Here, both the recurrent prediction and compression model were able to reproduce the lack of generalising in the first task (Fig. 3.5E, F, the left column). In the second task, the second unsafe sound encouraged generalisation across duration (Fig. 3.5C, right) or across sweep direction (Fig. 3.5D, right) in the behavioural data. This generalisation is captured by both the prediction and compression models after second conditioning on duration, but for second conditioning on sweep direction the prediction model appears to capture the phenomenon far better than the compression model (Fig. 3.5E, right column). The

cochleagram-only model produces the same uniform response probabilities and does not predict behaviour accurately (Fig. 3.5E, F, bottom row).

3.2.6 Perception of Envelope Features

The 'envelope' of the sound stream is important for auditory perception, for example in speech processing (Rosen, 1992; Hickok and Poeppel, 2007). The experiments described so far focused on differences within the short frequency-modulation bouts presented always at 3 Hz (Fig. 3.1-5). Next, we investigated how the 'envelope', i.e. the contour, influenced generalisation patterns. Since amplitude modulations and the periodic envelope are important characteristics for acoustic communication signals (Leaver and Rauschecker, 2010), we trained and tested mice with 100% sinusoidal amplitude-modulated (AM) stimuli (differing in dimensions of carrier frequency and modulation rate, Fig. 3.6A).

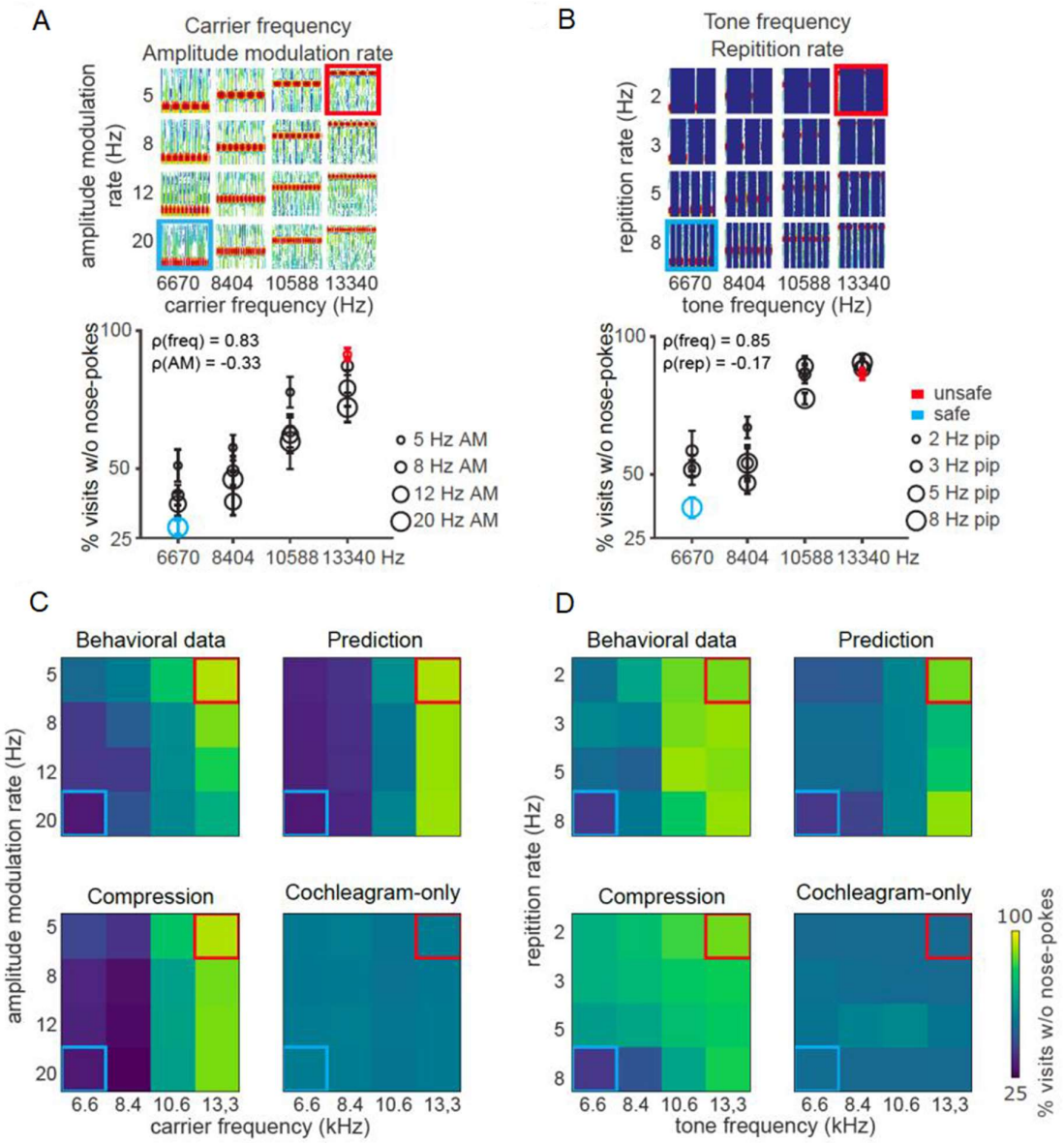


Figure 3.6. Bidirectional generalisation of periodic sound. (A) Top: Spectrograms of FM sweeps used in the task. **(A) Bottom:** Generalisation gradients plotting the response proportion to each of the stimuli presented in the task for the combination of tone frequency and AM sweep rate dimensions. **(B)** Same as **(A)** for the combination of tone frequency and repetition rate dimensions. For each figure, the safe and unsafe sounds are marked in blue and red, respectively. **(C-D)** Comparison between behavioural data and responses by the prediction, the compression and the cochleagram-only models for **(A-B)** respectively.

We chose well discriminable carrier frequencies that differed by 1 octave (Chen et al., 2019) and amplitude-modulations ranging between 5 and 20 Hz modulation rate (Kurt and Ehret, 2010) (ΔF > 1 octave). Fig. 3.6A shows responses to trained and novel sounds. We found that, for all mice, performance was mainly controlled by the carrier frequency and influenced by the amplitude modulation (the average correlation coefficient between frequency and behaviour, ρ(freq), was 0.83; between amplitude-modulation rate and behaviour, ρ(AM) = -0.33). A two-way ANCOVA revealed a significant main effect of carrier frequency ($F(1,124) = 168.76$, $p = 0$), a significant main effect of modulation rate ($F(1,124) = 26.38$, $p = 0$), but no interaction between the two dimensions ($F(1,124) = 0$, $p = 0.98$).

We then tested sounds differing in dimensions of tone frequency and repetition rate (Fig. 3.6B), using rates ranging between 2 and 8 Hz (ΔF of 2 octaves), since neurons in auditory cortex selectively prefer stimulus rates ranging from 2 to 15 Hz (Kilgard and Merzenich, 1999). Similar to amplitude-modulated stimulus, we found that the animals' performance was influenced by both dimensions (the average correlation coefficient between frequency and behaviour, ρ(freq), was 0.85; between repetition rate and behaviour, ρ(rep) = -0.17). A two-way ANCOVA revealed a significant main effect of tone frequency ($F(1,92) = 189.3$, $p = 0$), a significant effect of repetition rate ($F(1,92) = 6.72$, $p = 0.01$) and significant interaction between the two dimensions ($F(1,92) = 4.36$, $p = 0.04$). Thus, although tone frequency exerts a stronger control over behaviour, repetition rate influences in a frequency-dependent manner.

Next, we examined the model predictions for these tasks. The prediction and compression models are both able to qualitatively capture the responses fairly well (Fig. 3.6C, D). For stimulus set varying along carrier frequency and amplitude-modulation rate, they both tend to overly disregard the modulation rate as a variable for the behavioural outcome, which might be due to limitations of the model architectures for capturing modulation frequencies at a low rate. For the task involving tone frequency and repetition rate, the compression model tends to somewhat overestimate the fall-off with lower carrier frequency of the response to the novel sounds, while the prediction model underestimates them. The cochleagram-only model predicts approximately the same response for all sounds in both tasks, and thus fails to mimic the structure seen in the behavioural data (Fig. 3.6C, D).

3.2.7 Summary and Quantitative Comparison of Models

For an overview of the effect of the different dimensions on performance, we collated across tasks the correlation coefficients between each dimension tested and performance (Fig. 3.7A). The variability across individual animals was small, suggesting that the dimensional relevance reflects an innate neural representation in the brain. While frequency consistently exerted considerable control over the behaviour across all tests, direction and duration were more variable and their contribution to behavioural decisions depended on the second sound dimension they were tested with. Temporal features such as rate of frequency modulation, amplitude modulation, or repetition rate, had less control over behaviour in this type of go/no-go discrimination. Thus, the different features tested here, and in the ranges used, differed in their influence over the behaviour in a way that was dependent on the other

dimensions of the tested sound, revealing a hierarchy of influence dominated by frequency (Fig. 3.7B).

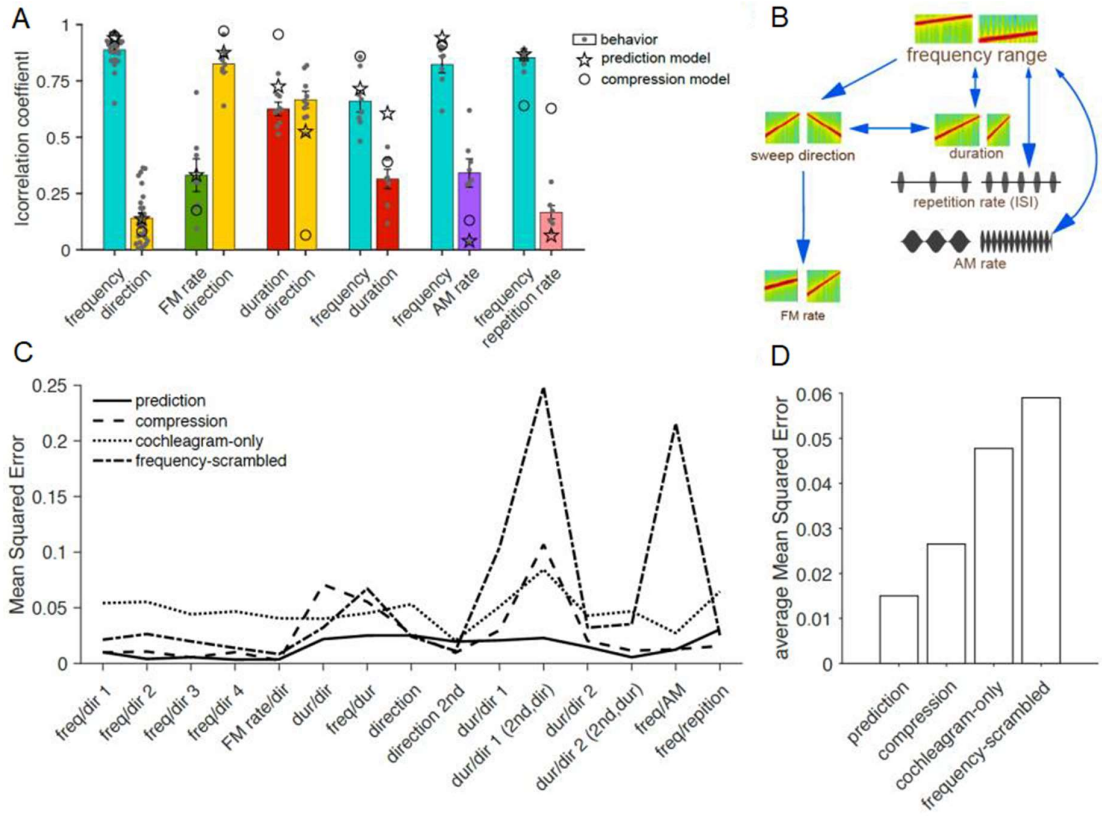


Figure 3.7. Summary and Quantitative comparison of models. (A) Bar plot shows the absolute value of the correlation coefficient between each dimension's tested variable space and performance of each mouse. Star and circle represent the absolute value of correlation coefficient between dimensions and prediction by prediction model and compression model, respectively. Error bars represent standard error. **(B)** Scheme shows the perceptual hierarchy inferred by the results. **(C)** For each task, the mean squared error between the mouse responses to the novel sounds and the responses of the prediction, compression, cochleagram-only models, and prediction model trained with scrambled frequency channels. **(D)** For each model type, the mean squared error from **(C)** averaged over all 15 tasks.

This hypothesis is also supported by the result of the prediction and compression models when we calculated the correlation coefficient values for the model results in the same way as we do for the data (Fig. 3.7A). The prediction model was most similar to the data. We also quantified the results of each model more directly for each task by measuring the mean squared error between the model responses for the novel sounds and the true response of the mice. In Fig. 3.7C, we plot the mean squared error for each task for the prediction model, the compression model and the cochleagram-only model. The cochleagram-only model performs worst in almost all cases indicating that the auditory-nerve level representation is not the space in which sound classes are discriminated. The compression and the prediction models both perform better than the cochleagram-only model, with the prediction model appearing to perform the best overall, performing better than the compression model for four tasks (narrow-frequency/direction, frequency/direction 40ms, frequency/duration and duration/direction 2nd), worse in two tasks (direction 2nd and frequency/repetition rate) and very similarly in the other nine tasks.

Figure 3.7C also displays another variation on the prediction model, the scrambled-frequency-channel prediction model. Here, we trained the encoding model using data where the cochleagram-activity was scrambled over frequency channels. This eliminated much of the natural statistical structure of the stimulus, embodying the hypothesis that natural statistics are unimportant for predicting the responses of the mice. The frequency scrambled model performed worse than the prediction and compression models, and also worse than the cochleagram-only model. While the temporal information has been retained, the frequency channels are scrambled, and the model learns representations that fail to capture the perceptual space that mice use. This provides further evidence that the generalisation

patterns of the mice, and the hierarchy of importance of sound dimensions in mouse perception, are a consequence of the encoding representation being optimised for representing natural stimuli.

In order to estimate the overall behavioural prediction capacity of each model type, we averaged the mean squared error of all predicted test sound responses for each model type (Fig. 3.7D). Overall, the prediction model does a better job at predicting the test responses (mean squared error: 0.015) compared to the compression model (mean squared error: 0.027), the frequency-scrambled model (mean squared error: 0.059) and the cochleagram-only model (mean squared error: 0.048).

3.3 Discussion

We set out to understand how sound dimensions are integrated in auditory perception and to what extent their differential control over behaviour can be explained by their relative relevance in natural sounds. The results have consequences for our understanding of perception and the underlying circuits. Using an automatic behavioural paradigm in mice, with a design reminiscent of the implicit bias test used in humans (Greenwald et al., 1998), we tested the natural perception of combinations of sound dimensions. Mice were trained to distinguish between safe and unsafe sounds that differed in two dimensions, for example frequency and sweep direction, and then tested whether they categorised the safety of novel sounds according to one or both dimensions. A complex pattern of behavioural generalisation over different tasks emerged. This revealed that the perception of acoustic dimensions is hierarchically organised and that this hierarchy determines how acoustic dimensions interact

with each other. Furthermore, our modelling suggests that sounds are judged to be different or similar within a representational space optimised for the prediction of the immediate future of natural sounds from their past.

3.3.1 Hierarchical Organisation of Acoustic Dimension Processing

Mice categorised novel sounds as 'unsafe' or 'safe' often on the basis of one of the two tested dimensions and ignored the other dimension. This suggests a hierarchy of relevance of the dimensions examined, such that frequency dominates behavioural decisions, followed by frequency sweep direction and duration in equal measure, while sweep rate was largely ignored (Fig. 3.7B). Mice can, however, be actively trained to discriminate along their non-preferred dimension, for example sweep direction.

Our findings are consistent with other behavioural studies that did not address the dimension question specifically. For example, rats discriminate along the frequency dimension more easily than along amplitude modulation rate (Kurt and Ehret, 2010) or loudness (Polley, 2006). In mice, maternal responsiveness to wiggling calls depends critically on both frequency and duration (Ehret and Riecke, 2002; Geissler and Ehret, 2002), which matches our inference that frequency range and duration were at high levels of the perceptual hierarchy.

3.3.2 Temporal Prediction and the Hierarchy of Stimulus Dimensions

We asked what determines the hierarchy of stimulus dimensions we observe - why do we observe the particular hierarchy that we do, and not, for example, sweep rate being more salient than sound frequency? We hypothesised that the hierarchy is a consequence of the statistics of the natural sounds experienced by an animal during its lifetime and evolution, which would be reflected in the neural representation in the brain (likely forebrain or midbrain). We hypothesised that it is in this representational space that sounds are categorised; for example, in this space, frequency differences could be more relevant than sweep-direction differences, and these in turn could be more relevant than sweep-rate differences.

Within this perspective, there are many possible frameworks which could determine the representational space learnt from natural sounds. One hypothesis is that the brain seeks to find a representation that compresses the incoming natural sounds, removing redundant information (Attneave, 1954; Barlow, 1959; Olshausen and Field, 1996, 1997; Rao and Ballard, 1999; Carlin and Elhilali, 2013). Another hypothesis is that the brain seeks to find a representation that enables the efficient prediction of immediate future natural sound input from recent past sound input (Bialek et al., 2001; Salisbury and Palmer, 2016; Chalk et al., 2018; Singer et al., 2018). To this end we built encoding models that learnt representations of natural sounds optimised for either hypothesis, compression or prediction. We then trained a decision model to estimate the responses of the mice to the safe and unsafe sounds, given the representations provided by these encoding models, and examined the responses

predicted for the novel sounds on which they had not been trained. The model predictions of the responses to the novel sounds were remarkably consistent with those measured in the experiment. We did not see this consistency when the decision model was applied directly on the cochlear representation without an intervening encoding model, or when the encoding model was trained on non-natural sounds. This suggests that the representational space within which the mice categorise sounds is optimised for natural stimuli.

Although both the compression model and the prediction model typically perform well, overall the prediction model performs slightly better, with the compression model performing poorly for some tasks. However, it should be born in mind that artificial neural network model solutions are somewhat sensitive to initial parameter values. To determine the robustness of our findings, it would be good in further work to optimise the same models multiple times with different initial random parameters settings and examine their consistency. With this caveat in mind, our results so far suggest that the representational space within which sound objects are judged to be similar or different may be a consequence of a neural representation optimised to represent the predictive features of natural sounds. For example, this suggests that sound frequency dominates over sweep direction in discrimination as past sound frequencies tend to be more informative about future sounds than the direction of frequency sweeps. Similarly, sweep direction is more relevant for discrimination than sweep modulation rate perhaps because the direction tends to be more informative for prediction than modulation rate. Hence, usefulness for prediction may at least partially determine the basis of sound categories.

Here, we have compared two unsupervised learning principles by which the brain may learn representations to guide behaviour; compression and prediction. The idea that the brain seeks to efficiently compress incoming stimuli has a long pedigree, and includes ideas such efficient coding and redundancy reduction (Attneave, 1954; Barlow, 1959), sparse coding (Olshausen and Field, 1997), and many versions of predictive coding (Rao and Ballard, 1999). The idea that brain seeks to predict future input, encoding only input which is temporally predictive rather than compressing all input, has been somewhat less explored, and includes ideas such as predictive information (Bialek et al., 2001; Salisbury and Palmer, 2016; Chalk et al., 2018) and temporal prediction (Singer et al., 2018, 2019) and also relates to slowness ideas (Stone and Harper, 1999; Berkes and Wiskott, 2005). Compression and prediction have typically been used to explain neural responses, rather than animal behaviour as we have done. Optimising models for compression has had some success in explaining certain spatial tuning properties in the visual cortex (Olshausen and Field, 1996, 1997; Rao and Ballard, 1999), as has prediction of future input (Singer et al., 2018, 2019). However, while both compression (Klein et al., 2003; Cusack and Carlyon, 2004; Zhao and Zhaoping, 2011; Carlson et al., 2012; Kozlov and Gentner, 2016) and prediction (Singer et al., 2018) have also had some success in explaining tuning aspects of neurons in the auditory midbrain or auditory cortex, it has been argued that prediction models perform somewhat better, particularly for temporal tuning properties (Singer et al., 2018).

We argue that the temporal prediction model performs well due to being optimised for natural stimuli for an appropriate normative principle. However, while this may be the normative reason, this improved performance will instantiate through some learned network. Hence, it is interesting to speculate on the proximal reason that the different models might

have performed differently on the basis of their receptive fields and network structure. The receptive fields for the prediction model were more A1-like, being a range of moderate breadths and lengths with a diversity of well-defined shapes. In contrast, the receptive fields for the compression model were more broad, ill-defined and speckly (Fig. 4.2, 4.3 and 4.5). We speculate that the prediction model best explained the behavioural responses to the novel sounds because its receptive fields varied the appropriate amount across the relevant dimensions, both within receptive fields and across units. For example, the receptive fields of the prediction model may differ more over sound frequency than they do over sweep direction to the degree most consistent with behaviour, and hence explain the pattern of behavioural response better than the other models.

The structure of the recurrency in the prediction normative models might also provide an advantage; despite having an input window of only 20ms, the recurrency allows the use of inputs from much further into the past to predict the immediate future. Hence, this might explain why there is some dependence of the behavioural response on repetition rate for the prediction and compression models (Fig. 3.6D). The recurrent weight matrix of the compression model is substantially more sparse than for the prediction model, which implies that the compression model does not make as much use of its recurrency to learn longer temporal dependencies (Fig. 4.7). The prediction model trained on frequency-scrambled sound performs the worst overall (Fig. 3.7), perhaps due to overly broad and otherwise inappropriate receptive fields, and the cochleagram-only model may have not performed well by not having the advantage of pre-processing by appropriately sized and shaped receptive fields to enable good generalization.

Temporal prediction and compression are not mutually exclusive, and indeed could be combined (e.g. PredNet (Lotter et al., 2020)) or perhaps differentially applied in different sub-systems. Furthermore, while it has been argued that unsupervised processes are important in guiding sensory representation, reinforcement signals may also influence them, rather than just affecting the decision processes as we assume in our models. Hence, while we have demonstrated a method to ascertain the models and principles governing the representations that guide sensory categorisation, and have found a prediction models to perform best, there are of course - as always - more prediction, compression and other models that could be explored.

3.3.3 Neurophysiological Considerations of Dimension Perception

Is the hierarchical relevance of acoustic features reflected in the functional and anatomical organisation of the auditory system? The lack of hierarchy in our encoding models suggests that an anatomical hierarchy is not required. The notion that sensory systems might be organised hierarchically, i.e. into a series of stages that progressively increase abstract representations, has been a popular hypothesis for decades (Chechik et al., 2006). Hierarchical processing of sensory information has been intensively studied in the visual modality (Hochstein and Ahissar, 2002; Grill-Spector and Malach, 2004; Hegdé, 2008; Markov and Kennedy, 2013). The influential reverse hierarchy theory proposed by Ahissar and Hochstein (Ahissar and Hochstein, 1997) posits that high-level abstract inputs are prioritised for perception along the sensory hierarchy and that this hierarchy maps onto an anatomical one. The acoustic dimensions used in this study do not map well onto an anatomical hierarchy

of sequential processing across different auditory stations. Neural selectivity to frequency is observed across all auditory stations, from the cochlea to the cortex, and is organised tonotopically (Brugge, 1992). Neural selectivity for sound duration, sweep direction, sweep rate, repetition rate, and modulation rate all appear as early as the auditory (Brand et al., 2000; Langner et al., 2002; Hage and Ehret, 2003; Kuo and Wu, 2012; Clayton et al., 2021). Thus, the observed hierarchical organisation for the dimensions used in this study may not result from sequential feature extraction along different auditory stations, but from the dynamic representation of each auditory feature in an early station, probably the auditory midbrain or cortex.

The distribution of response selectivity within stations in the auditory system might nonetheless influence the functional hierarchy we observed. For example, while almost all neurons show frequency selectivity, direction or modulation rate selectivity have been found in subpopulations of neurons in auditory cortex and midbrain (Heil et al., 1992a; Hage and Ehret, 2003). Both these dimensions are in fact somewhat embedded in the tonotopic map (Heil et al., 1992a; Zhang et al., 2003; Kuo and Wu, 2012). It is therefore plausible that a given station in the auditory system processes the direction and rate dimensions of a specific sound, albeit by neurons that are selective for a specific range of frequencies. This is consistent with our finding that the learning of non-preferred dimension discrimination is frequency-dependent, i.e. direction discrimination does not generalise well outside the trained frequency range.

To our knowledge, our experiment constitutes the first successful use of multidimensional generalisation to infer untrained dimensional integration and the hierarchy of auditory information processing. These results yield important insights into the processes underlying auditory object processing and categorisation. Tapping into the untrained behavioural response of mice to sounds combining different acoustic dimensions, we found that these dimensions are hierarchically organised in the control they exert over behaviour. This emergent hierarchy is consistent with the predictive value of these dimensions in natural sounds as demonstrated by models. Overall, and inferring beyond the auditory modality, the data support the more general idea that both the sensitivity of physiological responses and behavioural decisions to natural stimuli have evolved to optimally use the predictive value of the different stimulus dimensions in the natural environment.

