CITY UNIVERSITY OF HONG KONG

香港城市大學

Neural Correlates of Mnemonic and Predictive Representations in the Auditory Cortex

聽覺皮層中助記符和預測表示的神經相關性

Submitted to

Department of Neuroscience

神經科學系

in Fulfillment of the Requirements

for the Degree of Doctor of Philosophy

哲學博士學位

by

Cappotto Drew

September 2022

二零二二年九月

Abstract

In order for a sensory system to display adaptive behavior in response to external stimuli, it is advantageous to employ mechanisms capable of learning and encoding the probabilities present in ongoing stimulus streams. Because real-world stimuli tend to contain repetitive features, sensory systems have evolved to become highly sensitive to such regularities in order to detect deviations, while simultaneously "tuning out" stimuli features which do not require constant attendance owing to those regularities. This manner of efficient sequence processing relies on the ability to encode stimuli into short term memory, form predictions about upcoming stimulus features based on those that come before, and to make comparisons in order to update those assumptions when a deviant event is detected.

In this thesis, I investigate neural correlates during auditory sequence processing, used as a platform to probe and decode auditory sensory memories, predictions, and implicit learning in the auditory system. In the first experiment, I show that auditory sensory memory contents can be decoded from electrophysiological signals recorded in awake humans and anesthetized rats using homologous methods, suggesting that the mechanisms of sensory memory encoding are evolutionarily conserved across species. In the second experiment, I show that mnemonic and predictive representations of auditory stimuli can be simultaneously decoded from neural activity in anesthetized rats at overlapping latencies, but based on largely uncorrelated data features. Predictive representations are dynamically updated over the course of stimulation, suggesting a gradual formation of prediction. In the third experiment, conducted in awake humans, I show that neural correlates of prediction errors to unexpected sound contents are modulated by time-based predictions in a contextually congruent manner, such that local

(vs. global) time-based predictions amplify prediction errors to unexpected sequence elements (vs. chunks). These modulations were shared between contextual levels in terms of the spatiotemporal distribution of neural activity, suggesting the brain integrates different predictions with a high degree of contextual specificity, but in a shared and distributed cortical network.

The experiments comprising this thesis explore phenomena that are integral to our understanding of cognitive processing and the mechanisms by which we interface with the outside world. As external stimuli contain incessant streams of complex regularities, the brain must find ways to parse meaningful information in the most efficient manner. The mechanisms responsible for this process rely on the brain's intrinsic ability to learn such regularities, form a model allowing it to predict what events are likely to occur, and encode features into memory for comparison in order to update that model when deviants are detected. The following chapters detail the background of these mechanisms and three experiments which probe the resultant phenomena.

Qualifying Panel and Examination Panel

CITY UNIVERSITY OF HONG KONG
Qualifying Panel and Examination Panel

Surname: CAPPOTTO
First Name: Drew
Degree: Doctor of Philosophy
College/Department: Department of Neuroscience

The Qualifying Panel of the above student is composed of:

Supervisor(s):

Prof. Jan SCHNUPP	Department of Neuroscience, City University of Hong Kong

Qualifying Panel Member(s):

Dr. LAU Chun Yue Geoffrey	Department of Neuroscience, City University of Hong Kong
Dr. YANG Sungchil	Department of Neuroscience, City University of Hong Kong

This thesis has been examined and approved by the following examiners:

Prof. Jan SCHNUPP	Department of Neuroscience, City University of Hong Kong
Dr. WONG Wing Kuen	Department of Social and Behavioural Sciences, City University of Hong Kong
Prof. CHAIT Maria	The Ear Institute, University College London
Prof. WANG William Shi Yuan	Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University

Declaration

I, Drew Cappotto, declare that this thesis entitled “Neural Correlates of Mnemonic and Predictive Representations in the Auditory Cortex” represents my original work and the contents of this thesis have never been submitted to this University or other institutions for a degree or any other qualifications in the form of thesis or other report, excluding content originating in collaborative manuscripts, as acknowledged in the relevant statements on co-authorships. The thesis corresponds to a compendium of scientific articles already published or under review for publication. The articles are listed below:

Study Ⅰ

Decoding the Content of Auditory Sensory Memory Across Species

Authors: Drew Cappotto, Ryszard Auksztulewicz, HiJee Kang, David Poeppel, Lucia Melloni, Jan Schnupp

Journal: Cerebral Cortex, Volume 31, Issue 7, July 2021, Pages 3226–3236, https://doi.org/10.1093/cercor/bhab002

Study Ⅱ

Simultaneous Mnemonic and Predictive Representations in the Auditory Cortex

Authors: Drew Cappotto, HiJee Kang, Kongyan Li, Lucia Melloni, Jan Schnupp, Ryszard Auksztulewicz

Journal: Current Biology. Published April 28, 2022. https://doi.org/10.1016/j.cub.2022.04.022

Study Ⅲ

"What" and "When" Predictions Modulate Auditory Processing in a Contextually Specific Manner

Authors: Drew Cappotto, Dan Lou, Hiu Wai Lai, Fei Peng, Lucia Melloni, Jan Schnupp, Ryszard Auksztulewicz

Journal: *Currently under revision*

List of Abbreviations

AC : auditory cortex
ASM : auditory sensory memory
DMTS : delayed match-to-sample tasks
DSS : dynamic separation of sources
ECoG : electrocorticography
EEG : electroencephalogram
ERP : event-related potentials
FRA : frequency response areas
FDR : false-discovery-rate
FWE : family-wise error
GLM : general linear model
IFG : inferior frontal gyrus
i.p. : intraperitoneal
ITPC : inter-trial phase coherence
LSPL : left superior parietal lobule
MMR : mismatch response
MMN : mismatch negativity
RM ANOVA : repeated measures analysis of variance
SNR : signal-to-noise ratio
SPL : sound pressure level
SSA : stimulus specific adaptation
STG : superior temporal gyrus
STM: short term memory
WM : working memory

List of Figures

List of Tables

Table of Contents

Chapter 1. Introduction

1.1 Sequence Learning

As an integral component of adaptive behavior, the ability to perceive and learn regular patterns present in the environment is a trait shared across species, suggestive of a mechanism so useful that it has been evolutionarily conserved (Henin et al., 2021; Kikuchi et al., 2018, 2017). While sequence learning takes place across sensory domains, the focus of this thesis is on the brain's ability to learn and process regularities from sequences of sounds. Interestingly, research has demonstrated that the ability to recognize sequential patterns to an extent persists across differing states of attention and wakefulness (Denham and Winkler, 2020; Tivadar et al., 2021) - that is to say, such regularities can be learned as a result of the statistical regularities present in a given stimulus stream. Previous studies have established neural markers corresponding to the detection of deviant elements in sequences learned under both active and passive exposure - e.g., during attended and unattended stimuli (Tivadar et al., 2021). Importantly, such markers have also been observed in different forms of wakefulness - e.g., fully awake (during active and passive exposure) and in various states of consciousness (anesthesia, coma, sleep) (Tivadar et al., 2021).

Auditory sequences can contain multiple types of regularities, which are in turn encoded and processed by the auditory system in multiple ways. The neural representations of these regularities have been proposed to fall into a taxonomy wherein five mechanisms are responsible for encoding the corresponding regularities within sequence streams (Dehaene et al., 2015). In the aforementioned framework, the first type of sequence processing would rely on the processing of timing and transition information,

where the identify and onset of subsequent tokens within a sequence are established, while chunking occurs when several tokens are grouped into a single unit which can then be stored or manipulated as a whole. Ordinal knowledge would entail the understanding of the order in which individual tokens belong within a given sequence (e.g. first, third, last). A further level of abstraction is described as “algebraic patterns”, wherein relationships between chunked tokens are processed – for instance, AAB and XXY contain similar patters of two identical items followed by a third item which is different. Nested tree structures are a higher level of abstraction still, relying on symbolic rules such as those present in language processing in the form of grammar and meaning. Owing to the complexities that can be contained within sequences, hierarchal processing as a cognitive method allows for the organization of information into hierarchical structures, or so-called “chunks of chunks” wherein layers of sequence features can be more efficiently processed (Dehaene et al., 2015). Intracranial recordings in humans have revealed separable neural patterns and anatomical segregation in hierarchical processing, with low-level processing associated with sensory areas and higher-level processing associated with frontal and anterior lobes (Henin et al., 2021). This distributed network of hierarchical sequence processing is not limited to the sensory cortices, and the role of the hippocampus in mnemonic and temporal processing during sequence processing is well established (Bellmund et al., 2020).

In the context of sequence processing, “mnemonic representations” entail a memory of past sequence elements, independent of the currently processed element, while “predictive representations” entail a specific prediction of which sequence element is expected at a given time - as well as the comparison of this prediction with the currently

processed stimulus. These representations are intrinsically linked, as prior mnemonic representations service the accrual of information across previously learned patterns, while predictive representations can be seen as a form of memory retrieval used to predict the current or future sensory states (Baumgarten et al., 2021). Indeed, sequence learning has been shown to mediate predictive mechanisms in sensory cortices (Luft et al., 2015) and associative regions (Baumgarten et al., 2021) by reactivating sensory representations. Recent studies have successfully paired concepts of statistical learning and predictive coding by investigating neural correlates of melodic expectation to naturalistic music, observing that neural responses to statistically less likely notes elicit markers consistent with their level of statistical predictability (Di Liberto et al., 2020). Human fMRI studies in the visual domain have further established the role of temporal regularity in sequence learning and their resultant effects on the decodability of predictable stimuli (Luft et al., 2015) and a recent psychophysics study has shown that human observers can not only automatically extract implicit information about recurring stimulus sequences, but that they can also store this information in long-term memory (Bianco et al., 2020). The neural mechanisms thought to be responsible for predictive and mnemonic processing in these contexts, as well as their associated neural markers, will be discussed in the following sections.

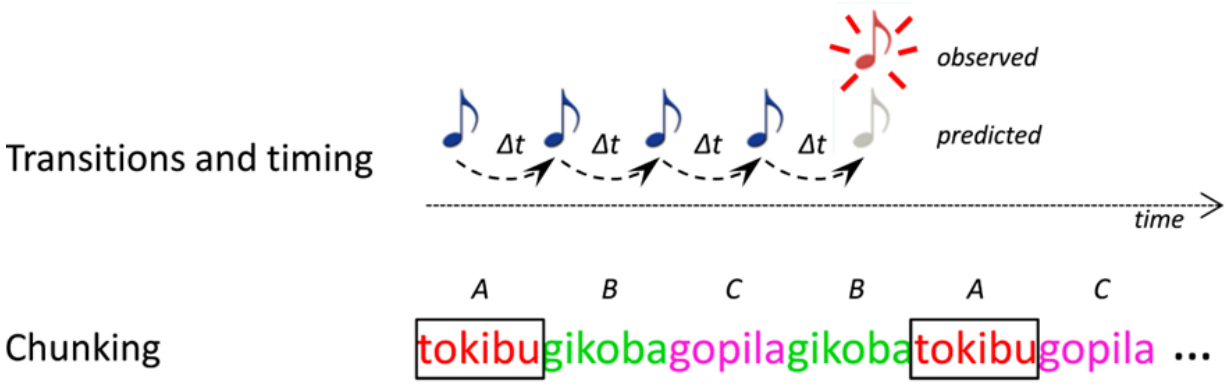

Figure 1.1 Illustration of a partial taxonomy of sequence knowledge. (Top) Transitions between specific items at a specific time delay. (Bottom) A sequence of "chunks". Adapted from Dehaene et al., 2015.

1.2 Auditory Sensory Memory Encoding

When the brain is exposed to sensory stimuli, either as part of a repetitive series or as an isolated event, a passive "buffer" exists within the sensory modality allowing the stimulus to be automatically stored before that buffer is overwritten (Spector, 2011). Although the length of this buffer in the auditory system has been broadly defined in the literature (generally accepted as under 5 seconds with some estimates upwards of 10 seconds) we can conceive of auditory sensory memory as a temporary store of auditory stimulus that operates without the need for cognitive maintenance (Nees, 2016), and as such one that persists in different attentional states (Pasternak and Greenlee, 2005). This buffer allows the opportunity for subsequent processing, such as being moved into longer-term storage or active maintenance. Indeed, such a capacity is of great practical benefit, as the auditory system experiences stimulus momentarily - individual sound events do not persist in the environment in the same way as objects viewed by the visual system, which can be re-scanned by the retina. Auditory sensory memory (ASM), as such, is a vital component of sensory processing with implications across the scope of this thesis.

Owing to the loosely defined nature of its time windows, the terms ASM, working memory (WM), and more broadly short-term memory (STM) are often used interchangeably when speaking about the maintenance of mnemonic information in temporary storage, with the practical differences largely coming down to retention time scale and active manipulation of memory contents.

Classical psychometric studies have measured sensory memory storage primarily through the use of delayed match-to-sample tasks (DMTS), wherein a sensory item is presented and, after a period of time, the participant is asked to match their memory of the presented item against a provided sample (Daniel et al., 2016; Nees, 2016) to establish psychometric functions of memory retention. Sustained neural activity in the prefrontal cortex (PFC) during retention periods has been observed in WM studies, with numerous theoretical models having been proposed to account for the role of such activity in sensory memory maintenance (Stokes, 2015). For instance, “synaptic reverberation” has been proposed as the underlying mechanism responsible for WM retention, wherein neural activity is sustained by recurrent excitatory loops mediated by NMDA receptor dynamics (Wang, 2001). However, sensory memories are not always reflected in sustained neural activity, despite the ability for them to be retrieved after extended periods in DMTS tasks. A recent study has shown that memories held without conscious awareness did not elicit sustained neural activity until tasks requiring active manipulation of memory contents were performed, and numerous studies have observed similar phenomena (Stokes, 2015; Trübutschek et al., 2018). A contemporary theory posits that such memory items are stored in "activity-silent" states - that information is stored in such a way (e.g., encoded into synaptic weights, see Figure 1.2.2) that sustained neural activity

is not required for their retention (Stokes, 2015). Recent studies have demonstrated that the neural traces responsible for this retention can be reactivated through the use of sensory "impulses" consisting of broadband probe stimulus presented during the activity-silent period (Wolff et al., 2019, 2015).

While the PFC has been widely implicated in working memory maintenance, as evident from numerous lesion and imaging studies (Stokes 2015), sensory areas have also been found to play a crucial role in STM processing. A recent study found that optogenetic suppression of neurons in the auditory cortex (AC) of rodents early in the delay period impaired WM performance, while suppression in later delay periods did not (Yu et al., 2021)This is of particular interest, as it establishes the role of the AC in early encoding of auditory mnemonic representations before upstream processing in the PFC. Indeed, this finding is consistent with past human research demonstrating lesions in the AC resulted in impaired auditory discrimination and reduced mismatch negativity (MMN) amplitudes (Alain et al., 1998) when compared to healthy control subjects.

The latter finding indicates another method of tapping into ASM. In addition to DMTS paradigms, a traditional method to measure the neural correlates of ASM includes the use of so-called oddball paradigms, where the brain's response to a deviant and unexpected stimulus elicits a measurable "mismatch response" (MMR) (Winkler et al., 1993). The MMN, measured with electroencephalography (EEG), is a prime example of an MMR. This response is so fundamental to cognition that its absence is often used to predict the outcome of comatose patients (Morlet and Fischer, 2014). The results of previous MMR studies are largely compatible with silent-coding theories, as MMR has been postulated to result from the processing of deviant stimuli in comparison to an

existing ASM trace (Näätänen et al., 2005). In order for a stimulus to be deviant in this context, the brain must first form a prediction about what event was expected in order to make a comparison to the stimulus that actually occurred. The next section will therefore discuss the notion of predictive processing.

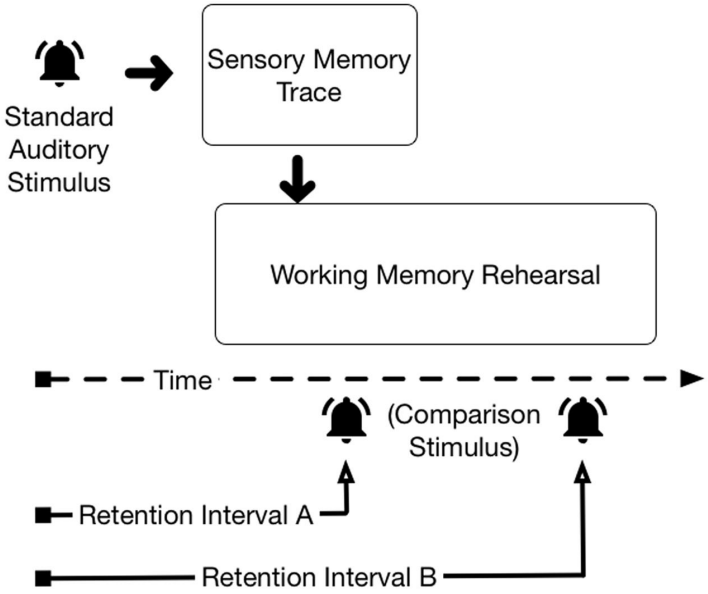

Figure 1.2.1 Hypothetical availability of an ASM trace for two retention intervals. Adapted from Nees, 2016.

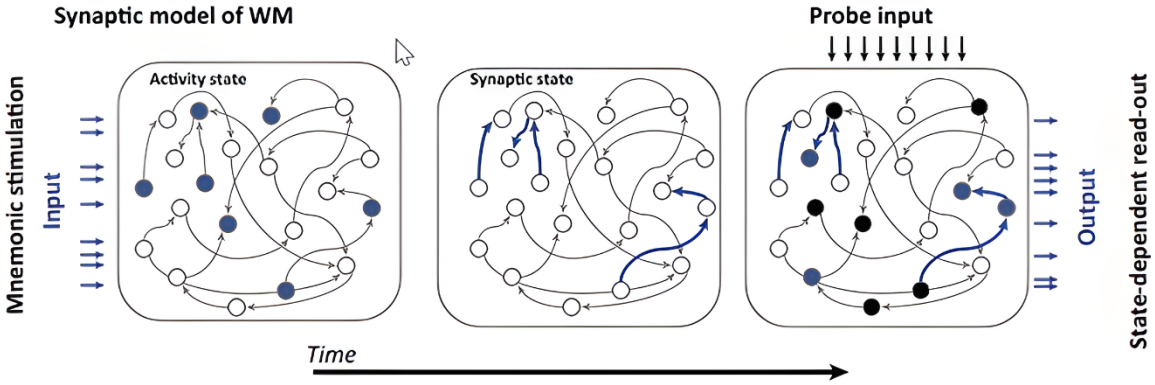

Figure 1.2.2 Schematic of the synaptic model of WM, adapted from Stokes, 2015. Task-relevant input (left-side horizontal arrows) drives a stimulus-specific activity state (filled circles) that in turn triggers a specific pattern of short-term synaptic plasticity between cells (bold arrows). Memory is read out from this synaptic

trace via the context-dependent response at retrieval (black filled circles). The probe-driven response will be patterned by the hidden state of synaptic efficacy, resulting in a discriminable output pattern (right-side horizontal arrows).

1.3 Predictive Coding

The presence of multiple stimulus streams in natural environments presents a unique challenge to our perceptual systems, as monitoring all of these streams in real time to extract useful information would require a tremendous amount of cognitive resources that our brains might otherwise use in service of other tasks. The predictive coding framework outlines one approach which our brains use to make this challenge more manageable (Friston, 2005; Heilbron and Chait, 2018). Stimulus streams often have consistent, repetitive, or otherwise predictable features – e.g. statistical regularities similar to those responsible for sequence learning. Sensory systems thus can make the parsing of incessant streams of sensory information more efficient by forming predictions based on regularities detected or rules discovered in the course of sensory experience. These predictions allow the brain to create a probable model of the outside world, which can be updated when errors are detected between the model predictions and external inputs (Fairhall et al., 2001; Friston et al., 2006; Rubin et al., 2016; Schröger et al., 2014). In the auditory system, explanations based on predictive coding have been applied to several phenomena, such as MMN responses, temporal expectation, and so-called omission responses (Denham and Winkler, 2020; Heilbron and Chait, 2018). In such contexts, experimental paradigms typically employ the use of repetitive stimulus streams which allow a build-up of predictions, based on memory of recent stimulation. Within the predictive coding framework, memory is intrinsically linked with predictive mechanisms in

the form of adaptive memory traces employed in downstream error correction (Wacongne et al., 2012). Interestingly, a recent study on auditory associative learning in awake mice showed that neural activity evoked by a predicted stimulus contains information both about its most likely predictor and its actual past, but that this information relies on dissociable neural codes, suggesting that mnemonic and predictive representations coexist within sensory cortices (Libby and Buschman, 2021).

The present literature on animal models of predictive processing is largely within the context of stimulus specific adaptation (SSA), and employs mismatch or omission designs reliant on repetition of the same stimulus token, making it difficult to separate predictive from simple adaptive mechanisms. A recent study (Parras et al., 2017) attempted to disentangle adaptive and predictive effects using an oddball paradigm in single-unit recordings from awake and anesthetized animals, demonstrating that predictive effects are organized hierarchically and suggestive of underlying MMN mechanisms. Although different models of the mechanisms underlying predictive coding exist, the most widely accepted model postulates that neurons responsible for making predictions reside in deep cortical layers, while neurons responsible for error detection reside in superficial cortical layers. Within this framework, information about predictions relies on backward connections and information about detected errors relies on forward connections, with the underlying predictive model being update by errors as a result of this cortical loop (Heilbron and Chait, 2018, see Figure 1.3). Further animal studies (Malmierca et al., 2019) have investigated neuronal pattern sensitivity with findings compatible with predictive coding frameworks reliant on temporal and spectral regularities entrained at the single-unit level, and have provided compelling evidence for a prediction-

error-signaling-based explanation of hierarchical deviance detection gradient between auditory subcortical and cortical regions, and prefrontal cortices (Casado-Román et al., 2020). However, studies employing animal models and different attention states to investigate predictive mechanisms beyond mismatch signaling have been largely lacking (Heilbron and Chait, 2018), and one goal of the research presented herein is to partially address that gap in the literature.

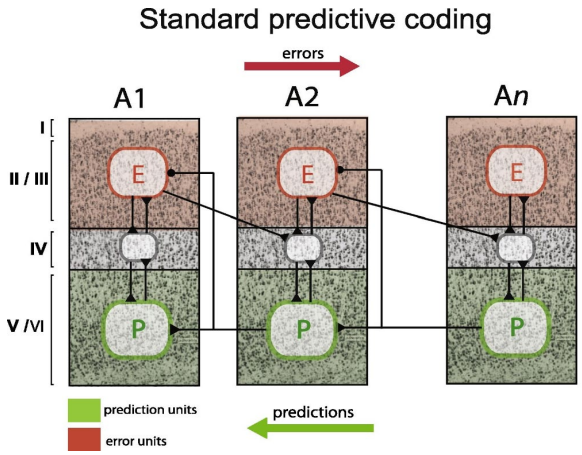

Figure 1.3 Arrangements of error and expectation neurons in the auditory cortex implied by the standard model of predictive coding. Adapted from Heilbron and Chait, 2018.

Oddball paradigms and MMR analyses have traditionally been employed to investigate mnemonic predictive processing. However, auditory streams contain information not only about the content of events, but also about the timing at which those events occur. Mnemonic and temporal predictions have been postulated as relying on dissociable neural correlates and partially separable underlying mechanisms (Friston and Buzsáki, 2016; Hsu et al., 2013), while both play a role in the modulation of stimulus-evoked activity in the superior temporal gyrus (Auksztulewicz et al., 2018). A recent human electrocorticography study found that temporal and mnemonic predictions engage overlapping but separable brain regions at different latencies, with computational modelling revealing increased plasticity in auditory regions during mnemonic processing and increased synaptic gain in motor regions during temporal processing (Auksztulewicz et al., 2019). It has also been proposed that interactions between mnemonic and temporal

predictions are inherent to the processing of musical sequences (Musacchia et al., 2014). In this context, it has been suggested that neural entrainment along the non-lemniscal (secondary) auditory pathway (sensitive to the rhythmic sequence structure) can modulate activity in the lemniscal (primary) pathway (encoding stimulus contents), including MMR processing. Such modulatory effects are perhaps unsurprising, as temporal information is essential to sequence processing and its underlying mnemonic and predictive mechanisms, while MMR remains a reliable neurological marker of sequence deviants.

1.4 Thesis Overview

This thesis investigates the mechanisms underlying the brain's ability to learn regularities in ongoing stimuli sequences, make predictions about future events, and encode features into memory. Experimental Chapters 2-4 convey research that I have undertaken throughout my PhD studies to investigate the individual phenomena that comprise these mechanisms.

Chapter 2: In the first study, a cross-species approach was employed to test whether auditory memory contents can be decoded from electrophysiological signals recorded in different species. Awake human volunteers (N=21) were exposed to auditory pure tone and noise burst stimuli during an auditory sensory memory task using EEG. In a closely matching paradigm, anesthetized female rats (N=5) were exposed to comparable stimuli while neural activity was recorded using electrocorticography (ECoG) from the auditory cortex. In both scenarios, acoustic frequency of recent tokens could be decoded from neural activity evoked by pure tones as well as that evoked by neutral frozen noise burst stimuli used to probe mnemonic representations held in silent-state

storage, suggesting that the mechanisms of sensory memory encoding are evolutionarily conserved across species.

Chapter 3: In the second study, neural activity elicited by repeated stimulus sequences was recorded using ECoG in the auditory cortex of anesthetized female rats (N=8), where events within the sequence were occasionally replaced with a broadband noise burst or omitted entirely. The results show that both stimulus history and predicted stimuli can be decoded from neural responses to broadband impulses, at overlapping latencies, but based on uncorrelated data features. The results also demonstrate that predictive representations are dynamically updated over the course of stimulation in a manner consistent with standard learning curves, suggesting these mechanisms are retained independently of attentional state.

Chapter 4: In the third study, we disambiguate neural correlates of “what” and “when” predictions by independently manipulating the predictability of temporal onset and acoustic contents at two contextual levels (single stimuli and stimulus pairs). Healthy volunteers (N=20) performed a repetition detection task while we recorded their neural activity using EEG. The results reveal that “what” and “when” predictions interactively modulated stimulus-evoked response amplitude in a contextually congruent manner, such that faster “when” predictions modulated the amplitude of mismatch responses to unexpected single stimuli, while slower “when” predictions modulated the amplitude of mismatch responses to unexpected stimulus pairs. We also find that the neural effects of these modulations were shared between the two contextual levels of prediction signaling in terms of the spatiotemporal distribution of EEG signals. Furthermore, by analyzing the entrainment of low frequency neural activity to the stimulus stream, we found evidence

for a gradual increase of entrainment to slow temporal predictions (regarding the timing of stimulus pairs).