3.4 Materials and Methods

3.4.1 Experiments

3.4.1.1 Animals

A total of 103 female C57BL/6JOlaHsd (Janvier, France) mice were used. All mice were 5-6 weeks old at the beginning of the experiment. Animals were housed in groups in a temperature-controlled environment (21 ± 1°C) on a 12 h light/dark schedule (7am/7pm) with access to food and water *ad libitum*.

All animal experiments were approved by the local Animal Care and Use Committee (LAVES, Niedersächsisches Landesamt für Verbraucherschutz und Lebensmittelsicherheit,

Oldenburg, Germany) in accordance with the German Animal Protection Law. Project license number 33.14-42502-04-10/0288 and 33.19-42502-04-11/0658

3.4.1.2 Apparatus: the Audiobox

The Audiobox was a device developed for auditory research from the Intellicage (TSE, Germany). The Audiobox served both as living quarters for the mice and as their testing arena. Each animal was individually identifiable through an implanted transponder, allowing the automatic measure of specific behaviours of individual animals. The mice lived in the apparatus in groups of 7 to 10 for several days. Mice were supplied water and food *ad libitum* but are required to discriminate and respond to sounds when they go to the drinking 'chamber'.

At least one day before experimentation, each mouse was lightly anaesthetised with Avertin i.p. (0.1ml/10g) or isoflurane and a sterile transponder (PeddyMark, 12 mm × 2 mm or 8 mm × 1.4 mm ISO microchips, 0.1 gr in weight, UK) was implanted subcutaneously in the upper back. Histoacryl (B. Braun) was used to close the small hole left on the skin by the transponder injection.

The Audiobox was placed in a temperature regulated room under a 12 h dark/light schedule. The apparatus consisted of three parts: a home cage, a drinking 'chamber', and a long connecting corridor. The home cage served as the living quarter, where the mice have access to food *ad libitum*. Water was delivered in the drinking 'chamber', which was positioned inside a sound-attenuated box. The presence of the mouse in the 'chamber', a

'visit', was detected by an antenna located at the entrance of the chamber. The antenna read the unique transponder carried by each mouse as it enters the chamber. The mouse identification was then used to select the correct acoustic stimulus. A heat sensor within the chamber sensed the continued presence of the mouse. Once in the 'chamber', specific behaviours (nose-poking and licking) could be detected through other infrared sensors. All behavioural data was logged for each mouse individually. Access to water was controlled by opening or closing of the doors behind the nose-poking ports. Air puffs were delivered through an automated valve which is placed on the ceiling of the 'chamber'. A loudspeaker (22TAF/G, Seas Prestige) was positioned above the 'chamber', for the presentation of the sound stimuli. During experimentation, cages and apparatus were cleaned once a week by the experimenter.

3.4.1.3 Sounds

Sounds were generated using MATLAB (Mathworks) at a sampling rate of 48 kHz and written into computer files. Intensities were calibrated for frequencies between 1 and 18 kHz with a Brüel & Kjær (4939 ¼" free field) microphone.

Mice were trained with pairs of either frequency-modulated (FM) sweeps, amplitude-modulation (AM) sounds, or pure tone pips. For 2-dimensional discrimination task, sound pairs used in training as safe and conditioned differed in two out of four chosen dimensions. For the FM sweep sounds these dimensions were frequency range, duration, sweep direction of modulation and velocity of the sweep. For example, when the two dimensions used were the frequency range and the sweep direction, the safe sound could be an upward sweep in

the high frequency range while the unsafe sound would be a downward sweep in the low frequency range. Frequency was modulated exponentially from low to high frequencies (upward sweep) or from high to low frequencies (downward sweep). Sweeps had a duration of 20 ms (default) or 40 ms, including 5 ms rise/decay, and one of four modulation velocities (50, 62.5, 87.5 or 100 octave/sec; with 50 octaves/s as default). The sweeps were presented at roving-intensities (70 dB SPL ± 3 dB SPL). For the AM stimuli, tested dimensions were carrier frequency and modulation rate. The AM sounds had 100% sinusoidal modulation and had one of four carrier frequencies (6670, 8404, 10588 or 13340 Hz), as well as one of four modulation rates (5, 8, 12 or 20 Hz). For pure tone pips, the dimension of tone frequency and repetition rate were tested. Similar to the AM stimuli, the pure tones pips had one of four frequencies (6670, 8404, 10588 or 13340 Hz) and one of four repetition rates (2, 3, 5 or 8 Hz). Pure tones had a length of 20ms, with a 5ms rise/fall ramp and were presented with intensities that roved between 67 and 73 dB SPL. For single dimensional discrimination, sounds pairs differed in sweep direction, duration or frequency were used for training.

Our choice of sounds was based on the existing literature. Rodents perceive differences in both sound frequency and sweep direction (Wetzel et al., 1998; Hage and Ehret, 2003; Zhang et al., 2003; Mercado et al., 2005). Mice show increased sensitivity for frequencies between 4 and 16 kHz (Zhang et al., 2003), and easily discriminate frequency distances of at least 0.20 octave (de Hoz and Nelken, 2014; Chen et al., 2019), therefore, we tested frequency ranges between 3 and 18 kHz such that the centre frequency differed by more than 0.60 octave. The choice of a 50 octave/s velocity of modulation was based on both physiological (Zhang et al., 2003) and behavioural (Gaese et al., 2006) data.

The frequency-only discrimination data were from mice trained with upward frequency sweeps of the same range used in the task with 2 dimensions (3 to 6 kHz vs. 9 to 18 kHz). For direction discrimination, the mice described in Fig. 3.4A were used. These are mice that were trained to discriminate between sweeps of identical frequency range (5-10 kHz), but opposite direction of modulation. Data on duration-only discrimination were obtained from mice trained to discriminate FM upward sweeps (4 to 16kHz) that differed only in duration (20ms vs. 80ms).

3.4.1.4 Discrimination Task

As described in a previous study (Chen et al., 2019), mice were trained to discriminate two sounds, the 'safe' and 'unsafe' sound, in the Audiobox. Throughout the duration of the experiment, one sound (i.e. 9 to 18 kHz FMs) was always 'safe', meaning that access to water during these visits was granted upon nose-poke. For the first 5 days, only the safe tone was played in each visit. The doors giving access to the water within the chamber were open on the first day of training and closed thereafter. A nose-poke from the mouse opened the door and allowed access to water. Another sound, an 'unsafe' one (i.e. 6 to 3 kHz FMs), was introduced in a small percentage of 'unsafe visits' and a nose-poke during these visits was associated with an air puff and no access to water. Most of the visits, however, remained safe (safe sound and access to water) as before. The probability of unsafe visits was 9.1% for the first 2 days, increased to 16.7% for the next 2 days, then stayed at 28.6% until they showed steady discrimination performance for at least 3 consecutive days (Fig. 3.1F).

Mice that failed to learn the task, i.e. had no differential responses to the safe and the unsafe tone for more than 2 days, were excluded from the analysis. In total, 12 out of 105 mice were excluded.

3.4.1.5 Multidimensional Generalisation Measurement

During generalisation testing, visits consisted of 55.6% safe visits, 22.2% conditioned visits and 22.2% novel visits, in which a novel sound was presented. Nose poking during the presentation of the novel tone resulted in opening of the doors that gave access to water, meaning that novel visits were never accompanied by an air-puff, meaning they were safe in nature.

Since the safe and unsafe sound differed in two dimensions, novel sounds represented all possible combinations of values used (see Sounds) along each dimension. For example, when using the 9 kHz to 18 kHz upward FM sound as the safe sound and the 6 kHz to 3 kHz downward FM sound as the unsafe sound, tested stimuli (including the safe and unsafe sounds) resulted from factorially combining 4 different frequency ranges with 2 different sweep directions.

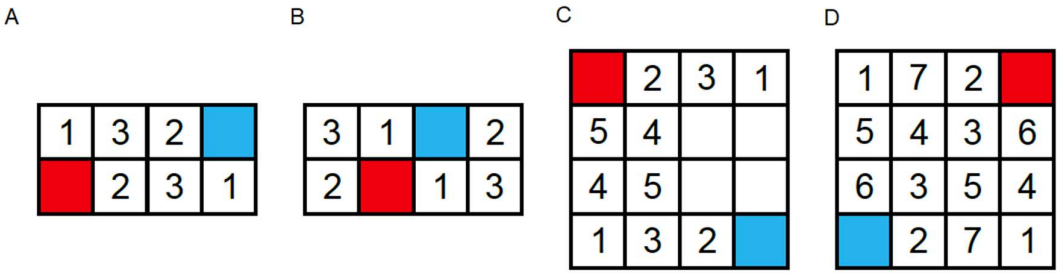


Figure 3.8. Order of presented novel stimuli. (A-B) For tasks using 2 dimensions with 2 (e.g. direction) and 4 (e.g. frequency range) values respectively. The order in which novel sounds varying in either or both dimensions were tested was as shown. **(C-D)** For tasks testing 2 dimensions with 4 values each, such as the combination of frequency and duration, novel sounds were tested in the order shown.

On each testing day, only two novel sounds were presented (in 11.1% of total visits each) in addition to the safe and unsafe sounds. Each novel sound was tested for 4 days (~50 visits). Novel stimuli were presented in the order presented in Fig. 3.8.

3.4.1.6 Analysis of Performance in the Audiobox

Data were analysed using in-house scripts developed in MATLAB (Mathworks). Performance traces for different stimuli were calculated by averaging the fraction of visits without nose-pokes over a 24-hour window. Discrimination performance was quantified by the standard measures from signal detection theory, the discriminability (d'). It was calculated with the assumption that the decision variables for the safe and unsafe tone have a Gaussian distribution around their corresponding means and have comparable variances. The d' value provides the standardised separation between the mean of the signal present distribution and the signal absent distribution. It is calculated as:

$$d' = Z(HR) - Z(FAR)$$

where $Z(p)$, $p \in [0,1]$, is the inverse of the cumulative Gaussian distribution, HR is the hit rate, where a hit is the correct avoidance of a nose-poke in a conditioned visit, and FAR is the false alarm rate, where a false alarm is the avoidance of a nose-poke in a safe visit. Since d'

cannot be calculated when either the false alarms reach levels of 100% or 0%, we used 95% and 5%, respectively, in the few cases where this happened for these calculations. This manipulation reduced d' slightly, and therefore our d' estimates are conservative.

3.4.1.7 Statistical Analysis

Group comparisons were made using multiple-way ANCOVAs with type I sum of squares after testing for normality distribution using the Shapiro-Wilk test. Samples that failed the normality test were compared using Wilcoxon signed rank test. Multiple comparisons were adjusted by Bonferroni correction. For analysis of data consisting of two groups we used either paired t-tests for within-subject repeated measurements or unpaired t-tests otherwise. For data consisting of more than two groups or multiple parameters, we used repeated-measures ANCOVA. All multiple comparisons used critical values from a t-distribution, adjusted by Bonferroni correction with an alpha level set to 0.05. Means are expressed ± SEM. Statistical significance was considered if $p < 0.05$.

3.4.2 Models

3.4.2.1 Introduction

We developed models to predict the responses of the mice to the novel sounds in the different tasks. There were three stages (sub-models) composing the full models, each stage providing input for the subsequent stage.

1. A cochleagram – a simple model of the cochlea.

2. A normative encoding model – representing the processing of the cochlea input by the auditory nervous system, to the level of the midbrain or cortex. This provides a neural representation.

3. A decision model – representing the transformation from the neural representation to the animal behavioural response; the percentage change of not poking its nose.

There were four types of full model. These only differed in their encoding model (stage 2). The cochleagrams (stage 1) and the decision model (stage 3) had same form in all models. The models were:

1. Prediction model. This had a recurrent temporal prediction model as its encoding model, pre-trained to predict the immediate future of natural sounds.

2. Compression model. This had a sparse recurrent autoencoder as its encoding model, pre-trained to compress natural sounds.

3. Cochleagram-only model. This had no encoding model, with the decision model mapping straight from the cochleagram to the behavioural response.

4. Frequency-scrambled model. This was the prediction model, but pre-trained to predict the immediate future of natural sounds whose frequency information has been scrambled.

Preparing and using a full model consisted of the following process:

1. We trained the normative encoding model (if present) using a database of natural sounds. We then fixed the weights of the encoding model.

2. For each task, we trained a separate decision model to predict the behavioural responses for the safe and unsafe sounds, using the pre-trained encoding models' activity to those sounds as input to the decision model. We then fixed the weights of the decision model for each task. This provides a separate full model for each task that only differed in the parameters of their decision models.

3. For each task, a full model outputs a prediction of the probability of visits without nose-poking for each novel sound.

Below we detail each stage of the full models, including the different types of normative encoding model.

3.4.2.2 The Cochleagram Model

Sounds were converted into cochleagrams using a cochlear model, the spec-Hill model(Rahman et al., 2020). This very simple model has been demonstrated to provide good input to cortical models for predicting neural responses in primary auditory cortex(Rahman et al., 2020). This cochleagram model was identical to the previously published version (Rahman et al., 2020) except for the sound frequency range and the time-bin size - we used 5 ms time-bins. Briefly, a spectrogram was produced from the sound waveform by taking the power spectrum using 10-ms Hanning windows, overlapping by 5 ms. The power of adjacent frequency channels was summed using overlapping triangular windows (using code adapted from melbank.m, http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html), with 32 log-spaced center frequencies ranging from 500 Hz to 22,627 Hz. The resulting power in each time-frequency bin was then converted to log values, then a threshold was applied, and then put through a Hill function to provide the cochleagram activity.

3.4.2.3 The Normative Encoding Models

3.4.2.3.1 Training and Validation Data for All Normative Encoding Models

We used an auditory dataset composed of natural sounds: birdsong and other animal calls, environmental sounds, and human speech by children and adults as well as baby vocalizations. Sound recordings of birdsong and mammal and insect calls were drawn from the Macauley Library at the Cornell Lab of Ornithology[2] (https://www.macaulaylibrary.org/),

[2] [i] Specifically, we use the following recordings from the Macauley Library: ML3429, ML3891, ML53181, ML55348, ML55350, ML59291, ML61494, ML62970, ML67895, ML70539, ML72995, ML80731,

and the British Library Sounds Environment & Nature Collection (https://sounds.bl.uk/Environment). We also used a corpus of ferret vocalizations recorded in our laboratory, which was supplemented by a recording found on https://freesound.org/people/J.Zazvurek/sounds/155115/. Recordings of adult human speech were taken from http://databases.forensic-voice-comparison.net/ (Morrison et al., 2012). We also used some human speech and environmental sounds, such as snapping twigs, which had been recorded in our laboratory in an anechoic chamber. Recordings of children speaking were from the CHILDES database (MacWhinney, 2000; Beckman and Edwards, 2010) from TalkBank (talkbank.org). Baby vocalizations were from the OxVoc database (Parsons et al., 2014).

All recordings were sampled at or resampled to 44,100 Hz, and converted into cochleagrams. The cochleagrams were divided into training and validation sets such that the proportions of different types of sound were approximately consistent in both datasets. Both sets were normalised to the mean and standard deviation of the training set. Gaussian white noise, with a signal to noise ratio of 6dB, was added to the cochleagram activity of both data sets. Adding such noise for temporal prediction models has been found to produce tuning properties that slightly better resemble the biology (Singer et al., 2020), and can be interpreted as internal neural noise. All cochleagrams were then split into one-second-long

ML100723, ML100797, ML100857, ML111595, ML116302, ML116303, ML126289, ML129248, ML130909, ML132511, ML132526, ML136504, Ml164696, ML171372, ML191165, ML191178, ML191290, ML197064, ML199078, ML201215, ML205481, ML205774, ML207517, ML206448, ML207181, ML210645, ML213055, ML213356, ML217313, ML217490, ML220049, ML233387, ML516692, ML527184 and ML527292

snippets. Then to remove uninformative snippets, those resulting snippets for which the root mean square of their cochleagram activity (over frequency channels and time-bins) was below twice the standard deviation of the added noise (0.356) were removed. This gave a total of 9,050 snippets in the training set and 1,129 snippets in the validation set. We denote the cochleagram activity for a given snippet by $s_f[t]$ where f is the frequency channel and t is the time-bin. There were f = 1 to F = 32 frequency channels and t = 1 to T = 200 time-bins in the snippet.

3.4.2.3.2 Temporal Prediction Model

The temporal prediction model was optimised to predict the immediate future activity of cochleagrams based on their recent past activity. It was based on the standard temporal prediction model (Singer et al., 2018) but with the addition of recurrent connections. This recurrent temporal prediction model was a recurrent neural network (RNN), and consisted of an input layer, a single recurrent hidden layer, and an output layer. The input layer activity was the activity in all 32 frequency channels of the immediate past 4 time-bins (20 ms) of the cochleagram. The output layer predicted the activity in the single upcoming future time-bin (5 ms) of the cochleagram.

The input layer of the recurrent temporal prediction model had i = 1 to I = DF = 128 units. This represents the past d = 1 to D = 4 time-bins (20 ms) for all f = 1 to F = 32 frequency channels of the cochleagram. Thus, the activity of the input layer was the activity of the cochleagram in a window of the immediate past, but flattened out into a vector. Hence the activity of input unit $u_i[t] = s_f[t - d + 1]$, where $i = (d - 1)f + f$.

Likewise, the output layer had k=1 to K= QF linear units. This represents all f'= 1 to F= 32 frequency channels for the q =1 to Q future time-bins to be predicted. As we noted above, in the figures Q= 1 time-bin (5ms). Thus, the target predicted by the output layer was the activity of the cochleagram in a window of the immediate future, but flattened out into a vector. Hence, the activity of output unit $v_k[t]$ is trying to predict a target $y_k[t] = s_f[t+q]$ where $k = (q-1)f + f$.

The input layer was connected to a recurrent hidden layer with j= 1 to J= 1600 hidden units, which had the rectified linear function as their activation function:

$$\mathrm{g}(x) = \max(0, x)$$

For a given cochleagram snippet, the activity $a_j[t]$ of hidden unit j at time step t was given by the following equation:

$$a_j[t] = \mathrm{g}\left(b_j + \sum_{i=1}^{I} w_{ji} u_i[t] + \sum_{j'=1}^{J} w_{jj'}\, a_{j'}[t-1] \right)$$

where b_j is the bias value for hidden unit j, w_{ji} the weight from the input unit i to the hidden unit j, and $w_{jj'}$ is the recurrent weight from hidden unit j' to hidden unit j. Also, $a_{j'}[t-1]$ is the activity of hidden unit j' at time step t - 1.

In turn, the activity $v_k[t]$ of output unit k at time t is given by:

$$v_k[t] = b_k + \sum_{j=1}^{J} w_{kj} a_j[t]$$

where b_k is the bias value for output unit k, w_{kj} is the weight from hidden unit j to output unit k.

The trainable parameters were the input weight matrix w_{ji}, the recurrent weight matrix $w_{jj'}$, the hidden unit bias vector b_j, the output weight matrix w_{kj} and the output unit bias vector b_k. These parameters were optimised to perform temporal prediction on the training set by minimising the mean squared error with L1 regularisation on the parameters. The equation for the objective function that is minimised is:

$$E = \frac{1}{NCK}\sum_{n=1}^{N}\sum_{t=D}^{T-Q}\sum_{k=1}^{K}(v_{kn}[t] - y_{kn}[t])^2 + \lambda\left(\sum_{j=1}^{J}\sum_{i=1}^{I}|w_{ji}| + \sum_{j=1}^{J}\sum_{j'=1}^{J}|w_{jj'}| + \sum_{k=1}^{K}\sum_{j=1}^{J}|w_{kj}|\right)$$

where we have introduced the additional subscript n, where n is the snippet number, and N = 9,050 is the number of snippets in the training set. Thus, $v_{kn}[t]$ is the prediction $v_k[t]$ for snippet *n* and $y_{kn}[t]$ is the target $y_k[t]$ for snippet n. Here, C is the number of valid time-bins accounting for the spans of the input and target windows, $C = T - Q - D + 1$. The L1 regularisation strength is governed by the hyperparameter λ.

The training was performed using stochastic gradient descent on mini-batches of 100 training examples and the Adaptive Moment Estimation (ADAM) update function (Kingma and Ba, 2015). We also decreased the learning rate for ADAM in a step-wise manner throughout training. We trained the model for 1000 epochs, decreasing the initial set learning rate for ADAM (0.001) by a factor of 10 every 150 or 200 epochs. Typically, the objective function reached a steady state within 500 epochs. All weights were initialised using the Glorot uniform initialiser (Glorot and Bengio, 2010) except for the hidden-to-hidden weight matrix which was initialised using an identity matrix multiplied by 0.01. The bias vectors for

both the hidden and the output layer were initialised to 0. To provide stability, the weights w_{ji}, $w_{jj'}$, w_{kj} were further regularised by using the method of constraining the incoming weight norms to a maximum of 1 throughout training (Srebro et al., 2005).

We did a hyperparameter search for the model using different values of λ: 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7} and 0. We chose the model with the λ-value that resulted in the lowest mean squared prediction error on the held-out validation set. For the Q = 1, this was $\lambda = 10^{-5}$.

3.4.2.3.3 Compression Model

The compression model is optimised to encode a compressed representation of the input data, where the output layer reconstructs the original input from the compressed encoding. The same snippet of the cochleagram is used for both the input and the output, which had the same input dimensions as for the recurrent temporal prediction model, 4 time-bins of the cochleagram with 32 frequency channels per time-bin.

The network architecture was identical to the recurrent temporal prediction model apart from the output layer, which had the same number of linear units (I = K = 128 input/output units = 32 frequency channels × 4 time-bin) as the input layer.

The output was described by:

$$v_k[t] = b_k + \sum_{j=1}^{J} w_{kj} a_j[t]$$

where $v_k[t]$ is the estimate of the input at time step *t* for output unit *k*.

The cost function is similar to the recurrent temporal prediction model apart from the target values being the current input rather than the future of the cochleagram. Hence, both the input and target are represented by $u_i[t]$. The cost was calculated as the mean squared error between the input and model reconstruction, with L1 regularisation on the parameters and the hidden unit activity to give a compressed representation.

$$E = \frac{1}{KCN}\sum_{k=1}^{K}\sum_{t=D}^{T}\sum_{n=1}^{N}(v_{kn}[t] - u_{in}[t])^2 + \lambda\left(\sum_{j=1}^{J}\sum_{i=1}^{I}|w_{ji}| + \sum_{j=1}^{J}\sum_{j\prime=1}^{J}|w_{jj\prime}| + \sum_{k=1}^{K}\sum_{j=1}^{J}|w_{kj}|\right) + \lambda_a\sum_{j=1}^{J}\sum_{t=D}^{T}\sum_{n=1}^{N}|a_{jn}[t]|$$

where $u_{in}[t]$ the activity of input unit i at time t for training example n, λ_a is the regularisation parameter for the hidden unit activity, and $C = T - D + 1$.

Training, initialisation, and weight norm constraining were done exactly as with the temporal prediction model, and the objective function also typically reached steady state within about 500 epochs. Unlike the temporal prediction model, there is no quantitative method for choosing λ and λ_a, so we relied on a qualitative approach that made the sparse autoencoder comparable to the temporal prediction model. To do this, we did a hyperparameter search for the sparse autoencoder model using different values of λ: 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7} and 0, while setting λ_a to 0. We chose the model with representations that qualitatively most closely resembled the receptive fields of auditory cortical neurons and had a similar number of active units to the temporal prediction model

(Q = 1). For the main sparse autoencoder model (D = 4), this was a λ-value of 10^{-5}. Using this λ value, we then added regularisation to the hidden unit activity and did a hyperparameter search over λ_a: 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}, 10^{-9}, 10^{-10}, and 0. Again, we then chose the resulting model that formed representations qualitatively most similar the receptive fields of auditory cortical neurons and with a similar number of active units to the temporal prediction model (Q = 1). For the main sparse autoencoder model (D = 4), this was a λ_a-value of 10^{-7}.

3.4.2.4 Frequency Scrambled Encoding Model

This was the temporal prediction model given above, but when training this encoding model, we scrambled the frequency channels of each natural sound cochleagram. More precisely, for each one second cochleagram snippet, the 32 frequency channels were randomly reordered. This random reordering was different and independent for each cochleagram snippet. The model was then trained in the exact same way as the temporal prediction model above, and it also typically settled within 500 epochs. The behavioural task stimuli (see below) were not frequency scrambled.

3.4.2.5 Measuring Encoding Model Receptive Fields

For the temporal prediction model (Q = 1) and the sparse autoencoder model (D = 4), we measured the spectro-temporal receptive fields (STRFs) of their hidden units (Fig. 3.2H). To do this we took the trained model and gave it directly as input cochleagrams of 500 second duration whose activity at each frequency and time-bin were independently drawn from a

Gaussian distribution with a mean of 0 and a standard deviation of 1. We then used the response of each model unit to this stimulus to calculate its STRF using an elastic-net regression model (Willmore et al., 2016; Willmore, 2020). An elastic-net regression model is a model that applies both L1 and L2 regularisation to its weights, and includes an additional hyperparameter that specifies the ratio between the L1 and L2 term in the cost function.