Chapter 2. Decoding The Content of Auditory Sensory Memory Across Species

2.1 Abstract

In contrast to classical views of working memory maintenance, recent research investigating activity-silent neural states has demonstrated that persistent neural activity in sensory cortices is not necessary for active maintenance of information in working memory. Previous studies in humans have measured putative memory representations indirectly, by decoding memory contents from neural activity evoked by a neutral impulse stimulus. However, it is unclear whether memory contents can also be decoded in different species and attentional conditions. Here, we employ a cross-species approach to test whether auditory memory contents can be decoded from electrophysiological signals recorded in different species. Awake human volunteers (N=21) were exposed to auditory pure tone and noise burst stimuli during an auditory sensory memory task (ASM task) using electroencephalography. In a closely matching paradigm, anesthetized female rats (N=5) were exposed to comparable stimuli while neural activity was recorded using electrocorticography from the auditory cortex. In both species, acoustic frequency could be decoded from neural activity evoked by pure tones as well as neutral frozen noise burst stimuli. This finding demonstrates that memory contents can be decoded in different species and different states using homologous methods, suggesting that the mechanisms of sensory memory encoding are evolutionarily conserved across species.

2.2 Introduction

As a crucial component of adaptive intelligence, working memory (WM) allows for the temporary retention of information, breaking the immediacy of sensations and

allowing for actions that are not reflexive. Sensory WM refers to an organism's ability to retain information from a specific sensory modality (Spector, 2011). Auditory sensory memory (ASM) is a low-level subset of auditory WM wherein features of acoustic events are automatically maintained for a short period of time without the need for active cognitive retention (Pasternak and Greenlee, 2005). In comparison to auditory WM, which is actively maintained over longer periods of time, ASM can be thought of as a passively retained "buffer" that decays over time and is "overwritten" by new auditory sensory input (Pasternak and Greenlee, 2005), serving as a temporary store before relevant information can be moved into higher-level memory systems when required. The auditory system cannot "re-hear" acoustic events, and as such automatic retention of such events is essential for higher-level cognitive processes (e.g., active maintenance of auditory WM or long-term storage). ASM is a vital, low-level, function of the memory system upon which our understanding of higher-level auditory memory functions is built.

Early findings demonstrated that maintenance of WM was accompanied with persistent neural activity in frontal areas (Huang et al., 2016; Tark and Curtis, 2009). Recent studies have demonstrated that WM is not always reflected in sustained neural activity. This has led to new conceptualizations as to how WM maintenance takes place in the brain (Fries, 2005; Kamiński and Rutishauser, 2019; Mongillo et al., 2008; Stokes, 2015). One possibility is that WM is instantiated by "activity-silent" neural states, whereby sensory cortical regions do not show sustained activity during the retention period despite clearly observed activity during encoding and recall periods (Stokes, 2015). It has been shown that those activity-silent neural states can be probed by measuring "impulse responses" to a standardized broad-band probe stimulus during the activity-silent period

and decoding the resultant neural response to make inferences about its contents (Wolff et al., 2019, 2015).

Existing research on silent-coding of WM has focused on higher-level auditory processes such as active retention of auditory features during WM maintenance, or lower-level processes in the visual system (Wolff et al., 2020, 2015). Despite cross-species studies enjoying the benefits of established research methods developed and refined for each species (Mishra and Gazzaley, 2016), additional research in neural correlates of WM, particularly those which employ paradigms for activity-silent coding, focus either on the prefrontal cortex or employ single-species and single-conscious-state models (Bigelow et al., 2014; Constantinidis and Procyk, 2004; Murray et al., 2017; Spaak et al., 2017; Stokes, 2015). Thus, whether the mechanism subserving WM is preserved across species, conscious states and hierarchical levels remains unknown. To begin to address this, here we capitalize on cross-species investigations, and on understanding low-level memory processes which we see as integral to understanding higher-level functions in the auditory WM system and, as such, focus on silent-state activity during the auditory sensory memory period. We use a multivariate decoding method to analyze data acquired from EEG recordings in awake humans and ECoG recordings in anesthetized rats, decoding stimulus features from neural activity evoked by both the stimuli and frozen noise bursts presented during the ASM period. Such an approach allows us to bring invasive techniques to bear that help with precise localization, and allow for the investigation of causal mechanisms using methods that aren't available in human subjects alone. By testing whether the neural response to frozen noise bursts contain information about the stimulus feature held in ASM, we hope to establish grounding for

new models of ASM research by demonstrating the efficacy of cross-species and cross-attention-state approaches, elaborating on existing decoding research in a field whose limits have yet to be defined.

2.3. Methods and Materials

2.3.1 Human Electroencephalography

Participant Sample:

Participants (N = 21, 12 male, 9 female, median age = 25, age range = 22 - 50) volunteered to take part in the study upon written consent. The work was conducted in accordance with protocols approved by the Human Subjects Ethics Sub-Committee of the City University of Hong Kong. All subjects were self-reported as healthy with no hearing impairment.

Behavioral Paradigm and Stimulus Design

Stimuli were presented in 10 separate blocks, where participants responded to stimulus pairs. Prior to task blocks, participants were given the opportunity to modify the playback volume from its default level of 83 dB SPL if they judged it to be too loud or too soft, with levels adjusted by the researcher within +/- 5 dB SPL to a comfortable setting for each participant.

During blocks, participants were presented with a pair of pure tones separated by an auditory burst of frozen pink noise (a full-bandwidth noise signal with equal power in proportionately wide bands and a power density decreasing at 3 dB per octave) (Figure 2.1A). Tones preceding the frozen noise burst (T1) were randomly drawn from a set of

six semitones, from a half-octave chromatic scale, starting at 440 Hz. Tones following the frozen noise burst (T2) were detuned per trial by picking at random one of 10 possible frequencies from a set that extended one semitone above to one semitone below T1 in steps of 22.22% of a semitone. The same frozen noise burst stimulus was used across trials and participants. Each of the three events within the stimulus sequence were 200 ms in duration, with pure tones tapered by 5 ms linear on and off ramps. Stimulus events within each sequence were separated by randomly-selected gaps of silence, uniformly distributions of both gaps across trials. Gaps between T1 and frozen noise burst ranged between 0.6-1.6 s, and the gaps between frozen noise burst and T2 ranged between 0.3-0.8 s, consistent with the lower range of time intervals used in ASM paradigms (Nees, 2016). Gap durations, tone frequencies, and detuning values were assigned at random for each trial, with T2 always being a detuning of T1, for an average of 100 presentations for each T1 frequency per subject. A 600 ms wait time was employed after a “start trial” button press, and before the presentation of stimuli, in order to separate T2 traces and motor activity from the neural response to T1.

During blocks, a black screen with white fixation cross was presented. Participants were instructed to press the “Start” button on a USB joypad to begin each trial and tasked with identifying if T2 was a higher or lower frequency than T1 by pressing buttons labeled either “Higher” or “Lower”. A short break at the participant’s discretion was given every 60 trials, with the percentage of correct responses over the last 60 trials displayed on screen during the break. Note that our primary research question focused on “silent” neural activity related to the echoic memory of T1 during the period in which the frozen noise burst was presented, and the task and performance feedback served merely to

incentivize participants to continue with the task through the blocks with similar levels of attention and engagement throughout. Although task presence can be thought to introduce an element of active retention (thus bringing the paradigm into the purview of WM), we are differentiating ASM vs WM in the context of time scales rather than task requirements. No feedback was given on a trial-by-trial basis, nor were rewards or punishment employed. In total, participants completed 600 task trials.

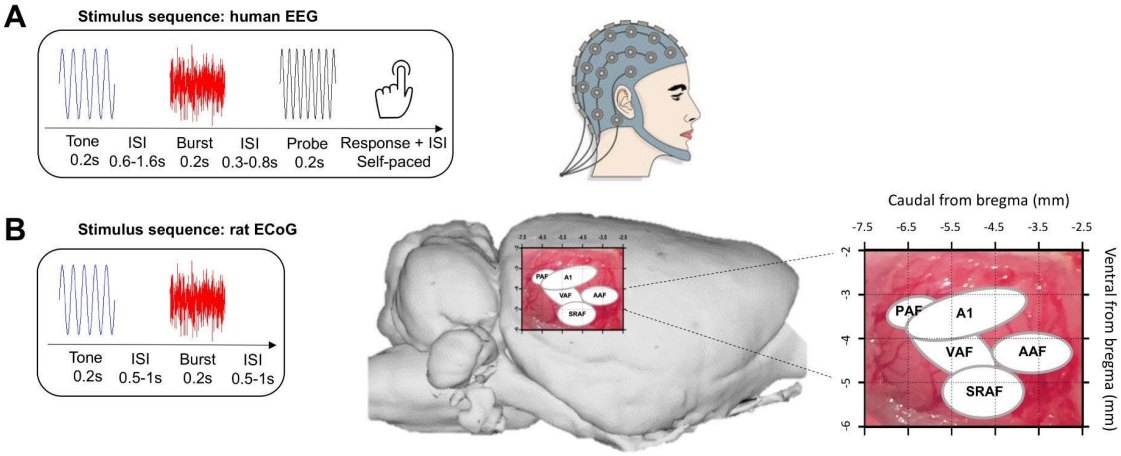

Figure 2.1 Stimuli and recording techniques. **(A)** Human volunteers were exposed to homologous stimulus streams in an auditory sensory memory task, in which they were asked to report whether the frequency of the sample tone was higher or lower than the frequency of the probe tone. **(B)** Rats were exposed to stimulus streams under anesthesia (adapted from Polley, 2007).

Neural Data Acquisition and Pre-processing

Neural data was collected via an ANT Neuro EEGo Sports 64 channel 10-20 electroencephalogram (EEG), grounded at the nasion and referenced to the Cpz electrode, at a sampling rate of 1000 Hz. Each participant completed 10 blocks (~1.5 hours), recorded in succession with short breaks. Participants were seated in a quiet room, fitted with Brainwavz B100 earphones, which delivered the audio stimuli via a MOTU Ultralite MK3 USB soundcard at 44.1 kHz, 16 bit. Data from all 21 subjects were included in EEG analysis, with 17 subjects being included in behavioral analysis due to inaccuracies in recording incorrect responses for the initial four subjects. EEG data was pre-processed using the SPM12 Toolbox (Wellcome Trust Centre for Neuroimaging, University College London; RRID: SCR_007037) for MATLAB (The MathWorks; RRID: SCR_001622). Continuous data were high-pass filtered at 0.2 Hz and downsampled (using antialiasing filtering) to reduce the source sampling rate to 300 Hz for computational efficiency. A notch filter was then applied between 48 Hz and 52 Hz before low-pass filtering at 90 Hz. All filters were 5th order zero-phase Butterworth. Eyeblink artefact detection was performed using channel Fpz for all but one subject (for whom Fz was substituted as a result of a bad Fpz channel on that subject's recording), and the eyeblink artefacts were removed by subtracting their two spatiotemporal principal components from all EEG channels (Ille et al., 2002). Data were then re-referenced to the average of all channels, segmented into epochs ranging from -100 ms before to 500 ms after stimulus onset for all stimulus events of interest (Sample Tone, frozen noise burst and Probe Tone), and denoised using the "Dynamic Separation of Sources" (DSS) algorithm (de Cheveigné and Simon, 2008). This denoising procedure is commonly

applied to maximize reproducibility of stimulus-evoked responses across trials, while preserving differences between responses evoked by different stimulus types (de Cheveigné and Parra, 2014; de Cheveigné and Simon, 2008). For each subject, epoched data were linearly detrended, and the first seven DSS components (constituting the most reproducible components, as determined based on data ranging from −100 to 500 ms relative to tone/frozen noise burst onset) were retained and used to project both the tone-evoked and frozen noise burst data back into sensor-space.

2.3.2 Animal Electrocorticography

Subjects, Experimental Apparatus and Surgical Procedures

Five adult female Wistar rats, acquired from the Chinese University of Hong Kong, were used as subjects. Naive rats aged between 16 and 24 weeks (median age = 20 weeks) with weights between 257 and 345 g (median weight = 285 g) were tested for normal hearing (click auditory brainstem response thresholds < 20 dB) and received no prior stimulus exposure. A mixture of ketamine (80 mg/kg, intraperitoneal injection; i.p) and xylazine (12 mg/kg, i.p) was used to induce anesthesia at the outset of the experiment. Dexamethasone (0.2 mg/kg, i.p) was delivered before surgery as an anti-inflammatory. Anesthesia was maintained throughout the experiment via urethane injections (0.75 mg/kg, i.p) one hour after the initial dose of ketamine and xylazine with supplementary doses (0.2 - 0.5 ml) delivered based on the presence of a withdrawal reflex when the animals' toes were pinched. Based on previous rodent studies (Malmierca et al., 2019), this protocol allowed for faster induction of anesthesia via the initial administration of ketamine and xylazine, while avoiding subsequent NMDA-specific

inhibitory effects of ketamine through the use of urethane to maintain anesthesia for electrocorticography (ECoG) recordings. The anesthetized animal was placed in a stereotaxic frame to allow hollow ear bars to be placed for sound delivery and fix the animals' head for craniotomy. Body temperature was maintained at 36 ± 1°C with an electric heating pad throughout the procedure and monitored via rectal thermometer. During surgery the skin was cut and muscle tissue over the temporal lobe of the skull was removed to allow for a unilateral craniotomy exposing a 5×4 mm region over the right primary auditory cortex, leaving the dura intact. The cranial window started 2.5 mm posterior from the Bregma, and ventral from the temporal edge of the lateral skull surface, in order to locate the auditory cortex during craniotomy (Figure 2.1B). The ECoG array was placed over the exposed cortex and a cotton roll was placed between the skin and the array to keep impedance low and the array securely in place.

Correct placement of the ECoG array was verified by recording a set of Frequency Response Areas (FRAs) from each site by collecting responses to 100 ms pure tones varying in sound level (30 - 80 dB SPL) and frequency (500 - 32k Hz, ¼ octave steps). Each tone was presented 10 times, in a randomly interleaved fashion, with an onset-to-onset ISI of 500 ms. FRA maps for each ECoG array placement were used to verify the placement of the array was similar across subjects. Note also that the spatial PCA analysis (described below) which underpins our analysis was performed separately for each subject, which minimizes any effects of the array misalignment between subjects.

Experimental Paradigm and Stimulus Design

For the main experiment, the stimulus sequence closely matched the sequence administered in the human study but was adjusted for the rat hearing range and passive delivery. Audio sequences consisted of tones followed by the same frozen pink noise bursts used in the human paradigm, each separated by gaps of silence with a duration chosen randomly from the interval 0.5-1 s. Tones were randomly drawn from a set of six frequencies spaced seven semitones apart, beginning with 1100 Hz. The lower limit of 1100 Hz ensured that all tones were well above the lower limit of the rat's frequency range, and the seven semitone spacing (just over half an octave) was chosen to ensure that the tones should relatively easily discriminable for the rat's auditory system, but we also wanted to avoid frequency steps corresponding to integer number of octaves so that all tones differed not just in pitch height but also in pitch chroma. To our knowledge there is currently no evidence that the rat auditory system is sensitive to pitch chroma or designed to perceive octave equivalence, but ensuring that there could be no "issues chroma confusion" if they did was an easy precaution to take. Each tone was presented binaurally in a random order 50 times each per block, with the two animals exposed to two blocks and the remaining exposed to three blocks. Both tones and noise bursts were 200 ms in duration, with tones tapered with 5 ms cosine on/off ramps (Figure 2.1B).

Neural Data Acquisition and Pre-processing

ECoG signals were acquired at a sampling rate of 24,414 Hz using a 8 x 8 Viventi ECoG electrode array (Woods et al., 2018) with 400 μm electrode spacing, three ground channels located in the corners of the array, and a common reference. The array was connected to a Tucker Davis Technologies (TDT) PZ5 neurodigitizer and recorded via a

RZ2 processor (controlled by BrainWare software). Acoustic stimuli were delivered by a TDT RZ6 multiprocessor at a playback sampling rate of 48,828 Hz. To extract neural activity evoked by acoustic stimuli, the recorded electrode signals were first low-pass filtered using a cutoff frequency of 90 Hz using a 5th order Butterworth filter, and downsampled to 300 Hz. We decided to analyze low-frequency (local field) potentials rather than high-gamma-band activity, as they provide a closer homologue to human EEG signals. As for the human EEG data, the pre-processed signals were then re-referenced to the average of all channels, as commonly used in ECoG studies (Ball et al., 2009), and segmented by extracting 600 ms long voltage traces from −100 ms to +500 ms relative to the onset of each tone or frozen noise burst stimulus. The epoched traces were baseline-corrected by subtraction of the mean pre-stimulus voltage values, and linearly detrended (Salisbury, 2012).

2.3.3 Univariate analysis: summarizing tone-evoked and frozen noise burst-evoked activity

As an initial step, EEG and ECoG data were subject to univariate analyses, to assess whether tone frequency modulated tone- and frozen noise burst-evoked activity on a channel-by-channel basis. Epoched data were averaged across trials, separately for each tone frequency. First, to visualize the evoked responses, event-related potentials (ERPs) were concatenated across tone frequencies and participants/animals, resulting in two-dimensional matrices with single channels along one dimension and concatenated time points, tone frequencies, and participants/animals along the second dimension. These matrices were then subject to principal component analysis using singular value decomposition, resulting in spatial principal components (describing channel

topographies) and temporal principal components (describing voltage time-series concatenated across tone frequencies and participants/animals), sorted by the ratio of explained variance. The top principal components explaining 95% of the original variance were summarized by calculating their weighted average, weighted by the proportion of variance explained. The resulting summarized voltage time-series were then averaged per tone frequency across participants/animals. In an identical procedure, frozen noise burst-evoked single-trial data were averaged across trials, separately for each preceding tone frequency, and subject to principal component analysis as described above.

Next, to test whether any time points and channels show significant amplitude correlations with tone frequency, single-participant ERP data in the original sensor space (i.e., prior to the principal component analysis, which was only used for visualization purposes) were converted into three-dimensional images (two spatial dimensions and one temporal dimension) and entered into a general linear model (GLM), separately for each species (humans, rats) and stimulus type (pure tone, neutral frozen noise burst). Each GLM was based on a flexible factorial design with one random factor (participant / rat) and one fixed factor (tone frequency / preceding tone frequency). A parametric linear contrast across six frequencies was designed to test for the effect of tone frequency on ERP amplitude. The resulting statistical parametric maps were thresholded at $p<0.05$ (two-tailed) and corrected for multiple comparisons across spatiotemporal voxels at a family-wise error (FWE)-corrected $p = 0.05$ (cluster-level) (Kilner et al., 2005).

The human EEG data were additionally source-localized, to infer the most probable cortical sources contributing to the sensor-level effects. Specifically, since we observed a significant univariate effect of tone frequency on the amplitude of tone-evoked

responses (but not noise-evoked responses; see Results), we focused on estimating the source activity underlying the sensor-level effects of tone frequency on tone-evoked responses. To this end, we used the multiple-sparse-priors approach to source localization under group constraints (Litvak and Friston, 2008). For each tone frequency and participant, the entire time epoch (from 100 ms before to 500 ms after tone onset) of sensor-level tone-evoked responses over all EEG channels were subject to source localisation. The resulting source estimates within the time window in which we observed significant results (113-260 ms relative to tone onset; see Results) were converted into 3D images (in MNI space), smoothed with a 6×6×6 mm Gaussian kernel, and entered into a general linear model with one within-subjects factor (Tone Frequency) and one between-subjects factor (Participant). Following estimation of the general linear model, we obtained statistical parametric maps for parametric linear contrasts between which were then thresholded at $p < 0.05$ (two-tailed, uncorrected). Significant effects were inferred at a cluster-level $p < 0.05$ (FWE, small-volume corrected), correcting for multiple comparisons across voxels under random field theory assumptions (Kilner et al., 2005). Sources were labeled using the Neuromorphometrics atlas, as implemented in SPM12.

2.3.4 Multivariate analysis: decoding sensory and mnemonic tone frequency information

To test whether information about tone frequency can be decoded from the pattern of tone-evoked and frozen noise burst-evoked activity observed across multiple channels and time points, we subjected the data to multivariate analyses. To this end, we adapted methods established in previous research on multivariate EEG decoding of visual stimulus orientation during visual WM tasks (van Ede et al., 2018; Wolff et al., 2017, 2015)

and similar approaches in decoding active retention of auditory stimuli (i.e. pure tones) during WM (Wolff et al., 2019).

A multivariate decoding method was employed in analyzing data acquired from both species to decode the frequency of the preceding T1 stimulus from neural activity evoked by the frozen noise bursts which did not carry any overt information about the sample tone given that the noise tokens used were always identical and were presented well after sample tone-evoked responses returned to baseline (i > 600 ms following the offset of sample tone in humans and 500 ms in anesthetized rats). Channels with an average signal-to-noise ratio (SNR; defined as the ratio of root-mean-square values of post-stimulus and pre-stimulus amplitudes) lower than 8 dB (Alaerts et al., 2009) were discarded from analysis. This resulted in discarding 3.17% ± 1.53% EEG channels (mean ± SD) from subsequent multivariate decoding. All ECoG channels in all rats fulfilled the SNR criterion and were used in subsequent decoding. Prior to decoding, single trial tone-evoked responses were sorted by tone frequency, and single-trial frozen noise burst-evoked responses were sorted by preceding tone frequency.

We sought to determine whether activity evoked by the sample tone (probing the sensory trace), and/or by the frozen noise burst (probing ASM contents), contained information about the sample tone feature (Figure 2.2). To estimate decoding time-courses, we adopted a sliding window approach, integrating over the relative voltage changes within a 100 ms window of each time-point (Wolff et al., 2019). This approach is a direct replication of previously established multivariate decoding methods (Wolff et al., 2020). Furthermore, pooling information over multiple time-points (in addition to multiple channels) in a multivariate manner has been shown to boost decoding accuracy

(Grootswagers et al., 2017; Nemrodov et al., 2018). To this end, per channel and trial, the time segments within 100 ms of each analyzed time-point were down-sampled by binning the data over 10 ms bins, resulting in a vector of 10 average voltage values per channel. Next, the data were de-meaned by removing the channel-specific average voltage over the entire 100 ms time window from each channel and time bin. This step ensured that the multivariate analysis approach was optimized for decoding transient activation patterns (voltage fluctuations around a zero mean) at the expense of more stationary neural processes (overall differences in mean voltage) (Wolff et al., 2019). The vectors of binned single-trial temporal data were then concatenated across channels for subsequent leave-one-out cross-validation decoding. As a multivariate decoding metric, we used the Mahalanobis distance (De Maesschalck et al., 2000), taking advantage of the potentially monotonic relation between tone frequency and neural activity (Auksztulewicz et al., 2019; Wolff et al., 2019). In other words, responses to similar tones are expected to yield low Mahalanobis distance metrics, while responses to more dissimilar tones are expected to yield larger Mahalanobis distance metrics. In a leave-one-out cross-validation approach (which has been shown to be optimal for EEG decoding (Grootswagers et al., 2017) per trial, we calculated 6 pairwise distances between EEG/ECoG amplitude fluctuations measured in a given test trial and mean vectors of EEG/ECoG amplitude fluctuations averaged for each of the 6 tone frequencies in the remaining trials. The Mahalanobis distances were computed using the shrinkage-estimator covariance obtained from all trials excluding the test trial (Ledoit and Wolf, 2004). This approach, combining Mahalanobis distance with Ledoit–Wolf shrinkage, has been previously shown to outperform other correlation-based methods of measuring dissimilarity between brain

states (Bobadilla-Suarez et al., 2019). Mahalanobis distance-based decoding has also been shown to be more reliable and less biased than linear classifiers and simple correlation-based metrics (Walther et al., 2016).

The single-trial relative Mahalanobis distance estimates were then averaged across trials per tone frequency (for tone-evoked responses) or preceding tone frequency (for frozen noise burst-evoked responses), resulting in a 6 x 6 distance matrix for each analyzed time point. Overall decoding quality was quantified by comparing the estimated distance matrices with an “ideal decoding” distance matrix, with the lowest distance values along the diagonal and linearly increasing distance values along the off-diagonal. To obtain an easily interpretable measure of decoding quality, for each participant/animal and time point (from 50 ms before to 450 ms after tone/frozen noise burst onset), we normalized the observed and ideal decoding matrices by de-meaning and dividing the entire matrix by its maximum absolute value, and calculated the linear regression slope coefficient between the estimated distance matrix and the ideal distance matrix. Following data normalization, the resulting regression coefficients ranged between -1 (below-chance decoding) and 1 (ideal decoding), and formed decoding time-series which effectively summarized, per time point, how well the observed decoding matrices approximate the ideal decoding matrix. These decoding time-series were then smoothed with a Gaussian smoothing kernel (s.d. = 16 ms; Wolff et al., 2019) and averaged across participants/animals.

Furthermore, to quantify the comparison between decoding based on human EEG and rat EcoG, we performed a representational similarity analysis on the estimated distance matrices at six different time points (from 50 ms before to 450 ms after stimulus

onset, in 100 ms steps) for the tones and bursts. Specifically, for both human EEG and rat ECoG data, we calculated pairwise Pearson correlation coefficients (Walther et al., 2016) between 12 distance matrices (obtained for tones and bursts, at 6 time points). This resulted in 12x12 distance correlation matrices which summarized how similar the multivariate decoding of tone frequency is across time points as well as between tone-evoked and burst-evoked responses.

To establish the null distribution for statistical testing, we used a permutation-based approach, such that in each permutation the single-trial relative distance metrics were randomly reassigned stimulus labels. The resulting reshuffled single-trial decoding estimates were averaged across trials to obtain surrogate distance matrices. These distance matrices were then normalized and subject to linear regression, smoothing over time, and averaging across participants/rats, as described above. This procedure was repeated 10,000 times to obtain a null distribution of decoding estimates for each time point. Per time point, p-values quantifying the significance of observing above-chance decoding were calculated by counting the proportion of surrogate decoding estimates exceeding the observed decoding estimate. Across time points, p values were corrected using a false-discovery-rate (FDR) approach at an FDR = 0.05 (Benjamini and Hochberg, 1995). This procedure allowed for implementing exactly the same statistical procedures for both EEG and ECoG datasets.