Units with a response of zero throughout this stimulus were declared non-active units, having no STRF; only active units are displayed in the figures. The temporal prediction model (Q = 1) had 89 active units and the sparse autoencoder model (D = 4) had 146 active units.

3.4.2.6 Decision Models

3.4.2.6.1 Behavioural Stimulus Set

The sound stimuli for the behavioural discrimination task was processed in a similar manner as for the sounds used for the encoding models. The behavioural task sound snippets used as input to the models consisted of ~2 second snippets of the repeated sounds played to the mice in the Audiobox. We cosine ramped the first 5 ms of each snippet stimulus set to avoid spectral splatter. All snippets were resampled to 44,100 Hz and converted into cochleagrams. For stimuli with roving amplitude levels, we used the average of the cochleagrams at the different levels.

3.4.2.6.2 Decision Model

For each discrimination task, the decision model was trained with the corresponding behavioural responses of the safe and unsafe trials used as targets, while the associated safe and unsafe sounds were used as the training input. The behavioural responses consisted of an array of zeros and ones with each entry r_m indicating whether a mouse did not poke its nose for a particular trial m (r_m = 0: poke, r_m =1: no-poke) to the sound used in that trial, be it a safe sound or an unsafe sound. Mouse identity was ignored in training the decision model, and the responses of all the mice in the group for that task were collated together.

The safe and unsafe sounds were converted to cochleagrams, then put through a normative encoding model which had been pre-trained with natural sounds and then had its weights fixed. The encoding model hidden unit activity provided the input to the decision model. The decision model consisted of a logistic regression from the encoding model hidden unit activity throughout the ~2 second snippet to the behavioural response to a single trial. The decision model was then trained to estimate the behavioural response to the full set of trials for the safe and unsafe sounds.

The behavioural response prediction, p, for the output unit for a given trial is provided by:

$$p = \mathrm{h}\left(b_o + \sum_{j=1}^{J} \sum_{t=1}^{T} w_{jt} a_j[t] \right)$$

where w_{jt} is the weight from the activity of encoding-model hidden unit j at time t in the snippet onto the single output unit, $a_j[t]$ is the activity of encoding-model hidden unit j at

time t in response to the sound snippet, be it the safe or unsafe sound. Also, J is the number of hidden units in the encoding model (J= 1600), T is the duration of the snippet in time-bins (T= ~400) and b_o is the bias. A sigmoid activation function was used for the output unit:

$$\mathrm{h}(x) = \frac{1}{1 + e^{-x}}$$

The trainable parameters for this model were w_{jt} and b_o, and we fix all the weights and biases associated with the pre-trained normative model. We used the binary cross entropy as the cost function of the decision model, alongside L1 regularisation of the decision model weights:

$$E = \sum_{m=1}^{M} -r_m log_2(p_m) - (1 - r_m) log_2(1 - p_m) \; + \lambda \sum_{j=1}^{J} \sum_{t=1}^{T} \left| w_{jt} \right|$$

where r_m is the behavioural response for trial m and p_m is the model's response prediction for trial m. Also, λ is the regularisation strength and M is the number of trials.

The training was performed using stochastic gradient descent on mini-batches of 100 training examples and the Adaptive Moment Estimation (ADAM) update function (Kingma and Ba, 2015). We also decreased the learning rate for ADAM in a step-wise manner throughout training. We trained the model for 150 epochs, decreasing the initial set learning rate for ADAM (0.001) by a factor of 10 every 15 epochs. Typically, the objective function reached a steady state within 50 epochs. The weights w_{jt} were initialised using the Glorot uniform initialiser (Glorot and Bengio, 2010) and b_o was initialised to 0. For the behavioural tasks that had a second conditioning step to probe generalisation, w_{jt} and b_o were initialised to the values learned by the model for the first conditioning.

For each encoding model type, we then performed the following process. For each task, we fit the decision model for a range of λ-values: 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, and 0. Then for each λ-value, we used the model to predict generalisation responses to the novel sounds used in the task. For each novel sound, this was done by producing its cochleagram and running it through the encoding model. The encoding model activities were then in turn put through the trained decision model to provide p, the model prediction of average response of the mice, that is the average probability of no-nose-poke in response to the novel sound. After getting a p for each novel sound, we then measured the accuracy of the predictions over all the novel sounds in the task. We did this by measuring the mean squared error (MSE) between the model predictions and the actual responses. This then provided a plot of MSE as a function of λ for each task. We cannot simply choose the λ that gives us the lowest MSE on the novel sounds, as this would be akin to using the test dataset to choose the hyperparameters. Instead, for each task examined, we averaged together the plots of all the other tasks (except second conditioning tasks), to provide a plot of the average MSE as a function of λ for all tasks except the one being examined, and the λ with the lowest average MSE on this plot was then chosen. This λ-value was then used for the task examined to provide the figures and the summary assessments associated with that task. For the discrimination tasks with a second conditioning phase, the same λ-value found for the first conditioning was used.

3.4.2.6.3 Cochleagram-Only Model

The decision model for the cochleagram-only model was exactly the same as for the other full models except that the input to the decision model was the cochleagram rather than the encoding model hidden unit activity. Hence, the decision model is given by:

$$p = \mathrm{h}\left(b_o + \sum_{f=1}^{F} \sum_{t=1}^{T} w_{ft} s_f[t] \right)$$

where w_{ft} is the weight from frequency channel f onto the sole output unit m at time step t, and $s_f[t]$ is the value of the cochleagram for frequency channel f at time step t. This model used the same cost function, training procedure, and hyperparameter selection procedure as the decision models that used the encoding model activity as input.

3.4.3 Modelling Implementation

All code was implemented in MATLAB and Python and run on NVIDIA GeForce GTX-1080 GPUs.

Chapter 4. Temporal Prediction and Primary Auditory Cortex

4.1 Introduction

In the previous chapter, we showed that normative encoding principles can form representations of sound from which the behaviour of mice performing an auditory discrimination task can be explained. In this chapter, we further probe the properties of these normative encoding models. We explore whether the recurrent connectivity of the models can provide insights into the functional connectivity between neurons in primary auditory cortex (A1), and how well the model representations can predict neural activity in response to natural sounds.

Neurons in the auditory cortex have more complex tuning properties than the preceding subcortical auditory neurons. This complex tuning is likely important for obtaining and analysing more abstract information from natural sounds (King et al., 2018). Despite their importance for auditory perception, the functions and computations of A1 neurons remain poorly understood.

The tuning properties and activity of cortical neurons are closely related to their connectivity, that is the circuits in which they are embedded (Cash and Yuste, 1999; London and Häusser, 2005; Sabatini and Svoboda, 2008; Yuste, 2015). Locally, the connectivity of A1 neurons is dependent on distance – A1 neurons that are close to each other are more likely

to be connected, and no local connections are formed beyond a 300 µm soma-to-soma distance (Levy and Reyes, 2012). The spatial organisation of A1 is commonly characterised by the frequency tuning of A1 neurons, which can be approximately described by organisation along a tonotopic axis (however, see Kanold et al., 2014; Gaucher et al., 2020). Hence, both tonotopy and local connectivity suggest that neurons with similar frequency tuning might be more likely to connect to one another. Another study indicates that synaptic inputs to the same dendritic tree on an A1 neurons also comes from neurons that are heterogenous in their frequency tuning (Chen et al., 2011), but with some tendency for the inputs to have similar frequency tuning. However, beyond these two studies little is known about functional connectivity, i.e. the relationship between tuning properties and connectivity, in A1, unlike in V1 (Ko et al., 2011; Cossell et al., 2015; Iacaruso et al., 2017). We will investigate whether our normative models can explain what little is known of functional connectivity in A1.

Another way to expand our knowledge about A1 is by building models that can predict the responses of A1 neurons to natural sounds. A complete understanding of the sensory systems must include the ability to accurately predict sensory neural responses to arbitrary stimuli, particularly to natural sounds (Carandini et al., 2005). Such a model of A1 responses to natural sounds has been challenging to achieve due to the nonlinear computations and complex selectivity for auditory features of A1 (Rabinowitz et al., 2011). The responses of A1 neurons to natural sounds are only partially captured by the linear models known as spectro-temporal receptive fields (STRFs), or by linear-nonlinear (LN) models, which append a static nonlinearity to an STRF, and are better captured by more elaborate models (Harper et al., 2016; Meyer et al., 2017). The prediction of A1 neural responses to natural sounds has largely focused on fitting models directly to neural responses. However, in recent times,

neuroscientists have employed normative task-driven models to try and explain processing in V4 and IT (Cadieu et al., 2014; Yamins et al., 2014) and V1 (Cadena et al., 2019) of the visual system, as well as auditory cortex in humans (Kell et al., 2018). The idea is that an accurate model of sensory processing must eventually be able to solve ethological tasks enabled by the sensory systems, and that these models will contain features that are similar to the brain (Yamins and DiCarlo, 2016). The intermediate layers of a task-optimised model can then be mapped to sensory neurons to predict their response to arbitrary stimuli. This approach is also known as transfer learning. Using this method, we examined how well our normative models can explain A1 responses to natural sounds.

In this chapter, we investigated whether temporal prediction can improve our understanding of the two above-mentioned aspects of A1 organisation. First, we asked whether the recurrent temporal prediction model can capture what is known about the functional connectivity in A1. More specifically, we interpreted the recurrent weight matrix of the model as a connectivity map of A1, and compared our model findings with experimental data on the functional specificity of local synaptic connectivity (Chen et al., 2011). We found that the temporal prediction model captured fairly well the known functional connectivity of A1 neurons. Second, we asked if the same normative model can accurately predict A1 neuron responses to natural sounds. We used the activity of the hidden units of the normative model as input to an LN model. We then fitted the LN model to estimate single-unit responses recorded in A1 of awake ferrets listening passively to natural sounds (Espejo et al., 2019). We then evaluated the ability of the fitted LN models, using the normative models as a preprocessing stage, to predict neural responses on a held-out test set. Our temporal prediction model was moderately effective at predicting A1 responses.

4.2 Results

We first trained a temporal prediction model and an autoencoder model using a large corpus of natural sounds. The temporal prediction and autoencoder models were identical to the encoding models used in the previous chapter. To summarise, we converted a dataset containing natural sounds into a power cochleagram representation with 32 frequency channels log-spaced from 50 to 22,627 Hz in 5 ms time-bins and compressed using an exponential function combined with the Hill function (Rahman et al., 2020). We then added Gaussian noise with a signal-to-noise ratio of 6 dB. For the temporal prediction model, we then trained a recurrent neural network to do temporal prediction on the cochleagrams for a corpus of sound that was deemed representative of the natural soundscape (Fig. 4.1A). The neural network was optimised to predict the immediately upcoming future time-bin (5 ms) given the four most recent time-bins (20 ms) of the cochleagram. The recurrent neural network had one fully connected recurrent hidden layer consisting of 1600 hidden units, and the input and output layers represent the past and future of the cochleagram respectively. Using the same corpus of natural sounds, we also trained an autoencoder model with similar network architecture to recreate the original input, providing a compressed representation (Fig. 4.1B).

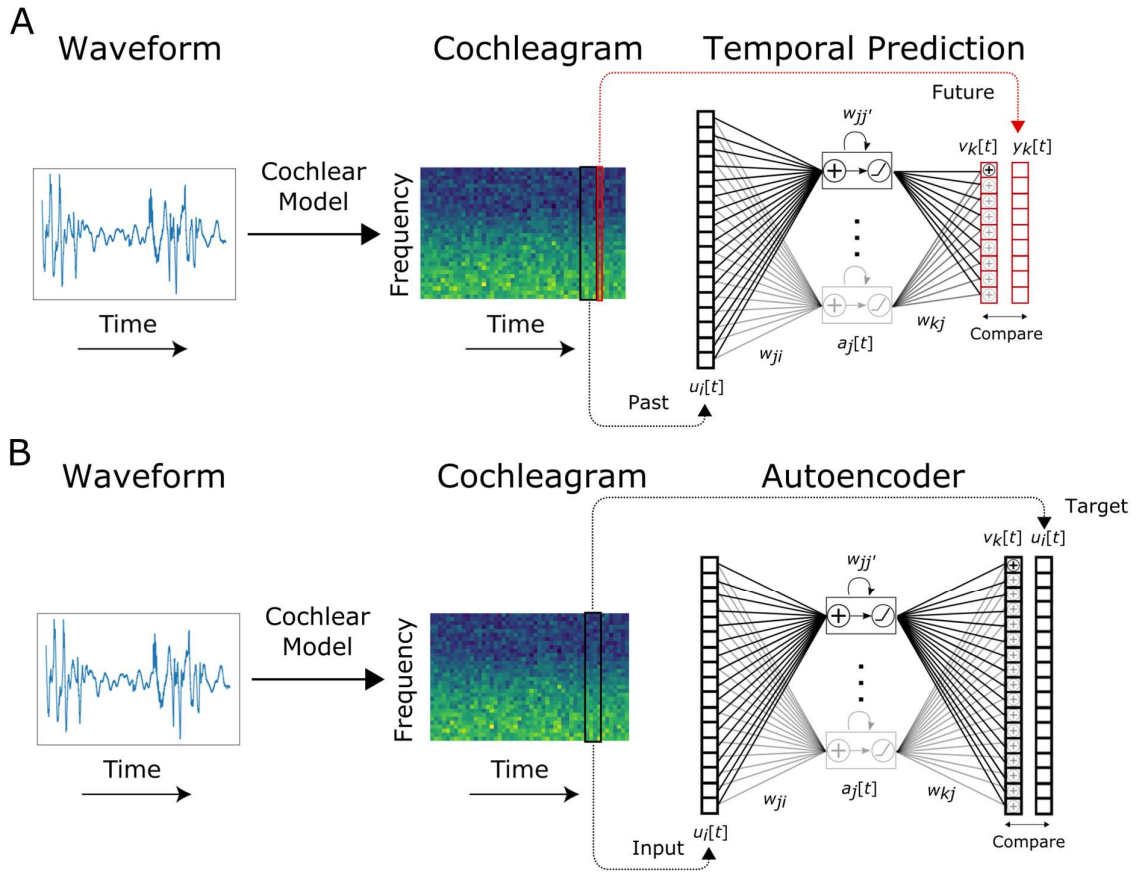


Figure 4.1. Schematic of the recurrent temporal prediction and autoencoder models. (A) The temporal prediction model. Raw sound waves of natural sound were converted to cochleagrams, and a sliding window was applied so that 20 ms (4 time-bins) of the cochleagram served as the input (past, black) from which the temporal prediction model is trained to predict the immediate future time-bin (5 ms) (future, red). **(B)** The autoencoder model is trained to recreate the input 20 ms (4 time-bins) from a compressed representation of that same input.

4.2.1 Receptive Fields and Frequency Tuning

First, we investigated the representations learned by the hidden units of the trained models. Since the hidden units are recurrently connected and nonlinear, we cannot simply use their weight matrices to assess their representations. We therefore used a 'virtual physiology'

technique, where we measured the responses of each hidden unit when Gaussian noise was presented to the networks. Based on these responses, we estimated the STRFs of the hidden units using elastic-net regression (Willmore, 2020). In the analysis below, we only show active hidden units, which we define as units that respond to Gaussian noise.

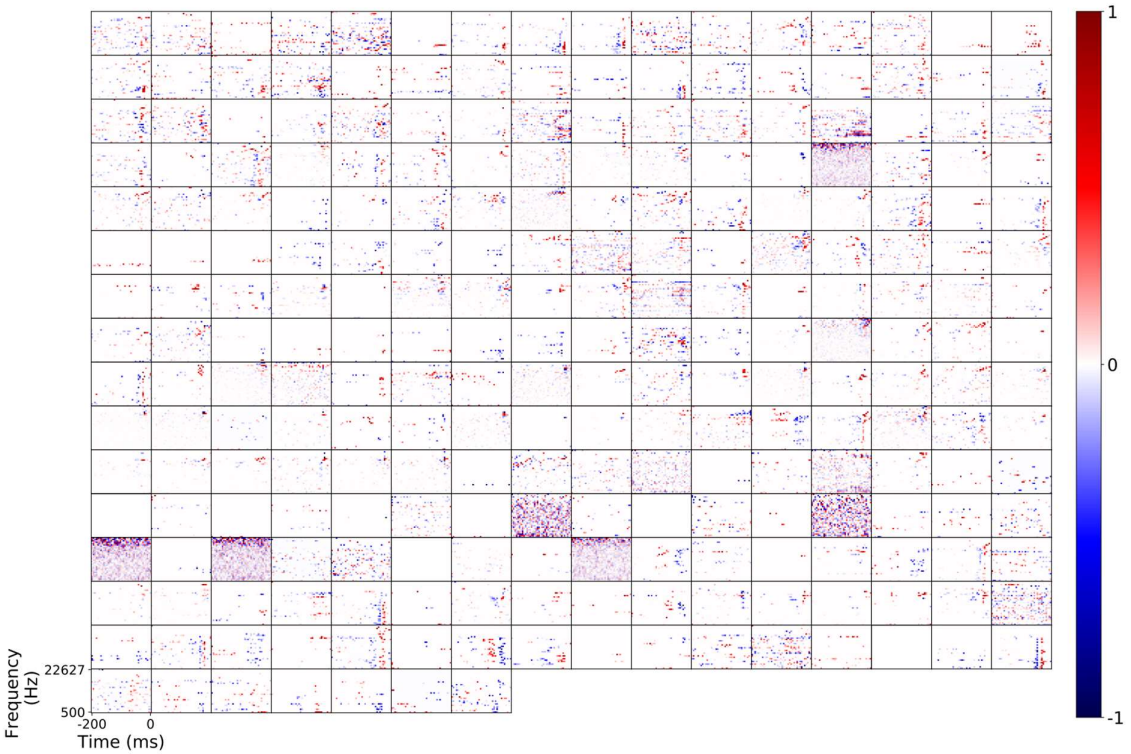


Figure 4.2. Receptive fields of neurons in primary auditory cortex (A1). Spectro-temporal receptive fields of 252 single-units from ferrets. Red denotes excitation, and blue inhibition. For visualisation, each STRF was normalised so that its maximum absolute value was 1. Single-unit responses were recorded in A1 of awake ferrets listening passively to natural sounds (Espejo et al., 2019). STRFs were obtained using an elastic-net regression model mapping the natural sounds over a 200 ms window to the neural responses.

The STRFs of real A1 neurons are diverse, showing both narrow and broadband tuning (Fig. 4.2). Many units display quite punctate tuning with sparse and sharp representations.

Others exhibit more broadband tuning with lagging or flanking inhibition. Some show leading inhibition or no lagging inhibition. A few have ill-defined structure, with inhibitory and excitatory spots throughout the receptive field. In general, the STRFs have stronger weights closer to the present (the right-hand side of the STRF).

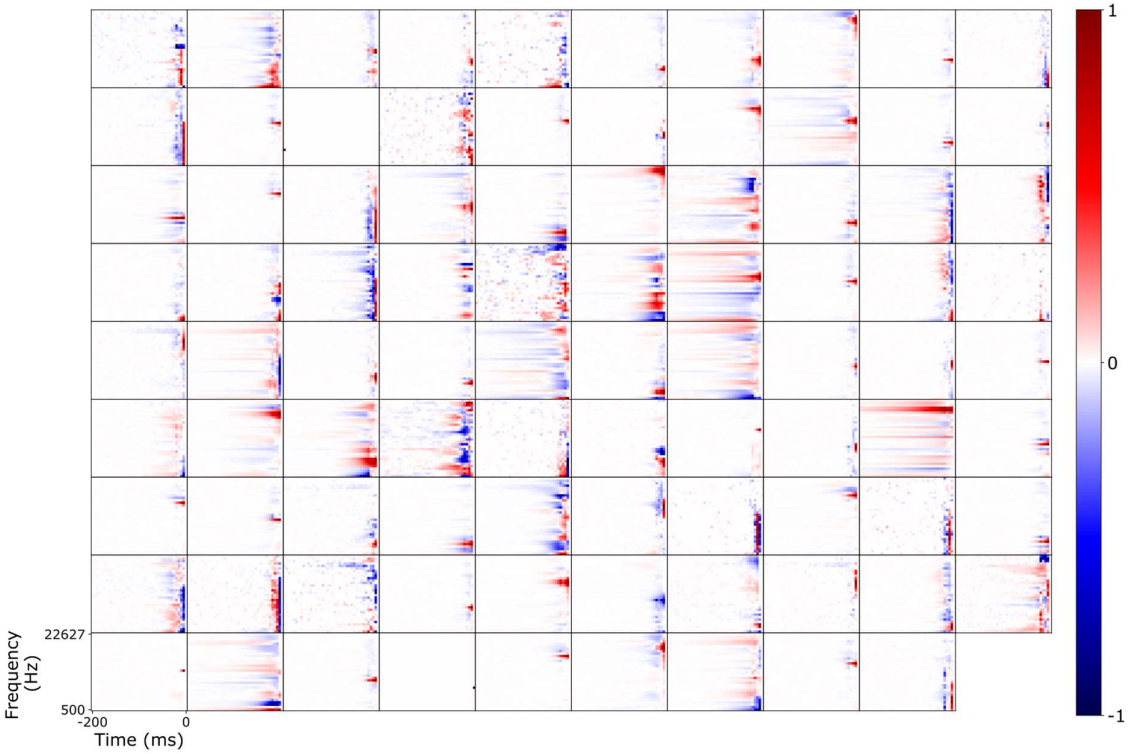


Figure 4.3. Receptive fields of the temporal prediction model. Spectro-temporal receptive fields of all active units of the temporal prediction model. Red denotes excitation, and blue inhibition. For visualisation, each STRF was normalised so that its maximum absolute value was 1. STRFs were obtained using an elastic-net regression model mapping Gaussian noise over a 200 ms window to hidden unit responses.

Qualitatively, the trained temporal prediction model contains hidden units whose STRFs resemble those of A1 neurons (Fig. 4.2 and 4.3). Like the A1 neurons, the hidden unit STRFs have diverse tuning properties. In frequency space, the bandwidths of the STRFs varied

across the extremes from narrow to broadband characteristics. We also observe flanking and lagging inhibition within the ensemble of units. Some units show leading inhibition or no lagging inhibition. These are all features observed in A1 STRFs (Rahman et al., 2019, 2020), as we can see in Fig 4.2. Some model STRFs are more complex, with temporally consecutive bands of excitation and inhibition across frequency channels, while some exhibit more ill-defined structure. Such multi-peaked tuning has been observed in a proportion of A1 neurons (Gaucher et al., 2020). In general, the STRFs of the model units have stronger weights closer to the present. However, some units are driven by input throughout large parts of the 200 ms input time-window. Both of these trends have been found in A1 neurons (DeCharms et al., 1998).

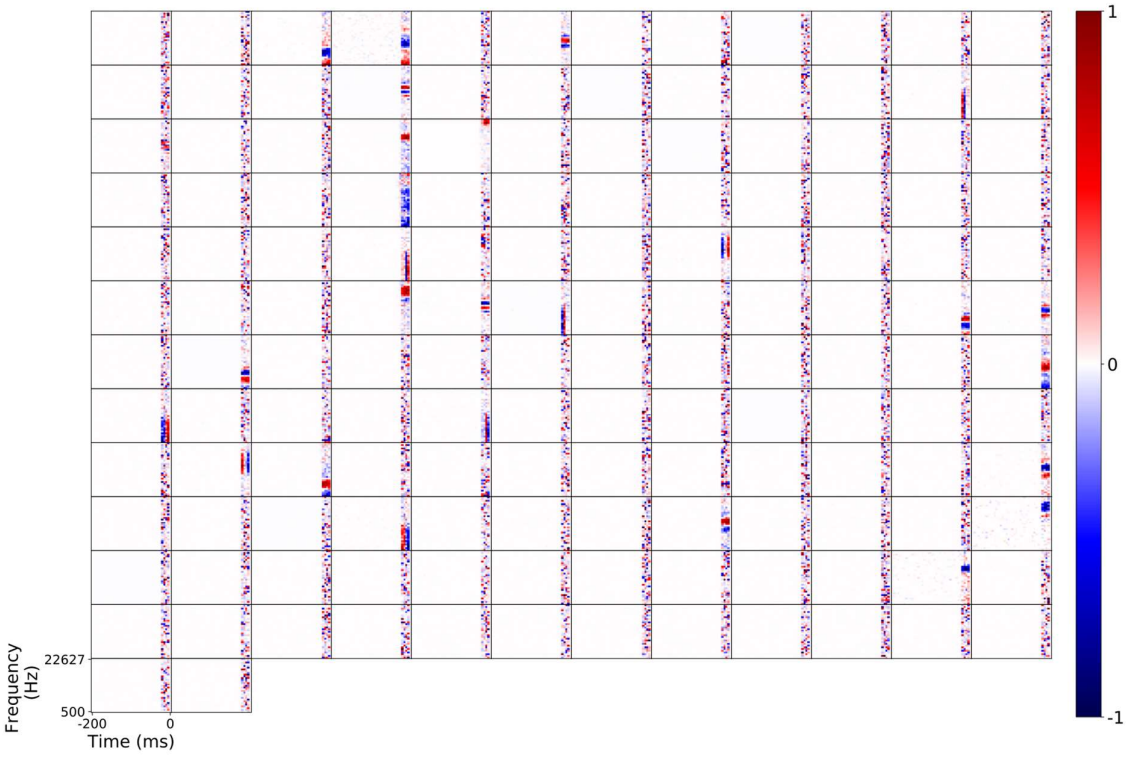


Figure 4.4. Receptive fields of autoencoder model. Spectro-temporal receptive fields of all active units of the autoencoder model. Red denotes excitation, and blue inhibition. For visualisation, each

STRF was normalised so that its maximum absolute value was 1. STRFs were obtained using an elastic-net regression model mapping Gaussian noise over a 200 ms window to hidden unit responses.

In general, the autoencoder STRFs were less similar than the temporal prediction model STRFs to A1 STRFs (Fig. 4.2 and 4.4). The autoencoder units tended to exhibit more broadband STRFs and mostly showed a salt-and-pepper pattern of excitatory and inhibitory spots across both frequency and temporal channels, which was only rarely seen in the A1 data. A few units did resemble more typical A1 neurons by having a more localised representation in frequency space. Temporally, the representation of virtually all units was localised to the most recent four time-bins. This suggests that the network recreates the 20 ms window of the stimulus by mapping only the input to the compressed representation, and indicates that the network did not make use of much past information to aid the compression.

We define the best frequency (BF) of a hidden unit as the frequency channel in its STRF that contains the largest power summed over time. This is a non-standard way of estimating BF, but we wanted to be able to assign BFs to diverse units, for example those units with largely inhibitory weights. The distribution of best frequencies for real A1 neurons exhibit a wide range in tuning, with a bias towards the higher frequencies (Fig. 4.5A). This may be a consequence of the wide spectral hearing range for ferrets, which is approximately 40 Hz to 40 kHz (Kavanagh and Kelly, 1988), and the high frequency tuning bias may be due to catching the low frequency tails of neurons tuned to frequencies above 22 kHz.

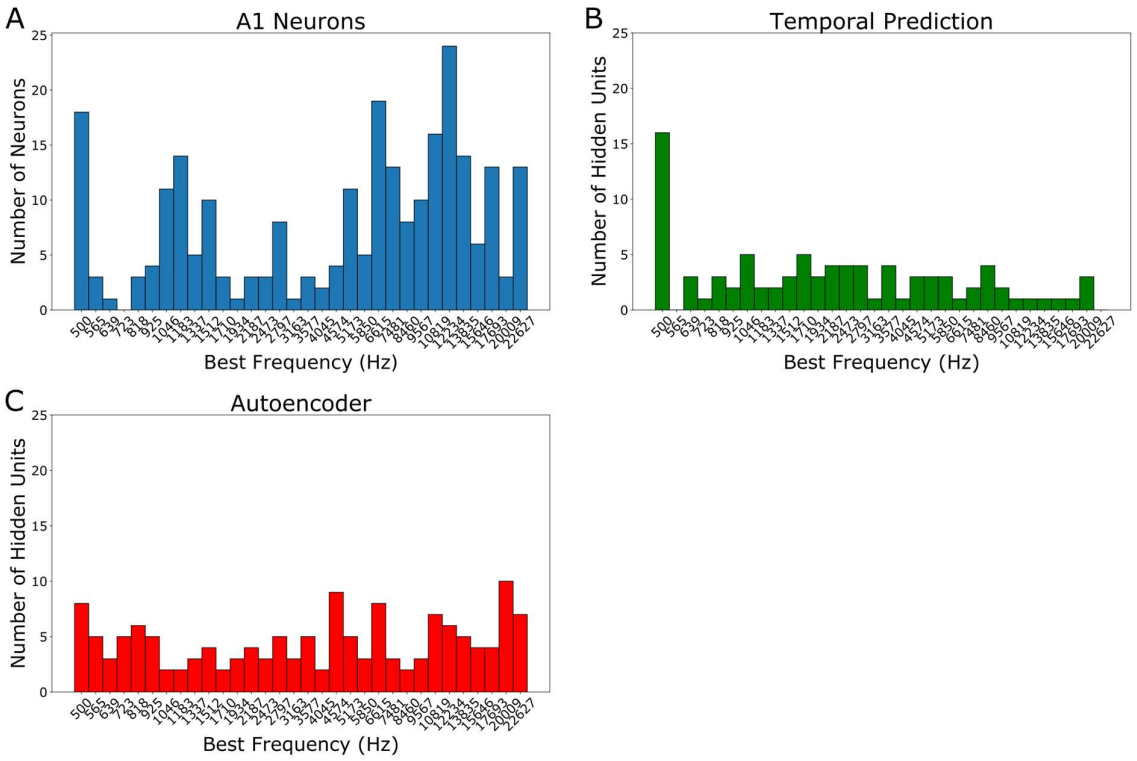


Figure 4.5. The best frequency distributions of model units and real A1 neurons. The best frequency of each unit was defined as the frequency channel of the receptive field with the largest accumulative power over time. **(A)** The distribution of best frequency units of A1 neurons from awake ferrets. **(B)** The distribution of best frequency units of the temporal prediction model. **(C)** The distribution of best frequency units of the autoencoder model.

For the temporal prediction model, the distribution of BFs of the hidden units covered the majority of the frequency range. The distribution of best frequencies was largely flat, but had a dominance of units tuned to 500 Hz, which might be attributed to an edge effect. The flat distribution over the logarithmic frequency space produces a more pronounced representation of lower frequencies compared to higher frequencies, which matches the tonotopy established in the cochlea that is inherited by A1 (Fig. 4.5B). The BFs of the

autoencoder units should not be given too much importance as most of the units are broadband, but are shown here for completeness (Fig. 4.5C).

4.2.2 Functional Connectivity

We asked if the recurrent temporal prediction model can capture what is known about functional connectivity in A1. Local connections are more common for neurons in close proximity and do not occur beyond 300 μm (Levy and Reyes, 2012). Given the tonotopy of A1, we expect similarly frequency-tuned neurons to be more likely to form local connections. We investigated the features of the recurrent weight matrix, and compared this matrix to the local recurrent connectivity of A1. Each hidden unit could be interpreted as a group or population of functionally similar A1 neurons, with positive and negative weights interpreted respectively as proportional to the potency or probability of excitatory and inhibitory connections between them.

The recurrent weight matrix of the temporal prediction model shows a complex pattern of connectivity. The matrix displays a prominent excitatory diagonal, which indicates a strong degree of excitatory recurrent connections from each hidden unit to itself (Fig. 4.6). Amongst the 500 Hz BF units, we observe extensive inter-unit communication as demonstrated by the excitatory and inhibitory connections in this group. In fact, many units excite and inhibit the activity of 500 Hz units. On the opposite end of the frequency spectrum, we find a 20 kHz tuned hidden unit that is both widely inhibited by lower BF hidden units and inhibits a wide range of units. The recurrent weight matrix of the autoencoder demonstrates a much sparser connectivity compared to the temporal prediction model, with no excitatory

diagonal. We see that a few units receive input from most units and some units project to most units, as indicated by the horizontal and vertical lines in the matrix (Fig. 4.7).

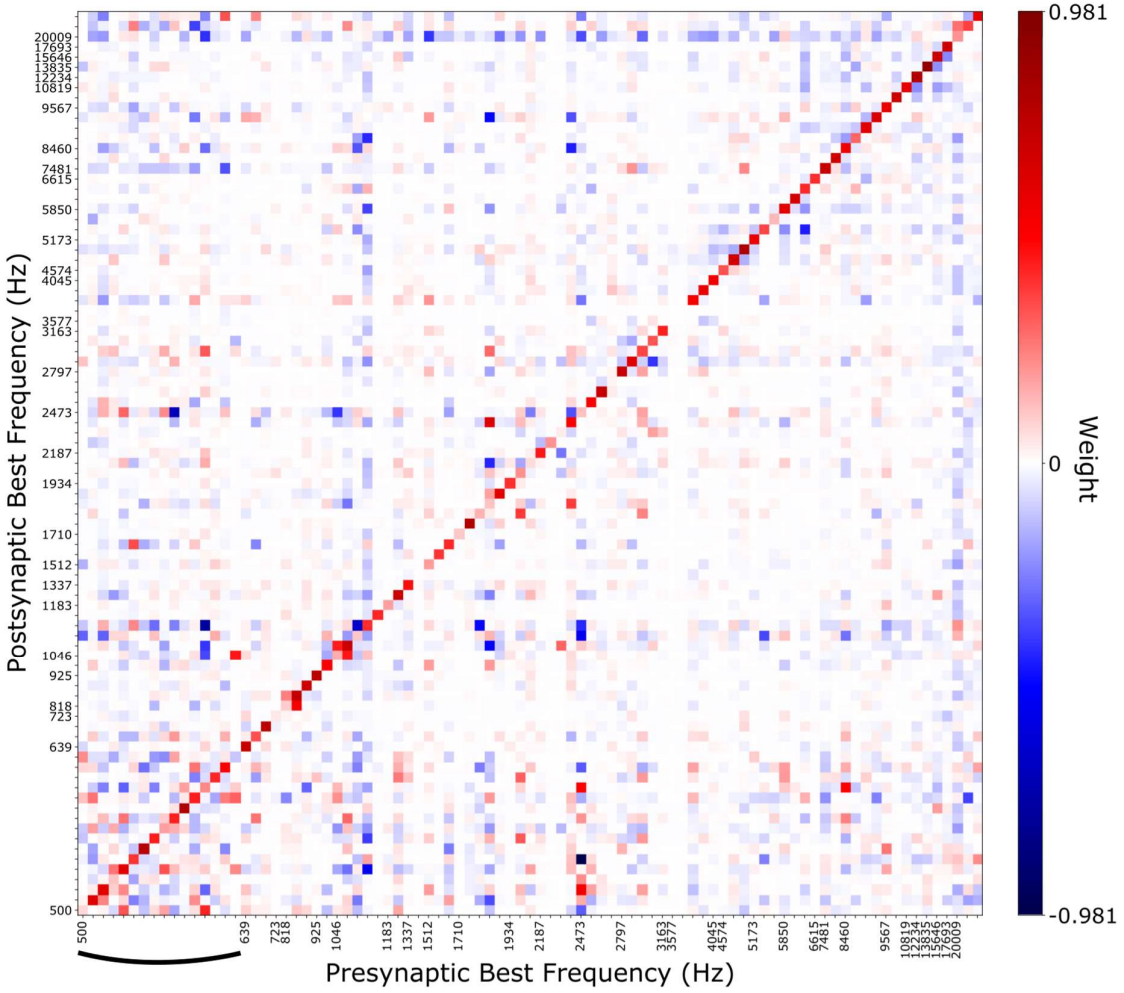


Figure 4.6. The recurrent weight matrix of the temporal prediction model. The recurrent weight matrix was interpreted as the model's prediction of A1 local connectivity, where positive and negative weights are excitatory and inhibitory synapses, respectively. Each number on the X-axis shows that the numbered column and columns to the right have that frequency value until the next number is reached. The bracket on the X-axis illustrates this by showing all units that have a best frequency of 500 Hz. The Y-axis follows the same form as the X-axis.

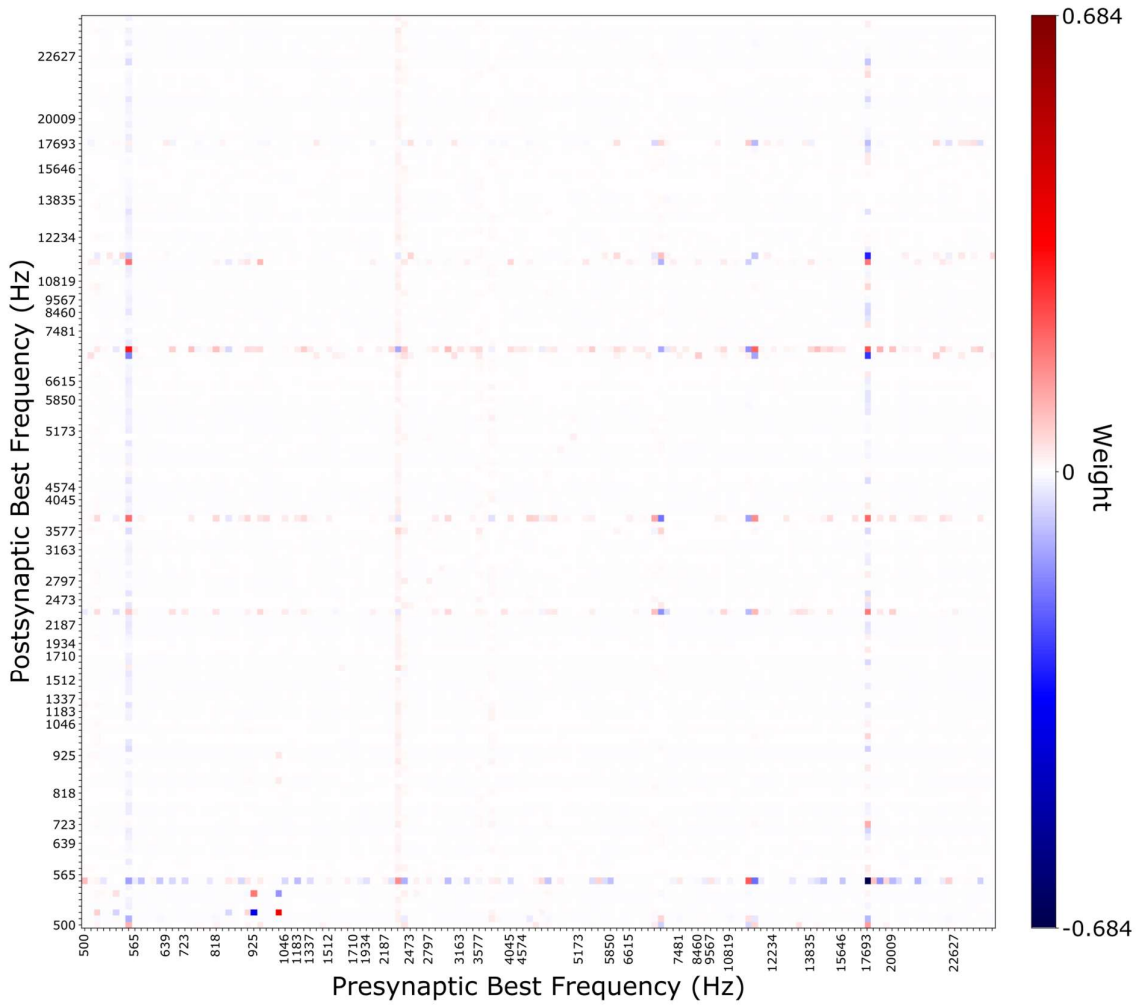


Figure 4.7. The recurrent weight matrix of the autoencoder model. The recurrent weight matrix was interpreted as the model's prediction of A1 local connectivity, where positive and negative weights are excitatory and inhibitory synapses, respectively. Each number on the X-axis shows that the numbered column and columns to the right have that frequency value until the next number is reached. The Y-axis follows the same form as the X-axis.

Compared to primary visual cortex (V1), there has only been limited investigation of functional connectivity in A1. One experimental study used calcium imaging on layer 2/3 pyramidal neurons in mouse A1 to determine the BFs of neighbouring individual dendritic spines (Chen et al., 2011). The BFs of the spines indicated the BFs of two different neurons synapsing on the imaged neuron. We will compare these results to our model. The calcium currents in the study were NMDA-dependent, so we limited the scope of our analysis to the excitatory connections of the model. To compare the model results with the experimental data, we first grouped model units with similar BFs (see Methods). The experimental data was grouped accordingly (Fig. 4.8A). The grouped weight matrix of the model can be interpreted as proportional to the probability of neurons in a certain BF group forming excitatory connections to neurons in another BF group. From this, we can calculate what the model expects to be the probability that a pair of afferents to a receiving neuron will each have specified BFs (see Methods) (Fig. 4.8B and C). This calculation requires the distribution of BF of all model units (Fig. 4.5B and C), and the assumption that the BFs of neighbouring spines on the dendritic tree of real A1 neurons are independent of each other. We performed the same analysis for the autoencoder model (Fig. 4.8C).

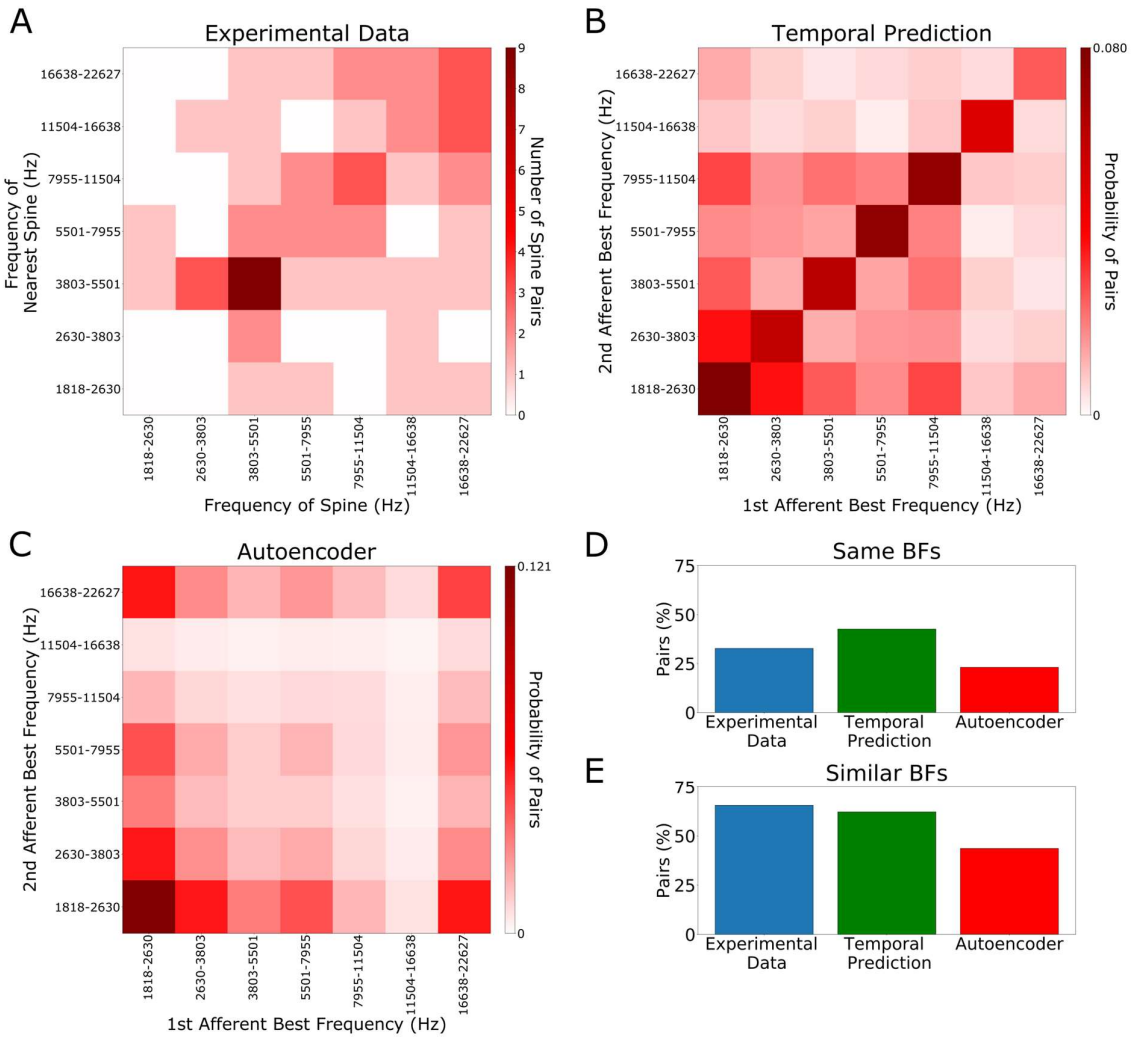


Figure 4.8. Functional connectivity of model units and real A1 neurons. (A) Best frequencies (BFs) of neighbouring spine pairs of pyramidal layer 2/3 neurons in mouse A1 (Chen et al., 2011). The BF of a dendritic spine is indicative of the BF the afferent neuron that makes a bouton on the spine. The colour denotes the number of spine pairs observed as a function of their BFs. 33% of spine pairs shared the same BF. 66% of spine pairs had the same or adjacent BFs. **(B)** We estimated the temporal prediction model's prediction of the probability of spine pairs, as a function of the BF of each spine in the pair. To do this we assumed that the model's excitatory connection strength was proportional to the probability of a bouton-to-spine connection being present. Over the BF range shown, 43% of pairs shared the same BF, and 62% had the same or adjacent BFs. **(C)** Bouton pair probability as function of the BFs of the afferents for the autoencoder model. Over the BF range shown, 23% of pairs shared the

same BF, and 43% had the same or adjacent BFs. **(D)** The proportion of pairs with the same BF for the models and real A1. **(E)** The proportion of pairs with the same or adjacent BFs for the models and real A1.

For the temporal prediction model, we observe a trend in the model connections where model units with the same or similar frequency tuning are more likely to connect to the same 'postsynaptic' unit (Fig. 4.8B). The probability of two 'presynaptic' units with the same BFs synapsing onto the same 'postsynaptic' neuron constitutes 43% of the total probability distribution of afferent BF pairings in the presented frequency range (Fig 4.8D). When we also included pairs with adjacent BFs, the probability increased to 62% (Fig 4.8E). The corresponding values from the experimental study were 33% and 66%, respectively (Fig 4.8A, D, and E). Overall, the connection probabilities of the temporal prediction model (Figure 4.8B) and the connection probabilities of the experimental data (Figure 4.8A) showed some correlation, with a correlation coefficient of 0.272. Thus, we conclude that a recurrent temporal prediction model can, to some extent, capture experimental findings on functional connectivity in A1.

The autoencoder model did not have the diffuse diagonal connectivity pattern seen in the experimental data (Fig. 4.8C). 23% of the afferent pairs in the investigated frequency range shared the same BFs, increasing to 44% when including adjacent BF pairs. This is lower than the values of 33% and 66%, respectively, in the experimental data. The model connectivity pattern was anti-correlated with the experimental connectivity patterns (-0.209)

(Chen et al., 2011). Thus, the normative principle of compression, in the form that we applied it, does not appear to be able to explain functional connectivity in A1.

4.2.3 Predicting A1 Neuron Responses to Natural Sounds

We then investigated the potential of the normative encoding models to predict and explain A1 neural responses to natural sounds. To do this, we prepended the normative models to the standard linear-nonlinear (LN) model as an intermediate processing stage between the cochleagram and the LN model, and then fitted the weights of the LN model to predict the responses to natural sounds of A1 neurons (Fig. 4.9). This method of assessing normative models has been used previously in the visual system (Cadena et al., 2019). We used single-unit recordings from A1 of awake ferrets passively listening to natural sounds (Espejo et al., 2019). We only used units with a firing rate of at least 1 Hz, and a noise ratio below 40. This restricted the dataset to 248 single units. The dataset consists of 18 natural sounds and corresponding neural response recordings. The LN models we used consisted of a linear mapping followed by a half-wave rectified sigmoid.