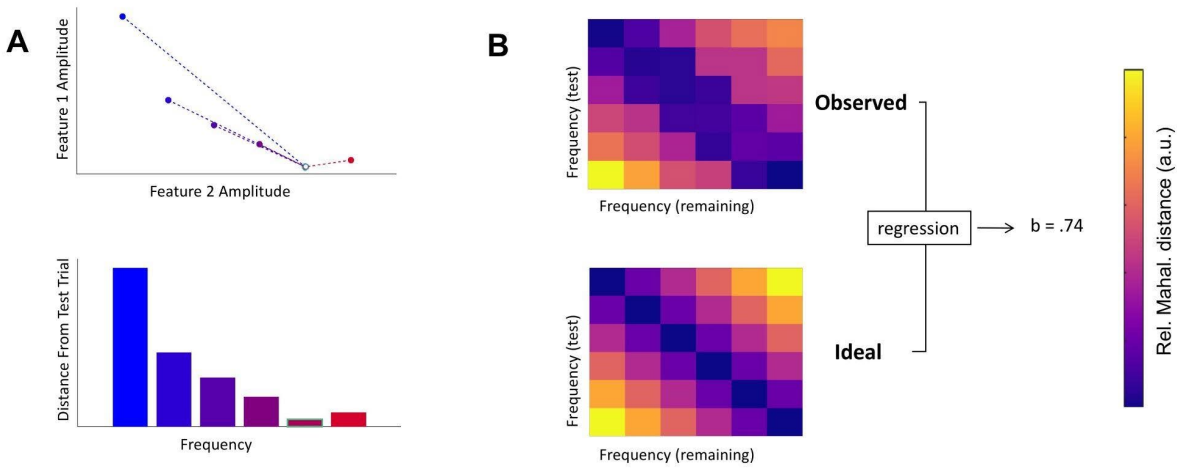

Figure 2.2 Decoding Method. **(A)** Decoding methods were based on estimating multivariate Mahalanobis distance between EEG/ECoG feature amplitudes in a given (test) trial and average amplitudes calculated for all 6 frequencies features, respectively (excluding the test trial). The top panel presents EEG/ECoG feature amplitudes for two example features (empty circle, test trial; solid circles, ERPs calculated from the remaining trials; acoustic frequencies are color coded). Dashed lines on the top panel and bars on the bottom panel represent the multivariate distance between amplitudes observed in the test trial and the remaining trials. **(B)** Frequency-tuning matrices summarizing the population-level tuning curves, were obtained after averaging across trials, per frequency, resulting in a 6 × 6 similarity matrix between all tone frequencies (each row represents the distance of all test trials of a given frequency to the remaining trials sorted per frequency and is shown in columns). The observed frequency-tuning matrices (top, example from one participant) were regressed with the "ideal" tuning matrix (bottom), which consisted of the difference (in Hz) between pairs of tone frequencies. This regression coefficient provided a summary statistic that reflects decoding quality (i.e., how closely the relative dissimilarity between tone-evoked neural responses; "observed" in the figure) corresponds to the relative dissimilarity between tone frequencies ("ideal" in the figure).

2.4 Results

Behavioral Results

Performance across all human subjects yielded an average 79% accuracy rate (SEM = 3.49%). A repeated measures analysis of variance (RM ANOVA) was conducted on the dependent variable accuracy, with within-subject factors of probe divisions, and with a random factor of participants. Another set of RM ANOVA was performed on the hit

rates with within-subject factors of ISIs. ISIs were categorized into 10 time windows for analysis, with the windows derived from equal range divisions between the smallest and largest ISI times. RM ANOVAs revealed no effect between ISI time windows and performance ($p > 0.05$), and a significant effect on task performance related to the distance of T2 frequency detuning relative to T1 ($F(1,16) = 518.45$, $p < 0.001$), with smaller detuning values resulting in more incorrect responses. These results indicate that the task was sufficiently difficult to keep subjects engaged and memory items reliably retained during trials.

Univariate analysis: single-channel correlations with tone frequency

Our univariate analysis compared the averaged ERPs per T1 frequency, as well as averaged ERPs for the frozen noise bursts that followed given T1 frequency values. Using the family-wise error (FWE) corrected univariate tests outlined in the methods we observed a significant effect of tone frequency on tone-evoked responses in both human (EEG) and rat (ECoG) datasets (Figure 2.3). In human EEG, a single cluster of amplitudes correlated with tone frequency, extending over bilateral anterior and right temporal channels and ranging between 113 and 260 ms after stimulus onset (p FWE = 0.021, Tmax = 3.27). Using a source localization procedure (see Methods), we inferred the most likely cortical sources contributing to these source-level effects, which were localized in the right superior temporal gyrus (MNI coordinates: [48 -6 -16]; cluster-level p FWE = 0.007; Fmax = 21.01; Zmax = 3.93) and in the right middle/inferior frontal gyrus (MNI coordinates: [38 50 -2]; cluster-level p FWE = 0.022; Fmax = 10.84; Zmax = 2.87). Similarly, in rat ECoG, a single cluster of amplitudes with a broad spatial distribution and a temporal range of 63-160 ms post-onset was observed (p FWE = 0.012, Tmax = 7.21).

In contrast, frozen noise burst-evoked responses did not correlate with preceding tone frequency (human EEG and rat ECoG: all clusters $p > 0.8$), indicating that univariate analyses are not sufficient to decode frequency labels in either EEG or ECoG based on ERP amplitude of frozen noise burst-evoked responses.

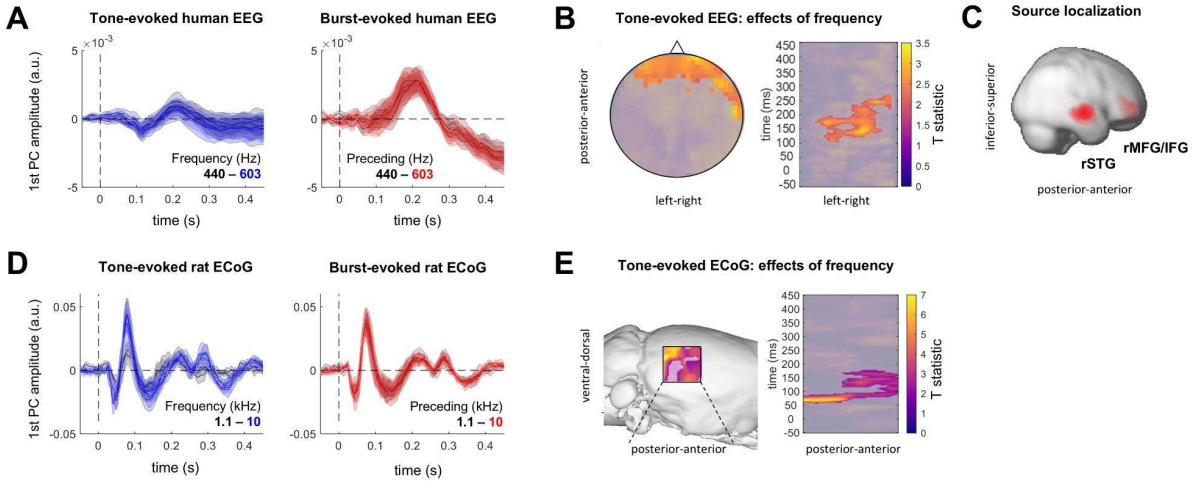

Figure 2.3 Univariate analyses. **(AD)** In Humans and Rats tones and frozen noise bursts evoked robust neural activity; different frequencies are represented as individual traces, from lowest frequencies (black traces) to highest frequencies (blue/red traces). Shaded areas denote SEM across subjects. **(BE)** In Humans and Rats tone-evoked activity correlated with tone frequency (parametric contrast T values; highlighted clusters: pFDR<.05). However, no significant effects of tone frequency were observed in univariate analyses of frozen noise burst-evoked activity. **(C)** Source localization of the univariate effect of tone frequency on tone-evoked EEG responses. Significant sources of activity, whose activity was parametrically related to tone frequency, were identified in the right superior temporal gyrus (rSTG) and in the right middle/inferior frontal gyrus (rMFG/IFG).

Multivariate analysis: decoding tone frequency from transient response patterns

Our multivariate analysis computed distance matrices for neural responses evoked by tones of different frequencies, or by neutral frozen noise bursts preceded by tones of given frequencies. These matrices were compared to an “ideal decoding” distance matrix

to quantify overall decoding quality. This analysis revealed that, similar to the univariate analysis, tone frequency was reflected in tone-evoked response amplitudes (Figure 2.4). In human EEG data, significant decoding (p FDR<0.05) was observed between 10 and 450 ms relative to tone onset (all betas > 0.041, peak beta = 0.175; all p < 0.027), while in rat ECoG data, significant decoding was observed between -23 and 413 ms relative to tone onset (all betas > 0.041, peak beta = 0.709; all p < 0.033). Please note that each decoding estimate for a given time point is based on data pooled over a 100 ms time window centered around that time point; hence, the exact latency of decoding onsets should be treated with ±50 ms precision. Taken together, tone frequency could be robustly decoded from tone-evoked activity in both humans and rats.

However, unlike in the univariate analysis, tone frequency was also reflected in subsequent frozen noise burst-evoked response amplitudes. Decoding of previously-heard T1 frequency from frozen noise burst EPRs was present in both EEG and ECoG. Significant decoding in EEG occured in our analysis between 247 and 343 ms relative to frozen noise burst onset (all betas > 0.036, peak beta = 0.070; all p < 0.027, FDR-corrected). In ECoG data, significant decoding was present in three time windows (early cluster: 13-160 ms post-frozen noise burst, all betas > 0.071, peak beta = 0.192; middle cluster: 226-303 ms post-frozen noise burst, all betas > 0.075, peak beta = 0.110; late cluster: 343-400 ms post-frozen noise burst, all betas > 0.062, peak beta = 0.118; all p < 0.033, FDR-corrected). Given the relatively low number of rats, we inspected individual rats' decoding peaks to exclude the possibility that the three significant clusters result from individual differences in peak latencies across rats. For each of the identified clusters, the majority of individual rats had at least one decoding peak within a given

cluster (cluster 1: 4/5 rats; cluster 2: 4/5 rats; cluster 3: 3/5 rats). Thus, in both species, the neural response elicited by the auditory frozen noise burst contained statistically significant information about the previously-heard stimuli retained in the sensory memory hold.

We have further quantified the similarity in decoding matrices obtained for tone-evoked and burst-evoked responses (Figure 2.4CF) in a representational similarity analysis (Kriegeskorte et al., 2008). This has revealed that the decoding correlation patterns were qualitatively similar across both species - i.e., significant correlations were observed in both species, both across time points within tone-evoked and burst-evoked responses, as well as between tone-evoked and burst-evoked responses. However, the decoding matrices based on rat ECoG data were relatively more similar across time points and between tone-evoked and burst-evoked responses than the decoding matrices based on human EEG data. Specifically, the decoding matrices based on rat ECoG data were highly correlated across all post-stimulus time points for both tone-evoked responses (all pairwise rho > 0.9, all $p < 0.001$; 50-450 ms after tone onset) and burst-evoked responses (all pairwise rho > 0.6, all $p < 0.001$; 50-350 ms after burst onset), as well as between tone-evoked and burst-evoked responses (all pairwise rho > 0.7, all $p < 0.001$; between 50-450 ms after tone onset and 50-350 ms after burst onset). In contrast, the decoding matrices based on tone-evoked human EEG responses were only highly correlated across neighboring time points (rho = 0.959, $p < 0.001$ for 50/150 ms after tone onset; rho = 0.913, $p < 0.001$ for 250/350 ms after tone onset) and, to a smaller extent, across more distant time points (all remaining pairwise rho > 0.396, $p < 0.018$). However, they were less consistently correlated across time points for burst-evoked responses

(maximum rho = 0.5828, p < 0.001), and between tone-evoked and burst-evoked responses (maximum rho = 0.7820, p < 0.001).

In a control analysis, to ensure that frequency decoding based on noise burst-evoked responses is not driven by trials presented at short ISIs (and possibly contaminated by the neural response evoked by the preceding tone), we entered the single-trial epoched data into a linear regression and, per channel and time point, calculated the residual after regressing out the ISI preceding the noise burst onset from single-trial amplitude values. These residuals were then used to obtain decoding estimates for both human EEG and rat ECoG data, as described above. In both cases, the decoding results were virtually identical as in the original analysis, with all previously reported clusters of significant decoding also showing statistical significance in the control analysis, and no additional clusters appearing in the control analysis. Therefore, it is unlikely that trial-by-trial differences in ISI between tone and noise burst could have contributed to the decoding results reported above.

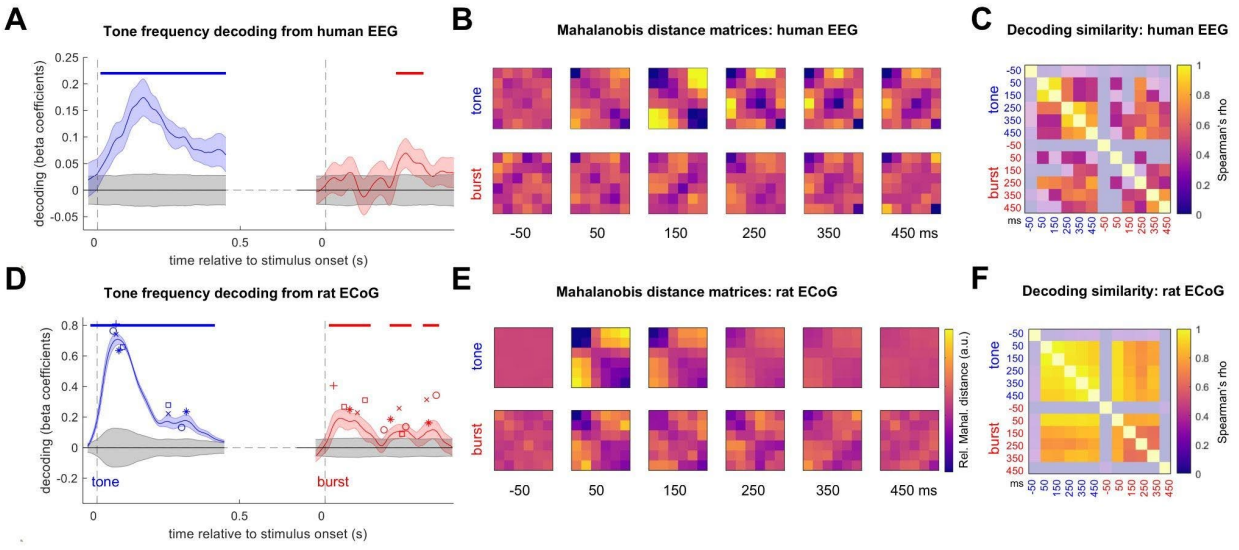

Figure 2.4 Multivariate analyses. **(AD)** In Humans and Rats, tone frequency could be decoded from both tone-evoked (blue) and frozen noise burst-evoked activity (red). Red/Blue shaded areas: SEMs across

subjects. Grey shaded areas: 95% confidence intervals of the null distribution of decoding time-series, reflecting the range of values for which decoding could have been observed by chance. Horizontal bars: pFDR<.05. Individual markers in (D) represent individual rats' decoding peaks. **(BE)** Relative Mahalanobis distance matrices per time point, forming the basis for decoding (beta coefficients) in (AD). **(CF)** Pairwise similarity (Pearson correlation coefficients) of the Mahalanobis distance matrices. Saturated colors mark $p < 0.05$.

2.5 Discussion

As established in previous research, the neural response to a sensory frozen noise burst contains information about the contents of WM held in the activity-silent period (Wolff et al., 2019, 2015). We elaborated on these findings by designing a task that does not require active retention and uses smaller time intervals to place stimuli in the range of lower-level auditory sensory memory, applying this technique in a cross-species approach. We demonstrate that stimulus feature can be decoded from the evoked response to stimuli events using a univariate analysis, where ERP amplitude modulates parametrically with stimulus value in both anesthetized rat and awake human subjects, consistent with existing research conducted in awake humans (Auksztulewicz et al., 2019; Wolff et al., 2015). It is worth noting that the significant decodability that is visible in the baseline of our stimulus decoding (Figure 2.4) can be attributed to the length of the sliding time window in decoding based on spatiotemporal patterns of transient responses. Our use of smaller intervals between frequencies in human trials (six tones, one semitone apart) further demonstrates that this technique is robust enough to decode more subtle differences in auditory stimuli than demonstrated in previous literature (Wolff et al., 2019), and comparable results in anesthetized rats under passive stimuli exposure serve as a counterpoint to a WM interpretation of behavioral paradigms employing active tasks while operating within the ASM temporal range.

In our human EEG data, the observed univariate relationship between ERP amplitude and tone frequency could be source-localized to the right higher-order auditory cortex (superior temporal gyrus) and frontal regions (middle/inferior frontal gyrus). These findings are consistent with the right-lateralisation of spectral (vs. temporal) auditory processing in the superior temporal gyrus (Britton et al., 2009; Poeppel, 2003; Schönwiesner et al., 2005) and with the parametric encoding of tone frequency in the right frontal cortex during memory tasks (Spitzer and Blankenburg, 2012). Importantly, the fact that sensor-level EEG effects of tone frequency could also be source-localized to auditory regions (in addition to frontal regions) make the human EEG results more comparable to the rat ECoG data, which were based on signals recorded only over auditory regions.

In addition to decoding a stimulus feature from the neural response to stimulus events, we also demonstrate that equivalent paradigms can be used to decode stimulus feature from neural responses to uncorrelated auditory frozen noise bursts (2019; Wolff et al., 2020, 2015). Interestingly, no relationship was found between EEG/ECoG response amplitudes to stimulus events and frozen noise bursts of the same preceding tone feature in univariate analysis, illustrating the need for multivariate decoding methods. The results of our multivariate EEG decoding show significant bursts later in the time course (400-500 ms) than found in similar research (Wolff et al., 2019), suggesting that periods of decodability may be task-dependent. Our study employed a significantly more narrow range of token values than previously attempted, possibly testing the limits of this particular decoding method in the context of silent-state neural encoding of memory tokens. Although the longer latency of frequency decoding from noise bursts was

surprising, both noise-evoked and tone-evoked ERPs (Figure 2.3C) were typical, showing comparable latencies and peaks, with tone-evoked decoding (Figure 2.4D) also appearing at the expected latency. Additionally, late reactivation of latent WM traces have also been shown by some earlier studies (Wolff et al., 2015) and may be better understood in the context of template matching, with late reactivation required for comparison in the EEG task (Myers et al., 2015; Wolff et al., 2015)

Traditionally, neural correlates of ASM in humans have been investigated in oddball paradigms yielding mismatch negativity (MMN) responses (Winkler et al., 1993). MMN responses to deviant stimuli during passive auditory oddball paradigms can also be observed in the absence of consciousness, e.g. in some comatose patients (Morlet and Fischer, 2014). Interestingly, previous MMN studies also fall in line with silent-coding theories, as MMN responses have been postulated to result from deviant stimuli in comparison to an existing ASM trace (Näätänen et al., 2005). Delayed match to sample (DMTS) tasks have also been considered a reliable method of investigating both sensory and WM (Daniel et al., 2016), and classical studies have employed DMTS paradigms to establish psychometric functions of ASM retention and decay periods (Nees, 2016). In addition to their usefulness in behavioral studies, DMTS paradigms have been employed in modern human WM research (Myers et al., 2015; Wolff et al., 2019, 2015) to investigate neural components of sensory and WM traces. Interestingly, despite the use of rats in auditory research, many of which use similar DMTS tasks with jittered time intervals, there is sparse literature on ASM periods in the rat (O'Connor and Ison, 1991).

In this study, we sought to fill several gaps in the existing literature by employing an ASM task to investigate neural correlates of ASM across species using contemporary

decoding methods. In the broader scope of sensory memory research, our work employs a useful new tool in observing and analyzing neural phenomena. Existing tools, such as MMN and stimulus specific adaptation (SSA) access this information indirectly by assessing the modulation of a neural response to a particular repeated stimulus (Carbajal and Malmierca, 2018). As demonstrated in previous studies (Costa-Faidella et al., 2011), repetition effects can be observed in time intervals coinciding with windows typical of both ASM and SSA, making such findings somewhat ambiguous given the presence of adaptation effects across multiple timescales in the auditory system. Similarly, multiple time scales of adaptation corresponding to stimulus duration have been observed in single-unit cortical recordings in anesthetized cats (Ulanovsky et al., 2004), and recent work has demonstrated topographically-organized tone selectivity in SSA across multi-unit cortical recordings in the anesthetized rat (Nieto-Diego and Malmierca, 2016). Taken together, this may provide a possible explanation for our observed decodability using a multivariate approach, and the lack thereof using a univariate approach, as the recording methods employed in our study measure much larger neural populations than those possibly responsible for underlying SSA selectivity, and as a result lack the sensitivity to measure the more fine-grained patterns of tone selectivity accessible in single-unit recordings (Natan et al., 2017; Nieto-Diego and Malmierca, 2016; Ulanovsky et al., 2004). As such, our decoding approach may provide an invaluable tool in assessing this phenomenon, that is applicable to both invasive recordings in animal models and non-invasive human models.

Our cross-species approach is, to the authors' knowledge, the first attempt at decoding auditory memory traces in different species using the same analysis method

and comparable stimuli. While the decoding matrices obtained from rat ECoG data were more strongly correlated across time points and between tone-evoked and burst-evoked responses (Figure 2.4CF), qualitatively similar correlation patterns were observed for decoding matrices obtained from human EEG data. Taken together, these findings suggest that the neural encoding of sensory memories is a general mechanism that has been evolutionarily maintained across species - a prospect that is also supported by previous MMN research using rat models. One such study observed a mismatch response from epidural potentials in anesthetized rats when presented with deviant tones in an oddball paradigm (Astikainen et al., 2011). Additional studies have yielded similar findings in awake and anesthetized rats using similar methods (Nakamura et al., 2011). As the ability for an organism to quickly differentiate between acoustic changes in its environment offers a potential benefit to its survival, such findings support the notion of ASM as an evolutionarily-conserved adaptive trait. Our findings, paired with those previously mentioned and the limited behavioral studies available on rat ASM, further suggest the suitability of the rat in establishing animal models for research in central auditory processing.

In contrast to previous auditory studies requiring human participants to attend to a memory item (Wolff et al., 2019), our results demonstrate that active maintenance is not required for this approach to work, placing our findings in the purview of existing human ASM research relying on MMN responses, which have been shown to be conserved across conscious states (Morlet and Fischer, 2014; Winkler et al., 1993). Of significance to the field, our findings suggest that animal models may provide an acceptable proxy for human sensory memory research, offering the benefit of significant decodability and

higher signal-to-noise ratios from electrocorticography not feasible in human subjects, with implications for research across conscious states. Future studies could capitalize on our findings, possibly applying these methods to asleep or unconscious humans and awake rats. Given the key differences between ASM and WM (e.g. ASM as an automatic process that is present across attentive states and shorter time scales than the higher-level WM system), future studies could also apply our approaches to paradigms that manipulate WM contents or investigate their efficacy in WM retention intervals. While at shorter time windows, such as those employed in our study, ASM and SSA are partially overlapping (Costa-Faidella et al., 2011), future research should also seek to establish if the observed effects differ between active memory processes and passive adaptation. Furthermore, applying the decoding methods to additional research in anesthetized rats would also prove a logical extension, as areas such as longer time scales for retention or manipulation of passively maintained memory items remain largely unexplored in this context.

Chapter 3. Simultaneous Mnemonic and Predictive Representations in the Auditory Cortex

3.1 Summary

Recent studies have shown that stimulus history can be decoded via the use of broadband sensory impulses to reactivate mnemonic representations (Cappotto et al., 2021, Stokes, 2015; Wolff et al., 2019, 2015). However, memory of previous stimuli can also be used to form sensory predictions about upcoming stimuli (Barron et al., 2020; Rust and Palmer, 2021). Predictive mechanisms allow the brain to create a probable model of the outside world, which can be updated when errors are detected between the model predictions and external inputs (Fairhall et al., 2001; Friston et al., 2006; Rubin et al., 2016; Schröger et al., 2014). Direct recordings in the auditory cortex of awake mice established neural mechanisms for how encoding mechanisms might handle working memory and predictive processes without “overwriting” recent sensory events in instances where predictive mechanisms are triggered by oddballs within a sequence (Libby and Buschman, 2021). However, it remains unclear whether mnemonic and predictive information can be decoded from cortical activity simultaneously during passive, implicit sequence processing, even in anesthetized models. Here, we recorded neural activity elicited by repeated stimulus sequences using electrocorticography (ECoG) in the auditory cortex of anesthetized rats, where events within the sequence (referred to henceforth as “vowels”, for simplicity) were occasionally replaced with a broadband noise burst or omitted entirely. We show that both stimulus history and predicted stimuli can be decoded from neural responses to broadband impulses, at

overlapping latencies, but based on independent and uncorrelated data features. We also demonstrate that predictive representations are dynamically updated over the course of stimulation.

3.2 Results

In the present experiment, we adapt recent techniques for decoding auditory working memory traces (Cappotto et al., 2021; Wolff et al., 2019, 2015) to simultaneously probe both memory and predictive processes. ECoG was recorded from the auditory cortex (AC; Figure S3.1A) of anesthetized rats (N=8) while repeated stimulus streams of vowels were presented, with vowels occasionally omitted or replaced with a broadband noise burst (Figure 3.1A). Two types of blocks were employed. In “predictable” blocks, vowels were grouped into one of six triplets (AAO, AOO, AAI, AII, OOI, OII) with each triplet presented at least 25 times in a given block of identical triplets before being replaced with another triplet (see STAR Methods). In control blocks, we presented the vowels in a pseudo-randomized order while keeping the position of bursts and omissions fixed (relative to their corresponding predictable block), to tap into mnemonic processing without predictive components (Figure 3.1B). In both types of blocks, 5% of vowels were replaced with omissions and 5% with bursts.

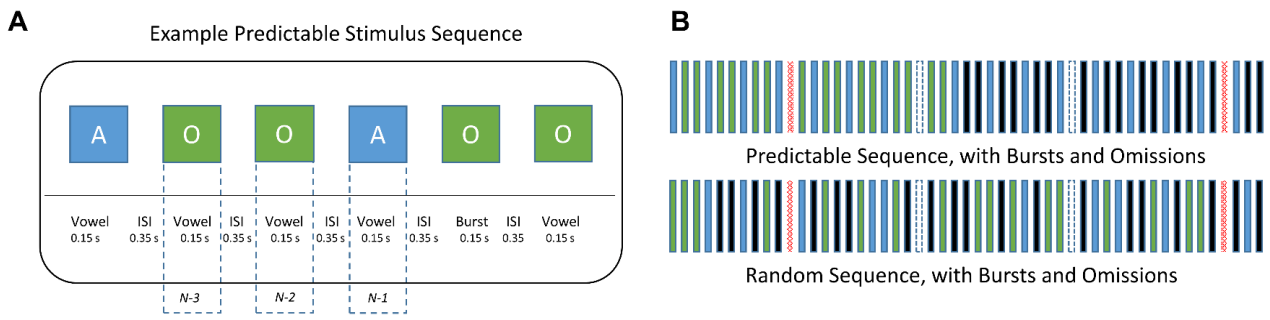

Figure 3.1 Stimulus sequences. **(A)** An example of an AOO predictable stimulus sequence, where one vowel of the triplet has been randomly substituted by a noise burst (or, not shown, alternately omitted entirely) following a minimum of three triplet repetitions. In paired random blocks, the relative position of the burst/omission substitution remains unchanged, while the surrounding vowels are randomized. Vowel positions relative to the burst/omission are denoted as N-1, N-2, and N-3. **(B)** Segment of an example predictable sequence, in which vowel tokens are omitted or replaced with a noise burst after 3 repetitions (top) and the randomized version of that sequence where vowel tokens from the full sequence are presented pseudo-randomly while burst and omission tokens remained in the same relative positions (bottom).