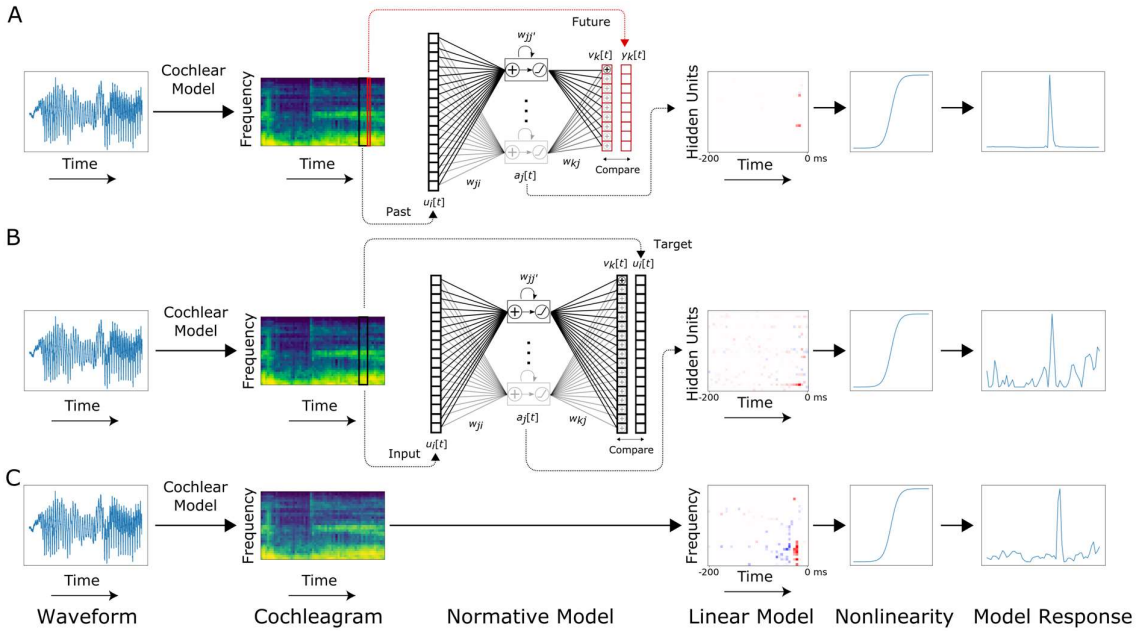


Figure 4.9. Model diagrams of the neural response models. First, temporal prediction and autoencoder normative models were pretrained on a large natural sound corpus as previously described, and then their weights were fixed. Next, for a different set of fewer natural sounds for which A1 neural responses had been recorded, the raw sound waves were converted to cochleagrams. Then the cochleagrams were preprocessed by either the temporal prediction model, the autoencoder model or were not preprocessed to provide input to an LN model. The LN model consisted of a linear stage followed by a sigmoid nonlinearity. LN models were then fitted to estimate neural responses for the three different kinds of preprocessing. **(A)** The temporal prediction model with appended LN model. To pretrain the temporal prediction model, raw sound waves from the large natural sound corpus were converted to cochleagrams, and a sliding window was applied so that 20 ms (4 time-bins) of the cochleagram served as the input (past, black) from which the temporal prediction model is trained to predict the immediate future time-bin (5 ms) (future, red). **(B)** The autoencoder model with appended LN model. The autoencoder model was pretrained to recreate the input 20 ms (4 time-bins) from a compressed representation of that same input. **(C)** The conventional LN model without input preprocessed by a normative model.

First, we took the already-trained normative models and fixed their weights. Then we ran the 18 natural sounds in the experimental dataset through a cochlear model to produce cochleagrams, which we then put through the normative models. Next, for each recorded neuron we fitted 3 LN models: one mapping from the temporal prediction normative model output to the neural response, one mapping from the compression normative model output to the neural response, and a conventional LN model, which was fitted to map directly from the cochleagrams to the neural response (Fig. 4.9). We used either the past 3 or 40 time-bins (15 ms and 200 ms, respectively) of either the cochleagram or model representation as input to the LN models. The LN model was fitted using 14 of the sounds, and the remaining 4 sounds were used as a held-out test set. Finally, we evaluated the resulting neural models on the held-out test set. We use the normalized correlation coefficient, CC_{norm}, as our evaluation metric (Hsu et al., 2004; Schoppe et al., 2016).

The distribution over neurons of prediction performance (CC_{norm}) on the test set for each model, and the average prediction performance of each model, for the 200ms input window LN model, are presented in figure 4.10. For each model, the CC_{norm}values span a wide range. The cochleagram LN model ranges from -0.399 to 0.715 (Fig. 4.10A), whereas the autoencoder spans from -0.118 to 0.74 and the temporal prediction model from -0.117 to 0.706 (Fig. 4.10B and C, respectively). Overall, we find that the autoencoder LN model produces the best neural predictions on average (average CC_{norm}= 0.253), followed by the temporal prediction model (average CC_{norm}= 0.236), and finally the cochleagram LN model (average $CC_{norm}$of 0.226) (Fig. 4.10D).

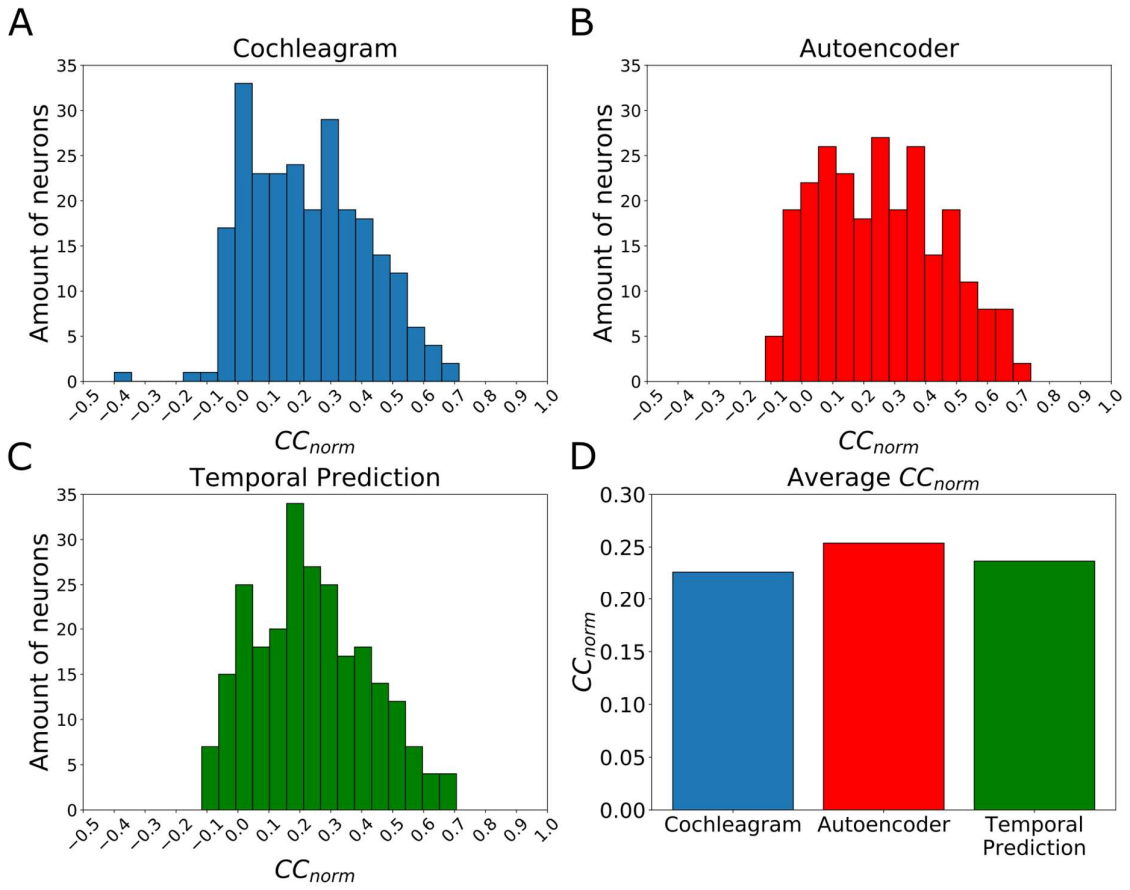


Figure 4.10. Neuron model performance (CC_{norm}) **on the test set using a 200 ms input window for the LN model. (A)** Cochleagram model. **(B)** Autoencoder model. **(C)** Temporal prediction model. **(D)** Average CC_{norm} for the neuron models.

However, using a 200 ms window might not provide a fair comparison of the models, as it enables the LN model to capture the temporal profile of any tuning, potentially obscuring any temporal tuning advantages brought by the normative model. Hence, we narrowed the time window to a smaller time window of 15ms (3 time-bins). This window is just long enough to allow the LN model to account for axonal time delay to the cortex (Bizley et al., 2005), which is not accounted for in the normative models, but is not long enough for the LN-model alone to capture the full temporal tuning profile.

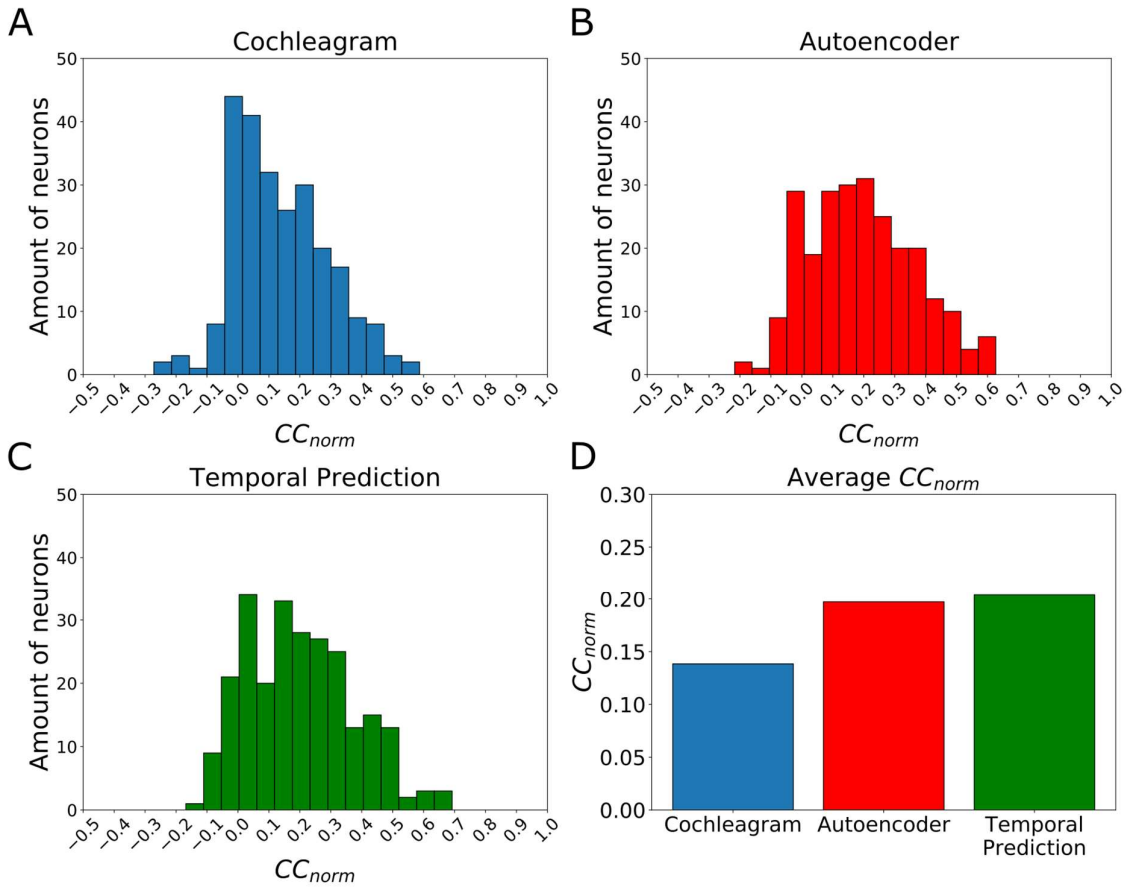


Figure 4.11. Neuron model performance (CC_{norm}) on the test set with a shorter input time-window (15 ms) for the LN model. (A) Cochleagram model. **(B)** Autoencoder model. **(C)** Temporal prediction model. **(D)** Average $CC_{norm}$for the neuron models.

We re-trained the same models, but used 3 time-bins (15 ms) of the cochleagram or the normative models as input (Fig. 4.11). The cochleagram model ranges from -0.271 to 0.587, the autoencoder from -0.217 to 0.627, and the temporal prediction model from -0.168 to 0.692 (Fig. 4.12A, B, and C). Overall, the temporal prediction model was the best model for predicting neural responses on a shorter time window (average CC_{norm}= 0.203) (Fig. 4.11D). The autoencoder model had a similar performance (average CC_{norm}= 0.197), while the cochleagram model performance decreased considerably (average CC_{norm}= 0.138).

However, the performance of all neural models degraded when using the 15 ms time window rather than using the 200 ms time window.

4.3 Discussion

We optimised normative recurrent neural network models to perform temporal prediction and compression on a natural sound dataset and evaluated their ability to reflect the biology of A1. The STRFs of the temporal prediction model capture the diversity of real STRFs from A1 neurons (Fig. 4.2 and 4.3), while all autoencoder units exhibit a speckled broadband STRF that is unlike that of real A1 neurons (Fig. 4.4). The functional connectivity suggested by the temporal prediction model reflects experimental studies on dendritic spines of pyramidal layer 2/3 neurons from mouse A1 (Chen et al., 2011), while the autoencoder model fails to replicate that study (Fig. 4.8). Normative models affixed to an LN model were moderately better able to predict neural responses than an LN model. The autoencoder normative had the best performance (Fig. 4.10) when the LN model time window was 200 ms, while the temporal prediction model had a slightly better performance over the autoencoder when using a more time-restricted LN-model window of 15 ms (Fig. 4.11).

4.3.1 Receptive Fields and Frequency Tuning

We find that the temporal prediction model settles on STRFs similar to the STRFs found in experimental A1 studies. Our finding reinforces a previous study, which demonstrated that a similar non-recurrent neural network optimised for temporal prediction on cochleagrams also settled on A1-like representations (Singer et al., 2018). A diverse range of features seen in the

STRFs of A1 neurons were also seen in the temporal prediction model. This included units with narrow and broadband tuning, lagging inhibition and no-lagging inhibition, and leading inhibition and flanking inhibition. The autoencoder units were less similar to A1 neuron STRFs, often being broadband, speckly and limited to a specific short time-window. The temporal limitation of the autoencoder model is most likely due to the restricted use of the recurrency as demonstrated by its recurrent weight matrix (Fig. 4.7). The prediction model, in contrast, has a more extensively connected recurrent weight matrix, which most likely results in its more temporally extended tuning. The analysis of the BFs of the autoencoder units was complicated by the majority being broadband and speckly. An alternative approach to measure BF would be to simply input pure tones at different frequencies to the network and find the frequency that gave the highest response. Quantitative assessments of the STRF similarities between the data and the models can also be investigated using methods such as dimensionality reduction or tuning width measurements as done in Singer et al. (2018).

4.3.2 Functional Connectivity

We find that a recurrent neural network trained on temporal prediction of cochleagrams captures aspects of the known local connectivity in A1 in its recurrent weight matrix, while the connectivity suggested by the autoencoder fails to mirror experimental findings. We acknowledge that the BFs of the autoencoder are difficult to assess due to their broadband receptive fields with speckly excitation and inhibition, and this will affect the interpretation of the functional connectivity matrix. However, we also stress that the absence of a strong diagonal in the recurrent connectivity matrix and its sparsity in concert imply that the autoencoder is unable to replicate experimental findings; A1 neurons receive afferents from

neurons with a wide frequency range of BFs and yet with an increased odds of receiving from units with similar BFs (Chen et al., 2011; Levy and Reyes, 2012).

The spatial organisation and functional properties of presynaptic input to dendrites play an important role in information processing (Cash and Yuste, 1999; London and Häusser, 2005; Sabatini and Svoboda, 2008; Yuste, 2015). Functional connectivity of A1 neurons is likely a major factor in forming their response properties (Ferster and Lindström, 1985; Douglas et al., 1995), but has so far been little explored. Functional connectivity has been more extensively investigated in V1. For example, in V1 of mice, the probability of layer 2/3 pyramidal neurons forming connections with each other tends to be positively correlated with the similarity of their preferred stimuli (Ko et al., 2011; Cossel et al. 2015). Specifically, V1 neurons that are similarly tuned to oriented stimuli are more likely to be connected (Ko et al., 2011), and V1 neurons with receptive fields that are co-oriented and co-axially aligned are also more likely to form connections, suggesting that the synaptic inputs to V1 neurons are strongly driven by elongated lines (Iacaruso et al., 2017). These findings indicate that the functional connectivity in V1 may be of great relevance for processing the elements of visual objects. Hence, this emphasizes that determining how the synaptic connections between A1 neurons relate to their tuning properties is of high value in understanding the operation of A1.

In our rate-based network, units can be seen as a crude approximation of a high-rate spiking neuron or the collective rate of a group of similarly tuned neurons. Self-excitatory connections in the recurrent weight matrix might correspond to a range of biological elements; the integrative nature of the neuronal membrane, alterations of ion channel

dynamics, autapses, or connections between the similarly tuned neurons of a group. Model units also send both excitatory and inhibitory afferents, which violates Dale's principle that states that individual neurons release a single neurotransmitter (Eccles, 1986). The connections represented by the recurrent matrix weights might not necessarily correspond with unisynaptic connections in the biology, but with disynaptic or higher order; e.g. an inhibitory weight might represent lateral inhibition via an inhibitory neuron.

4.3.3 Normative Encoding Models

The ability of the normative encoding models to predict A1 responses to natural sounds is quite modest, and calls for more exploration. An important factor is the choice of an appropriate time window for the LN model. Initially we used 200 ms of stimuli history as this has been shown to best for neural response predictions in A1 using standard LN models (Rahman et al., 2019). However, an input time window of this order plays to the strength of models that focus on capturing frequency tuning rather than temporal tuning, as the weights from the different times of the LN model can capture the temporal tuning. The notable strength of the temporal prediction model is its capacity to capture the temporal properties of neural tuning, and this may be lost with a 200 ms window. To account for this we also used a shorter stimulus history of 15 ms, a window that just spans the range of the shorter axonal delays to auditory cortex (Bizley et al., 2005). In this case, the cochleagram LN model performance deteriorates, while the encoding models are more robust to the constrained input window, with the temporal prediction model becoming the best at predicting neural responses.

A moderate body of work already exists concerning normative principles applied to A1 receptive field modelling. One study found that a neural model optimised for efficient coding produced a receptive field that was sharply frequency-tuned with flanking and lagging inhibition, and with a preference for recent stimulus (Zhao and Zhaoping, 2011). However, the single typical STRF this approach produces does not account for the diversity of A1 STRFs, as broadly tuned STRFs were absent. Furthermore, the temporal extent of the model receptive field did not exceed 15 ms unlike many real A1 STRFs (Rahman et al., 2019, 2020). Other studies optimised neural models for sparse coding (Blättler and Hahnloser, 2011; Carlson et al., 2012; Brito and Gerstner, 2016; Młynarski and McDermott, 2017).

Sparse coding is another approach that involves training models to recreate the input stimulus under a sparsity constraint. A sparsity constraint constrains the proportion of active units to a small fraction of all units. While sparse coding models recreate some frequency tuning aspects of A1 receptive fields (Carlson et al., 2012), they have so far fared less well at capturing temporal tuning (Singer et al., 2018). Another problem with many of these sparse coding models lies in their need for an iterative settling process to converge on a sparse representation (Olshausen and Field, 1996, 1997; Carlson et al., 2012), which makes such an approach unsuitable for the highly dynamic temporal profile of sound. However, sparse coding has been more successful in explaining spatial aspects of visual receptive fields (Olshausen and Field, 1996, 1997). Some models related to sparse coding use a sparsity constraint in conjunction with high dispersal, in which units are being constrained to have equal activation levels (Ngiam et al., 2011; Kozlov and Gentner, 2016) and maximal decorrelation between units. These models do not require the above-mentioned iterative

settling process (Klein et al., 2003). They can also capture some frequency tuning aspects of A1 receptive fields, but perform poorly at capturing temporal tuning.

Another principle, temporal slowness, minimises the correlation between the activity of different model units, while constraining the activity of units to vary slowly over time (Carlin and Elhilali, 2013). This approach can also be used to capture some frequency tuning properties of STRFs in central auditory areas, but also struggles to capture temporal tuning aspects of STRFs. One general limitation of almost all the above approaches lies in their failure to capture the preference seen in real A1 STRFs for more recent stimuli, with the power in the real STRFs typically sharply rising near the present and then more slowly decaying away into the past. Indeed, some receptive fields found by these models fairly evenly spanned the entire temporal window of the STRF. In contrast, the STRFs learnt by temporal prediction models better accounted for the asymmetry and greater power near the present seen in real A1 STRFs (Singer et al., 2018). This is likely because more recent stimuli are generally more predictive than stimulus further in the past, which establishes a temporally asymmetric profile in the model receptive fields.

4.3.4 Model Limitations and Future Investigations

There are limitations to our approach which are important for comparison to studies of functional connectivity in A1. Our definition of best frequency for the model units is determined by the frequency channel with the largest power. A more standard way of determining best frequency would be to find the response to isolated pure tones at a range of frequencies and pick the one that gave the largest response. Variations between different

definitions of the best frequency of the model units will affect the analysis of its functional connectivity. Furthermore, the biological interpretation of the recurrent weight matrix is challenging due to the fundamental differences between artificial neural networks and biological neurons, as the weights of the recurrent matrix does not have a straightforward biological interpretation as mentioned in section 4.3.2. These considerations might compromise the ability of the normative models to explain complex details such as functional connectivity. The ability of the encoding models to capture A1 neural activity was also very modest as the CC_{norm} values were quite low. This may be due to the sparsity constraint, which produces fewer units. A further step to improve model performance might be to apply dropout regularisation to our networks to produce more active hidden units, which would give the neural encoding model a smoother span of inputs. Also, as previously mentioned, our model does not adhere to Dale's principle, that neurons have only excitatory or inhibitory outputs. Hence, a more biology-informed network would be a sensible next step to model both neural responses and connectivity.

Another approach would use some biological data to nudge a normative model towards a more biologically realistic optimum (Młynarski et al., 2021). Furthermore, we could explore alternative evaluation metrics. The metric we used relies on the correlation coefficient between the model and neural response (Schoppe et al., 2016). An alternative to explore might be representational similarity analysis, which evaluates how the representational dissimilarities of the experimental conditions compare with the ones predicted by the model (Kriegeskorte et al., 2008; Diedrichsen et al., 2021). As the normative

models are the same between Chapters 3 and 4, we go into more detail on further developments of these models and their testing in the Discussion chapter, Chapter 5.

4.4 Methods

4.4.1 The Cochleagram and Normative Models

The cochleagram model and normative models, the temporal prediction model and the autoencoder model, were the same as for Chapter 3. Refer to the Methods section in Chapter 3 for details.

4.4.2 Determining Model Spectro-Temporal Receptive Fields and Best Frequencies

For the temporal prediction model (Q = 1) and the sparse autoencoder model (D = 4), we measured the spectro-temporal receptive fields (STRFs) of their hidden units (Fig. 4.3 and Fig. 4.4). We did this by using 500 s of cochleagram input where every value was drawn independently from a Gaussian distribution, with a mean of 0 and standard deviation of 1. This cochleagram input was run through the networks, and the input was then mapped to the response of each model hidden unit using an elastic-net regression model (Willmore et al., 2016; Willmore, 2020). Cross validation was used to set the regularisation strength. We used a time history of 15 ms and 200 ms, hence h = 1 to H = 3 or 50.

$$z[t] = \sum_{h}^{H} \sum_{f}^{F} w_{fh} u_{fh}[t] + b$$

Here, the input $u_{fh}[t] = s_f[t - h + 1]$, where s_f represents the power of the cochleagram s for frequency band f at time t for history h = 1 to H time-bins. $z[t]$ is the linear mapping from the model activity to the neural response $r[t]$. The free parameters are w_{fh}, which represents the STRF and regression coefficients for the corresponding model activity snippet, and b is a bias term.

Units which had a response of zero throughout this stimulus were declared non-active unit, having no STRF, and only active units were displayed in the figures. The temporal prediction model (Q = 1) had 89 active units and the sparse autoencoder model (D = 4) had 146 active units.

For the real data we measured the STRFs of each neuron using the same regression model as above, but mapping from cochleagrams of the natural sound input to the neural response peri-stimulus time histogram binned in 5 ms time-bins. The natural sound inputs and neural responses came from Espejo et al. (2019).

4.4.3 Functional Connectivity

4.4.3.1 A1 Data

We used experimental data from a study where the BFs of neighbouring dendritic spines of pyramidal layer 2/3 neurons in A1 of mice were determined using two-photon imaging (Chen et al., 2011). In particular, we used the data from Figure 3f of the paper. For a more detailed methodology, refer to the paper.

4.4.3.2 Functional Connectivity Analysis

The study by (Chen et al., 2011) performed calcium imaging on the NDMA-dependent calcium current in adjacent pairs of spines on the dendrites of neurons in mouse A1. The frequency tuning properties of each spine in a pair provided an indication of the frequency tuning of a corresponding pair of excitatory neurons pre-synaptic to the recorded neuron. They found that while the apparent frequency tuning of the pre-synaptic neuron pairs could differ greatly, there was a tendency for them to show similar frequency tuning (Figure 4.4). We examined the recurrent connectivity matrices of our normative models for similar patterns of connectivity.

To do this, we took the recurrent weight matrix of the normative model, and set the negative weights to zero, as Chen et al. (2011) were only concerned with excitatory connections. We calculated the BFs of all the units in the model and summed all weights for each pre- and postsynaptic pairing of BFs, providing a matrix of connection weight as function of the BF of the pre- and postsynaptic unit. We then aggregated the BFs into groups of 3 adjacent BFs, summing together the weights in these groups, to provide a more coarsely grained matrix of connection weight as function of the BF of the pre- and postsynaptic unit. The grouping of 500 and 565 Hz only contained 2 BFs.