3.2.2 Univariate analyses: only vowel-evoked activity differentiates between vowels

To test whether vowel identity influences average neural activity, we tested for the effects of vowel (A, I, or O) and block (predictable vs. random) on vowel-evoked ECoG activity (event-related potentials). We observed that vowel-evoked activity differentiates between the three vowels, both in predictable blocks (Figure S3.2A; 13-260 ms; F_{max} = 58.56; $p_{FWE} < 0.001$) and in random blocks (Figure S3.2B; 13-207 ms; F_{max} = 58.21; $p_{FWE} < 0.001$). The main effect of block (predictable vs. random) on vowel-evoked activity was not significant (all $p_{FWE} > 0.05$).

We then tested whether burst-evoked and/or omission-evoked activity also differentiates between the (preceding) vowels at different “positions” in the sequence,

relative to the burst/omission (N-1 position: the immediately preceding vowel, N-2 position: two stimuli before the burst/omission, N-3 position: three stimuli before the burst/omission). This analysis revealed that, similarly to the vowel-evoked responses, burst-evoked responses did not significantly differentiate between predictable and random blocks (Figure S3.2C; all $p_{FWE} > 0.05$). However, unlike the vowel-evoked activity (which was modulated by vowel identity), noise burst-evoked activity was not significantly modulated by (preceding) vowel identity when neural activity was analyzed in a mass-univariate manner. Specifically, neither the effect of the immediately preceding vowel on burst responses (N-1: all $p_{FWE} > 0.05$; Figure S3.2D), nor of the previous vowels (N-2, N-3: all $p_{FWE} > 0.05$) were significant.

Omission-evoked responses peaked relatively early (83-93 ms), with a rising activity visible already prior to expected stimulus onset, possibly marking the offset response to the interrupted stimulus train rather than a true omission (Chien et al., 2019). Nevertheless, just like burst-evoked activity, omission-related activity was also not significantly modulated by block type (Figure S3.2E; all $p_{FWE} > 0.05$) or preceding vowel identity (Figure S3.2F; N-1, N-2, N-3: all $p_{FWE} > 0.05$).

3.2.3 Multivariate analysis: specific decoding boost for predictable vowels

Although in the univariate analysis burst-evoked activity did not differentiate between preceding vowels, based on our previous study (Cappotto, et al., 2021) we hypothesized that preceding stimuli can be decoded in a multivariate analysis. Specifically, by analyzing the spatiotemporal pattern of activity evoked by noise bursts, which did not carry overt information about the preceding vowels given that noise tokens

were always identical and presented after vowel-evoked responses had returned to baseline (400 ms after stimulus offset), we sought to determine if activity evoked by noise bursts contained information about the preceding vowels (separately for N-1, N-2, and N-3 vowels). This analysis revealed significant decoding of vowels up to N-3 in predictable blocks and up to N-2 in random blocks (Figure 3.2A; Table S3.1). Overall, immediately preceding stimuli could be decoded better than previous stimuli (Table S3.2), but not as well as currently processed stimuli (Table S3.3).

Crucially, if burst-evoked activity can reactivate not only mnemonic representations (irrespective of the currently processed stimulus), but also predictive representations (tokens which would have been predicted but are replaced by a noise burst), we would expect a specific decoding improvement for N-3 (but not N-2 or N-1) vowels presented in predictable blocks vs. random blocks. The decoding results were consistent with this hypothesis. Specifically, decoding was significantly improved for the N-3 vowels presented in predictable blocks relative to the random blocks (paired t-test, early cluster: 77-103 ms, $t_{max} = 3.45$, cluster-level $p_{FWE} = 0.010$; late cluster: 227-270 ms, $t_{max} = 3.79$, cluster-level $p_{FWE} < 0.001$; Figure 3.2A), suggesting that we could access a predictive representation of the vowel replaced by a noise burst. In a follow-up analysis using representational dissimilarity matrices, we found that this predictive representation contained information not only about the specific N-3 vowel replaced by the burst, but also about the entire triplet preceding the burst (Figure S3).

While mnemonic and predictive representations cou3.ld be decoded based on burst-evoked activity, decoding stimulus history based on omission-evoked activity did not yield any significant results (Figure 3.2B; all $p_{FWE} > 0.05$). This suggests that, at least

in this experimental protocol (vowel triplets) and in ECoG recorded under anesthesia, a stronger activation of the network (e.g., burst presentation) is necessary to make mnemonic and/or predictive representations observable.

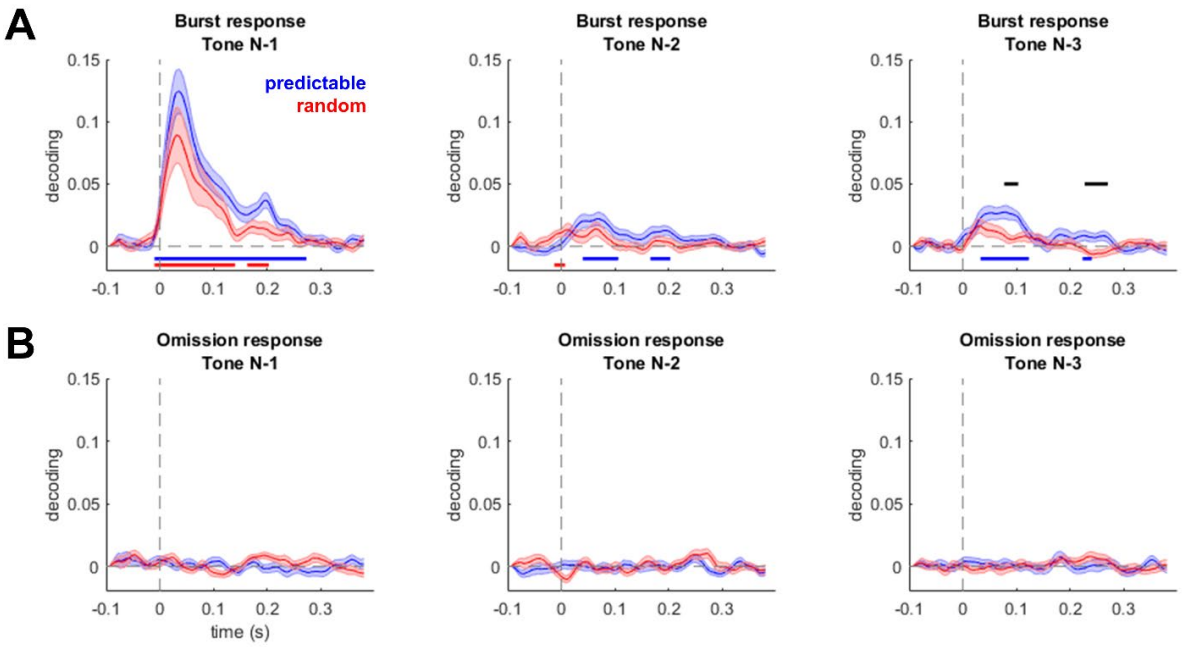

Figure 3.2 Multivariate analyses. **(A)** Time courses of decoding of preceding vowels based on burst-evoked activity. Left / middle / right panel: decoding N-1 / N-2 / N-3 vowel (blue: predictable blocks; red: random blocks; shaded area: SEM across recording sessions; blue/red horizontal line: decoding in predictable/random blocks significantly different from zero, $p_{FWE} < 0.05$; black horizontal line: decoding significantly different between predictable and random blocks, $p_{FWE} < 0.05$). Shaded area: SEM across recording sessions. See also Tables S1-S3. **(B)** Decoding based on omission responses. Legend as above.

3.2.4 Multivariate analysis: decoding of predicted vowels gradually improves over time

Having established that decoding of the predicted vowel (N-3) shows a specific improvement in predictable vs. random blocks, we sought to determine whether this boost shows features of a predictive representation. We reasoned that, in predictable blocks,

predictions should be learned over time, and consequently the decoding of the N-3 vowel should gradually improve within and across blocks containing identical triplets (Figure 3.3AE). To test this, we performed a linear regression analysis on single-trial decoding estimates, using two "learning" regressors - one quantifying possible gradual improvements of decoding within each sequence containing identical triplets (within blocks), and one quantifying possible gradual improvements of decoding over the course of the entire recording session (across blocks). We treated the random blocks as a control for passage of time (including gradual suppression of activity due to habituation, short-term plasticity to repeated presentations of stimuli, and changes in stimulus-related and baseline activity due to prolonged anesthesia) since no learning was expected in this condition. This analysis revealed that, for the early time window in which we observed a decoding boost in the predictable vs. random condition (77-103 ms), the "within blocks" learning effect was significantly higher in the predictable than in random blocks (Wilcoxon sign rank test, $Z_{21} = 2.485$, $p = 0.013$; Figure 3.3BCD), although significance testing of regression coefficients within conditions against zero did not yield significant effects (predictable: $Z_{21} = 1.477$, $p = 0.139$; random: $Z_{21} = -1.825$, $p = 0.068$). No significant learning effects across blocks were observed for the early time window (all $p > 0.5$). Conversely, for the later time window in which we observed a decoding boost (227-270 ms), the "across blocks" learning effect (Figure 3.3FGH) showed borderline significance in the predictable condition against zero ($Z_{21} = 2.033$, $p = 0.042$; uncorrected) but not in the random condition ($Z_{21} = 0.122$, $p = 0.903$), although a direct comparison of learning coefficients between conditions did not yield a significant effect ($Z_{21} = 1.303$, $p = 0.192$). The "within blocks" learning effect did not yield any significant effects in the later time

window (all p > 0.5). An additional analysis of N-2 and N-1 stimuli decoding revealed neither significant learning at either time scale, nor a significant difference in learning coefficients between predictable and random blocks (all p > 0.1). Taken together, these results provide evidence that the early N-3 decoding in predictable blocks improves at faster time scales (within blocks) relative to random blocks, but the evidence for any decoding improvement at longer time scales (across blocks) is weak.

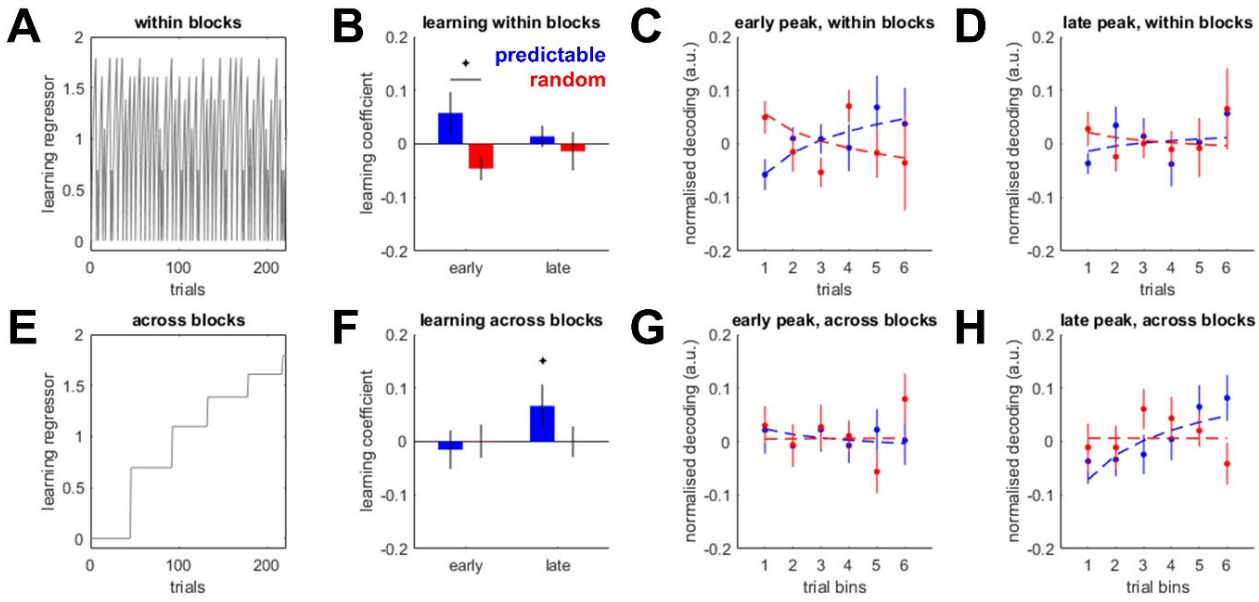

Figure 3.3 Learning effects. **(A)** A trial-by-trial regressor of learning within blocks (faster time scale) was quantified as the (log) burst number in a block of identical triplets. **(B)** Regression coefficients ("within blocks" learning) for two time windows with significant N-3 decoding boost (see Figure 3.2A, right). Error bars denote SEM across recording sessions. Asterisk denotes a significant Wilcoxon sign rank test. **(C)** Normalized decoding per trial within a block of identical triplets: early time window. Error bars denote SEM across recording sessions. **(D)** Normalized decoding per trial within a block of identical triplets: late time window. **(E)** A trial-by-trial regressor of learning across blocks (slower time scale) was quantified as the block number in a recording session, binned into six bins. **(F, G, H)** Learning effects across blocks, figure legend as in (B, C, D).

3.2.5 Multivariate analysis: predictive and mnemonic representations rely on uncorrelated data features

While the decoding boost observed for the N-3 vowel in predictable blocks, and its gradual improvement over time, bear the hallmarks of a predictive representation, we have also accessed mnemonic representations by decoding previous vowels (N-1 and N-2) in random blocks. To test whether the decoding of predictive and mnemonic representations rely on the same data features, we performed three further analyses. First, we repeated decoding using a searchlight, where each decoding estimate was based on a subset of channels. While no significant N-3 decoding was found in random blocks based on all channels and correcting for multiple comparisons across time points, a searchlight could in principle uncover channels more sensitive to N-3 vowel identity. We then correlated the spatial maps of decoding estimates between the predictable and random blocks. We reasoned that if predictive and mnemonic representations rely on similar data features, the N-3 maps should be correlated across blocks. This analysis revealed significant correlations between spatial decoding maps in predictable and random blocks only for the N-1 vowel (Figure 3.4A; 33-70 ms; $t_{max} = 5.72$; cluster-level $p_{FWE} < 0.001$), but not for the earlier vowels (N-2, N-3: all $p_{FWE} > 0.05$). Specifically, while for the N-1 vowel the spatial maps of decoding obtained in predictable and random blocks were similar (t-test Bayes Factor: 865.33, indicating extremely strong evidence for a correlation) and showed the strongest contribution of the anterior/inferior channels, for the N-3 vowel they were more orthogonal (t-test Bayes Factor: 0.2569, indicating moderate evidence against correlation; cf. N-2: Bayes Factor 0.3558), showing an inferior-superior gradient in predictable blocks and an anterior-posterior gradient in

random blocks (Figure 3.4BC). This contrasted with correlations between decoding maps obtained for odd vs. even trials, which were significant for each vowel position (N-1: r_{max} = 0.27; N-2: r_{max} = 0.13; N-3: r_{max} = 0.11; all significant at $p_{FWE} < 0.05$ correcting across time points). While the latter correlation coefficients were moderate to low, likely due to a decreased signal-to-noise ratio as a result of splitting the dataset in half, this finding suggests that N-3 decoding maps are relatively stable across trials (odd vs. even) but uncorrelated across conditions (predictable vs. random).

Second, we repeated the decoding of vowels in each position, this time training on trials drawn from one type of blocks (e.g., random) and testing on trials from the other type of blocks (e.g., predictable). This analysis (Figure S3.4A) revealed that only N-1 decoding generalized across block types (train on random, test on predictable: T_{max} = 12.92, $p_{FWE} < 0.001$; train on predictable, test on random: T_{max} = 13.39, $p_{FWE} < 0.001$), with no differences observed between blocks (all paired t-test $p_{FWE} > 0.05$). Conversely, for N-2 and N-3 decoding, no significant cross-block decoding was observed in either direction (all $p_{FWE} > 0.05$).

Third, we performed a cross-temporal generalization analysis (Figure 3.4DE), training on one vowel position (e.g., N-1) and testing on another (e.g., N-3). This analysis revealed that while decoding generalizes across time points within each vowel position (e.g., training on neural activity 100 ms and testing on 150 ms after vowel onset; cf. Cappotto et al., 2021), it does not generalize across vowel positions (e.g., training on N-1 and testing on N-3), except for a temporally limited interference effect between N-1 and N-2 vowels (Table S4).

These results suggest that the decoding boost observed for N-3 vowels in predictable blocks (reflecting a predictive representation) relies on data features that are specific to these blocks, and are not generalizable to the random blocks or to other vowels.

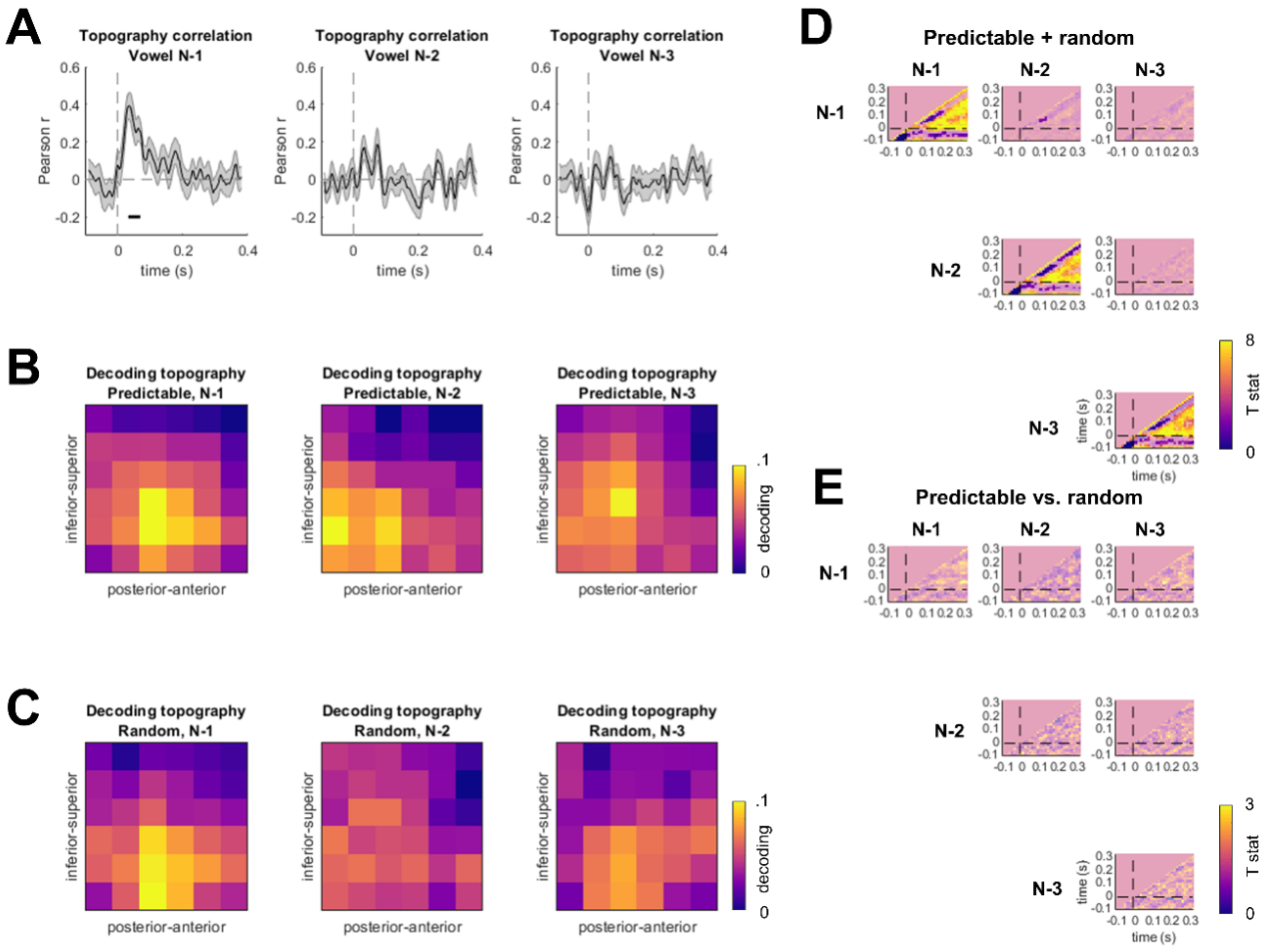

Figure 3.4 Spatial topography of predictive and mnemonic representations. **(A)** Time courses of correlation coefficients between decoding topographies in predictable vs. random blocks. Left / middle / right panel: decoding N-1 / N-2 / N-3 vowel (shaded area: SEM across recording sessions; black horizontal line: correlation coefficients significantly different from zero, $p_{FWE} < 0.05$). **(B)** Decoding topographies based on the 0-100 ms decoding time window, predictable blocks. Left / middle / right panel: decoding N-1 / N-2 / N-3 vowel. **(C)** Decoding topographies based on the 0-100 ms decoding time window, random blocks. Figure legend as in (B). **(D)** Cross-temporal generalization averaged across conditions (predictable + random). Rows: test data; columns: remaining data used for estimating decoding matrices. Each panel shows a cross-temporal decoding matrix with each time point representing decoding based on the Mahalanobis

distance between a particular vowel position (N-1, N-2, N-3) and latency of neural activity and another vowel position and latency of neural activity. Unmasked areas represent significant cross-temporal decoding generalization at $p_{FWE} < 0.05$, cluster-level corrected. Only one (symmetric) side of the diagonal is plotted. **(E)** Cross-temporal generalization: differences between conditions (predictable vs. random). Figure legend as in (D). See also Figure S3.4A and Table S3.4.

3.3 Discussion

In the present study, we demonstrated that stimulus history (sensory memory traces of token values up to N-3) can be decoded from neural responses to broadband noise bursts in both repeated triplet and randomized blocks, expanding on previous research (Cappotto et al., 2021; Wolff et al., 2019, 2015). Crucially, we also provide evidence for the decoding of predictive mechanisms by linking increased N-3 decodability to predictable blocks, further established through the presence of learning effects as the number of triplet pattern repeats increases. This demonstrates that neural responses to noise bursts tap into predictive mechanisms, establishing a novel method for decoding both phenomena simultaneously and independent of attentional tasks. Our results suggest that mnemonic and predictive decoding rely on largely uncorrelated data features - specifically, decoding N-3 stimuli in predictable blocks cannot be generalized to decoding other stimuli in the same blocks, or to the data features present in random blocks.

Previous work has established the use of broadband noise impulses in decoding sensory memory tokens (Wolff et al., 2019, 2015) mediated by mechanisms that function under anesthesia in animal models (Cappotto et al., 2021). Here, we expand on these findings by decoding further stimulus history, showing that it is possible to decode

memory representations of both sequences and individual tokens up to N-3. We also expand on another recent study (Luo et al., 2021) showing that sequence contents can be preferentially decoded from auditory cortical activity in rat models, but that this decoding benefit is only observed for rats with prior training. Similarly, previous work in the visual system of awake mice found that prior training elicits predictive representations that can be decoded (Gavornik and Bear, 2014). Unlike these studies, which used several interleaved sequences in a continuous stream during prior exposure blocks, we used a protocol in which a sequence (triplet) was repeated and then replaced, without prior training. This suggests that for such repetitive sequences, decoding can be achieved in naive and anesthetized rats. In contrast to the previous study (Gavornik and Bear, 2014), our results did not reveal any significant decoding on the omission responses. One possible explanation is that, in anesthetized brains which had not undergone prior training, predictive representations require a stronger activation (e.g., broadband noise bursts) to become observable than would be the case for awake brains of trained participants. In both the present and previous studies (Cappotto, et al. 2021), we have demonstrated that univariate analysis was not sufficient to decode memory tokens and multivariate methods provided significant decoding.

The present literature on animal models of predictive processing is largely within the context of stimulus-specific adaptation (SSA), making it difficult to separate predictive from adaptive mechanisms. Our findings in the AC are not likely to be explained by a simple SSA explanation, given that we observed the decodability of randomly substituted tokens within repeated sequences as well as within non-repeating triplets. If adaptation were responsible for decodability, this effect would be unlikely to increase with overall

triplet repetition, as pattern sensitivity and resulting deviance detection has been shown to rely on hierarchical and contextual error detection (Casado-Román et al., 2020). Our results, suggesting that decoding N-3 tone identity and triplet identity may occur at different latencies (Figure S3.3), are also consistent with the latter hypothesis, as they suggest that predictive processing of single elements might be more short-lived than the encoding of entire sequences.

Importantly, by contrasting responses to noise bursts in predictable vs. random sequences, we tapped into both predictive and mnemonic representations. This goes beyond recent findings in humans showing that predictive neural activity can be explained by memory of past stimuli, but which could not access mnemonic representations independently of predictive processing (Baumgarten et al., 2021). Interestingly, a recent study on auditory associative learning in awake mice showed that neural activity evoked by a predicted stimulus contains information both about its most likely predictor and its actual past, but that this information relies on orthogonal neural codes, suggesting that mnemonic and predictive representations coexist within sensory cortices (Libby and Buschman, 2021). Although our paradigm did not test for this explicitly, our observation of uncorrelated data features enabling decoding in predictable vs. random blocks, and a lack of decoding generalization across blocks and across vowels (N-1 vs. N-3), suggests that such mechanisms are not dependent on active processes, and they can also be observed indirectly over broad neural populations. It is important to note that the spatial resolution of ECoG makes it difficult to identify discrete neural populations due to changes in spatio-temporal representation, and finer recording techniques with single cell resolution would be required to accurately discern if mnemonic and predictive

representations decoded in our paradigm rely on unique neural populations or are multiplexed within the same population.

Importantly, we also establish that the decodability of predictable N-3 tokens gradually increases with repeated triplet presentations (relative to random blocks), implicating passive learning effects as a measure of predictive mechanisms. Recent studies have successfully paired concepts of statistical learning and predictive coding by investigating neural correlates of melodic expectation to naturalistic music, observing that neural responses to less statistically-likely notes elicit markers consistent with their level of statistical (Di Liberto et al., 2020). Human fMRI studies in the visual domain have further established the role of temporal regularity in sequence learning and their resultant effects on the decodability of predictable (Luft et al., 2015). However, studies employing animal models and different attention states to investigate predictive mechanisms have been lacking (Heilbron and Chait, 2018). Although further investigations would be required to clearly verify the role of learning effects at multiple time scales, our results provide an indication of prediction formation at a relatively fast time scale (prediction updating following a presentation of a new triplet).

While it is intrinsically interesting that anesthesia did not abolish the emergence of predictive representations in our study, one must acknowledge that this raises questions about the extent to which our results are representative of neural functions in a normal, awake state. Different types of anesthetic agents (e.g., ketamine, equithesin, pentobarbital) have been shown to affect various features of neuronal activity, such as spontaneous rate, response threshold, or oscillations, in the auditory cortex to a greater or lesser extent (Cheung et al., 2001; Gaese and Ostwald, 2001; Zurita et al., 1994).

However, experiments under anesthesia are still considered an efficient and useful tool in identifying neural mechanisms when carefully controlled. We selected urethane as our main agent for controlling the anesthesia, as it has been widely used for memory-related studies as an agent with minimal effect on spectral tuning, neural discriminability, and information processing (e.g. Astikainen et al., 2011; Capsius and Leppelsack, 1996; Ruusuvirta et al., 1998; Schumacher et al., 2011). Importantly, a comparable hierarchical gradient across subcortical and cortical regions observed for prediction error signaling between urethane-anesthetized and awake rodents (Parras et al., 2017) supports the notion of preserved predictive processing even under anesthesia.

In summary, the present study observed concurrent mnemonic and predictable representations under anesthesia, indicating mechanisms at work in passive preparations and thus providing a new model for investigating simultaneous memory and predictive mechanisms independent of attentional state.

3.4 STAR Methods

3.4.1 Key Resources

Resource	Source	Identifier
Experimental models: Organisms/strains		
Adult female Wistar rats	Chinese University of Hong Kong	RGD_13525002
Software and algorithms		
MATLAB	Mathworks	SCR_001622
Python	Python	SCR_008394
SPM12	University College London	SCR_007037
Deposited data		
Code and Processed Data	Zenodo	https://doi.org/10.5281/zenodo.6407267

3.4.2 Experimental Model and Subject Details

Subjects

Eight young adult female Wistar rats, acquired from the Chinese University of Hong Kong, were used in the experiment. The rats were “naive”, i.e. had no experience or

training with the stimulus sets prior to recording, were aged between 8 and 13 weeks (median age = 10.5 weeks), and weighed between 216 and 289 g (median weight = 238 g). Normal hearing was ascertained by measuring auditory brainstem response at thresholds < 20 dB sound pressure level (SPL) to broadband click trains.

Anesthesia and Surgical Procedures

Anesthesia was induced with an intraperitoneal (i.p.) injection of ketamine (80 mg/kg) and xylazine (12 mg/kg), and maintained throughout the experiment via 20% urethane injections. A first dose of 0.25 ml/kg of the urethane solution was administered one hour after the induction with ketamine and xylazine, and further 0.25 ml/kg doses were delivered as required, based on periodic assessments of anesthesia depth via the toe pinch withdrawal reflex. Dexamethasone (0.2 mg/kg, i.p.) was delivered before surgery as an anti-inflammatory. This protocol, based on previous rodent studies (Cappotto et al., 2021; Malmierca et al., 2019), allowed for fast induction of anesthesia via the initial administration of ketamine and xylazine, while avoiding later NMDA-specific inhibitory effects of ketamine through the use of urethane to maintain anesthesia for ECoG recordings. The anesthetized animal was placed in a stereotaxic frame, and the animal's head was fixed with hollow ear bars to allow sound delivery. An isothermal heating pad and a rectal thermometer were used to maintain body temperature at 36 ± 1°C throughout the experiment. The skin and muscle tissue over the temporal lobe of the skull were removed, and a craniotomy was performed to expose a 5×4 mm region over the right AC, leaving the dura intact. The anterior edge of the craniotomy was 2.5 mm

posterior from Bregma, and the dorsal edge was 2 mm ventral from Bregma (Figure S3.1A).

Experimental Apparatus

The ECoG array was placed on the exposed cortex and a cotton roll was placed between the remaining skin and the array to hold the array securely in place and ensure a stable, low impedance contact between the recording sites and the dura. A hole was drilled through the skull anterior to the Bregma on the animal's left to place a small stainless steel screw which served as ground and reference electrode for the electrode array and headstage amplifier. Correct placement of the ECoG array was verified by recording a set of Frequency Response Areas (FRAs; Figure S3.1B) from each site by collecting responses to 100 ms pure tones varying in sound level (30 - 80 dB SPL) and frequency (500 - 32,000 Hz, ¼ octave steps). Each tone was presented 10 times, in a randomly interleaved fashion, with an onset-to-onset ISI of 500 ms.

3.5 Methods Details

3.5.1 Stimulus Design

The artificial vowels were generated using custom Python scripts. Consecutive vowels were separated by 350 ms of silence (500 ms onset to onset ISI). We deemed artificial vowels preferable to tones as they activate larger parts of the tonotopic array and they resemble many types of natural sounds, including many vertebrate vocalizations or insect sounds, making them arguably more ecologically valid than pure tones. These

generated pulse trains which were subsequently passed through a cascade of two 2nd-order Butterworth bandpass filters with a bandwidth equal to 20% of the center (formant) frequency (scipy.signal functions butter() and lfilter()). The formant frequencies for these artificial vowels were chosen to lie between 900 and 9000 Hz to bring them well into the auditory range of rats, and the fundamental frequencies (F0s) of the vowels were relatively low, between 260 and 420 Hz, to generate a large number of closely stacked harmonics under each formant. Stimulus sequences consisted of combinations of three possible artificial vowels, one we refer to as “A” with formants and 3000 and 5400 Hz and an F0 of 420 Hz, an “O” with formants 900 and 2700 Hz and F0 260 Hz, and an “I” with formants 1050 and 9000 Hz and F0 300 Hz. On occasion, as described further below, one of the vowels in the sequence could be replaced by either a 150 ms frozen pink noise burst computed according to the algorithm described in https://github.com/python-acoustics/python-acoustics/blob/master/acoustics/generator.py, or by a silent pause. The artificial vowel and pink noise tokens were loaded onto a Tucker Davis Technologies (TDT) RZ6 digital sound processor which was programmed using custom written software to present the tokens in a predefined order at a sample rate of 48,828 Hz through headphone drivers connected to the hollow ear bars via 3D printed adapters.

3.5.2 Experimental Paradigm

Two types of blocks were employed. In “predictable” blocks, vowels were grouped into triplets, which repeated at least 25 times (range 25-100, mean 30) before being replaced with another triplet (e.g., AOOAOOAOO…AAIAAIAAI…). In “random” blocks, vowels were presented in a random order, while keeping the base frequency of each

vowel constant and comparable to the predictable block (e.g., AOIOIAIOAOOAIIA…). Each session contained ~72 such blocks (amounting to a total of 2100 triplets per session), presented in a different order per session. Triplets were selected to prevent redundant combinations from occurring during presentation (e.g., AOO, OAO, and OOA would result in identical sequences with different starting points, and thus only AOO was used). The triplets were then concatenated to form the long stimulus sequences presented in the experimental sessions. In these sequences, 5% of stimulus events were replaced with omissions, and 5% were similarly replaced with a burst of pink noise. The vowels that were replaced with noise bursts or silent pauses were chosen pseudo-randomly, subject to the constraint that a minimum of three repetitions of a given triplet had to have occurred before a vowel could be replaced. In a control condition ("random" sessions), vowels were presented randomly, rather than in predefined triplets. The positions of omissions and noise bursts within the stimulus sequences were kept the same across the predictable and random blocks.

3.5.3 Neural data acquisition and pre-processing

An 8 x 8 Viventi ECoG electrode array with 400 μm electrode spacing (Woods et al., 2018) was used to acquire ECoG recordings, employing three ground channels located in the corners of the array, and a common reference. A (TDT) PZ5 neurodigitizer was used to record signals from the array via a RZ2 processor. FRA responses were recorded with BrainWare software at a sampling rate of 24,414 Hz, and responses to the vowel sequences were recorded using custom Python code at a sampling rate of 6104 Hz. The recorded electrode signals were first low-pass filtered at a cutoff frequency of 90

Hz using a 5th order Butterworth filter, and downsampled to 300 Hz to extract neural activity evoked by acoustic stimuli. The pre-processed signals were re-referenced to the average of all channels (Ball et al., 2009), and segmented by extracting 500 ms long voltage traces from −100 ms to +400 ms relative to the onset of each token. Epoched traces were baseline-corrected by subtraction of the mean pre-stimulus voltage values, and linearly detrended (Salisbury, 2012).

3.6 Quantification and statistical analysis

3.6.1 Univariate analysis: summarizing vowel-evoked, omission-evoked, and frozen noise burst-evoked activity

Univariate analysis was performed to assess whether vowel types (A, I, O) modulated vowel-evoked, burst-evoked, and omission-evoked activity on a channel-by-channel basis (Figure S3.2). Additionally, in the analysis of burst-evoked and omission-evoked activity, we tested whether it is modulated by the preceding sounds at different “positions” relative to the burst/omission (N-1 position: the immediately preceding vowel, N-2 position: two stimuli before the burst/omission, N-3 position: three stimuli before the burst/omission). Epoched data were separated per vowel, position, and condition, and then averaged across trials. First, to visualize the evoked responses, trial-averaged ECoG responses were concatenated across sound types/positions/conditions/animals, resulting in 2 two-dimensional matrices per condition with single channels along one dimension and concatenated time points along the second dimension. A principal component analysis using singular value decomposition was performed on the resulting matrices. The output provided spatial principal components describing channel topographies, and

temporal principal components describing voltage time-series concatenated across vowels/positions and animals, sorted by the ratio of explained variance. A weighted average was calculated to summarize the top principal components explaining 95% of the original variance, weighted by the proportion of variance explained. These resulting voltage time-series were averaged per vowel across animals. Frozen noise burst-evoked and omission-evoked single-trial data were similarly averaged across trials, separately for each preceding vowel and position, and subject to the same principal component analysis described above.

The above principal component analysis was used only for the purposes of visualizing the data. In order to test if any time points and channels showed significant amplitude modulations by vowel (in case of vowel-evoked responses) or preceding vowel in each position (in case of burst-evoked and omission-evoked responses), single-subject trial-average ECoG data in the original electrode grid were converted into three-dimensional matrices containing two spatial dimensions and one temporal dimension. These matrices were then converted to 3D images and entered into a repeated-measures ANOVA with one within-subjects factor (vowel; three levels) and one repeated-measures factor (rat), implemented in SPM12 (University College London) as a general linear model (GLM). This was done separately for each stimulus type (vowel-evoked responses, burst-evoked responses, and omission-evoked responses). The effects of preceding vowels on burst-evoked and omission-evoked responses were analyzed in separate ANOVAs per position. To test for the effect of vowel on evoked activity amplitude, an omnibus F test across 3 vowels was used. The resulting statistical parametric maps were thresholded at $p < 0.005$ (two-tailed) and corrected for multiple comparisons across spatiotemporal

voxels at a family-wise error (FWE)-corrected p_{FWE} = 0.05 (cluster-level) (Kilner et al., 2005).

3.6.2 Univariate analysis: oscillatory activity

To test whether sequence processing is associated with spectral peaks in the neural response spectrum at the syllable and triplet rate (Henin et al., 2021), we analyzed phase coherence of neural activity (Figure S3.4B). Specifically, for each rat and recording session, we split the continuous single-channel ECoG data into 175 chunks of 12 triplets, and, for each chunk, calculated the Fourier spectrum of neural activity measured during that chunk. Inter-trial phase coherence (ITPC) was calculated according to the following equation (Ding and Simon, 2013):

$$ITPC_f = \left(\left[\Sigma^N cos\phi_f\right]^2 + \left[\Sigma^N sin\phi_f\right]^2\right)/N,$$

where φ_f denotes the Fourier phase at a given frequency f and N = 175 chunks. In the initial univariate analysis, phase coherence estimates were averaged across channels. To test for the presence of statistically significant phase coherence peaks, coherence values at the token rate (2 Hz) and triplet rate (0.667 Hz) were compared against the mean of coherence values at their respective neighboring frequencies (single token rate: 1.944 and 2.056 Hz; triplet rate: 0.611 and 0.722 Hz) using Wilcoxon's signed rank tests.

3.6.3 Multivariate analysis: decoding sensory, mnemonic, and predicted vowel information

Data were subjected to multivariate analyses to test if information about vowel type could be decoded from the pattern of burst-evoked and omission-evoked activity

observed across multiple channels and time points. To this end, we adapted methods established in previous multivariate decoding research, which has demonstrated decodability in similar data and experimental contexts (Cappotto et al., 2021; Myers et al., 2015; van Ede et al., 2018; Wolff et al., 2019, 2017).

Prior to decoding, single-trial omission or frozen noise burst-evoked responses were sorted by the preceding vowel, separately for each vowel position. While the randomized order of vowel presentation in relation to noise bursts (see Experimental Paradigm and Stimulus Design) effectively equalized the ratio of vowels presented at each position, we imposed an additional constraint on trial selection to ensure that decoding N-3 vowels is not confounded by the vowels presented immediately before the noise burst (N-1). Specifically, in decoding N-3 stimuli relative to noise burst X, we excluded trials for which N-3 and N-1 were identical (e.g., AAOAAX was included, since vowel N-1 corresponds to A and N-3 to O; however, AAOAXO was excluded, since both vowels N-1 and N-3 correspond to A). To equalize the number of trials across decoding conditions, the same constraint was imposed on N-2 stimuli (excluding trials for which N-2 and N-1 were identical) and on random blocks.

Decoding time-courses were estimated using a sliding window approach (Cappotto et al., 2021; Wolff et al., 2019), pooling information over multiple time-points and channels to boost decoding accuracy (Grootswagers et al., 2017; Nemrodov et al., 2018). Specifically, for each channel, trial, and time point, we first pooled voltage values within a 50 ms window relative to a given time point. Then, a vector of 5 average voltage values was calculated per channel and trial by downsampling the voltage values over 10 ms bins. In other words, a single vector of multivariate data corresponding to the test trial

(multiple channels x 5 time points within a 50 ms window, concatenated into a long vector) is compared against three vectors (one per vowel), each of exactly the same length as for the test trial but based on the remaining trials. The data were then de-meaned to remove the channel-specific average voltage over the entire 50 ms time window from each channel and time bin, ensuring that the multivariate analysis approach was optimized for decoding transient activation patterns (Cappotto et al., 2021; Wolff et al., 2019). For the subsequent leave-one-out cross-validation decoding, the vectors of binned single-trial temporal data were then concatenated across channels. We used the Mahalanobis distance (De Maesschalck et al., 2000) as a multivariate decoding metric to take advantage of the potentially monotonic relation between vowel category and neural activity (Auksztulewicz et al., 2019; Cappotto et al., 2021; Wolff et al., 2019). Responses to dissimilar vowels are expected to yield large Mahalanobis distance metrics, while responses to similar vowels are expected to yield low Mahalanobis distance metrics. Having been shown to be optimal for decoding (Grootswagers et al., 2017), a leave-one-out cross-validation approach was used per trial, wherein we calculated 3 pairwise distances between ECoG amplitude fluctuations measured in a given test trial and mean vectors of ECoG amplitude fluctuations averaged for each of the 3 vowels/positions in the remaining trials. A shrinkage-estimator covariance obtained from all trials, excluding the test trial, was used to compute the Mahalanobis distances (Ledoit and Wolf, 2004). Combining Mahalanobis distance with Ledoit–Wolf shrinkage has been shown to have performance advantages over other correlation-based methods of measuring brain-state dissimilarity (Bobadilla-Suarez et al., 2019), while Mahalanobis distance-based decoding

has known advantages over linear classifiers and simple correlation-based metrics (Walther et al., 2016).

Single-trial relative Mahalanobis distance estimates were averaged across trials, resulting in a 3 x 3 distance matrix for each rat, time point, relative vowel position (N-1, N-2, N-3), and substitution type (noise vs. omission). To obtain overall decoding quality traces, the 3 x 3 distance matrices were subject to a subtraction of the averaged off-diagonal elements (mean distance between vowels) from the averaged diagonal elements (mean distance within vowels). The resulting decoding time-series were entered into a 2x3 repeated-measures ANOVA with within-subjects factors Block (predictable vs. random) and Position (N-1, N-2, N-3), separately for the two substitution types (noise vs. omission). The resulting statistical parametric maps were thresholded at $p < 0.005$ (uncorrected). Across time points, p values were corrected using a family-wise error approach at a cluster-level $p_{FWE} = 0.05$ (Kilner et al., 2005).

We reasoned that significant decoding of the N-3 vowel in the predictable blocks, but not in the random blocks, would reveal predictive representations of the expected vowel. However, such representations may be formed both on an element-by-element basis (e.g., when hearing AOOAOOAOO, an “A” may be predicted because one is heard every 3 tokens), and also for an entire triplet (e.g., when hearing AOOAOOAOOX, "X" might also reactivate a representation of the AOO context). In a follow-up analysis, we wanted to test whether bursts/omissions reactivate representations containing (1) information about the entire preceding triplet, or (2) specific information about the N-3 vowel, independent of the rest of the triplet. To this end, we ran an additional decoding analysis, this time using a 18 x 18 stimulus matrix (corresponding to 18 possible triplets,

with 3 phase shifts for each of the 6 unique triplets; e.g., for a unique triplet AAO, the three phase shifts would correspond to AAO, AOA, and OAA), yielding 18 x 18 Mahalanobis distance matrices. This analysis focused on the predictable blocks only and zoomed into two time clusters in which we observed significant N-3 vowel decoding (see Results). To quantify the decoding of the entire triplet, we subtracted the mean of all off-diagonal elements of the 18 x 18 stimulus matrix from the mean of all diagonal elements (Figure S3.3A). To quantify the decoding of information about the N-3 vowel independent of the entire triplet identity, we subtracted the mean of those elements of the 18 x 18 stimulus matrix which did not share the first vowel from the mean of those elements of the matrix which did share the first vowel (excluding the diagonal elements, corresponding to identical triplets). The decoding estimates based on these representational dissimilarity matrices were subject to one-sample t-tests (two-tailed) across recording sessions (see Figure S3.3BC for results).

Since vowel decoding was relatively weaker for N-3 and N-2 vowels (see Results; Figure 3.2A), we have performed an additional analysis aiming at verifying whether spatial maps of decoding sensitivity can be reasonably established for these vowel positions. To this end, we performed an additional analysis in which we repeated the spatial correlation analysis, but rather than correlating predictable and random blocks, we correlated decoding based on odd vs. even trials within each block.

In an additional analysis, since we observed univariate differences in vowel-evoked responses (see Results), we tested whether decoding primarily relies on those channels that are also associated with sensory encoding of vowels. To this end, we repeated the decoding analysis for two subsets of channels - those which strongly

differentiated between vowels (with the corresponding F statistic of the main effect of vowel on the vowel-evoked responses higher than the median across channels) and those which differentiated weakly between vowels (F statistic below median across channels). The resulting decoding time-series were compared between the two groups of channels using a series of paired t-tests, correcting for multiple comparisons across time points at a family-wise error (FWE)-corrected $p_{FWE} = 0.05$ (cluster-level) (Kilner et al., 2005).

While we did not observe univariate differences in spectral peaks at the single vowel rate between condition (and we did not observe peaks at the triplet level overall; see Results), in a further analysis we also tested whether decoding might rely on those channels which show relatively higher triplet-rate peaks than other channels. Again, we repeated the decoding analysis for two subsets of channels, this time splitting them based on the single-channel phase coherence estimates for the single vowel rate (2 Hz; above/below median). The two resulting decoding time-series were compared using a series of paired t-tests, correcting for multiple comparisons as above.

For completeness, we also performed the decoding analysis on the vowel-evoked responses themselves (see Table S3.3 for results). While, given that vowel-evoked responses showed univariate effects of vowel identity, multivariate decoding was expected to be significant, we could use this analysis to compare the magnitude of decoding mnemonic information (N-1) based on burst-evoked responses, relative to decoding of vowel identity (N) based on vowel-evoked responses.

3.6.4 Multivariate analysis: learning effect on decoding

Another question we wanted to address is whether any decoding benefit we might observe in the predictable stimulus condition reflects predictive neural processing. In particular, we hypothesized that, if the decoding boost in predictable blocks is related to predictive processing, it should gradually build up, as the auditory system needs time to detect repeating patterns and learn to use them for predictions of which sound token is expected when. This can occur at two time scales: first, decoding can improve with each subsequent vowel token embedded in a block of identical triplets (reflecting learning within blocks); second, decoding can improve over subsequent blocks (reflecting learning across blocks). To test these hypotheses, we constructed two trial-by-trial learning regressors - a “within blocks” regressor quantifying the vowel position within a block of identical triplets, and an “across blocks” regressor quantifying which block of a particular triplet it is within the entire recording session. To facilitate comparisons between the two regressors, the “within blocks” regressor only included vowel position from 1 (first burst within a sequence) to 6 (sixth burst), while the “across blocks" regressor was binned into 6 bins of 2 blocks in each bin (e.g., bin 1 contained the first 2 blocks of a particular triplet, while bin 6 contained the last 2 blocks of the same triplet). Both regressors were log-transformed to increase the relative effect of the first bursts/sessions relative to the last bursts/sessions (HiJee et al., 2021). We then repeated the decoding analysis of the N-3 vowel and, per recording session and condition, performed a multiple linear regression with a constant term and the two learning regressors on single-trial decoding estimates. Specifically, for both of the time clusters in which we identified significant differences between decoding in predictable vs. random blocks, we selected the single-trial peak

decoding within a given time cluster, and then normalized (z-scored) the trial-by-trial peaks per rat, recording session, and condition. This resulted in 8 sets of learning coefficients: (1) for predictable vs. random conditions, (2) quantifying learning within vs. across blocks, (3) estimated for early vs. late time window. The resulting regression coefficients (betas) were tested for significant differences between predictable and random blocks (treated as a baseline condition) using Wilcoxon sign rank test. While we hypothesized that learning effects should be specific to N-3 stimuli, in an additional analysis we also tested for the same learning effects on the decoding of N-2 and N-1 stimuli.

3.6.5 Multivariate analysis: similarity between predictive and mnemonic representations

To test whether the predictive and mnemonic representations are shared, we quantified the spatial correlation of decoding topographies between predictable and random blocks. Our reasoning was that, if predictive and mnemonic representations are shared, decoding topographies should be similar between predictable and random blocks. On the other hand, if predictive and mnemonic representations are independent, the decoding topographies should be different between the two types of blocks. To this end, we repeated the decoding analysis, this time using a searchlight approach. Specifically, rather than using all channels for decoding, we used subsets of channels, with each subset forming a 3x3 grid. Different subsets overlapped by 1 row or column, resulting in 36 (6x6) decoding estimates based on the 3x3 grids, separately for each recording session, condition, and time point. We then correlated the spatial maps obtained for predictable and random blocks, separately for each recording session and

time point. The resulting Pearson correlation coefficients were entered into a series of one-sample t-tests, correcting for multiple comparisons across time points at $p_{FWE} = 0.05$ (Kilner et al., 2005).

3.6.6 Multivariate analysis: cross-temporal generalization

In a further analysis, we tested whether decoding a particular vowel generalizes across time points (suggesting that the reinstated representations rely on a similar neural code, independent of the latency of measured neural activity) and/or across vowel positions (suggesting that decoding one triplet element relies on a similar neural code as decoding another triplet element). To this end, we performed a cross-temporal generalization analysis, in which we repeated our multivariate decoding analysis but with an important modification of the leave-one-out cross-validation approach. First, to quantify generalization across time points, in calculating the Mahalanobis distance we incrementally shifted the latency of the test data with respect to the remaining trials, in 16 ms time steps - such that decoding was trained on one latency but tested on another. As a result of this approach, rather than decoding time series, per recording session and condition (predictable vs. random) we obtained decoding matrices with each matrix element representing the Mahalanobis distance between data measured at two different latencies. Second, to quantify generalization across vowel positions, we allowed the test data labels to be replaced by labels corresponding to another vowel than the remaining trials. As a result of this approach, rather than obtaining 3 decoding matrices (one per vowel position), we obtained 6 decoding matrices with the 3 additional matrices representing the Mahalanobis distance between data measured at two different vowel

positions. The resulting decoding matrices were entered into a series of 6 GLMs (one per vowel position pair), each implementing a paired t-test between decoding estimates obtained for the predictable and random conditions. The resulting statistical parametric maps were thresholded at $p < 0.005$ (two-tailed) and corrected for multiple comparisons across spatiotemporal voxels at a family-wise error (FWE)-corrected $p_{FWE} = 0.05$ (cluster-level) (Kilner et al., 2005).

3.7 Supplemental Results

3.7.1 Auditory cortical activity recordings in a sequence learning paradigm

To investigate the decodability of mnemonic and predictive representations from AC activity, we combined ECoG recordings in young adult female Wistar rats (N = 8) with an auditory sequence learning paradigm. All rats were “naive”, i.e. had no experience with the stimulus sequences prior to recording, and were anesthetized before being implanted with ECoG electrode arrays over their auditory cortex (Figure S3.1A, see STAR Methods for details on experimental procedures). Frequency Response Area (FRA) maps for each animal were used to visually verify whether the placement of the array was consistent across subjects (Figure S3.1B).

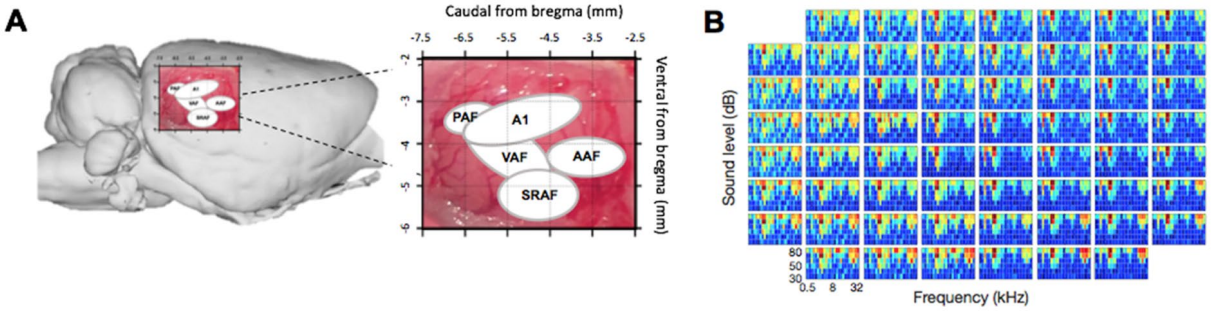

Figure S3.1. Electrocorticography methods and acoustic stimulation. **(A)** Electrode placement during electrocorticography. **(B)** An example Frequency Response Area map from one subject. These were used to confirm electrode array placement consistency over the AC across subjects.