Having acquired this weight matrix, we then interpreted it as proportional to the probability of a presynaptic neuron with a particular BF synapsing onto a postsynaptic neuron with a particular BF. More exactly, we define $P(A|C)$ as the probability of an excitatory

afferent synapsing from a neuron with a best frequency A to a postsynaptic neuron with best frequency C. We define a similar probability $P(B|C)$ for another presynaptic neuron with best frequency B. We assume that $P(A|C)$ is proportional to the corresponding value w_{AC} within column C in the coarse-grained recurrent weight matrix.

$$P(A|C) = \frac{w_{AC}}{\sum_A w_{AC}}$$

We define $P(C)$ as the probability of sampling a postsynaptic neuron with best frequency C from the neural population. For the models, we calculate this from the best frequency distribution plotted in figure 4.4A and 4.4B, after grouping the frequencies in the same way as for the weight matrix to provide the number of units $N(C)$ in each frequency bin C.

$$P(C) = \frac{N(C)}{\sum_C N(C)}$$

Assuming that $P(A|C)$ is independent of $P(B|C)$, we can then find the joint probability of a postsynaptic unit with best frequency C receiving afferents from presynaptic neurons with BF A and B, which we call $P(A, B, C)$.

$$P(A, B, C) = P(C) \times P(A|C) \times P(B|C)$$

Then,

$$P(A, B) = \sum_C P(A, B, C)$$

$P(A, B)$ can then be compared to the results of Chen et al. (2011). To facilitate the comparison, we binned the results of Figure 3f in Chen et al. (2011) using the same frequency bins we used for the models, taking the centre frequency of the bin as the middle of the three grouped frequencies, and the edge of the bins as the midpoint between the centre frequencies in logarithmic space.

4.4.4 A1 Neuron Modelling

4.4.4.1 Natural Sounds and Neural Recordings

We used a single-unit neural dataset recorded from the primary auditory cortex of six awake ferrets (Espejo et al., 2019). In summary, the ferrets were head-fixed and passively listened to a corpus of natural sounds. Extracellular recordings of the neural response to these sounds were performed using 1-4 independently positioned tungsten microelectrodes (FHC) from A1 or anterior auditory field (AAF). Recordings were performed in a sound-attenuating chamber (Gretch-Ken), and the sound was presented through free-field speakers (Manger W05, 50-35,000 Hz flat gain) located $\pm$ 30° azimuth and 80 cm distant from the ferrets. A standard reference was used to calibrate sound level (Brüel & Kjær 4191). A more detailed description of the experimental settings can be found in the paper. A more detailed description of the experimental settings can be found in the paper.

To enable a robust assessment of the peri-stimulus time histogram, we used the subset of the neural recordings where the natural sound was presented multiple times (10 or 20). This consisted of 18 sound clips of 4 seconds in length. We only used single-unit neural recordings with a firing rate above 1 Hz and a noise ratio below 40 (Linden et al., 2003; Rabinowitz et al., 2011). Of these recordings, 14 sound clips and their neural responses were used as our training set, and the remaining 4 clips and responses as a test set.

4.4.4.2 Neuron Models

We modelled the neural responses to natural sounds by using linear-nonlinear models (LN model), with either a normative model or the cochleagram providing input to the LN model. To do this, we first fit a linear model to the mapping from the input to the neural response. The neural response $r[t]$ comprised the spike counts in consecutive 5 ms bins, averaged over repeats, where t is the time-bin. For input $u_{jh}[t]$ to the linear model, we used either the activity of the hidden layer units from one of the normative models, or we used the cochleagram. More exactly, if the input was from a normative model, then $u_{jh}[t] = a_j[t - h + 1]$, where $a_j[t]$ is the activity of hidden unit *j*, and the input comes from some window of activity history from h = 1 to H time-bins into the past. If the input is the cochleagram, then $u_{jh}[t] = s_j[t - h + 1]$, where $j = f$ and f is the sound frequency band. The linear model is given by:

$$z[t] = \sum_{h}^{H} \sum_{j}^{J} w_{jh} u_{jh}[t] + b$$

where $z[t]$ is the linear model's estimate of neural response $r[t]$. We used a time history of 15 or 200 ms, hence H = 3 or 40, respectively. The free parameters that we fit were the weights w_{jh} and the bias term b. The linear model was fitted using Lasso regression with the following regularisation strengths: 1.00×10^{-1}, 2.00×10^{-2}, 1.17×10^{-2}, 6.84×10^{-3}, 4.00×10^{-3}, 2.34×10^{-3}, 1.37×10^{-3}, 8.00×10^{-4}, 4.68×10^{-4}, 2.74×10^{-4}, 1.60×10^{-4}, 9.36×10^{-5}, 5.41×10^{-5}, 3.20×10^{-5}, 6.40×10^{-6}, 1.28×10^{-6}, 2.56×10^{-7}, and 5.12×10^{-8}.

For each fit of the linear model (that is, for each regularisation strength), the resulting response was then put through a rectified sigmoid non-linearity, which was also fitted to

mapping from the linear model estimate $z[t]$ to the neural response $r[t]$. The sigmoid was given by:

$$z_{\text{nonlin}}[t] = \frac{\rho_1}{1 + e^{\frac{-(z[t]-\rho_3)}{\rho_2}}} + \rho_4$$

where the parameters ρ_1, ρ_2, and ρ_4 were initialised from exponential distributions with reasonable means measured from the input and neural response data. The parameter ρ_3 was initialised from a Gaussian distribution. For each fit of the linear model the sigmoid was intitialised and fitted multiple times and the fit with the higher correlation coefficient between $z_{\text{nonlin}}[t]$ and $r[t]$ was chosen.

The output of the sigmoid was followed by a rectified linearity to provide the final estimate of the neural response:

$$z_{\text{relu}}[t] = \max(0, z_{\text{nonlin}}[t])$$

Given the scarcity of our data, we trained our model using 7-fold cross validation. Twelve of the 14 sound snippets were used to fit the model, both the linear and sigmoid part, for a range of regularisation strengths. The remaining 2 sound snippets were used as a validation set to evaluate the fit for each regularisation strength, using the normalized correlation coefficient, CC_{norm}, as the measure of fit (Hsu et al., 2004; Schoppe et al., 2016). This was repeated for 7 folds, each time using a new validation set. The best performing regularisation strength was deemed the one with the best fit on average over all 7 validation sets. This regularisation value was then chosen for the final training round on all 14 sounds in the training set, and evaluated on the held out 4 sounds as the test set using CC_{norm}.

CC_{norm} is the correlation coefficient normalised to the theoretical optimum given the intrinsic noise of neurons, and is defined as:

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

where y is the neural firing rate, $\hat{y}$ is the predicted firing rate by the model, and SP is the signal power explained given by:

$$SP = \frac{Var(\sum_{n=1}^{N} R_n) - \sum_{n=1}^{N} Var(R_n)}{N(N-1)}$$

where R_n is the recorded neural activity of trial n, and N is the total number of trials.

4.4.5 Modelling Implementation

All work was done using custom code implemented in MATLAB and Python and run on NVIDIA GeForce GTX-1080 GPUs.

Chapter 5. Discussion

5.1 Summary of Results

This thesis investigated temporal prediction as a potential governing principle of the auditory system. First of all, we show that temporal prediction can explain several aspects of the most peripheral auditory neural structure: the cochlea. The tuning characteristics of the hidden units of the model resemble gammatone filters similar to the tuning properties of inner hair cells and auditory nerve fibres. Each hidden unit is sharply tuned in frequency, and all together the model units cover a wide span in frequency space with a similar logarithmic distribution to that found in the cochlea. Second, we demonstrate how a decision model based on the representations of a recurrent temporal prediction model is able to predict behaviour of mice performing several auditory discrimination tasks across two stimulus dimensions of sound. And, finally, we show that the same recurrent temporal prediction model somewhat replicates spectro-temporal receptive fields (STRFs) and local functional connectivity of pyramidal neurons in the auditory cortex of mice, and to some degree can predict A1 neural responses to natural sounds. Overall, we show that temporal prediction models can account for experimental findings in peripheral and central auditory structures. Our work replicates diverse aspects of the auditory system, such as physiological tuning, behaviour, anatomy and connectivity.

Throughout this thesis, we have compared our temporal prediction models to a corresponding autoencoder model to assess the role of compressed information. This compressive control fails to replicate cochlear features and spectro-temporal receptive fields

and functional connectivity in A1. However, its performance rivals the temporal prediction predictions for the auditory discrimination tasks and in predicting A1 neural responses to natural sounds.

5.2 Temporal Prediction as a Normative Principle

The work presented shows that the theory of temporal prediction may underlie features of the auditory system. The spontaneous emergence in temporal prediction models of a wide range of biological properties, without applying strict and explicit brain-like priors, is an argument in favour of temporal prediction.

A possible complication is that other non-normative constraints compete with the normative objective to determine the structure of biological representations. For example, the optimal theoretical solution might be unattainable biologically due to real-world limitations, and the brain may have arrived at an intermediate solution that, while effective, is sub-optimal. We might see this for our temporal prediction model of the cochlea, with the very fast onset of the impulses produced by the model being unrealisable in a biological system and so real cochlear gammatone filters have a slower onset.

Our work can be compared to other normative models of the auditory system. The temporal prediction model of the cochlea has some advantages over previous normative models that were optimised for efficient coding (Lewicki, 2002; Smith and Lewicki, 2006). These efficient coding models had gammatone filters that were symmetric (Lewicki, 2002), or required a temporally acausal iterative settling process to reach their representation of a

sound (Smith and Lewicki, 2006). Our model is in some respects more biologically plausible, as our model units display temporally asymmetric receptive fields – similar to the response properties of auditory nerve fibres – and acts causally on the sound without the need for an iterative settling process to arrive at its encoded representation.

The temporal prediction cortical model can be compared to a number of different normative models of spectro-temporal receptive fields, as we detail in discussion section 4.3.1 of Chapter 4 (Klein et al., 2003; Blättler and Hahnloser, 2011; Ngiam et al., 2011; Zhao and Zhaoping, 2011; Carlson et al., 2012; Carlin and Elhilali, 2013; Brito and Gerstner, 2016; Kozlov and Gentner, 2016; Młynarski and McDermott, 2017). One notable normative model is the work of Kell et al. (2018), who optimised a deep neural network for speech recognition and classification of music genres. They found that this model could explain aspects of both human behaviour and cortical responses. However, this approach is, in some sense, not a true normative model, but a phenomenological model of behaviour which provides a basis for comparison with neural activity. It provides an account of neural responses in terms of a behavioural task, but does not necessarily elucidate the principles which govern behaviour on that task. Thus, this approach remains a fundamentally descriptive model of the brain and behaviour, rather than directly shedding light on evolutionary principles that have guided the development of the brain.

5.3 Arguments Supporting Temporal Prediction in the Auditory System

The identification of normative principles that guide the functional organization of brain systems must consider: 1) the tasks an animal has to perform, 2) the biological constraint of neurons, such as wiring and metabolic restrictions, and 3) the natural environment of the animal (Simoncelli and Olshausen, 2001). Some normative theories emphasise certain of these aspects above others. Coding schemes such as efficient and sparse coding mainly address how to find a code that optimally represents the entirety of sensory stimuli in efficient ways, while assigning less importance to the behavioural relevance of encoded information. Temporal prediction addresses all of these considerations to some extent as we will explain below, and we argue that there are several *a priori* reasons why temporal prediction may be an appropriate normative principle for the brain.

First of all, temporal prediction is useful because actions in response to sensory input occur with some delay due to delays in sensory processing, action selection and motor activation. By having sensory representations that are optimised to predict the future of sensory input, the nervous system can obtain a better estimate of what the state of the world will be like when actions actually take place. For example, to catch a flying ball one needs to predict its trajectory, and not merely estimate its present location. In an evolutionary context, better estimations of the state of the future, even on the millisecond scale, will still provide an advantage in fitness to the organism.

Second, it has been argued that, by trying to find variables that efficiently map from the recent past to the immediate future of the input, temporal prediction may encode the underlying causes of sensory stimuli in the environment (Bialek et al., 2001). For example, one could speculate that to efficiently predict the future spectrogram of sounds at a busy social event, one might need to learn representations that separate out different speakers and capture some aspects of the dynamics of language. Perhaps, with a sufficiently sophisticated network, more advanced than we are currently using, simply optimising for prediction on diverse natural soundscapes may enable extraction of such underlying aspects of the sensory world.

Third, temporal prediction is a way for the brain to remove potentially superfluous information, namely information that is not predictive. Aspects of the world that do not influence the future will not likely be of relevance to the survival of an animal. Hence, temporal prediction helps winnow down the vast amount of incoming information to a smaller subset from within which the even smaller subset of behaviourally-relevant information can more easily be found. This is in contrast to efficient coding where, in many cases, the entire stimulus is represented, potentially including unimportant information. Indeed, a 'lossy' approach to sensory information processing, where only a selection of features of relevance is encoded and transmitted, has been argued from an evolutionary perspective to be more resource efficient in many environments (Marzen and DeDeo, 2017).

Fourth, an advantage of temporal prediction is that many of its implementations are fairly biologically realistic and operate causally in time and without the need for iterative

settling processes to find the encoded representation. In contrast, some models that implement sparse or efficient coding require an iterative process to find the most efficient code (Olshausen and Field, 1996; Rao and Ballard, 1999; Smith and Lewicki, 2006; Carlson et al., 2012). This is especially impractical in the auditory system due to the fast dynamics of sound.

Finally, an advantage of the temporal prediction framework is that it is an unsupervised learning process which can make use of ubiquitous training data, as all sensory information can serve as training data for the self-organization of the brain. While this is an advantage also held by other unsupervised approaches, such as efficient coding, it is not an advantage held by supervised learning or reinforcement learning. Supervised learning models, such as object or speech recognition models, have been used as normative models of sensory systems (Kell et al., 2018; Cadena et al., 2019). However, supervised learning requires extensive labelled data, and it is arguable how much external labelling of inputs, be it reinforcement signals or more precise labelling, is provided to animals as they develop. Certainly, there is substantially less labelled sensory input than unlabelled sensory input.

In addition to these *a priori* arguments, there is now substantial empirical evidence for temporal prediction models. Work that preceded this thesis has shown that temporal prediction can explain aspects of receptive fields across sensory modalities, and has replicated a number of receptive fields properties in the visual system and auditory system (Singer et al., 2018, 2019). Notably, these temporal prediction models capture the temporal asymmetries in real receptive fields, which most previous models have been unable to

explain. This thesis shows that temporal prediction can similarly account for temporal asymmetry in the responses of the auditory nerve (Chapter 2). Additionally, we have shown in Chapter 4 that temporal prediction can explain aspects of functional connectivity between neurons and also, to a lesser extent, neural responses to natural sounds. We also demonstrated in Chapter 3 that temporal prediction can to a degree explain behaviour associated with a sensory discrimination task, which suggests that temporally predictive information can serve as a useful representational space for perception. This is consistent with an argument which argues that insightful models of the brain must capture the flow of information in the causal direction, going from stimulus, to neural representation and eventually to behaviour (Kriegeskorte and Douglas, 2019). Hence, we have shown at a high level of abstraction that temporal prediction may be a useful principle for behavioural models of sensory decision making.

5.4 Implementation of Temporal Prediction in the Auditory System

An important issue raised by the temporal prediction hypothesis is the identification of processes that optimise the sensory systems to perform prediction. Crucially, are neural tuning properties optimised by evolutionary processes, or during the development of individual animals? We have shown that processing in the peripheral auditory system is consistent with some degree of optimisation for the extraction of predictive information from raw sound waves. However, it is unclear how the cochlea would come to perform temporal prediction. Studies on the cochlea indicate that its development is predominantly determined

genetically. Transient bilateral conductive hearing loss of gerbil pups did not impair the auditory periphery (Caras and Sanes, 2015). Displacing or completely removing the malleus bilaterally in gerbil pups resulted in a slight degradation of synapses between inner hair cells and auditory nerve fibres, but the sharpness of the cochlear filters remained intact, indicating undamaged outer hair cells (Ye et al., 2021). Thus, the biological realisation of the cochlea is largely hard-coded genetically, and optimisation for any normative principle governing it, such as temporal prediction, would likely occur by natural selection over evolutionary timescales.

The central auditory system is subject to experience-induced changes in sensory input, both throughout early development and adulthood (Dahmen and King, 2007). In contrast to the auditory periphery, conductive hearing loss has been shown to cause changes to the central auditory pathway (Tucci and Rubel, 1985; Chhan et al., 2017; Ye et al., 2021). Animal studies also point towards the existence of a critical period in the postnatal development of young animals wherein the sensory environment shapes the sensitivity to auditory features for neurons in the central auditory system (Sanes and Constantine-Paton, 1985; Zhang et al., 2001; Seidl and Grothe, 2005; De Villers-Sidani et al., 2007; Barkat et al., 2011; Caras and Sanes, 2015). Plasticity following this period is possible, but acts within the constraints set by the critical window (Knudsen, 2004). A1 is therefore highly plastic and subject to considerable shaping during development in response to the sensory environment. Thus, optimisation for temporal prediction in the central auditory system, if it occurs, is likely a consequence of plasticity during a critical period when the animal is young and learning in older animals, as well as perhaps some hard-wiring.

If the cortical optimisation for temporal prediction occurs by learning, this suggests that cortical neurons, at least at some stage of development, should show a representation of predictions or prediction errors. There is some evidence for this that comes from measurements of cortical responses to oddball stimulus sets, both non-invasive measurements in humans (Näätänen et al., 2007; Garrido et al., 2009; Auksztulewicz and Friston, 2015), and more precise invasive measurements in animals (Ulanovsky et al., 2003; Garrido et al., 2009; Rubin et al., 2016; Carbajal and Malmierca, 2018) which includes at subcortical levels (Parras et al., 2017). Oddball stimulus sets are regular trains of stimuli where a randomly-chosen small subset of the stimuli are different, provoking prediction error responses. Arguably, more powerful evidence comes from omission stimulus sets, where a randomly-chosen subset of stimuli is omitted, rather than being different, revealing 'responses' representing prediction or prediction error around the time of the omitted stimuli. Responses to omitted stimuli have been seen in both non-invasive measurements in humans (Yabe et al., 1997; Bendixen et al., 2009), and more precise invasive measurements in animals (Gao et al., 2009; Lakatos et al., 2013). However, while these studies are remarkable, the consensus is that further studies are needed and that the theories of predictive processing in the auditory system need substantial clarification (Heilbron and Chait, 2018; Denham and Winkler, 2020).

Furthermore, perhaps it is also possible that temporal prediction could be learnt in a manner that does not require prediction error neurons. Homeostatic and spike-time dependent plasticity rules have been argued to be sufficient for a recurrent neural network to perform gradient descent and learn to predict its stimuli (Galtier and Wainrib, 2013). Also,

individual neurons have been posited to be predictive units that "learn" to predict their own activity using a biologically plausible learning rule (Luczak et al., 2022).

Here, we have focused on prediction in the auditory system, but there is experimental evidence for the representation of predictive information in the retina (Palmer et al., 2015; Salisbury and Palmer, 2016), and many models of visual processing involve prediction, such as in temporal prediction or predictive coding (Rao and Ballard, 1999; Singer et al., 2018, 2019). Indeed, the importance of prediction is not limited to sensory systems; ideas involving prediction also have great relevance for other brain areas, such as those involved in motor control (Davidson and Wolpert, 2005). Whether the processes that implement predictive mechanisms across different brain systems have similarities, or are quite distinct, is an interesting question to consider.

5.5 Model Limitations and Future Investigations

While being a sensible starting point, there are a number of extensions to our model that are worth considering. To increase the biological fidelity of our findings, it might be sensible to incorporate more explicitly biological aspects into our models. Although artificial neural networks are simple to train, powerful, and widely used within and outside of neuroscience in recent time (Krizhevsky et al., 2012; Sussillo, 2014; Kriegeskorte, 2015; Yamins and DiCarlo, 2016; Barak, 2017; Richards et al., 2019), they are vastly simplified versions of their biological inspiration. Units in artificial neural networks are typically modelled as point neurons, without any explicit representations of neural features, such as spikes, dendritic computations or ion channel dynamics. Converting our models from rate-coded networks, where unit activity

represents firing rate, into spiking neural networks can serve as an initial step towards a more biologically realistic model. Artificial neural networks use backpropagation as a powerful learning algorithm, but the application of backpropagation to spiking neural networks is not straightforward, as the backwards pass would involve the un-differentiable activity of units with binary outputs (spike or no spike). Recent work has made improvements in the feasibility of training spiking neural networks, making this a more attractive research option (Huh and Sejnowski, 2018; Shrestha and Orchard, 2018; Zenke and Vogels, 2021). Furthermore, while backpropagation is the standard learning algorithm for artificial neural networks, the learning method in its conventional form seems unfeasible in the brain. A lot of attention has been dedicated to increasing the biological viability of backpropagation, which might also produce more biologically realistic models (Bengio et al., 2015, 2017; Lillicrap et al., 2016; Whittington and Bogacz, 2019).

Another aspect of our temporal prediction models that is not biological is that they do not obey Dale's law. This law states that a neuron releases the same neurotransmitter at all axon terminals and that neurons are either excitatory or inhibitory as a consequence (Eccles, 1986). Approximately 80% of neurons in the cerebral cortex are excitatory, while 20% are inhibitory (Wonders and Anderson, 2006; Marín, 2012). Implementing this ratio explicitly in our recurrent layer for our A1 models might improve its capacity for replicating biological findings, such as an improved ability to predict A1 responses to natural sounds. Another important aspect of neurons is their stochasticity; while we have explored this slightly by the noise applied to the input in our temporal prediction models, most notably in Chapter 2, there is much more that could be explored. It has been suggested that neural variability in sensory systems is a way of representing the uncertainty in the sensory environment and the

distribution of stimulus statistics (Hoyer and Hyvärinen, 2003; Orbán et al., 2016). Perhaps such an idea could have relevance for models positing the prediction of the distribution of possible futures by the brain.

The impulse responses of the cochlear temporal prediction units, presented in Chapter 2, qualitatively resembled the gammatone-filters of the cochlea, but had a sharper initial rise. The current model is a feedforward network without any interactions between model hidden units. In the current model, the hidden units represent an abstraction of hair cells and a distinct segment along the basilar membrane. Adding recurrency to the hidden layer might enable us to capture phenomena that involve interaction across the length of basilar membrane, such as two-tone suppression (Suga et al., 1963; Frishkopf, 1964). It might also mitigate the initial sharp rise of the feedforward impulse responses. Another avenue to explore for the cochlear temporal prediction model would be to include a latency period between the past and future input used for training and see if predicting further into the future and disregarding the immediately upcoming stimulus yields more gammatone-like representations. Applying more biologically realistic features to the model such as inner hair cell adaptation might also remove the initial click, and produce a model closer to the cochlea (Eatock, 2000).

In general, further investigating the temporal extent of the time windows used for the pairing of the past and the future in our temporal prediction models is a possibly fruitful avenue. This line of work has been explored for the recurrent temporal prediction model of A1 in Chapter 3, and would be good to do for the cochlear model. The cochlear model predicts

only 0.73 ms into the future, which is rather short. However, this modest future prediction may still have importance in an evolutionary context, where even a millisecond advantage in guiding action could make a difference to survival, notwithstanding the other advantages of temporal prediction that we outlined in Section 5.3, such as extracting underlying variables. Other combinations of the time windows used for the span of the past used for input and span of future that is predicted, might result in model properties that more accurately captures features of the auditory system.

Our recurrent temporal prediction model overall produced good predictions of the behaviour of mice performing an auditory discrimination task, but was less successful in predictions of A1 neural responses to natural sounds. As a first approach, we can apply dropout regularisation to our recurrent model (Hinton et al., 2012; Srivastava et al., 2014). This will produce more active units and result in a more smoothly distributed representation by the temporal prediction model. This could potentially improve the mapping between the natural sounds and measured neural responses. It might also be interesting to investigate how important the recurrency is for the behavioural decision model, and whether it is strictly needed for good estimates of auditory behaviour or neural activity.

Another interesting avenue to explore would be hierarchical temporal prediction models (Singer et al., 2019), where multiple temporal models are stacked with each layer of the temporal prediction model predicting the future hidden unit activity of the layer below. For example, currently we approximate the input to our A1 model with a cochleagram, but alternatively we could use our cochlear model as the initial input stack, and then train

subsequent stacks for temporal prediction on the units of the cochlear model. A hierarchical temporal prediction model of the visual system produced receptive fields that resembled real receptive fields at successive stages along the visual pathway (Singer et al., 2019) – from the retina, to simple cells, to complex cells, to tuning properties found in higher cortical middle temporal visual area. This suggest that such a modelling approach might recover tuning features along the auditory pathway.

Besides applying such adjustments to the model, it might be interesting to train our models with a dataset with different characteristics to the one we are currently using. Our models might recover receptive fields, hearing ranges and perceptual properties for a wide variety of species if we can acquire datasets that represent their natural soundscape.

While a particular normative theory such as temporal prediction offers a simple explanation and a convenient abstraction of a complicated system, the framework of normative models might be more nuanced. It has also been suggested that different brain areas are optimised for different cost functions, and that a normative goal within a neural system might change over time (Marblestone et al., 2016). Hence, brain systems might be the solution of optimisation for numerous principles. For example, some of our models are implemented with a sparsity constraint on the weights, which in some ways relates to aspects of other normative principles such as efficient coding.