3.7.2 Univariate analyses: only vowel-evoked activity differentiates between vowels

First, to test whether vowel identity influences mean neural activity in the AC at a coarse spatial resolution (i.e., forming smooth clusters of neighboring channels), we have performed a series of univariate analyses, testing for the effects of vowel (A, I, or O) and block (predictable vs. random) on vowel-evoked ECoG activity (event-related potentials). We observed that vowel-evoked activity does differentiate between the three vowels, both in predictable blocks (Figure S3.2A; 13-260 ms; $F_{max} = 58.56$; $p_{FWE} < 0.001$) and in random blocks (Figure S3.2B; 13-207 ms; $F_{max} = 58.21$; $p_{FWE} < 0.001$). The main effect of block on vowel-evoked activity was not significant (all $p_{FWE} > 0.05$).

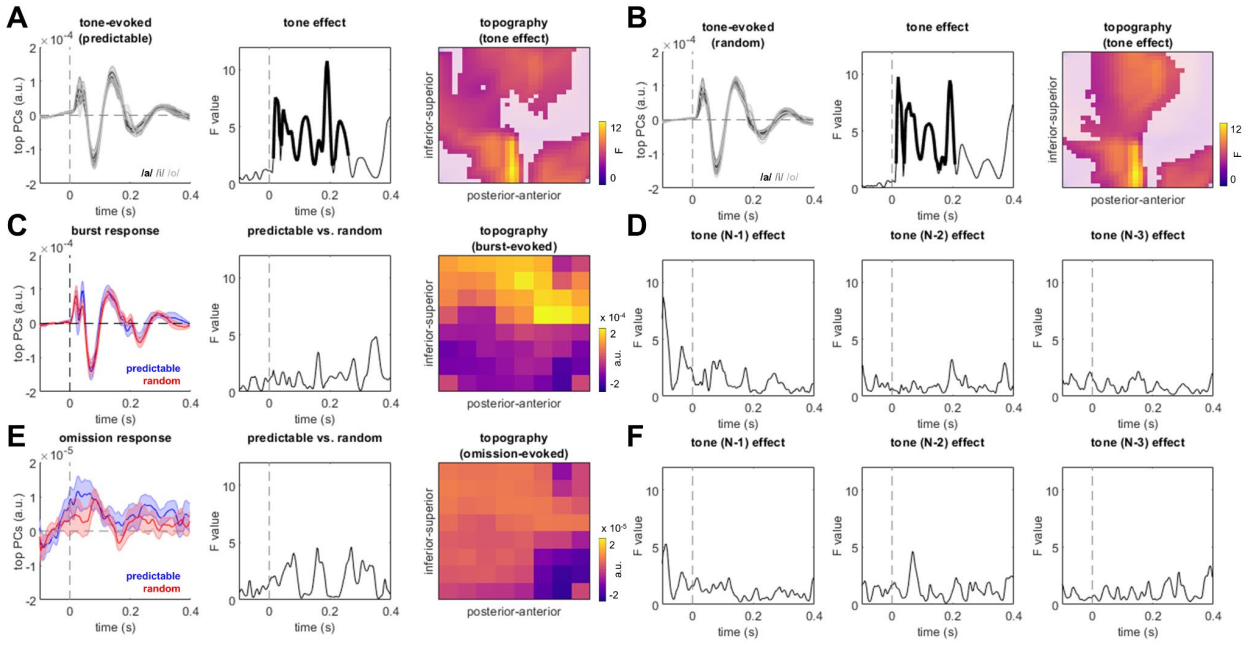

Figure S3.2. Univariate analyses. **(A)** vowel-evoked responses in predictable blocks. Left panel: Time courses of vowel-evoked responses, summarizing the top principal components explaining 95% of the

variance (shaded area: SEM across recording sessions). Middle panel: Time course of the main effect of vowel (bold: $p_{FWE} < 0.05$). Right panel: Topography of the main effect of vowel (unmasked area: $p_{FWE} < 0.05$). **(B)** vowel-evoked responses in random blocks. Figure legend as in (A). **(C)** Burst-evoked responses. Left panel: Time courses of noise burst-evoked responses, summarizing the top principal components explaining 95% of the variance (blue: predictable blocks; red: random blocks; shaded area: SEM across recording sessions). Middle panel: Time course of the main effect of predictability. No significant differences were observed between predictable and random blocks. Right panel: Topography of the noise burst-evoked responses, averaged across blocks and recording sessions, summarizing the top principal components explaining 95% of the variance. **(D)** Effects of preceding vowel on noise burst-evoked responses. Left/middle/right panel: Time courses of the main effect of the preceding vowels (N-1 / N-2 / N-3, relative to noise burst). No significant differences were observed. **(E)** Omission-evoked responses. Figure legend as in (C). No significant differences were observed between predictable and random blocks. **(F)** Effects of preceding vowel on omission-evoked responses. Figure legend as in (D). No significant differences were observed.

Having established that vowel-evoked activity differentiates between the three vowels, but not between experimental conditions (predictable vs. random), we then tested whether burst-evoked and/or omission-evoked activity also differentiates between the (preceding) vowels at different “positions” in the sequence, relative to the burst/omission (N-1 position: the immediately preceding vowel, N-2 position: two stimuli before the burst/omission, N-3 position: three stimuli before the burst/omission). This analysis revealed that, similarly to the vowel-evoked responses, burst-evoked responses did not significantly differentiate between predictable and random blocks (Figure S3.2C; all $p_{FWE} > 0.05$). However, unlike the vowel-evoked activity (which was modulated by vowel identity), noise burst-evoked activity was not significantly modulated by (preceding) vowel identity when neural activity was analyzed in a mass-univariate manner. Specifically, neither the effect of the immediately preceding vowel on burst responses (N-1: all $p_{FWE} > 0.05$; Figure S3.2D), nor of the previous vowels (N-2, N-3: all $p_{FWE} > 0.05$) were significant.

Omission-evoked responses peaked relatively early (83-93 ms), with a rising

activity visible already prior to expected stimulus onset, and thus possibly marking the offset response to the preceding interrupted stimulus train rather than a true omission (Chien et al., 2019). Nevertheless, just like burst-evoked activity, omission-related activity was also not significantly modulated by block type (Figure S3.1E; all $p_{FWE} > 0.05$) or preceding vowel identity (Figure S3.2F; N-1, N-2, N-3: all $p_{FWE} > 0.05$).

3.7.3 Multivariate analysis: specific decoding boost for predictable vowels

Although noise burst-evoked activity did not differentiate between preceding vowels when analyzed in a mass-univariate way, based on our previous study (Cappotto et al., 2021) we hypothesized that preceding stimuli can nevertheless be decoded in a multivariate analysis. Specifically, by analyzing the spatiotemporal pattern of activity evoked by noise bursts, which did not carry any overt information about the preceding vowels given that the employed noise tokens were always identical and presented after vowel-evoked responses had returned to baseline (400 ms after stimulus offset), we sought to determine if activity evoked by noise bursts contained information about the preceding vowels (separately for N-1, N-2, and N-3 vowels). This analysis revealed significant decoding of vowels up to N-3 (Figure 3.2A) in predictable blocks (N-1: -10-273 ms, $t_{max} = 13.36$, cluster-level $p_{FWE} < 0.001$; N-2, early cluster: 40-106 ms, $t_{max} = 3.83$, cluster-level $p_{FWE} < 0.001$; N-2, late cluster: 166-203 ms, $t_{max} = 3.23$, cluster-level $p_{FWE} = 0.001$; N-3, early cluster: 33-123 ms, $t_{max} = 5.44$, cluster-level $p_{FWE} < 0.001$; N-3, late cluster: 223-240 ms, $t_{max} = 2.98$, cluster-level $p_{FWE} = 0.044$). Decoding was significantly better for the immediately preceding (N-1) vowels than for the earlier vowels (N-1 vs. N-2: -6-236 ms, $t_{max} = 7.36$, cluster-level $p_{FWE} < 0.001$; N-1 vs. N-3: -3-216 ms, $t_{max} = 7.84$, cluster-level $p_{FWE} < 0.001$). Clusters of significant decoding extended into the baseline

are likely due to the sliding time window approach (50 ms) adopted in the multivariate analyses (see Methods).

In random blocks as well, significant decoding of preceding vowels was possible, but only the N-1 and N-2 vowels could be decoded (N-1, early cluster: -10-140 ms, t_{max} = 7.73, cluster-level p_{FWE} < 0.001; N-1, late cluster: 163-203 ms, t_{max} = 3.26, cluster-level p_{FWE} = 0.001; N-2: -13-7 ms, t_{max} = 3.03, cluster-level p_{FWE} = 0.029). As was the case for predictable blocks, in the random blocks decoding was also significantly better for the immediately preceding (N-1) vowels than for the earlier vowels (N-1 vs. N-2: 0-136 ms, t_{max} = 5.91, cluster-level p_{FWE} < 0.001; N-1 vs. N-3: -6-126 ms, t_{max} = 5.53, cluster-level p_{FWE} < 0.001).

For completeness, we also performed the decoding analysis on the vowel-evoked responses themselves. While, given that vowel-evoked responses showed univariate effects of vowel identity, multivariate decoding was expected to be significant, we could use this analysis to compare the magnitude of decoding mnemonic information (N-1) based on burst-evoked responses, relative to decoding of vowel identity (N) based on vowel-evoked responses. This analysis revealed significant decoding of vowel identity based on vowel-evoked responses during the entire post-stimulus time window (predictable: t_{max} = 12.85, cluster-level p_{FWE} < 0.001; random: t_{max} = 9.74, cluster-level p_{FWE} < 0.001) with no significant differences between blocks (all p_{FWE} > 0.05). Overall, current vowel decoding was almost an order of magnitude higher than the decoding of previous vowel identity based on burst-evoked responses (N vs. N-1: t_{max} = 11.75, cluster-level p_{FWE} < 0.001; mean decoding within 0-100 ms post-stimulus, mean ± SEM a.u.: N decoding 0.65 ± 0.07, N-1 decoding 0.07 ± 0.01). Taken together, using multivariate analyses, we could access mnemonic representations reactivated by a noise burst, up to 3 stimuli in the past in the predictable blocks and up to 2 stimuli in the past in the random

blocks. Overall, immediately preceding stimuli could be decoded better than previous stimuli, but not as well as currently processed stimuli.

Crucially, if burst-evoked activity can reactivate not only mnemonic representations (irrespective of the currently processed stimulus), but also predictive representations (regarding the stimulus which would have been predicted but is replaced by a noise burst), we would expect a specific decoding improvement for N-3 (but not N-2 or N-1) vowels presented in predictable blocks vs. random blocks. The decoding results were consistent with this hypothesis. Specifically, a significant decoding boost was only observed for the N-3 vowels presented in predictable blocks (paired t-test, predictable vs. random: early cluster: 77-103 ms, $t_{max} = 3.45$, cluster-level $p_{FWE} = 0.010$; late cluster: 227-270 ms, $t_{max} = 3.79$, cluster-level $p_{FWE} < 0.001$), suggesting that we could access a predictive representation of the vowel replaced by a noise burst.

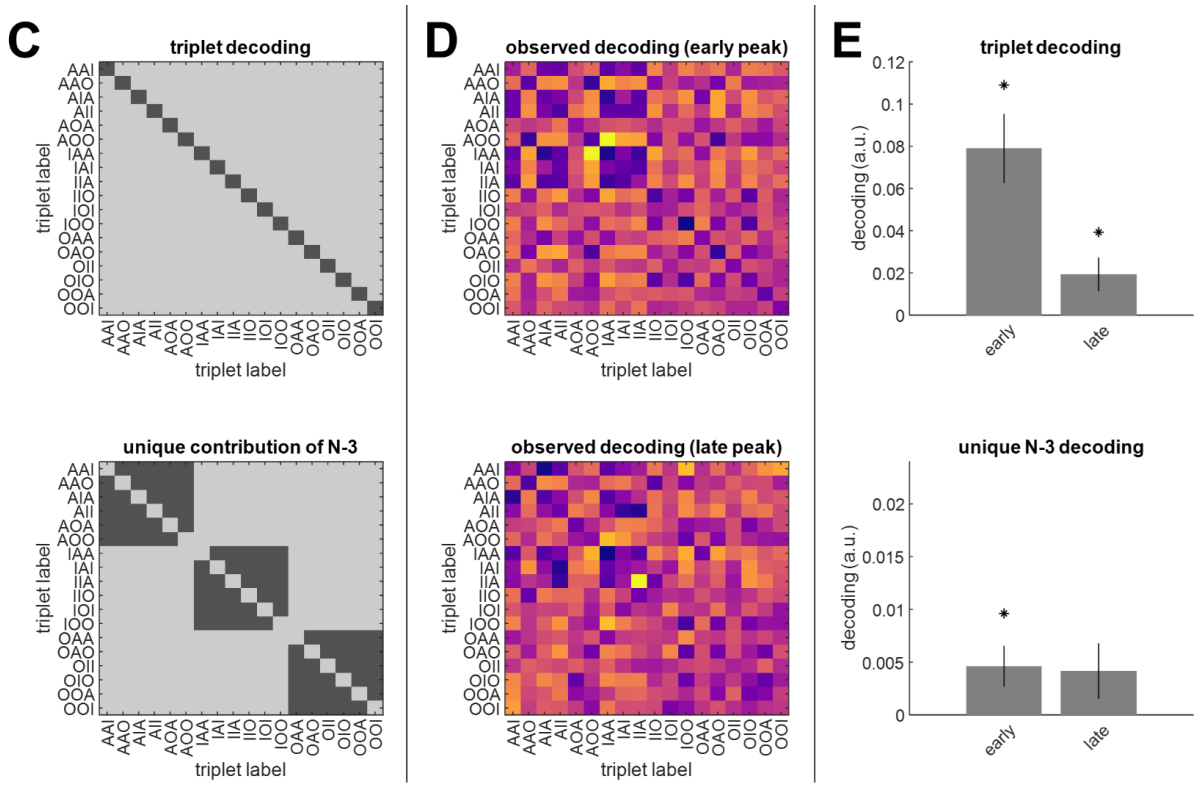

Figure S3.3. Multivariate analyses. . **(C)** Stimulus representational dissimilarity matrices used to quantify the decoding of entire triplets (upper panel) and the unique contribution of decoding the first element of each triplet (i.e., N-3 vowels) while excluding entire triplets along the diagonal (lower panel). Darker gray shows stimulus similarity, lighter gray shows stimulus dissimilarity. **(D)** Observed decoding matrices for the early peak (left panel) vs. late peak (right panel) observed in the N-3 decoding trace (A, right panel). "Warmer" colors denote larger Mahalanobis distance (a.u.). **(E)** Triplet decoding (left panel) and the unique contribution of decoding the first element of each triplet (N-3 vowel; right panel) for the early and late peaks (separate bars). Error bars denote SEM across recording sessions. Asterisks denote significance ($p < 0.05$).

While mnemonic and predictive representations could be decoded based on burst-evoked activity, decoding stimulus history based on omission-evoked activity did not yield any significant results (Figure S3.3B; all $p_{FWE} > 0.05$). This suggests that, at least in this experimental protocol (vowel triplets) and under anesthesia, a stronger activation of the network (e.g., burst presentation) is necessary to make mnemonic and/or predictive representations observable in this type of extracellular recordings.

The finding that burst-evoked responses in predictable blocks could be used to decode mnemonic representations about the past 3 triplet elements, as well as predictive representations about the N-3 vowel, entails a possibility that what is actually being represented is the entire triplet, rather than unique information about individual elements. In other words, predictions may be formed both on an element-by-element basis (e.g., when hearing AOOAOOAOO, an “A” may be predicted because one is heard every 3 tokens), and also for an entire triplet (e.g., when hearing AOOAOOAOOX, "X" might also reactivate a representation of the AOO context). To test this hypothesis, we repeated the decoding analysis, this time decoding stimulus information on a triplet-by-triplet level (Figure S3.3A). This analysis revealed that the entire triplet identity can indeed be decoded from burst-evoked responses, both for the early decoding cluster (77-103 ms; $t_{20} = 4.8118$; $p < 0.001$) and the later cluster (227-270 ms; $t_{20} = 2.4075$; $p = 0.0258$; Figure S3.2DE). However, information unique to the N-3 vowel but independent of triplet identity was also present in the early decoding cluster ($t_{20} = 2.3889$, $p = 0.0269$), but not in the later cluster ($t_{20} = 1.5881$; $p = 0.128$; Figure S3.3E). This suggests that a predictive representation of the expected triplet element is decodable shortly following the noise burst that replaces the expected vowel, while a representation of the entire triplet is present at a wider range of latencies following the noise burst.

In a control analysis (Figure S3.4A), we tested if vowel decoding relies on data features that can generalize across block types (predictable vs. random). To this end, we repeated the decoding of N-1, N-2, and N-3 vowels, but training on trials drawn from one type of blocks (e.g., random) and testing on trials drawn from the other type of blocks (e.g., predictable). This cross-block decoding analysis revealed that only N-1 decoding

generalized across blocks, with significant cross-block decoding observed for both types of blocks (train on random, test on predictable: Tmax = 12.92, pFWE < 0.001; train on predictable, test on random: Tmax = 13.39, pFWE < 0.001) and no differences observed between blocks (all paired t-test $p_{FWE} > 0.05$). Conversely, for N-2 and N-3 decoding, no significant cross-block decoding was observed in either direction (all $p_{FWE} > 0.05$).

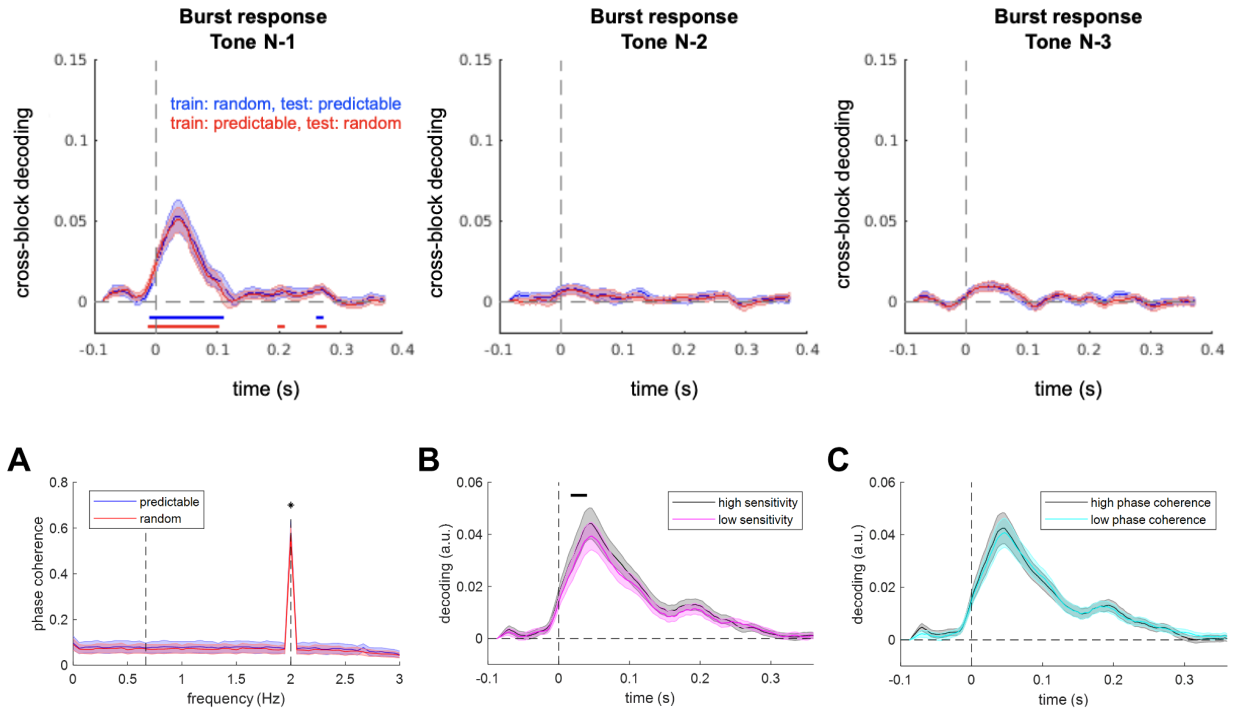

Figure S3.4 (top). (A) Cross-block decoding, training on trials drawn from one type of blocks (blue: random, red: predictable) and tested on trials from the other type of blocks (blue: predictable, red: random). Horizontal bars mark significant decoding (t-test against zero, pFWE < 0.05). Shaded area: SEM across recording sessions. **(bottom)** Spectral analysis and decoding based on channel subsets. **(B)** Phase coherence in predictable and random blocks. Asterisk denotes a significant syllable-rate peak (2 Hz). Shaded areas: SEM across recording sessions. **(C)** Decoding traces (averaged across conditions and vowel positions) based on channels showing high sensitivity to vowels (black line) vs. low sensitivity to vowels (magenta line). Horizontal bar denotes a significant difference between high vs. low sensitivity channels ($p_{FWE} < 0.05$). Shaded areas: SEM across recording sessions. **(D)** Decoding traces (averaged across conditions and vowel positions) based on channels showing high syllable-rate phase coherence (black line) vs. low syllable-rate phase coherence (cyan line). Shaded areas: SEM across recording sessions.

3.7.4 Univariate analyses: spectral peaks of neural activity observed for single vowel rate but not triplet rate

In a frequency-domain analysis of the ECoG responses, we tested whether sequence processing is associated with frequency peaks in the neural response spectrum at the vowel (2 Hz) and triplet (0.66 Hz) rate, as reported in ECoG studies in humans (Henin et al., 2021). To this end, we analyzed phase coherence of neural activity and observed robust spectral peaks at the single vowel rate (Wilcoxon's signed-rank test against neighboring frequency points: $Z = 5.6545$, $p < 0.001$; Figure S3.4B), but not at the triplet rate ($p = 0.4569$), consistent with a recent study in anesthetized rats (Luo et al., 2021). No differences in spectral peaks were observed between predictable and random blocks at either the single vowel rate ($p = 0.7943$) or the triplet rate ($p = 0.6639$).

3.7.5 Multivariate analysis: decoding using channel subsets

In two supplementary analyses, we tested whether decodability primarily relies on those channels which are also associated with sensory encoding of vowels, and/or on those channels which show robust phase coherence at the syllable rate. The first analysis revealed a significant main effect of channel selection (17-40 ms; Fmax = 10.45; $p_{FWE} < 0.001$; Figure S3.4C), suggesting that channels showing stronger differences between vowel-evoked responses also contribute more strongly to decoding vowel memory. No significant interactions between channel selection (N-1; N-2; N-3) and vowel position and/or condition (predictable vs. random) were observed ($p_{FWE} > 0.05$). The second analysis did not reveal any significant effect of channel selection (main and interaction

effects: $p_{FWE} > 0.05$; Figure S3.4D), suggesting that phase coherence at the syllable rate is not related to memory decoding.

3.7.6 Multivariate analysis: predictive and mnemonic representations rely on independent codes

Since vowel decoding was relatively weaker for N-3 and N-2 vowels (Figure 3.2), we have performed a control analysis aiming at verifying whether spatial maps of decoding sensitivity can be reasonably established for these vowel positions. To this end, we performed a control analysis in which we repeated the spatial correlation analysis, but rather than correlating predictable and random blocks, we correlated decoding based on odd vs. even trials within each block. We found that spatial correlations between split-halves were significant for all three vowel positions. While the mean correlation coefficients were overall moderate to low (N-1: $r_{max} = 0.27$; N-2: peak $r_{max} = 0.13$; N-3: peak $r_{max} = 0.11$), likely due to a decreased signal-to-noise ratio as a result of splitting the dataset in half (yielding ~33 trials per vowel), these peaks were significant for all three vowel positions after correcting for multiple comparisons across time points ($p_{FWE} < 0.05$). This finding suggests that, for the earlier vowel positions, spatial maps are relatively stable across trials (odd vs. even) but uncorrelated across conditions (predictable vs. random).

3.7.7 Multivariate analysis: cross-temporal generalization

In a further analysis, we tested whether decoding generalizes across time points within a single response (suggesting a similar neural code for decoding based on activity measured at different latencies) and/or across vowel positions (suggesting that decoding

one vowel, e.g. N-1, relies on a similar neural code as decoding another vowel, e.g. N-2). This analysis revealed that, per vowel position, decoding did generalize across time points (N-1: t_{max} = 20.21, p_{FWE} < 0.001; N-2: t_{max} = 16.05, p_{FWE} < 0.001; N-3: t_{max} = 13.64, p_{FWE} < 0.001; Figure S3.6D; cf. Cappotto et al., 2021). Interestingly, we also observed negative cross-generalization between N-1 and N-2 vowels, when averaging across predictable and random conditions, such that N-2 vowel decoding at 133-167 ms post-stimulus showed impaired (negative) decoding when trained on data corresponding to the N-1 vowel at 67-100 ms (F_{max} = 25.51, t_{min} = -5.05, p_{FWE} = 0.027), possibly reflecting an interference effect between N-1 and N-2 decoding. Beyond this finding, there were no significant cross-generalization clusters between the other vowel pairs (all pFWE > 0.05), as well as no differences between predictable and random blocks either in cross-temporal generalization across time points (all pFWE > 0.05; Figure S3.6E) or across vowel positions (all pFWE > 0.05).

Condition	**Dependent variable**	**Significant effect time range**	**t_{max}**	**Cluster-level p_{FWE}**
Predictable blocks	N-1 decoding	-10-273 ms	13.36	< 0.001
	N-2 decoding	40-106 ms	3.83	< 0.001
		166-203 ms	3.23	0.001
	N-3 decoding	33-123 ms	5.44	< 0.001
		223-240 ms	2.98	0.044
Random blocks	N-1 decoding	-10-140 ms	7.73	< 0.001

		163-203 ms	3.26	0.001
	N-2 decoding	-13-7 ms	3.03	0.029

Table S3.1. Multivariate analysis - decoding vs. baseline, related to Figure 3.2A. Decoding previous vowel identity based on burst-evoked responses: statistical results. Temporal clusters of significant decoding (one-sample t-tests against 0). Only significant effects shown. Clusters of significant decoding extended into the baseline are likely due to the sliding time window approach (50 ms) adopted in the multivariate analyses (see STAR Methods).

Condition	**Contrast**	**Significant effect time range**	**t_{max}**	**Cluster-level p_{FWE}**
Predictable blocks	N-1 vs. N-2	-6-236 ms	7.36	< 0.001
	N-1 vs. N-3	-3-216 ms	7.84	< 0.001
Random blocks	N-1 vs. N-2	0-136 ms	5.91	< 0.001
	N-1 vs. N-3	-6-126 ms	5.53	< 0.001

Table S3.2. Multivariate analysis - decoding differences between vowel positions, related to Figure 3.2A. Temporal clusters of significant differences in decoding between vowel positions (paired t-tests). Only significant effects shown. Clusters of significant decoding extended into the baseline are likely due to the sliding time window approach (50 ms) adopted in the multivariate analyses (see STAR Methods).

Effect	**Condition / contrast**	**Statistic**	**Cluster-level p_{FWE}**
Vowel decoding based on vowel-evoked activity	Predictable	t_{max} = 12.85	< 0.001
	Random	t_{max} = 9.74	< 0.001
	Predictable vs. random	n.s.	> 0.05
Vowel decoding vs.	N vs. N-1	t_{max} = 11.75	< 0.001

memory decoding			

Table S3.3. Multivariate analysis - vowel decoding based on vowel-evoked activity, related to Figure 3.2A. This analysis revealed significant decoding of vowel identity based on vowel-evoked responses during the entire post-stimulus time window. Overall, current vowel decoding was almost an order of magnitude higher than the decoding of previous vowel identity based on burst-evoked responses (mean decoding within 0-100 ms post-stimulus, mean ± SEM a.u.: N decoding 0.65 ± 0.07, N-1 decoding 0.07 ± 0.01).

Effect	**Condition / contrast**	**Statistic**	**Cluster-level p_{FWE}**
Generalization across time points	N-1	t_{max} = 20.21	< 0.001
	N-2	t_{max} = 16.05	< 0.001
	N-3	t_{max} = 13.64	< 0.001
Generalization across vowel positions	N-1 vs. N-2	F_{max} = 25.51, t_{min} = -5.05	0.027

Table S3.4. Multivariate analysis - cross-temporal generalization, related to Figure 3.4DE. For each vowel position, decoding did generalize across time points (cf. Cappotto et al., 2021). Negative cross-generalization was observed between N-1 and N-2 vowels, when averaging across predictable and random conditions, such that N-2 vowel decoding at 133-167 ms post-stimulus showed impaired (negative) decoding when trained on data corresponding to the N-1 vowel at 67-100 ms (Figure 3.4D), possibly reflecting an interference effect between N-1 and N-2 decoding. Beyond this finding, there were no significant cross-generalization clusters between the other vowel pairs (all pFWE > 0.05), as well as no differences between predictable and random blocks either in cross-temporal generalization across time points (all pFWE > 0.05) or across vowel positions (all pFWE > 0.05).

Chapter 4. "What" and "when" predictions modulate auditory processing in a contextually specific manner

4.1 Abstract

Extracting regularities from ongoing stimulus streams to form predictions is crucial for adaptive behavior. Such regularities exist in terms of the content of the stimuli (i.e., “what” it is) and their timing (i.e., “when” it will occur), both of which are known to interactively modulate sensory processing. In real-world stimulus streams, regularities also occur contextually - e.g. predictions of individual notes vs. melodic contour in music. However, it is unknown whether the brain integrates predictions in a contextually congruent manner (e.g., if slower “when” predictions selectively interact with complex “what” predictions), and whether integrating predictions of simple vs. complex features rely on dissociable neural correlates. To address these questions, our study employed “what” and “when” violations at different levels - single tones (elements) vs. tone pairs (chunks) - within the same stimulus stream, while neural activity was recorded using electroencephalogram (EEG) in participants (N=20) performing a repetition detection task. Our results reveal that “what” and “when” predictions interactively modulated stimulus-evoked response amplitude in a contextually congruent manner, but that these modulations were shared between contexts in terms of the spatiotemporal distribution of EEG signals. Effective connectivity analysis using dynamic causal modeling showed that the integration of “what” and “when” prediction selectively increased connectivity at relatively late cortical processing stages, between the superior temporal gyrus and the fronto-parietal network. Taken together, these results suggest that the brain integrates

different predictions with a high degree of contextual specificity, but in a shared and distributed cortical network.

4.2 Introduction

The ability to predict future events based on sensory information is an integral aspect of adaptive sensory processing. Real-world events are complex, consist of statistical regularities, and contain multiple features over which predictions can be formed (Dehaene et al., 2015). In the auditory domain, “what” and “when” predictions are present in virtually every stimulus stream, and their manipulation has been the foundation for numerous studies of predictive coding. “What” predictions are typically manipulated by introducing unexpected sensory deviants (oddballs), and comparing the neural responses to the unexpected vs. expected stimuli. In such oddball paradigms, the resulting classical mismatch response (MMR) is commonly interpreted as an error correction signal (Garrido et al., 2009). As opposed to “what” predictions, which often rely on MMR-based explanations, “when” predictions are typically explained by neural entrainment - phase alignment of neural activity to an external temporal structure (Auksztulewicz et al., 2019; Haegens and Zion Golumbic, 2018; Schroeder and Lakatos, 2009) (but see: (Doelling and Assaneo, 2021)). An influential study (Ding et al., 2016) has suggested that cortical activity can selectively entrain to contextual structures in linguistic sequences pursuant to levels of chunking. More recently, this finding has been extrapolated to artificial streams of auditory and visual stimuli (Henin et al., 2021).

Several studies have investigated predictions through independent manipulation of timing and content predictability, suggesting interactive and partly dissociable neural

correlates and putative underlying mechanisms (Arnal and Giraud, 2012; Auksztulewicz et al., 2018; Friston and Buzsáki, 2016; Kotz and Schwartze, 2010). MMR amplitudes are typically modulated by “when” predictions, such that deviant-evoked activity is higher when deviants are presented in temporally predictable (e.g., rhythmic/isochronous) sequences (Jalewa et al., 2021; Lumaca et al., 2019; Takegata and Morotomi, 1999; Todd et al., 2018; Yabe et al., 1997). In the auditory domain, such interactions have been suggested to rely on partially dissociable networks (Hsu, et al., 2013), while also jointly modulating stimulus-evoked activity in the superior temporal gyrus (Auksztulewicz et al., 2018). More generally, it has been proposed that interactions between “what” and “when” predictions are inherent to the processing of musical sequences (Musacchia et al., 2014). In this context, it has been suggested that neural entrainment along the non-lemniscal (secondary) auditory pathway (sensitive to the rhythmic sequence structure) can modulate activity in the lemniscal (primary) pathway (encoding stimulus contents), including MMR processing. Interestingly,

However, it is unknown if interactions between “what” predictions (in the lemniscal pathway) and “when” predictions (in the non-lemniscal pathway) are specific to differing contexts present in complex naturalistic stimuli such as speech or music (Hasson et al., 2015). In the case of naturalistic music stimuli, lower-level predictions can be formed about single notes within a sequence, while higher-level predictions can relate to the resulting melody contour, each occurring at their respective time scales. In principle, neural entrainment to a particular time scale might boost the processing of any stimuli presented in the expected time window (Auksztulewicz et al., 2019). However, if entrainment to slower (i.e., more global) temporal scales is functionally related to

chunking (Ding et al., 2016; Henin et al., 2021), it may show a specific modulation of the processing of stimulus chunks, rather than single elements. Thus, based on current hypotheses of neural entrainment and predictive processing, it is unclear if “when” predictions modulate the processing of stimulus contents (and the respective “what” predictions) in a contextually specific way - e.g. if temporal predictions amplify the processing of any stimuli presented at a preferred time window, or only those stimuli whose contents can be predicted at the corresponding time scale.

Here, we present streams of tones and independently manipulate content-based and time-based characteristics of the stream at two levels, while recording EEG in healthy volunteers. Temporal predictability was manipulated at slower (~2 Hz) and faster (~4 Hz) time scales, while acoustic deviants were introduced at lower (e.g. single tones) and higher (e.g.chunked tone pairs) levels, to evaluate the independent or interactive effect of “what” and “when” predictive processing across contexts.

4.3 Methods

EEG was recorded during an auditory repetition detection task in order to gauge (1) the effects of “when” predictions at higher and lower temporal scales on tone-evoked responses and on neural entrainment, as well as (2) the modulatory effect of “when” predictions on the neural signatures of higher and lower-level “what” predictions (MMRs). The use of musical sequences (ascending or descending musical scales) was chosen to reduce the influence of speech-specific processing on neural activity (e.g., modulation by language comprehension, speech-specific semantic and syntactic processing, etc.) and provide a better comparison to similar work in animal models (Jalewa et al., 2021). In the analysis, we focused on interactions between "what" and "when" predictions, specifically

testing whether MMRs are modulated by temporal predictability in a contextually specific way (such that slower “when” predictions selectively modulate MMRs to violations of higher-level “what” predictions). To explain the effects observed at the scalp level, we used source reconstruction and biophysically realistic computational modeling (dynamic causal modeling), which allowed us to infer the putative mechanisms of interactions between "what" and "when" predictions.

4.3.1 Participant sample

Participants (N=20, median age 21, range 19-25), 10 females, 10 males; 19 right-handed, 1 left-handed) volunteered to take part in the study upon written consent. The work was conducted in accordance with protocols approved by the Human Subjects Ethics Sub-Committee of the City University of Hong Kong. All participants self-reported normal hearing and no current or past neurological or psychiatric disorders.

4.3.2 Stimulus design and behavioral paradigm

An experimental paradigm was designed in which auditory sequences were manipulated with respect to “what” and “when” predictions at two contextual levels (“what” predictions of single tones vs. chunked tone pairs; “when” predictions at ~4 Hz vs. ~2 Hz), allowing for an analysis of their interactions at each level. To ensure that participants paid attention to stimulus sequences, the sequences contained very occasional repetitions, and participants were instructed to listen out for such repetitions (see below). The experimental manipulations of “what” and “when” predictions, however, were irrelevant to the task, such that neural responses to unexpected stimuli are not confounded by neural

activity related to target detection.

Auditory sequences were generated using Psychtoolbox for MATLAB (version 2021a) and delivered to participants fitted with Brainwavz B100 earphones via a TDT RZ6 multiprocessor at a playback sampling rate of 24414 Hz. Participants were seated in a sound-attenuated EEG booth. Visual stimuli (fixation cross) and instructions were presented on a 24-inch computer monitor and delivered using the Psychophysics Toolbox for MATLAB. Participants were asked to minimize movements and eye blinks and instructed to perform a tone repetition detection task, by pressing a keyboard button using their right index finger as soon as possible upon hearing an immediate tone repetition.

Stimuli were presented in sequences of 7 ascending or descending scales. Each scale was composed of 8 tones equally spaced on a logarithmic scale to form one octave. Thus, across 7 scales a total 56 tones were presented per sequence (Figure 4.1A, 4.1B). A trial was defined as the presentation of a sequence of 7 scales. Within a trial, all scales were either ascending or descending. The ascending and descending trials were presented in a random order. Each participant heard a total of 240 sequences (trials). The initial tone of each scale was randomly drawn from a frequency range 300-600 Hz. Each tone was generated by resynthesizing a virtual harp note F4 (played on virtualpiano.net), to match a fixed 166 ms duration and the fundamental frequency used at a given position in the scale. The tone manipulations were implemented in an open-source vocoder, STRAIGHT (Kawahara, 2006) for Matlab 2018b (MathWorks; RRID: SCR_001622). Tones were perceptually grouped into pairs by manipulating the intensity ratio of odd/even tones, with the even (2nd, 4th, 6th and 8th) tones within a scale presented 10 dB quieter relative to the odd-position tones (Kotz et al., 2018).

Manipulation of temporal predictability formed three conditions: in the fully-predictable (isochronous) condition, tones were presented with a fixed ISI (inter-stimulus interval) of 247 ms, resulting in all tones having predictable timing at both the slow time scale (chunks) and the fast time scale (elements). In the temporally-global (predictable slow, unpredictable fast) condition, the slow time scale was predictable (corresponding to a fixed pair onset asynchrony, i.e., a fixed 494 ms interval between the onsets of the odd, pair-initial tones) but the fast time scale was unpredictable (corresponding to a random onset of the even, pair-final tones, relative to the pair-initial tones). In this condition, the exact ISI of the pair-final tones was set by randomly drawing one value from the following 4 ISIs, relative to the standard 247 ms ISI: 33.3% shorter; 16.6% shorter; 16.6% longer; 33.3% longer. Finally, in the temporally-local (predictable fast, unpredictable slow) condition, the onset of stimuli at the fast time scale was predictable (corresponding to a fixed 247 ms ISI of the pair-final tones, relative to the pair-initial tones) but the slow time scale was unpredictable (corresponding to a random onset of the odd, pair-initial tones, relative to the expected 494 ms interval). In this condition, the exact ISI of the pair-initial tones was set by randomly drawing one value from the same 4 ISIs as above, and shifting the onset of the pair-initial tone by this value, relative to the expected 494 ms interval relative to the previous pair onset. A fixed inter-trial interval of 1 second was employed between the offset of the last tone of a 56-tone sequence and the onset of the first tone in the next sequence. The three timing conditions were administered in 12 blocks of 20 trials (4 blocks per condition). Blocks were pseudo-random in order, allowing no immediate repetitions of the same, timing, condition.

Content predictability was manipulated by altering the fundamental frequency of a

subset of tones within the scales, such that trials could contain an element deviant (i.e., a single deviant tone) or a “chunk” deviant (i.e., a deviant tone pair). The element deviants were introduced by replacing the final tone of a scale with an outlier frequency (i.e., a tone whose fundamental frequency was 20% lower/higher than the range of the entire scale). The chunk deviants were introduced by replacing the penultimate tone of a scale (i.e., the initial tone of the final pair, rendering the entire pair unpredictable) in the same manner.

To facilitate the extraction of statistical regularities in the sequences, in each trial, the first two scales were left unaltered. Two deviant tones were randomly placed within the subsequent 5 scales. Additionally, in 50% of the trials, a scale containing an immediate tone repetition was included in the last 5 scales. In subsequent EEG analysis, neural responses evoked by element and chunk deviants were compared with neural responses evoked by the respective standard tones, designated as the final (standard element) and penultimate (standard chunk) tones in two unaltered scales out of the final 5.

In total, 64.3% of the scales were left unaltered, 14.3% contained an element deviant, 14.3% contained a chunk deviant, and 7.1% contained a tone repetition. The global deviant probability equaled 3.57% of all tones, amounting to 80 deviant tones per deviant type (element, chunk) per temporal condition (fully predictable, temporally local, temporally global). To ensure that the EEG analysis is not confounded by differences in baseline duration between temporal conditions (e.g., element deviants preceded by shorter/longer ISIs in the temporally global condition than in the other two conditions), the ISIs preceding all deviant tones and designated standard tones were replaced by a fixed 247 ms ISI. Therefore, the temporal predictability manipulation was limited to tones

surrounding the analyzed tones and did not affect the exact timing of either deviants or standards.

Prior to experimental blocks, participants were exposed to a training session consisting of fully predictable sequences containing a tone repetition, to familiarize themselves with the task and stimuli. Participants performed training trials until they could detect tone repetition in 3 consecutive trials with reaction times shorter than 2 seconds. Then, during the actual experiment, participants received feedback on their mean accuracy and reaction time after each block of 20 trials. The data segments (scales) containing tone repetition were subsequently discarded from EEG analysis.

4.3.3 Behavioral analysis

Analysis was performed on the accuracy and reaction time data corresponding to participant responses during the repetition detection task. Reaction times longer than 2 seconds were excluded from analysis. Mean reaction times (from correct trials only) were log-transformed to approximate a normal distribution. Accuracy and mean reaction times were entered into separate repeated-measures ANOVAs with a within-subjects factor Time (fully predictable, temporally local, temporally global). Post-hoc comparisons were implemented using paired t-tests in MATLAB.

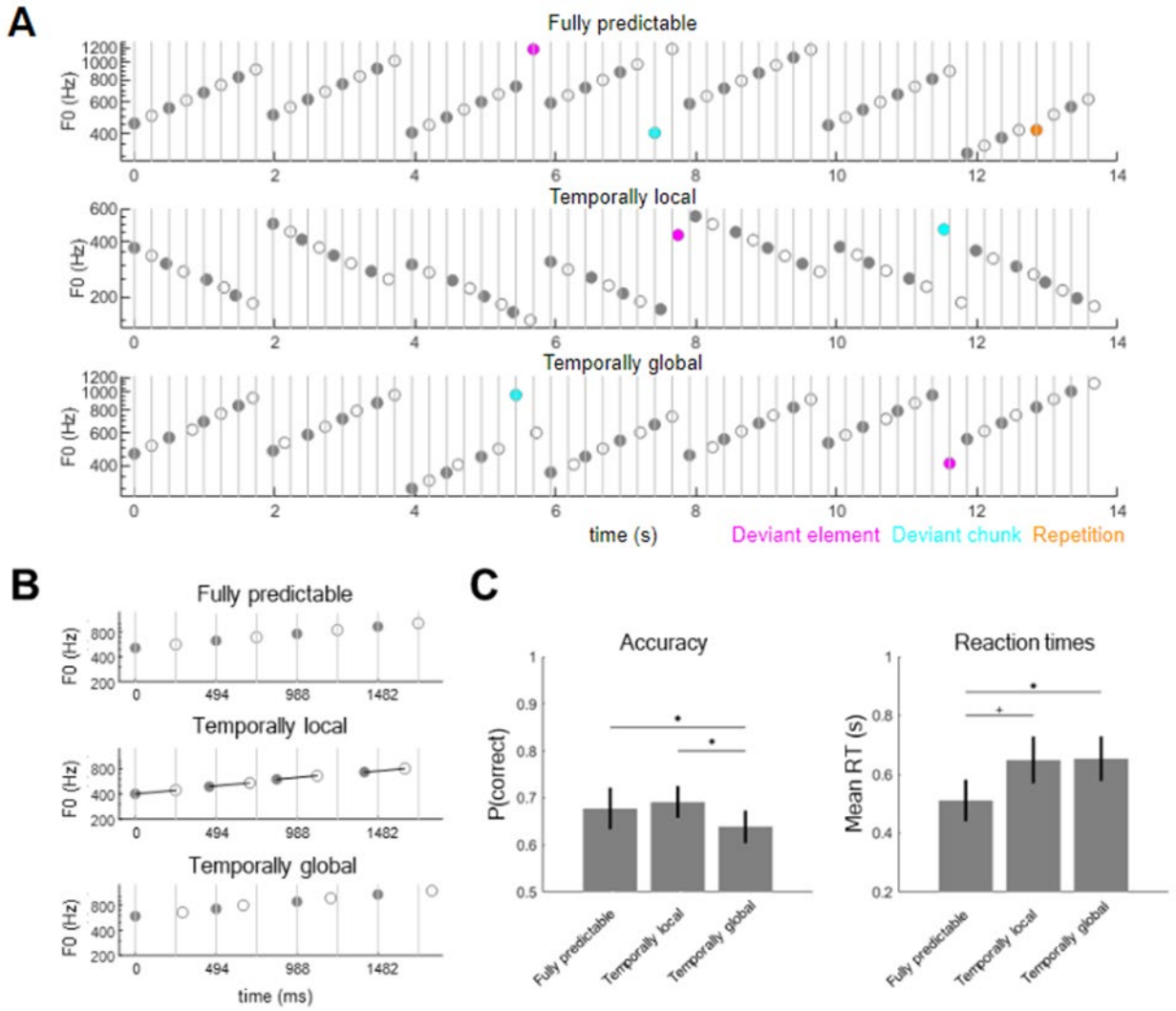

Figure 4.1 Experimental paradigm and behavioral results. **(A)** Participants listened to sequences of ascending (as represented on the figure) or descending scales of acoustic tones. Sequences were composed of tone pairs, where odd tones (gray circles) were louder than even tones (white circles). Participants performed a tone repetition detection task (orange circles: behavioral targets; presented in a subset of trials). Additionally, sequences could include deviant tones (magenta circles), in which one of the pair-final tones had an outlier fundamental frequency (F0), and deviant chunks (cyan circles), in which one of the pair-initial tones had an outlier F0. **(B)** Sequences were blocked into three temporal conditions: a fully-predictable condition (upper panel), in which ISI between tones was fixed at 0.247 s; a temporally-local condition (middle panel), in which the ISI between odd and even tones within pairs was fixed at 0.247 s but the ISI between odd tones (pair-initial tones) was jittered; and a temporally-global condition (lower panel), in which ISI between odd tones (pair-initial tones) was fixed at 0.494 s but the ISI between odd and even tones within pairs was jittered. **(C)** Behavioral results. Left panel: accuracy, right panel: reaction times. Error bars denote SEM across participants. Asterisks denote $p < 0.05$, plus symbol denotes a trend towards significance.

4.3.4 Neural data acquisition and pre-processing

EEG signals were collected using a 64-channel ANT Neuro EEGo Sports amplifier at a sampling rate of 1024 Hz with no online filters. The recorded data were pre-processed using the SPM12 Toolbox (version 7219; Wellcome Trust Centre for Neuroimaging, University College London; RRID: SCR_007037) for MATLAB (version R2018b). Continuous data were high-pass filtered at 0.1 Hz and notch filtered between 48 Hz and 52 Hz before being down-sampled to 300 Hz and subsequently low-pass filtered at 90 Hz. All filters were 5th order zero-phase Butterworth. Eyeblink artifacts were detected using channel Fpz and removed by subtracting the two top spatiotemporal principal components of eyeblink-evoked responses from all EEG channels (Ille et al., 2002). Cleaned signals were re-referenced to the average of all channels, as is recommended for source reconstruction and dynamic causal modeling (Litvak and Friston, 2008). The pre-processed data were analyzed separately in the frequency domain (phase coherence analysis) and in the time domain (event-related potentials; ERPs).

4.3.5 Phase coherence analysis

To test whether tone sequences are associated with dissociable spectral peaks in the neural responses at the element rate (4.048 Hz) and at the chunk rate (2.024 Hz), we analyzed the data in the frequency domain. Continuous data were segmented into epochs ranging from the onset to the offset of each trial (tone sequence). For each participant, channel, and sequence, we calculated the Fourier spectrum of EEG signals measured during that sequence. Based on previous literature, we then calculated the inter-trial phase coherence (ITPC), separately for each temporal condition (fully-predictable, temporally-local, temporally-global) according to the following equation (Ding and Simon,

2013) in order to infer phase consistency in each condition:

$$ITPC_f = \left(\left[\Sigma^N cos\phi_f\right]^2 + \left[\Sigma^N sin\phi_f\right]^2\right)/N,$$

where φ_f corresponds to the Fourier phase at a given frequency *f,* and *N* corresponds to the number of sequences (80 per condition). The same method was used to estimate the stimulus frequency spectrum by calculating the ITPC based on the raw stimulus waveform.

In the initial analysis, ITPC estimates were averaged across EEG channels. To test for the presence of statistically significant spectral peaks, ITPC values at the single-tone rate (4.048 Hz) and tone-pair rate (2.024 Hz) were compared against the mean of ITPC values at their respective neighboring frequencies (single-tone rate: 3.974 and 4.124 Hz; tone-pair rate: 1.949 and 2.099 Hz) using paired t-tests.

Furthermore, to test whether element-rate and chunk-rate spectral peaks observed at single EEG channels show modulations due to temporal predictability, spatial topography maps of single-channel ITPC estimates were converted to 2D images, smoothed with a 5 x 5 mm full-width-at-half-maximum (FWHM) Gaussian kernel, and entered into repeated-measures ANOVAs (separately for element-rate and chunk-rate estimates) with a within-subjects factor Time (fully predictable, temporally local, temporally global), implemented in SPM12 as a general linear model (GLM). To account for multiple comparisons and for ITC correlations across neighboring channels, statistical parametric maps were thresholded at $p < 0.001$ and corrected for multiple comparisons over space at a cluster-level $p_{FWE} < 0.05$ under random field theory assumptions (Kilner et al., 2005).

Finally, to test whether spectral signatures of temporal predictability are modulated

by experience with stimuli, we split the data into two halves (two consecutive bins of 40 trials), separately for each condition. Element-rate and chunk-rate ITPC estimates were averaged across EEG channels and compared separately for each of the two halves using repeated-measures ANOVAs with a within-subjects factor Time (fully predictable, temporally local, temporally global).

4.3.6 Event-related potentials

For the time-domain analysis, data were segmented into epochs ranging from -50 ms before to 247 ms after deviant/standard tone onset, baseline-corrected from -25 ms to 25 ms to prevent epoch contamination due to the temporally structured presentation (Fitzgerald et al., 2021), and denoised using the “Dynamic Separation of Sources” (DSS) algorithm (de Cheveigné and Simon, 2008). Condition-specific ERPs (corresponding to element/chunk deviants and the respective standards, presented in each of the three temporal conditions) were calculated using robust averaging across trials, as implemented in the SPM12 toolbox, and low-pass filtered at 48 Hz (5th order zero-phase Butterworth). The resulting ERPs were analyzed univariately to gauge the effects of “what” and “when” predictions on evoked responses. ERP data were converted to 3D images (2D: spatial topography; 1D: time), and the resulting images were spatially smoothed using a 5 x 5 mm FWHM Gaussian kernel. The smoothed images were entered into a general linear model (GLM) implementing a 3 x 3 repeated-measures ANOVA with a within-subject factors Contents (standard, deviant element, deviant chunk) and Time (fully predictable, temporally local, temporally global). Beyond testing for the two main effects and a general 3 x 3 interaction, we also designed a planned contrast quantifying the congruence effect (i.e., whether “when” predictions specifically modulate the

amplitude of mismatch signals evoked by deviants presented at a time scale congruent with “when” predictions, i.e., deviant elements in the temporally-local condition and deviant chunks in the temporally-global conditions). To this end, we tested for a 2 x 2 interaction between Contents (deviant element, deviant chunk) and Time (temporally local, temporally global). To account for multiple comparisons as well as for ERP amplitude correlations across neighboring channels and time points, statistical parametric maps were thresholded at $p < 0.001$ and corrected for multiple comparisons over space and time at a cluster-level $p_{FWE} < 0.05$ under random field theory assumptions (Kilner et al., 2005).

4.3.7 Brain-behavior correlations

To test whether the neural effects of “what” and/or “when” predictive processing correlate with each other, as well as with behavioral benefits of “when” predictions in the repetition detection task, we performed a correlation analysis across participants. Thus, for each participant, we calculated a single behavioral index (the difference between accuracy scores obtained in the temporally local vs. temporally global condition) and three statistically significant neural indices. The first neural index - the “congruence effect” - quantified the difference between deviant-evoked ERP amplitudes measured in the temporally congruent (deviant elements presented in the temporally local condition; deviant chunks presented in the temporally global condition) and incongruent (deviant chunks presented in the temporally local condition; deviant elements presented in the temporally global condition) conditions, averaged across electrodes in the significant cluster where we observed a significant congruence effect (i.e., a 2 x 2 interaction between “what” and “when” predictions; see Results and Figure 4.3C). The second neural

index - the “ITPC effect” - quantified the difference between the chunk-rate ITPC values obtained for temporally local vs. temporally global conditions in the second half of the experiment (see Results Figure 4.2D). The third neural index - the “mismatch effect” - quantified the difference between the absolute deviant-evoked and standard-evoked ERP amplitudes (averaged across significant channels and temporal conditions; Figure 4.3AB), since we hypothesized that performance in the repetition detection task might be related to overall deviance detection, we also included an index of “what” predictions. We then fitted a linear regression model with three predictors (i.e., the three neural indices) regressed against the behavioral accuracy index, and identified outlier participants using a threshold of Cook’s distance exceeding 5 times the mean. Correlations between all measures were quantified using Pearson’s *r* and corrected for multiple comparisons using Bonferroni correction, implementing a conservative correction given no a priori assumptions about the correlation coefficients.