It would also be interesting to explore the broader possible role of temporal prediction in the brain. For example, temporal prediction might also improve learning for a

reinforcement learning agent. At the moment, reinforcement learning algorithms have proven to be very proficient at solving problems and are able to outperform humans in highly complex games (Silver et al., 2016; Vinyals et al., 2019). However, these algorithms are computationally very expensive. Representation learning can often help reinforcement learning problems by transforming information about the environment into a more behaviourally relevant form and by ignoring redundant information (Bengio et al., 2013). Augmenting a reinforcement learning agent with temporal prediction might make learning more tractable and improve training time.

5.6 Experimental Predictions and Model Testing

The current temporal prediction models used in Chapters 2-4 made experimental predictions of neural tuning, functional connectivity, neural responses to natural sounds, and animal behaviour. We tested the predictions against some of the available data, however there remain a fair number of untested predictions. The new model developments outlined above in Section 5.5 may also enable further predictions to be made.

For the model in Chapter 2, the main experimental prediction we tested was the tuning sharpness of the auditory nerve. However, the model also makes predictions regarding the relationship of distance along the basilar membrane and the preferred frequency tuning of hair cells (Fig. 2.4). A more detailed recurrent version of the temporal prediction model in Chapter 2 could perhaps predict non-linear characteristics of the cochlea, such as two-tone inhibition characteristics, and perhaps predict some characteristics that have not yet been experimentally measured.

For the model in Chapter 3, the main experimental prediction that we tested was the response to novel sounds with characteristics between the two sounds that the mice were trained to judge as safe or unsafe. This model could potentially make predictions of other novel sounds in other sound spaces defined by whether they are safe/unsafe, including spaces composed of morphs between different natural sounds.

For the model in Chapter 4, the main experimental predictions that were compared to cortical data were spectro-temporal receptive fields (STRFs), functional connectivity and responses to natural sounds. We only qualitatively compared the model STRFs and the neural STRFs, but future work could also perform a more quantitative comparison, perhaps using forms of dimensionality reduction or tuning width measurements similar to Singer et al. (2018). Few studies have measured functional connectivity in auditory cortex, unlike visual cortex (Ko et al., 2011; Cossell et al., 2015; Iacaruso et al., 2017). The recurrent matrix in Chapter 4 could make further predictions regarding functional connectivity; for example, the matrix suggests that low frequency neurons should be more connected to other neurons than higher frequency units are. As more functional connectivity results are produced for A1, they can be tested against the model predictions. Finally, for the responses to natural sounds, new Bayesian frameworks for producing models are emerging, where the fitting procedure combines normative principles with the use of neural data (Młynarski et al., 2021). This is unlike the two-stage process we used where an LN model is appended to the already trained normative model. These new methods are more rigorous, and should avoid the problems we had in Chapter 4 in determining the appropriate time span of the LN model. The more

advanced models that we outlined above in Section 5.5, for example obeying Dale's law, or using dropout regularisation for a smoother representation, may even further improve the capacity of the models to make experimental predictions.

On top of the experimental predictions outlined above, there are also experimental predictions relating to the implementation of temporal prediction, as we outlined in section 5.4. For example, temporal prediction models, if they are optimised by learning, posit that there should be a representation of prediction error in auditory cortex.

5.7 Conclusion

Overall, we show that temporal prediction is a promising explanatory framework for the auditory system. Our approach relies on the rationale that if the brain has been refined by evolutionary or learning mechanisms to perform certain tasks, models optimised for these objectives should converge on similar representations and structures. The central challenge in this methodology is to identify the correct objective and appropriate implementation nuances needed to address biological constraints for the system described.

Based on the work presented in this thesis and previous studies, we argue that the physiology of the auditory system may be at least partially guided by optimisation for temporal prediction. We emphasise that our findings do not exclude other normative theories, as numerous studies exist in support of these alternative hypotheses in other brain areas. The brain might be optimised for several distinct cost functions in different brain

regions, and these cost functions might even change throughout time (Marblestone et al., 2016).

However, we argue that temporal prediction has many intrinsic advantages that makes it appealing for sensory neuroscience (see 5.2), and have shown that temporal prediction can explain features of the peripheral and central auditory system and of auditory behaviour. We hope other researchers might find value in our work, and that temporal prediction can help clarify our understanding of the brain.

Bibliography

Aertsen AMHJ, Olders JHJ, Johannesma PIM (1981) Spectro-temporal receptive fields of auditory neurons in the grassfrog - III. analysis of the stimulus-event relation for natural stimuli. Biol Cybern 39:195–209.

Ahissar M, Hochstein S (1997) Task difficulty and the specificity of perceptual learning. Nature 387:401–406.

Ahissar M, Hochstein S (2004) The reverse hierarchy theory of visual perceptual learning. Trends Cogn Sci 8:457–464.

Asokan MM, Williamson RS, Hancock KE, Polley DB (2018) Sensory overamplification in layer 5 auditory corticofugal projection neurons following cochlear nerve synaptic damage. Nat Commun 9:1–10 Available at: http://dx.doi.org/10.1038/s41467-018-04852-y.

Atencio CA, Sharpee TO (2017) Multidimensional receptive field processing by cat primary auditory cortical neurons. Neuroscience 359:130–141 Available at: http://dx.doi.org/10.1016/j.neuroscience.2017.07.003.

Atencio CA, Sharpee TO, Schreiner CE (2008) Cooperative Nonlinearities in Auditory Cortical Neurons. Neuron 58:956–966.

Atencio CA, Sharpee TO, Schreiner CE (2012) Receptive field dimensionality increases from the auditory midbrain to cortex. J Neurophysiol 107:2594–2603.

Attneave F (1954) Some informational aspects of visual perception. Psychol Rev 61:183–193.

Auksztulewicz R, Friston K (2015) Attentional enhancement of auditory mismatch responses: A DCM/MEG study. Cereb Cortex 25:4273–4283.

Bajo VM, King AJ (2012) Cortical modulation of auditory processing in the midbrain. Front Neural Circuits 6:1–12.

Bamiou DE, Sisodiya S, Musiek FE, Luxon LM (2007) The role of the interhemispheric pathway in hearing. Brain Res Rev 56:170–182.

Bandyopadhyay S, Shamma SA, Kanold PO (2010) Dichotomy of functional organization in the mouse auditory cortex. Nat Neurosci 13:361–368.

Barak O (2017) Recurrent neural networks as versatile tools of neuroscience research. Curr Opin Neurobiol 46:1–6 Available at: http://dx.doi.org/10.1016/j.conb.2017.06.003.

Barkat TR, Polley DB, Hensch TK (2011) A critical period for auditory thalamocortical connectivity. Nat Neurosci 14:1189–1996.

Barlow HB (1959) Sensory mechanisms, the reduction of redundancy, and intelligence. In: NPL Symposium on the Mechanization of Thought Process., pp 535–539.

Barlow HB (1961) Possible Principles Underlying the Transformations of Sensory Messages. Sens Commun:216–234.

Beckman M, Edwards J (2010) PhonBank Japanese Paidologos Corpus.

Bendixen A, Schröger E, Winkler I (2009) I heard that coming: Event-related potential evidence for stimulus-driven prediction in the auditory system. J Neurosci 29:8447–8451.

Bengio Y, Courville A, Vincent P (2013) Representation learning: A review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35:1798–1828.

Bengio Y, Lee D-H, Bornschein J, Mesnard T, Lin Z (2015) Towards Biologically Plausible Deep Learning. Available at: http://arxiv.org/abs/1502.04156.

Bengio Y, Mesnard T, Fischer A, Zhang S, Wu Y (2017) STDP-Compatible Approximation of Backpropagation in an Energy-Based Model. Neural Comput 29:555–577 Available at: https://doi.org/10.1162/NECO_a_00934.

Berkes P, Wiskott L (2005) Slow feature analysis yields a rich repertoire of complex cell properties. J Vis 5:9–9.

Bialek W, Nemenman I, Tishby N (2001) Predictability, Complexity, and Learning. Neural Comput 13:2409–2463.

Bishop C (2006) Pattern Recognition and Machine Learning, 1st ed. Springer-Verlag New York.

Bishop CM (1995) Training with Noise is Equivalent to Tikhonov Regularization. Neural Comput 7:108–116.

Bizley JK, Nodal FR, Nelken I, King AJ (2005) Functional organization of ferret auditory cortex. Cereb Cortex 15:1637–1653.

Blättler F, Hahnloser R (2011) An Efficient Coding Hypothesis Links Sparsity and Selectivity of Neural Responses. PLoS One 101:1435–1439.

Bourk TR, Mielcarz JP, Norris BE (1981) Tonotopic organization of the anteroventral cochlear nucleus of the cat. Hear Res 4:215–241.

Brand A, Urban A, Grothe B (2000) Duration tuning in the mouse auditory midbrain. J Neurophysiol 84:1790–1799.

Brito CSN, Gerstner W (2016) Nonlinear Hebbian Learning as a Unifying Principle in Receptive Field Formation. PLoS Comput Biol 12:1–24.

Brownell WE, Bader CR, Bertrand D, De Ribaupierre Y (1985) Evoked mechanical responses of isolated cochlear outer hair cells. Science (80-) 227:194–196.

Brugge JF (1992) An Overview of Central Auditory Processing. In, pp 1–33. Springer, New York, NY.

Cadena SA, Denfield GH, Walker EY, Gatys LA, Tolias AS, Bethge M, Ecker AS (2019) Deep convolutional models improve predictions of macaque V1 responses to natural images. PLoS Comput Biol 15:1–27.

Cadieu CF, Hong H, Yamins DLK, Pinto N, Ardila D, Solomon EA, Majaj NJ, DiCarlo JJ (2014) Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. PLoS Comput Biol 10.

Camalier CR, D'Angelo WR, Sterbing-D'Angelo SJ, De La Mothe LA, Hackett TA (2012) Neural latencies across auditory cortex of macaque support a dorsal stream supramodal timing advantage in primates. Proc Natl Acad Sci U S A 109:18168–18173.

Carandini M, Demb JB, Mante V, Tolhurst DJ, Dan Y, Olshausen BA, Gallant JL, Rust NC (2005) Do we know what the early visual system does? J Neurosci 25:10577–10597.

Caras ML, Sanes DH (2015) Sustained perceptual deficits from transient sensory deprivation. J Neurosci 35:10831–10842.

Carbajal G V., Malmierca MS (2018) The Neuronal Basis of Predictive Coding Along the Auditory Pathway: From the Subcortical Roots to Cortical Deviance Detection. Trends Hear 22:1–33.

Carlin MA, Elhilali M (2013) Sustained Firing of Model Central Auditory Neurons Yields a Discriminative Spectro-temporal Representation for Natural Sounds. PLoS Comput Biol 9.

Carlson NL, Ming VL, DeWeese MR (2012) Sparse codes for speech predict spectrotemporal receptive fields in the inferior colliculus. PLoS Comput Biol 8.

Carney LH, Yin TCT (1988) Temporal coding of resonances by low-frequency auditory nerve fibers: single-fiber responses and a population model. J Neurophysiol 60:1653–1677.

Cash S, Yuste R (1999) Linear summation of excitatory inputs by CA1 pyramidal neurons. Neuron 22:383–394.

Chalk M, Marre O, Tkačik G (2018) Toward a unified theory of efficient, predictive, and sparse coding. Proc Natl Acad Sci U S A 115:186–191.

Chambers AR, Hancock KE, Sen K, Polley DB (2014) Online stimulus optimization rapidly reveals multidimensional selectivity in auditory cortical neurons. J Neurosci 34:8963–8975.

Chase SM, Young ED (2008) Cues for sound localization are encoded in multiple aspects of spike trains in the inferior colliculus. J Neurophysiol 99:1672–1682.

Chechik G, Anderson MJ, Bar-Yosef O, Young ED, Tishby N, Nelken I (2006) Reduction of Information Redundancy in the Ascending Auditory Pathway. Neuron 51:359–368.

Chechik G, Nelken I (2012) Auditory abstraction from spectro-temporal features to coding auditory entities. Proc Natl Acad Sci U S A 109:18968–18973.

Chen C, Krueger-Burg D, De Hoz L (2019) Wide sensory filters underlie performance in memory-based discrimination and generalization. PLoS One 14:1–29.

Chen C, Trinh F, Harper N, de Hoz L (2021) Natural sound characteristics explain perceptual categorization. bioRxiv Neurosci Available at: http://biorxiv.org/cgi/content/short/2021.09.29.462467.

Chen X, Leischner U, Rochefort NL, Nelken I, Konnerth A (2011) Functional mapping of single spines in cortical neurons in vivo. Nature 475:501–505.

Chhan D, McKinnon ML, Rosowski JJ (2017) Identification of induced and naturally occurring conductive hearing loss in mice using bone conduction. Hear Res 346:45–54 Available at: http://dx.doi.org/10.1016/j.heares.2017.02.001.

Clayton KK, Williamson RS, Hancock KE, Tasaka G, Mizrahi A, Hackett TA, Polley DB (2021) Auditory Corticothalamic Neurons Are Recruited by Motor Preparatory Inputs. Curr Biol 31:310-321.e5.

Cossell L, Iacaruso MF, Muir DR, Houlton R, Sader EN, Ko H, Hofer SB, Mrsic-Flogel TD (2015) Functional organization of excitatory synaptic strength in primary visual cortex. Nature 518:399–403.

Cusack R, Carlyon R (2004) Auditory perceptual organization inside and outside the laboratory. In: Ecological Psychoacoustics (G NJ, ed), pp 15–48. Elsevier.

Dahmen JC, King AJ (2007) Learning to hear: plasticity of auditory cortical processing. Curr Opin Neurobiol 17:456–464.

Davidson PR, Wolpert DM (2005) Widespread access to predictive models in the motor system: A short review. J Neural Eng 2.

Davis KA, Ramachandran R, May BJ (2003) Auditory processing of spectral cues for sound localization in the inferior colliculus. JARO - J Assoc Res Otolaryngol 4:148–163.

de Boer E (1978) On cochlear encoding: Potentialities and limitations of the reverse-correlation technique. J Acoust Soc Am 63:115 Available at: http://scitation.aip.org/content/asa/journal/jasa/63/1/10.1121/1.381704.

de Hoz L, Nelken I (2014) Frequency Tuning in the Behaving Mouse: Different Bandwidths for Discrimination and Generalization Vicario DS, ed. PLoS One 9:e91676.

De Villers-Sidani E, Chang EF, Bao S, Merzenich MM (2007) Critical period window for spectral tuning defined in the primary auditory cortex (A1) in the rat. J Neurosci 27:180–189.

DeCharms RC, Blake DT, Merzenich MM (1998) Optimizing sound features for cortical neurons. Science (80-) 280:1439–1443.

Denham SL, Winkler I (2020) Predictive coding in auditory perception: challenges and unresolved questions. Eur J Neurosci 51:1151–1160.

DiCarlo JJ, Maunsell JHR (2005) Using neuronal latency to determine sensory-motor processing pathways in reaction time tasks. J Neurophysiol 93:2974–2986.

Diedrichsen J, Berlot E, Mur M, Schütt HH, Shahbazi M, Kriegeskorte N (2021) Comparing representational geometries using whitened unbiased-distance-matrix similarity. Neurons, Behav Data Anal Theory:1–25.

Dong DW, Atick JJ (1995) Statistics of natural time-varying images. Netw Comput Neural Syst 6:345–358 Available at: http://www.ncbi.nlm.nih.gov/pubmed/11014668.

Douglas RJ, Koch C, Mahowald M, Martin KAC, Suarez HH (1995) Recurrent excitation in neocortical circuits. Science (80-) 269:981–985.

Douglas RJ, Martin KAC (2004) Neuronal circuits of the neocortex. Annu Rev Neurosci 27:419–451.

Douglas RJ, Martin KAC, Whitteridge D (1989) A Canonical Microcircuit for Neocortex. Neural Comput 1:480–488.

Doya K (2008) Modulators of decision making. Nat Neurosci 11:410–416.

Eatock RA (2000) Adaptation in Hair Cells. Annu Rev Neurosci 23:285–314.

Eccles JC (1986) Chemical transmission and Dale's principle. Prog Brain Res 68:3–13.

Ehret G (1975) Frequency and intensity difference limens and nonlinearities in the ear of the housemouse (Mus musculus). J Comp Physiol 102:321–336.

Ehret G (1992) Categorical perception of mouse-pup ultrasounds in the temporal domain. Anim Behav 43:409–416.

Ehret G, Riecke S (2002) Mice and humans perceive multiharmonic communication sounds in the same way. Proc Natl Acad Sci U S A 99:479–482.

Eliasmith C, Anderson CH (1991) Neural Engineering.

Espejo ML, Schwartz ZP, David S V. (2019) Spectral tuning of adaptation supports coding of sensory context in auditory cortex.

Evans EF (1975) Handbook of Sensory Physiology.

Fay RR (1974) Auditory frequency discrimination in vertebrates. J Acoust Soc Am 56:206–209.

Fay RR, Popper AN (2000) Evolution of hearing in vertebrates: The inner ears and processing. Hear Res 149:1–10.

Ferster D, Lindström S (1985) Augmenting responses evoked in area 17 of the cat by intracortical axon collaterals of cortico-geniculate cells. J Physiol 367:217–232.

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

Florentine M (1983) Intensity discrimination as a function of level and frequency and its relation to high-frequency hearing. J Acoust Soc Am 74:1375–1379.

Földiák P (1991) Learning Invariance from Transformation Sequences. Neural Comput 3:194–200 Available at: https://doi.org/10.1162/neco.1991.3.2.194.

Fontaine B, Goodman DFM, Benichoux V, Brette R (2011) Brian Hears: Online Auditory Processing Using Vectorization Over Channels. Front Neuroinform 5:1–9 Available at: http://journal.frontiersin.org/article/10.3389/fninf.2011.00009/abstract.

Frishkopf LS (1964) Excitation and Inhibition of Primary Auditory Neurons in the Little Brown Bat Published by the Acoustical Society of America. 1016:1–2.

Friston K (2003) Learning and inference in the brain. Neural Networks 16:1325–1352.

Gaese BH, King I, Felsheim C, Ostwald J, von der Behrens W (2006) Discrimination of direction in fast frequency-modulated tones by rats. J Assoc Res Otolaryngol 7:48–58.

Galtier MN, Wainrib G (2013) A biological gradient descent for prediction through a combination of STDP and homeostatic plasticity. Neural Comput 25:2815–2832.

Games KD, Winer JA (1988) Layer V in rat auditory cortex: Projections to the inferior colliculus and contralateral cortex. Hear Res 34:1–25.

Gao L, Meng X, Ye C, Zhang H, Liu C, Dan Y, Poo MM, He J, Zhang X (2009) Entrainment of slow oscillations of auditory thalamic neurons by repetitive sound stimuli. J Neurosci 29:6013–6021.

Garrido MI, Kilner JM, Stephan KE, Friston KJ (2009) The mismatch negativity: A review of underlying mechanisms. Clin Neurophysiol 120:453–463 Available at: http://dx.doi.org/10.1016/j.clinph.2008.11.029.

Gaucher Q, Panniello M, Ivanov AZ, Dahmen JC, King AJ, Walker KMM (2020) Complexity of frequency receptive fields predicts tonotopic variability across species. eLife 9:e53462.

Geissler DB, Ehret G (2002) Time-critical integration of formants for perception of communication calls in mice. Proc Natl Acad Sci U S A 99:9021–9025.

Gilad A, Maor I, Mizrahi A (2020) Learning-related population dynamics in the auditory thalamus. eLife 9:e56307.

Gillespie PG, Müller U (2009) Mechanotransduction by Hair Cells: Models, Molecules, and Mechanisms. Cell 139:33–44.

Glorot X, Bengio Y (2010) Understanding the difficulty of training deep feedforward neural networks. J Mach Learn Res 9:249–256.

Goblick TJ, Pfeiffer RR (1969) Time-Domain Measurements of Cochlear Nonlinearities Using Combination Click Stimuli. J Acoust Soc Am 46:924–938 Available at: http://asa.scitation.org/doi/10.1121/1.1911812.

Goldberg JM, Brown PB (1968) Functional organization of the dog superior olivary complex: an anatomical and electrophysiological study. J Neurophysiol 31:639–656.

Goodfellow I, Bengio Y, Courville A (2016) Deep Learning. MIT Press. Available at: http://www.deeplearningbook.org.

Greenwald AG, McGhee DE, Schwartz JL (1998) Measuring individual differences in implicit cognition: the implicit association test. J Pers Soc Psychol 74:1464–1480.

Greenwood DD (1961) Critical Bandwidth and the Frequency Coordinates of the Basilar Membrane. J Acoust Soc Am 33:1344–1356.

Greenwood DD (1990) A cochlear frequency-position function for several species—29 years later. J Acoust Soc Am 87:2592–2605.

Greenwood DD (1996) Comparing octaves, frequency ranges, and cochlear-map curvature across species. Hear Res 94:157–162.

Grill-Spector K, Malach R (2004) The human visual cortex. Annu Rev Neurosci 27:649–677.

Guinan JJ, Norris BE, Guinan SS (1972) Single auditory units in the superior olivary complex: II: Locations of unit categories and tonotopic organization. Int J Neurosci 4:147–166.

Hage SR, Ehret G (2003) Mapping responses to frequency sweeps and tones in the inferior colliculus of house mice. Eur J Neurosci 18:2301–2312.

Hage SR, Ehret GG, Nter Ehret GÈ, Ehret GG (2003) Mapping responses to frequency sweeps and tones in the inferior colliculus of house mice. Eur J Neurosci 18:2301–2312.

Harper NS, Schoppe O, Willmore BDB, Cui Z, Schnupp JWH, King AJ (2016) Network Receptive Field Modeling Reveals Extensive Integration and Multi-feature Selectivity in Auditory Cortical Neurons. PLoS Comput Biol 12:1–30 Available at: http://dx.doi.org/10.1371/journal.pcbi.1005113.

Harris KD, Mrsic-Flogel TD (2013) Cortical connectivity and sensory coding. Nature 503:51–58.

Hateren J van, Ruderman D (1998) Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. R Soc:6 Available at: http://rspb.royalsocietypublishing.org/content/265/1412/2315.short.

Heeger DJ (2017) Theory of cortical function. Proc Natl Acad Sci U S A 114:1773–1782.

Heffner H, Masterton B (1980) Hearing in Glires: Domestic rabbit, cotton rat, feral house mouse, and kangaroo rat. J Acoust Soc Am 68:1584–1599.

Hegdé J (2008) Time course of visual perception: Coarse-to-fine processing and beyond. Prog Neurobiol 84:405–439.

Heil P, Langner G, Scheich H (1992a) Processing of frequency-modulated stimuli in the chick auditory cortex analogue: evidence for topographic representations and possible mechanisms of rate and directional sensitivity. J Comp Physiol A 171:583–600.

Heil P, Rajan R, Irvine DRF (1992b) Sensitivity of neurons in cat primary auditory cortex to tones and frequency-modulated stimuli. II: Organization of response properties along the “isofrequency” dimension. Hear Res 63:135–156.

Heilbron M, Chait M (2018) Great Expectations: Is there Evidence for Predictive Coding in Auditory Cortex? Neuroscience 389:54–73 Available at: https://doi.org/10.1016/j.neuroscience.2017.07.061.

Hickok G, Poeppel D (2007) The cortical organization of speech processing. Nat Rev Neurosci 8:393–402.

Hinton GE, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov RR (2012) Improving neural networks by preventing co-adaptation of feature detectors. :1–18 Available at: http://arxiv.org/abs/1207.0580.

Hochstein S, Ahissar M (2002) View from the top: Hierarchies and reverse hierarchies in the visual system. Neuron 36:791–804.

Hoyer PO, Hyvärinen A (2003) Interpreting neural response variability as Monte Carlo sampling of the posterior. Adv Neural Inf Process Syst.

Hsu A, Borst A, Theunissen F (2004) Quantifying variability in neural responses and its application for the validation of model predictions. Netw Comput Neural Syst 15:91–109.

Huang CL, Winer JA (2000) Auditory thalamocortical projections in the cat: Laminar and areal patterns of input. J Comp Neurol 427:302–331.

Huang Y, Rao RPN (2011) Predictive coding. Wiley Interdiscip Rev Cogn Sci 2:580–593.

Hudspeth AJ (2014) Integrating the active process of hair cells with cochlear function. Nat Rev Neurosci 15:600–614.

Huh D, Sejnowski TJ (2018) Gradient descent for spiking neural networks. Adv Neural Inf Process Syst 2018-Decem:1433–1443.

Hyvärinen A, Hurri J, Väyrynen J (2004) A unifying framework for natural image statistics: Spatiotemporal activity bubbles. Neurocomputing 58–60:801–806.

Iacaruso MF, Gasler IT, Hofer SB (2017) Synaptic organization of visual space in primary visual cortex. Nature 547:449–452 Available at: http://dx.doi.org/10.1038/nature23019.

Imig TJ, Adrián HO (1977) Binaural columns in the primary field (A1) of cat auditory cortex. Brain Res 138:241–257.

Imig TJ, Brugge JF (1978) Sources and terminations of callosal axons related to binaural and frequency maps in primary auditory cortex of the cat. J Comp Neurol 182:637–660.

Innocenti GM (1980) The primary visual pathway through the corpus callosum: morphological and functional aspects in the cat. Arch Ital Biol 118:124–188 Available at: http://europepmc.org/abstract/MED/6162430.

Ito S, Si Y, Feldheim DA, Litke AM (2020) Spectral cues are necessary to encode azimuthal auditory space in the mouse superior colliculus. Nat Commun 11:1–12 Available at: http://dx.doi.org/10.1038/s41467-020-14897-7.

Kanold PO, Nelken I, Polley DB (2014) Local versus global scales of organization in auditory cortex. Trends Neurosci 37:502–510 Available at: http://dx.doi.org/10.1016/j.tins.2014.06.003.

Kavanagh GL, Kelly JB (1988) Hearing in the ferret (Mustela putorius): Effects of primary auditory cortical lesions on thresholds for pure tone detection. J Neurophysiol 60:879–888.

Kell AJE, Yamins DLK, Shook EN, Norman-Haignere S V., McDermott JH (2018) A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. Neuron 98:630-644.e16 Available at: https://doi.org/10.1016/j.neuron.2018.03.044.

Kilgard MP, Merzenich MM (1999) Distributed representation of spectral and temporal information

in rat primary auditory cortex. Hear Res 134:16–28.

Kim H, Sul JH, Huh N, Lee D, Jung MW (2009) Role of striatum in updating values of chosen actions. J Neurosci 29:14701–14712.

King AJ, Hutchings ME (1987) Spatial response properties of acoustically responsive neurons in the superior colliculus of the ferret: A map of auditory space. J Neurophysiol 57:596–624.

King AJ, Teki S, Willmore BDB (2018) Recent advances in understanding the auditory cortex. F1000Research 7.

Kingma DP, Ba J (2015) Adam: A Method for Stochastic Optimization. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), pp 1–15 Available at: http://arxiv.org/abs/1412.6980.

Klein DJ, König P, Körding KP (2003) Sparse spectrotemporal coding of sounds. EURASIP J Appl Signal Processing 2003:659–667.

Knudsen EI (2004) Sensitive periods in the development of the brain and behavior. J Cogn Neurosci 16:1412–1425.

Ko H, Hofer SB, Pichler B, Buchanan KA, Sjöström PJ, Mrsic-Flogel TD (2011) Functional specificity of local synaptic connections in neocortical networks. Nature 473:87–91.

Kobrina A, Toal KL, Dent ML (2018) Intensity difference limens in adult CBA/CaJ mice (Mus musculus). Behav Processes 148:46–48.

Komune N, Yagmurlu K, Matsuo S, Miki K, Abe H, Rhoton AL (2015) Auditory brainstem implantation: Anatomy and approaches. Oper Neurosurg 11:306–320.

Kozlov AS, Gentner TQ (2016) Central auditory neurons have composite receptive fields. Proc Natl Acad Sci U S A 113:1441–1446 Available at: http://www.pnas.org/lookup/doi/10.1073/pnas.1506903113.

Kriegeskorte N (2015) Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. Annu Rev Vis Sci 1:417–446.

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

Kriegeskorte N, Mur M, Bandettini P (2008) Representational similarity analysis - connecting the branches of systems neuroscience. Front Syst Neurosci 2:1–28.

Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet Classification with Deep Convolutional Neural Networks. Adv Neural Inf Process Syst 25 25.

Kuo RI, Wu GK (2012) The Generation of Direction Selectivity in the Auditory System. Neuron 73:1016–1027.

Kurt S, Ehret G (2010) Auditory discrimination learning and knowledge transfer in mice depends on task difficulty. Proc Natl Acad Sci U S A 107:8481–8485.

Lakatos P, Musacchia G, O'Connel MN, Falchier AY, Javitt DC, Schroeder CE (2013) The Spectrotemporal Filter Mechanism of Auditory Selective Attention. Neuron 77:750–761 Available at: http://dx.doi.org/10.1016/j.neuron.2012.11.034.

Langner G, Albert M, Briede T (2002) Temporal and spatial coding of periodicity information in the inferior colliculus of awake chinchilla (Chinchilla laniger). Hear Res 168:110–130.

Lau B, Glimcher PW (2005) Dynamic Response-By-Response Models of Matching Behavior in Rhesus Monkeys. J Exp Anal Behav 84:555–579.

Leaver AM, Rauschecker JP (2010) Cortical Representation of Natural Complex Sounds: Effects of Acoustic Features and Auditory Object Category. J Neurosci 30:7604–7612.

LeCun Y, Bottou L, Orr GB, Müller K-R (1998) Efficient BackProp. In: Neural Networks: Tricks of the Trade, pp 9–50. Springer, Berlin, Heidelberg.

Levy RB, Reyes AD (2012) Spatial Profile of Excitatory and Inhibitory Synaptic Connectivity in Mouse Primary Auditory Cortex. 32:5609–5619.

Lewicki MS (2002) Efficient coding of natural sounds. Nat Neurosci 5:356–363.

Liberman MC (1982) The cochlear frequency map for the cat: Labeling auditory-nerve fibers of known characteristic frequency. J Acoust Soc Am 72:1441–1449.

Lillicrap TP, Cownden D, Tweed DB, Akerman CJ (2016) Random synaptic feedback weights support error backpropagation for deep learning. Nat Commun 7:1–10 Available at: http://dx.doi.org/10.1038/ncomms13276.

Linden JF, Liu RC, Sahani M, Schreiner CE, Merzenich MM (2003) Spectrotemporal structure of receptive fields in areas AI and AAF of mouse auditory cortex. J Neurophysiol 90:2660–2675.

Liu W, Atturo F, Aldaya R, Santi P, Cureoglu S, Obwegeser S, Glueckert R, Pfaller K, Schrott-Fischer A, Rask-Andersen H (2015) Macromolecular organization and fine structure of the human basilar membrane - RELEVANCE for cochlear implantation. Cell Tissue Res 360:245–262.

Lohse M, Dahmen JC, Bajo VM, King AJ (2021) Subcortical circuits mediate communication between primary sensory cortical areas in mice. Nat Commun 12:1–14 Available at: http://dx.doi.org/10.1038/s41467-021-24200-x.

London M, Häusser M (2005) Dendritic computation. Annu Rev Neurosci 28:503–532.

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

Luczak A, McNaughton BL, Kubo Y (2022) Neurons learn by predicting future activity. Nat Mach Intell 4:62–72.

Machens CK, Wehr MS, Zador AM (2004) Linearity of Cortical Receptive Fields Measured with Natural Sounds. J Neurosci 24:1089–1100.

MacWhinney B (2000) The CHILDES Project: Tools for analyzing talk., 3rd ed. Mahwah, NJ: Lawrence Erlbaum Associates.

Manzoni T, Barbaresi P, Bellardinelli E, Caminiti R (1980) Callosal projections from the two body midlines. Exp Brain Res 39:1–9.

Marblestone AH, Wayne G, Kording KP (2016) Toward an integration of deep learning and neuroscience. Front Comput Neurosci 10:1–41.

Marín O (2012) Interneuron dysfunction in psychiatric disorders. Nat Rev Neurosci 13:107–120.

Markov NT, Kennedy H (2013) The importance of being hierarchical. Curr Opin Neurobiol 23:187–194.

Marzen SE, DeDeo S (2017) The evolution of lossy compression. J R Soc Interface 14 Available at:

http://arxiv.org/abs/1506.06138%0Ahttp://dx.doi.org/10.1098/rsif.2017.0166.

Maunsell JHR, Gibson JR (1992) Visual response latencies in striate cortex of the macaque monkey. J Neurophysiol 68:1332–1344.

Mercado E, Orduña I, Nowak JM (2005) Auditory Categorization of Complex Sounds by Rats (Rattus norvegicus). J Comp Psychol 119:90–98.

Merzenich MM, Knight PL, Roth GL (1975) Representation of cochlea within primary auditory cortex in the cat. J Neurophysiol 38:231–249.

Meyer AF, Williamson RS, Linden JF, Sahani M (2017) Models of neuronal stimulus-response functions: Elaboration, estimation, and evaluation. Front Syst Neurosci 10:1–25.

Middlebrooks JC, Dykes RW, Merzenich MM (1980) Binaural response-specific bands in primary auditory cortex (AI) of the cat: Topographical organization orthogonal to isofrequency contours. Brain Res 181:31–48.

Miller KD (2016) Canonical computations of cerebral cortex. Curr Opin Neurobiol 37:75–84 Available at: http://dx.doi.org/10.1016/j.conb.2016.01.008.

Młynarski W, Hledík M, Sokolowski TR, Tkačik G (2021) Statistical analysis and optimality of neural systems. Neuron 109:1227-1241.e5.

Młynarski W, McDermott JH (2017) Learning Mid-Level Auditory Codes from Natural Sound Statistics. :1–34 Available at: http://arxiv.org/abs/1701.07138.

Moore BCJ (1986) Parallels betwen frequency selectivity measured psychophysically and in cochlear mechanics. Scand Audiol 15:139–152.

Morrison GS, Rose P, Zhang C (2012) Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice. Aust J Forensic Sci 44:155–167.

Moser T, Beutner D (2000) Kinetics of exocytosis and endocytosis at the cochlear inner hair cell afferent synapse of the mouse. Proc Natl Acad Sci U S A 97:883–888.

Näätänen R, Paavilainen P, Rinne T, Alho K (2007) The mismatch negativity (MMN) in basic research of central auditory processing: A review. Clin Neurophysiol 118:2544–2590.

Nabelek I, Hirsh IJ (1969) On the Discrimination of Frequency Transitions. J Acoust Soc Am 45:1510–1519.

Ngiam J, Koh PW, Chen Z, Bhaskar S, Ng AY (2011) Sparse filtering. Adv Neural Inf Process Syst 24 25th Annu Conf Neural Inf Process Syst 2011, NIPS 2011:1–9.

Nijhawan R (1994) Motion extrapolation in catching. Nature 370:256–257.

Ohl FW, Scheich H, Freeman WJ (2001) Change in pattern of ongoing cortical activity with auditory category learning. Nature 412:733–736.

Olshausen BA, Field DJ (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381:607–609 Available at: http://www.nature.com/doifinder/10.1038/381607a0.

Olshausen BA, Field DJ (1997) Sparse coding with an overcomplete basis set: A strategy employed by V1? Vision Res 37:3311–3325.

Olshausen BA, Field DJ (2004) Sparse coding of sensory inputs. Curr Opin Neurobiol 14:481–487.

Orbán G, Berkes P, Fiser J, Lengyel M (2016) Neural Variability and Sampling-Based Probabilistic

Representations in the Visual Cortex. Neuron 92:530–543.

Osen KK (1970) Course and termination of the primary afferents in the cochlear nuclei of the cat. An experimental anatomical study. Arch Ital Biol 108:21–51.

Palmer SE, Marre O, Berry MJ, Bialek W (2015) Predictive information in a sensory population. Proc Natl Acad Sci U S A 112:6908–6913 Available at: http://www.pnas.org/lookup/doi/10.1073/pnas.1506855112.

Panniello M, King AJ, Dahmen JC, Walker KMM (2018) Local and global spatial organization of interaural level difference and frequency preferences in auditory cortex. Cereb Cortex 28:350–369.

Parras GG, Nieto-Diego J, Carbajal G V., Valdés-Baizabal C, Escera C, Malmierca MS (2017) Neurons along the auditory pathway exhibit a hierarchical organization of prediction error. Nat Commun 8 Available at: http://dx.doi.org/10.1038/s41467-017-02038-6.

Parsons CE, Young KS, Craske MG, Stein AL, Kringelbach ML (2014) Introducing the oxford vocal (OxVoc) sounds database: A validated set of non-acted affective sounds from human infants, adults, and domestic animals. Front Psychol 5:1–10.

Plato (1925) Plato in twelve volumes: Vol. 9. Cambridge, MA, US: Harvard University Press.

Polley DB (2006) Perceptual Learning Directs Auditory Cortical Map Reorganization through Top-Down Influences. J Neurosci 26:4970–4982.

Rabinowitz NC, Willmore BDB, Schnupp JWH, King AJ (2011) Contrast Gain Control in Auditory Cortex. Neuron 70:1178–1191 Available at: http://dx.doi.org/10.1016/j.neuron.2011.04.030.

Rahman M, Willmore BDB, King AJ, Harper NS (2019) A dynamic network model of temporal receptive fields in primary auditory cortex. PLoS Comput Biol 15.

Rahman M, Willmore BDB, King AJ, Harper NS (2020) Simple transformations capture auditory input to cortex. Proc Natl Acad Sci U S A 117:28442–28451.

Raman IM, Zhang S, Trussell LO (1994) Pathway-specific variants of AMPA receptors and their contribution to neuronal signaling. J Neurosci 14:4998–5010.

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

Ratcliff R (1978) A theory of memory retrieval. Psychol Rev 85:59–108.

Raufer S, Guinan JJ, Nakajima HH (2019) Cochlear partition anatomy and motion in humans differ from the classic view of mammals. Proc Natl Acad Sci U S A 116:13977–13982.

Rayleigh, Lord (1907) XII. On our perception of sound direction. London, Edinburgh, Dublin Philos Mag J Sci 13:214–232 Available at: https://doi.org/10.1080/14786440709463595.

Rees A, Møller AR (1983) Responses of neurons in the inferior colliculus of the rat to AM and FM tones. Hear Res 10:301–330.

Rhode WS (1971) Observations of the Vibration of the Basilar Membrane in Squirrel Monkeys using the Mössbauer Technique. J Acoust Soc Am 49:1218–1231 Available at: http://asa.scitation.org/doi/10.1121/1.1912485.

Richards BA et al. (2019) A deep learning framework for neuroscience. Nat Neurosci 22:1761–1770 Available at: http://dx.doi.org/10.1038/s41593-019-0520-2.

Robles L, Ruggero MA (2001) Mechanics of the mammalian cochlea. Physiol Rev 81:1305–1352.

Rohrbaugh M, Brennan JF, Riccio DC (1971) Control of two-way shuttle avoidance in rats by auditory frequency and intensity. J Comp Physiol Psychol 75:324–330.

Rosen S (1992) Temporal information in speech: acoustic, auditory and linguistic aspects. Philos Trans R Soc Lond B Biol Sci 336:367–373.

Rubin J, Ulanovsky N, Nelken I, Tishby N (2016) The Representation of Prediction Error in Auditory Cortex. PLoS Comput Biol 12:1–28.

Ruderman DL, Bialek W (1994) Statistics of natural images: Scaling in the woods. Phys Rev Lett 73:814–817.

Ruggero MA, Rich NC, Recio A, Narayan SS, Robles L (1997) Basilar-membrane responses to tones at the base of the chinchilla cochlea. J Acoust Soc Am 101:2151–2163 Available at: http://asa.scitation.org/doi/10.1121/1.418265.

Rybalko N, Šuta D, Nwabueze-Ogbo F, Syka J (2006) Effect of auditory cortex lesions on the discrimination of frequency-modulated tones in rats. Eur J Neurosci 23:1614–1622.

Ryugo DK, Parks TN (2003) Primary innervation of the avian and mammalian cochlear nucleus. Brain Res Bull 60:435–456.

Sabatini BL, Svoboda K (2008) Ca2+ Signaling in dendritic spines. Struct Funct Organ Synap:441–458.

Salisbury JM, Palmer SE (2016) Optimal Prediction in the Retina and Natural Motion Statistics. J Stat Phys 162:1309–1323.

Sanes DH, Constantine-Paton M (1985) The sharpening of frequency tuning curves requires patterned activity during development in the mouse, Mus musculus. J Neurosci 5:1152–1166.

Schmolesky MT, Wang Y, Hanes DP, Thompson KG, Leutgeb S, Schall JD, Leventhal AG (1998) Signal timing access the macaque visual system. J Neurophysiol 79:3272–3278.

Schoppe O, Harper NS, Willmore BDB, King AJ, Schnupp JWH (2016) Measuring the Performance of Neural Models. Front Comput Neurosci 10:1–11 Available at: http://journal.frontiersin.org/article/10.3389/fncom.2016.00010.

Schouten MEH (1985) Identification and discrimination of sweep tones. Percept Psychophys 37:369–376.

Schulze H, Scheich H (1999) Discrimination learning of amplitude modulated tones in Mongolian gerbils. Neurosci Lett 261:13–16.

Seidl AH, Grothe B (2005) Development of sound localization mechanisms in the Mongolian gerbil is shaped by early acoustic experience. J Neurophysiol 94:1028–1036.

Sharpee TO, Rust NC, Bialek W (2004) Analyzing Neural Responses to Natural Signals. Neural Comput 250:223–250.

Shrestha SB, Orchard G (2018) Slayer: Spike layer error reassignment in time. Adv Neural Inf Process Syst 2018-Decem:1412–1421.

Sietsma J, Dow RJF (1991) Creating artificial neural networks that generalize. Neural Networks 4:67–79.

Silver D et al. (2016) Mastering the game of Go with deep neural networks and tree search. Nature 529:484–489 Available at: http://dx.doi.org/10.1038/nature16961.

Simoncelli EP, Olshausen BA (2001) Natural Image Statistics and Neural Representation. Annu Rev Neurosci 24:1193–1216.

Singer Y, Teramoto Y, Willmore BD, Schnupp JW, King AJ, Harper NS (2018) Sensory cortex is optimized for prediction of future input. eLife 7:e31557.

Singer Y, Willmore BDB, King AJ, Harper NS (2019) Hierarchical temporal prediction captures motion processing from retina to higher visual cortex. bioRxiv:575464 Available at: http://dx.doi.org/10.1101/575464.

Slattery WH, Middlebrooks JC (1994) Monaural sound localization: Acute versus chronic unilateral impairment. Hear Res 75:38–46.

Sloas DC, Zhuo R, Xue H, Chambers AR, Kolaczyk E, Polley DB, Sen K (2016) Interactions across multiple stimulus dimensions in primary auditory cortex. eNeuro 3:1–7.

Smith EC, Lewicki MS (2006) Efficient auditory coding. Nature 439:978–982 Available at: http://www.ncbi.nlm.nih.gov/pubmed/16495999.

Smith PH, Populin LC (2001) Fundamental differences between the thalamocortical recipient layers of the cat auditory and visual cortices. J Comp Neurol 436:508–519.

Srebro N, Srebro N, Shraibman A (2005) Rank, trace-norm and max-norm. Proc 18th Annu Conf Learn Theory 14:545--560.

Srinivasan MV, Laughlin SB, Dubs A (1982) Predictive coding: a fresh view of inhibition in the retina. Proc R Soc B 216:427–459.

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J Mach Learn Res 15:1929–1958.

Stone J V, Harper N (1999) Temporal constraints on visual learning: a computational model. Perception 28:1089–1104.

Suga N, Katsuki Y, Medical T (1963) Discharge pattern and inhibition of primary auditory nerve fibers in the monkey. J Neurophysiol 27:768–787.

Sussillo D (2014) Neural circuits as computational dynamical systems. Curr Opin Neurobiol 25:156–163 Available at: http://dx.doi.org/10.1016/j.conb.2014.01.008.

Syka J, Rybalko N, Brožek G, Jilek M (1996) Auditory frequency and intensity discrimination in pigmented rats. Hear Res 100:107–113.

Terman M (1970) Discrinimation of auditory intensities by rats 1. J Exp Anal Behav 13:145–160.

Theunissen FE, Sen K, Doupe AJ (2000) Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. J Neurosci 20:2315–2331.

Tucci DL, Rubel EW (1985) Afferent Influences on Brain Stem Auditory Nuclei of the Chicken: Effects of Conductive and Sensorineural Hearing Loss on N. Magnocellularis. J Comp Neurol 238:371–381.

Ulanovsky N, Las L, Nelken I (2003) Processing of low-probability sounds by cortical neurons. Nat Neurosci 6:391–398.

Usher M, McClelland JL (2001) The time course of perceptual choice: The leaky, competing accumulator model. Psychol Rev 108:550–592.

Vinyals O et al. (2019) Grandmaster level in StarCraft II using multi-agent reinforcement learning.

Nature 575:350–354 Available at: http://dx.doi.org/10.1038/s41586-019-1724-z.

Von Békésy G (1960) Experiments in hearing. Oxford, England: McGraw Hill.

Vyazovska O V. (2021) The effect of dimensional reinforcement prediction on discrimination of compound visual stimuli by pigeons. Anim Cogn:1–10.

Weinberger NM (2011) The medial geniculate, not the amygdala, as the root of auditory fear conditioning. Hear Res 274:61–74 Available at: http://dx.doi.org/10.1016/j.heares.2010.03.093.

Wen B, Wang GI, Dean I, Delgutte B (2012) Time course of dynamic range adaptation in the auditory nerve. J Neurophysiol 108:69–82.

Wessberg J, Stambaugh CR, Kralik JD, Beck PD, Laubach M, Chapin JK, Kim J, Biggs SJ, Srinivasan M a, Nicolelis M a (2000) Real-time prediction of hand trajectory by ensembles of cortical neurons in primates. Nature 408:361–365.

Wetzel W, Wagner T, Ohl FW, Scheich H (1998) Categorical discrimination of direction in frequency-modulated tones by Mongolian gerbils. Behav Brain Res 91:29–39.

Whittington JCR, Bogacz R (2019) Theories of Error Back-Propagation in the Brain. Trends Cogn Sci 23:235–250 Available at: https://doi.org/10.1016/j.tics.2018.12.005.

Willmore B (2020) Benlib-py. GitHub.

Willmore BDB, Schoppe O, King AJ, Schnupp JWH, Harper NS (2016) Incorporating midbrain adaptation to mean sound level improves models of auditory cortical processing. J Neurosci 36:280–289.

Winer JA (1992) The Functional Architecture of the Medial Geniculate Body and the Primary Auditory Cortex BT - The Mammalian Auditory Pathway: Neuroanatomy. In (Webster DB, Popper AN, Fay RR, eds), pp 222–409. New York, NY: Springer New York. Available at: https://doi.org/10.1007/978-1-4612-4416-5_6.

Winer JA (2006) Decoding the auditory corticofugal systems. Hear Res 212:1–8.

Winer JA, Prieto JJ (2001) Layer V in cat primary auditory cortex (AI): Cellular architecture and identification of projection neurons. J Comp Neurol 434:379–412.

Wiskott L, Sejnowski TJ (2002) Slow feature analysis: Unsupervised learning of invariances. Neural Comput 14:715–770.

Wonders CP, Anderson SA (2006) The origin and specification of cortical interneurons. Nat Rev Neurosci 7:687–696.

Yabe H, Tervaniemi M, Reinikainen K, Näätänen R (1997) Temporal window of integration revealed by MMN to sound omission. Neuroreport 8:1971–1974.

Yamins DLK, DiCarlo JJ (2016) Using goal-driven deep learning models to understand sensory cortex. Nat Neurosci 19:356–365.

Yamins DLK, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc Natl Acad Sci U S A 111:8619–8624.

Ye Y, Ihlefeld A, Rosen MJ (2021) Conductive hearing loss during development does not appreciably alter the sharpness of cochlear tuning. Sci Rep 11:1–11.

Yin TCT (2002) Neural Mechanisms of Encoding Binaural Localization Cues in the Auditory Brainstem BT - Integrative Functions in the Mammalian Auditory Pathway. In (Oertel D, Fay RR, Popper AN, eds), pp 99–159. New York, NY: Springer New York. Available at: https://doi.org/10.1007/978-1-4757-3654-0_4.

Yuan K, Fink KL, Winer JA, Schreiner CE (2011) Local connection patterns of parvalbumin-positive inhibitory interneurons in rat primary auditory cortex. Hear Res 274:121–128 Available at: http://dx.doi.org/10.1016/j.heares.2010.06.014.

Yuste R (2015) From the neuron doctrine to neural networks. Nat Rev Neurosci 16:487–497 Available at: http://dx.doi.org/10.1038/nrn3962.

Zenke F, Vogels TP (2021) The Remarkable Robustness of Surrogate Gradient Learning for Instilling Complex Function in Spiking Neural Networks. Neural Comput 33:899–925.

Zhang LI, Bao S, Merzenich MM (2001) Persistent and specific influences of early acoustic environments on primary auditory cortex. Nat Neurosci 4:1123–1130.

Zhang LI, Tan AYY, Schreiner CE, Merzenich MM (2003) Topography and synaptic shaping of direction selectivity in primary auditory cortex. Nature 424:201–205.

Zhao L, Zhaoping L (2011) Understanding auditory spectro-temporal receptive fields and their changes with input statistics by efficient coding principles. PLoS Comput Biol 7.

Zilany MSA, Bruce IC, Nelson PC, Carney LH (2009) A phenomenological model of the synapse between the inner hair cell and auditory nerve: Long-term adaptation with power-law dynamics. J Acoust Soc Am 126:2390–2412.