4.3.8 Source reconstruction

Source reconstruction was performed under group constraints (Litvak and Friston, 2008) which allows for an estimation of source activity at a single-participant level under the assumption that activity is reconstructed in the same subset of sources for each participant. Sources were estimated using empirical Bayesian beamformer (Belardinelli et al., 2012; Little et al., 2018; Wipf and Nagarajan, 2009) based on the entire post-stimulus time window (0-247 ms). Since in the ERP analysis (see Results) we identified two principal findings - namely a difference between ERPs evoked by deviants and standards, and an interaction between deviant type and temporal condition - we focused on comparing source estimates corresponding to these effects. In the analysis of the

difference between deviants and standards, source estimates were extracted for the 173-223 ms time window, converted into 3D images consisting of 3 spatial dimensions and smoothed with a 10 x 10 x 10 mm FWHM Gaussian kernel. Smoothed images were then entered into a GLM implementing a 3 x 3 repeated-measures ANOVA with within-subjects factors of Content (standard, deviant element, deviant chunk) and Time (fully predictable, temporally local, temporally global). In the analysis of the interaction between deviant type and temporal condition, source estimates were extracted for the 130-180 ms and processed as above. Smoothed images were then entered into a GLM implementing a 2 x 2 repeated-measures ANOVA with within-subjects factors of Content (deviant element, deviant chunk) and Time (temporally local, temporally global). To account for multiple comparisons as well as for source estimate correlations across neighboring voxels, statistical parametric maps were thresholded and corrected for multiple comparisons over space at a cluster-level $p_{FWE} < 0.05$ under random field theory assumptions (Kilner et al., 2005). Source labels were assigned using the Neuromorphometrics probabilistic atlas, as implemented in SPM12.

4.3.9 Dynamic causal modeling

Dynamic causal modeling (DCM) was used to estimate source-level connectivity parameters associated with general mismatch processing (deviant vs. standard) and with the contextual interaction between “what” and “when” predictions (element deviant presented in the temporally-local condition, and chunk deviant presented in the temporally-global condition, vs. element deviant presented in the temporally global condition, and chunk deviant presented in the temporally-local condition). DCM is a type of an effective connectivity analysis based on a generative model, which maps the data

measured at the sensor level (here: EEG channels) to source-level activity. The generative model comprises a number of sources which represent distinct cortical regions, forming a sparse interconnected network. Activity in each source is explained by a set of neural populations, based on a canonical microcircuit (Bastos et al., 2012), and modeled using coupled differential equations that describe the changes in postsynaptic voltage and current in each population. Here, we used a microcircuit consisting of four populations (superficial and deep pyramidal cells, spiny stellate cells, and inhibitory interneurons), each having a distinct connectivity profile of ascending and descending extrinsic connectivity (linking different sources) and intrinsic connectivity (linking different populations within each source). The exact form of the canonical microcircuit and the connectivity profile was identical as in previous literature on the topic (Auksztulewicz et al., 2018; Auksztulewicz and Friston, 2015; Fitzgerald et al., 2021; Rosch et al., 2019; Todorovic and Auksztulewicz, 2021). Importantly for our study, a subset of intrinsic connections corresponds to self-connectivity parameters, describing the neural gain of each region. Both extrinsic connectivity and gain parameters were allowed to undergo condition-specific changes, modeling differences between experimental conditions (deviants vs. standards, and the hierarchical interaction between “what” and “when” predictions).

Here, we used DCM to reproduce the single-participant, condition-specific ERPs in the 0-247 ms range. Based on the source reconstruction (see Results) and previous literature (Garrido et al., 2009), we included six sources in the cortical network: bilateral primary auditory cortex (A1; Montreal Neurological Institute coordinates: left, [-42 -22 7]; right, [46 -14 8]), bilateral superior temporal gyrus (STG; left, [-60 -20 -8]; right, [59 -25

8]), right inferior frontal gyrus (IFG; [40 26 -6]), and left superior parietal lobule (SPL; [-26 -40 46]). To quantify model fits, we used the free-energy approximation to model evidence, penalized by model complexity. The analysis was conducted in a hierarchical manner-first, model parameters (including extrinsic and intrinsic connections, as well as their condition-specific changes) were optimized at the single participants' level, and then the significant parameters were inferred at the group level.

At the first level, models were fitted to single participants' ERP data over two factors: “what” predictions (all deviants vs. standards) and the contextual interaction between “what” and “when” predictions (element deviant presented in the temporally-local condition, and chunk deviant presented in the temporally-global condition, vs. element deviant presented in the temporally-global condition, and chunk deviant presented in the temporally-local condition). At this level, all extrinsic and intrinsic connections were allowed to be modulated by both factors, corresponding to a “full” model.

Since model inversion in DCM is susceptible to local maxima due to the inherently nonlinear nature of the models, the analysis at the second (group) level implemented parametric empirical Bayes (Friston et al., 2015). Therefore, group-level effects were inferred by (re)fitting the same “full” models to single participants’ data, under the assumption that model parameters should be normally distributed in the participant sample, and updating the posterior distribution of the parameter estimates. We used Bayesian model reduction (Friston and Penny, 2011) to compare the “full” models against a range of “reduced” models, in which some parameters were not permitted to be modulated by the experimental factors. Specifically, we designed a space of alternative models, such that each model allowed for a different subset of connections to contribute

to the observed ERPs. The model space examined each combination of modulations of (1) ascending connections (e.g., from A1 to STG), (2) descending connections (e.g., from STG to A1), (3) lateral connections (e.g., from left to right STG), and (4) intrinsic connections (i.e., gain parameters). This resulted in 256 models (16 models for each of the two factors). The free-energy approximation to log-model evidence was used to score each model. Since no single winning model was selected (see Results), Bayesian model averaging was used to obtain weighted averages of posterior parameter estimates, weighted by the log-evidence of each model. This procedure yielded Bayesian confidence intervals for each parameter, quantifying the uncertainty of parameter estimates. Parameters with 99.9% confidence intervals falling either side of zero (corresponding to $p < 0.001$) were selected as statistically significant.

4.4 Results

4.4.1 Behavioral results

Performance across all trials revealed significant differences in accuracy across conditions (main effect of Time: $F_{2,38} = 7.3530$, $p = 0.002$), corresponding to significantly lower accuracy in the temporally-global condition (mean ± SEM: 63.88% ± 3.65%) than in the fully-predictable (mean ± SEM: 67.75% ± 4.64%; t19 = -2.5272, $p = 0.0205$) and temporally-local conditions (mean ± SEM: 69.12% ± 3.55%; t19 = -5.984, $p < 0.001$) (Figure 4.1C). Reaction times also significantly differed across conditions ($F_{2,38} = 3.5543$, $p = 0.0385$), with post-hoc analysis revealing that reaction times were significantly faster in the fully-predictable condition (mean ± SEM: 511 ± 74 ms) than in the temporally-global condition (mean ± SEM: 653 ± 79 ms; t19 = 2.4089, $p = 0.0263$). The difference between

the fully-predictable condition and the temporally-local condition (mean ± SEM: 649 ± 83 ms) trended towards significance ($t19 = 2.0132$, $p = 0.0585$). No significant difference was observed between the temporally-global condition and the temporally-local condition ($p = 0.9013$).

4.4.2 Phase coherence analysis

In the EEG spectrum of inter-trial phase coherence (ITPC; averaged across conditions and channels), both element-rate peak (4.048 Hz) and chunk-rate peak (2.024 Hz) were observed, relative to neighboring frequency points (element-rate: $t_{19} = 6.8489$, $p < 0.001$; chunk-rate: $t_{19} = 3.6274$, $p = 0.0018$). The ITPC peak estimates differed between experimental conditions, reflecting differences in the stimulus spectrum. Specifically, the chunk-rate ITPC estimates were higher in the fully-predictable and temporally-global conditions than in the temporally-local conditions, and this effect was observed at most of the EEG channels ($F_{max} = 46.30$, $Z_{max} = 6.43$, $p_{FWE} < 0.001$; pairwise comparisons: fully-predictable vs. temporally-local, $T_{max} = 8.02$, $Z_{max} = 6.10$, $p_{FWE} < 0.001$; temporally-global vs. temporally-local, $T_{max} = 9.62$, $Z_{max} = 6.81$, $p_{FWE} < 0.001$; fully-predictable vs. temporally-global, all $p_{FWE} > 0.05$). On the other hand, the chunk-rate ITPC estimates were higher in the temporally-global condition than in the other two conditions, and this effect was observed over right lateral channels ($F_{max} = 7.45$, $Z_{max} = 2.90$, $p_{FWE} = 0.031$; pairwise comparisons: temporally-global vs. fully-predictable, $T_{max} = 3.81$, $Z_{max} = 3.48$, $p_{FWE} = 0.004$; temporally-global vs. temporally-local, $T_{max} = 3.83$, $Z_{max} = 3.50$, $p_{FWE} = 0.001$; fully-predictable vs. temporally-local, all $p_{FWE} > 0.05$). Interestingly, the chunk-rate differences between conditions built up during the experiment: they were absent during the first half of the experiment ($F_{2,59} = 1.0433$, $p = 0.3622$), and were only observed

during the second half of the experiment ($F_{2,59}$ = 3.8798, p = 0.0293). This was not the case for the element-rate differences between conditions, which were stable during the experiment (first half: $F_{2,59}$ = 26.1701, p < 0.001; second half: $F_{2,59}$ = 26.9480, p < 0.001).

The emergence of chunk-rate differences in ITPC over the course of the experiment was reflected in behavior. Specifically, RTs decreased for the second half of the experiment, relative to the first half, only for the temporally-global condition (Wilcoxon's signed rank test: Z_{19} = -2.0926, p = 0.0364) but not for the fully-predictable condition (Z_{19} = -1.6902, p = 0.0910) or the temporally-local condition (Z_{19} = -0.8213, p = 0.4115). No differences in accuracy were observed for any of the three conditions across the first and second halves of the experiment (all p > 0.05).

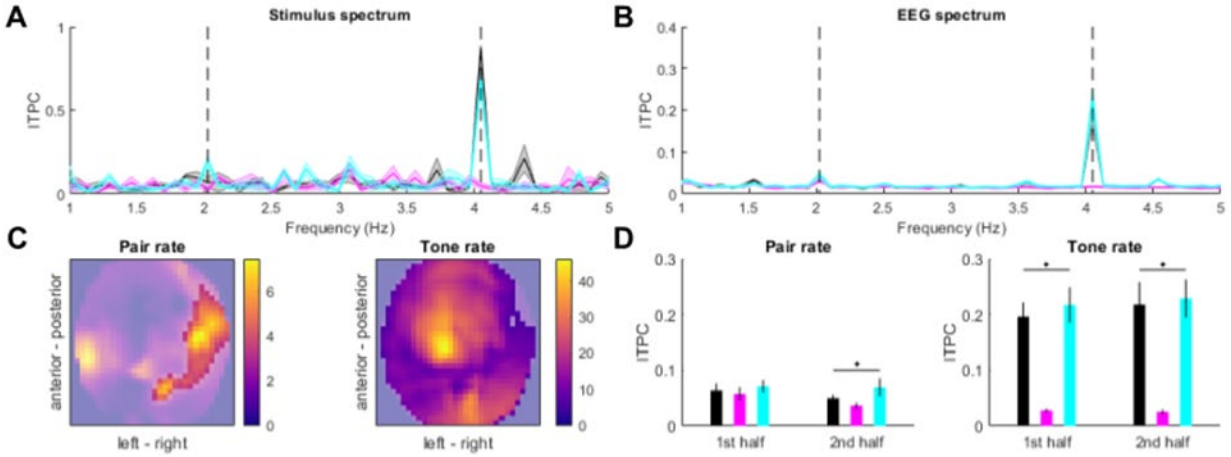

Figure 4.2 Spectral signatures of temporal predictability. **(A)** Inter-trial phase coherence (ITPC) in the stimulus spectrum. Black: fully-predictable, cyan: temporally-local, magenta: temporally-global. Chunk-rate (2.024 Hz) and element-rate (4.048 Hz) peaks are indicated by dashed vertical lines. **(B)** ITPC based on EEG activity (averaged across channels). Legend as above. Shaded areas indicate SEM across participants. **(C)** EEG topography maps of main effects of Condition (fully predictable vs. temporally local vs. temporally global) on the chunk-rate peak ITPC (left panel) and tone-rate peak ITPC (right panel). Statistical F values are represented on the color scale. Unmasked area corresponds to significant clusters (p_{FWE} < 0.05). **(D)** Chunk-rate (left panel) and element-rate (right panel) peak ITPC values plotted separately for the 1st half and 2nd half of the trials. Error bars denote SEM across participants.

4.4.3 Event-related potentials

To test for effects of “what” and “when” predictions on ERP amplitudes, we analyzed the data in the time domain. ERP amplitudes differed significantly between deviant and standard tones, pooled over temporal conditions (Figure 4.3A; posterior cluster: 173-223 ms, F_{max} = 53.94, Z_{max} = 6.68, p_{FWE} < 0.001; anterior cluster: 177-220 ms; F_{max} = 37.57; Z_{max} = 5.67; p_{FWE} < 0.001), corresponding to a typical anterior-posterior MMN topography after common-average referencing (Mahajan et al., 2017). When analyzing specific deviant types (element and chunk deviants vs. their respective standards), significant differences between deviants and standards were observed in both cases (element deviants vs. standards: posterior cluster, 173-223 ms, F_{max} = 41.50, Z_{max} = 5.94, p_{FWE} < 0.001; anterior cluster, 177-227 ms; F_{max} = 35.56; Z_{max} = 5.52; p_{FWE} < 0.001; chunk deviants vs. standards: posterior cluster, 170-220 ms, F_{max} = 45.63, Z_{max} = 6.20, p_{FWE} < 0.001; anterior cluster, 177-213 ms; F_{max} = 30.17; Z_{max} = 5.11; p_{FWE} < 0.001). No significant differences were observed between the two deviant types, pooling over temporal conditions (p_{FWE} > 0.05). Thus, the main effect of “what” predictions differentiated between deviants and standards, but not between deviant types.

In the analysis of the main effect of “when” predictions (pooled over deviants and standards), no significant differences between the three temporal conditions were revealed (all p_{FWE} > 0.05). Similarly, in the analysis of the interaction effect of “what” and “when” predictions (pooled over deviant types), no significant effects were revealed. Specifically, neither deviants nor standards showed significant ERP amplitude differences when presented in different temporal contexts (all p_{FWE} > 0.05). Thus, the overall temporal

structure of the sound sequences did not affect the element-evoked responses (averaged across deviants and standards) or the mismatch responses (differences between deviants and standards).

However, an analysis of the interaction between “what” and “when” predictions based on deviants presented in congruent temporal contexts (e.g. element deviants in the temporally-local condition) and those presented in non-temporally congruent contexts (e.g. element deviants in the temporally-global condition) revealed a significant interaction between deviant type and temporal condition (Figure 4.3B; left central-posterior cluster: 130-180 ms, F_{max} = 20.63, Z_{max} = 4.24, p_{FWE} = 0.044). Post-hoc analysis revealed that MMR amplitudes in temporally-local were significantly larger for deviant elements (mean ± SEM: -0.1640 ± 0.0942 μV) than for deviant chunks (mean ± SEM: 0.0091 ± 0.1010 μV; t_{19} = 2.2843, p = 0.0340, two-tailed). In the temporally-global condition, MMR amplitude was observed to be nominally larger for deviant chunks (mean ± SEM: -0.1725 ± 0.0851 μV) than for deviant elements (mean ± SEM: -0.0155 ± 0.1233 μV), although the effect did not reach significance (t_{19} = 1.9024, p = 0.0724, two-tailed). No significant interaction effects were revealed when comparing deviant types between the fully-predictable condition and either the temporally-global or the temporally-local conditions. Thus, we observed a specific increase in deviant ERP amplitude when this deviant was presented in a temporally congruent context.

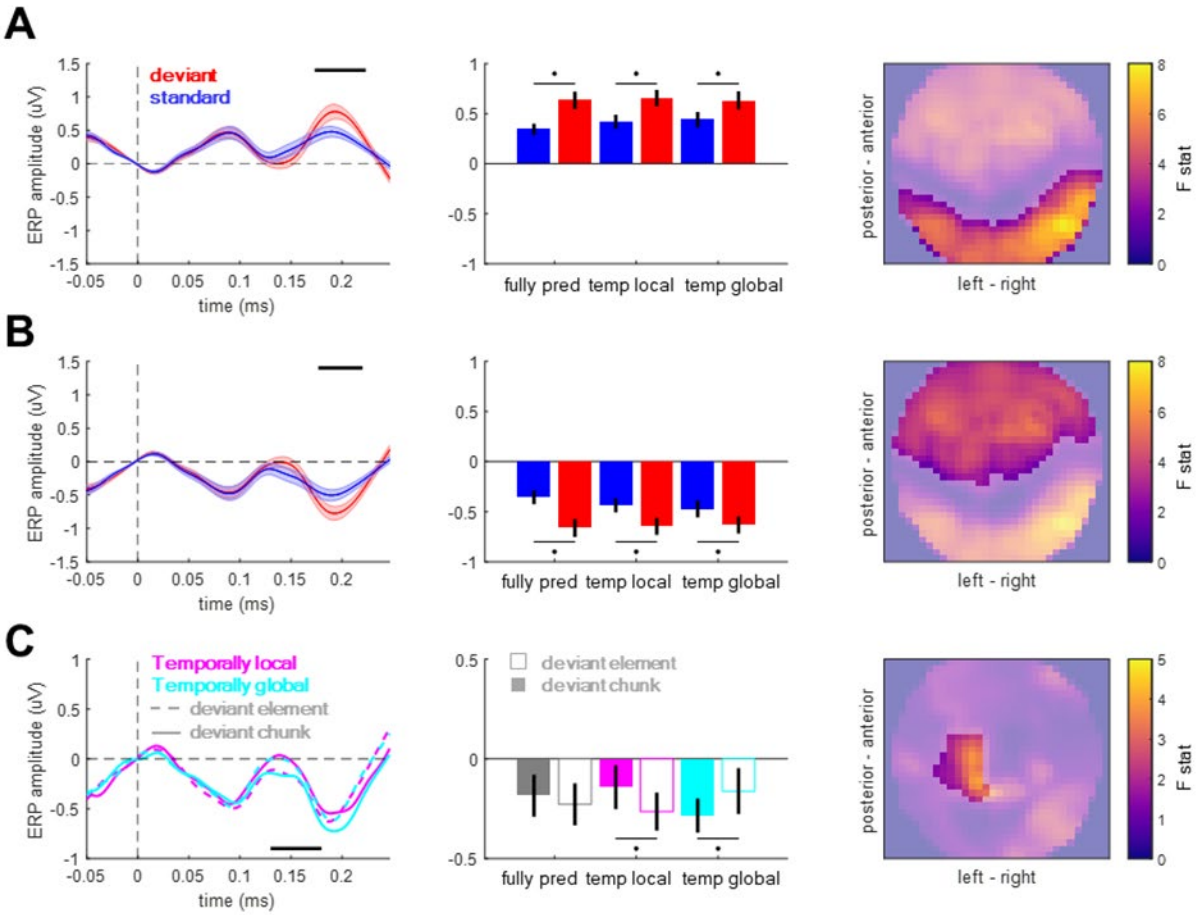

Figure 4.3 Event-related potentials. **(AB)** Main effect of content-based predictions (deviant vs. standard) in anterior (A) and posterior (B) clusters. Left panels: time courses of ERPs averaged over the spatial topography clusters shown in the right panels. Shaded area denotes SEM across participants. Black horizontal bar denotes $p_{FWE} < 0.05$. Middle panels: mean voltage values for standards (blue) and deviants (red). Right panels: spatial distribution of the main effect. Color bar: F value. **(C)** Contextual interaction between content-based predictions (deviant element vs. deviant chunk) and temporal predictions (temporally global vs. temporally local). Left panels: time courses of ERPs averaged over the spatial topography clusters shown in the right panels. Black horizontal bar denotes $p_{FWE} < 0.05$. Middle panels: mean voltage values for the six deviant conditions. Right panels: spatial distribution of the interaction effect. Color bar: F value.

4.4.4 Brain-behavior correlation analysis

Three neural predictors (the “congruence index”, quantifying the interactive effects of “what” and “when” predictions on ERPs; the “ITPC index”, quantifying the effect of “when” predictions on ITPC; and the “mismatch index”, quantifying the effect of “what”

predictions on ERPs) were tested as potential correlates of the behavioral benefits in the repetition detection task accuracy. We identified two outlier participants based on a linear regression model. Having excluded these two participants, we did not find any significant correlations between the neural indices and the behavioral index (Pearson's *r;* all $p > 0.2$), suggesting that behavior in the repetition detection task is not functionally related to ERP signatures of deviance detection. However, we did find a significant correlation between the congruence index and the ITPC index ($r = 0.6439$; $p = 0.0039$; corrected), such that the magnitude of the ERP difference between deviants presented in the temporally congruent vs. incongruent conditions positively correlated with the magnitude of the ITPC difference between temporally-local and temporally-global conditions.

4.4.5 Source reconstruction

To identify the most plausible sources underlying the observed ERP differences between deviants and standards, as well as the contextual interaction between deviant types and temporal conditions, we carried out a source reconstruction analysis (Figure 4.4). Overall, source reconstruction explained 76.43 ± 3.08% (mean ± SEM across participants) of sensor-level variance.

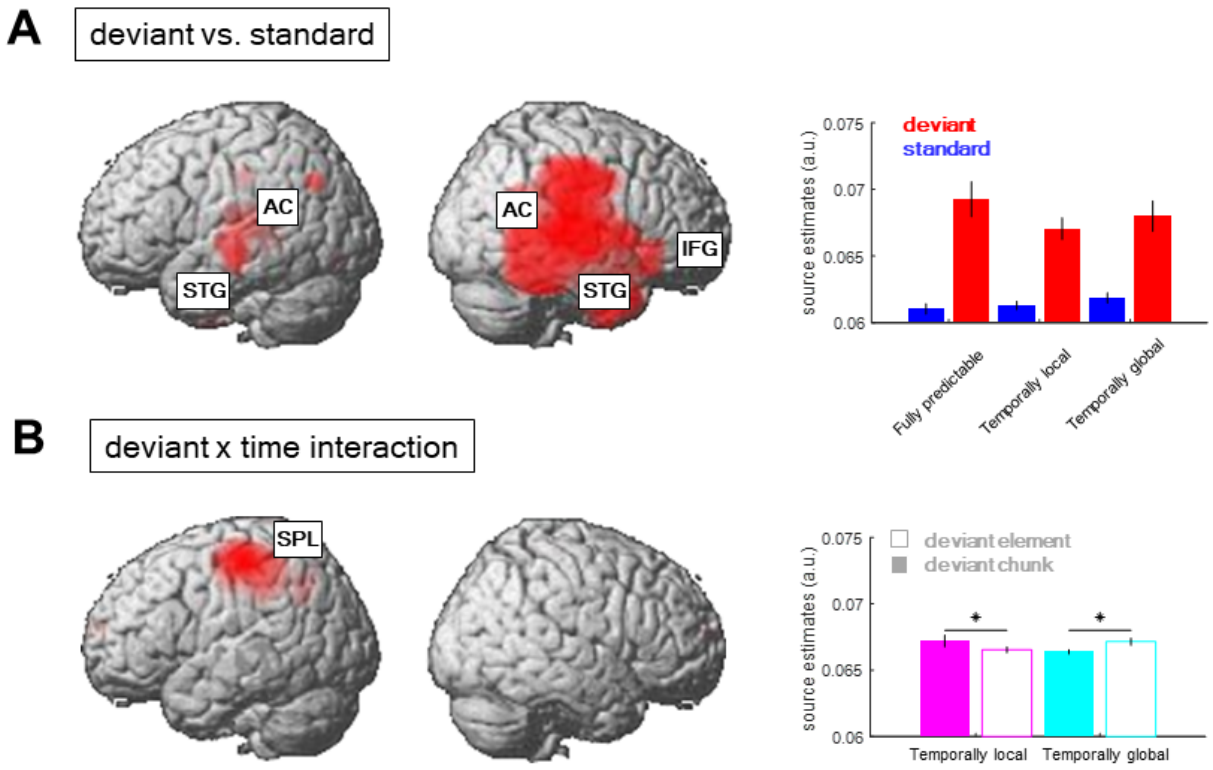

Figure 4.4 Source reconstruction. **(A)** Regions showing a significant main effect of content-based predictions (deviant vs. standard). Inset shows average source estimates per condition. Error bars denote SEM across participants. **(B)** Regions showing a significant contextual interaction effect between content-based predictions (deviant element vs. deviant chunk) and temporal predictions (temporally local vs. temporally global). Figure legend as in (A).

The difference between source estimates associated with deviants and standards was localized to a large network of regions (see Table 4.1 for full results), including bilateral auditory cortex (AC) and superior temporal gyri (STG) and the right inferior frontal gyrus (IFG). On the other hand, the interaction effect between deviant types and temporal conditions was localized to a spatially confined cluster in the left superior parietal lobule (SPL; see Table 4.1). A post-hoc analysis revealed that, in this cluster, element deviant responses presented in the temporally-local condition were associated with weaker source estimates than chunk deviant responses presented in the temporally-local condition (T_{max} = 3.67, Z_{max} = 3.46, p_{FWE} = 0.009, small-volume corrected). Similarly,

chunk deviant responses presented in the temporally-global condition were associated with weaker source estimates than element deviant responses presented in the temporally-global condition (T_{max} = 5.79, Z_{max} = 5.11, p_{FWE} = 0.003, small-volume corrected). Thus, while the deviant processing could be linked to a wide network of auditory and frontal regions, deviants presented in the corresponding temporal predictability conditions (e.g., element deviants in the temporally-local context) were associated with a relative decrease of left parietal activity.

Effect	**Cluster label**	**Peak MNI coords**	**F_{max}**	**Z_{max}**	**Voxel extent**	**p_{FWE}**
Deviant vs. standard	Right transverse temporal gyrus / auditory cortex (AC)	48 -20 12	53.99	4.83	2050 8	< 0.001
	Right superior temporal gyrus (STG)	44 -48 12	40.15	4.42		
	Right inferior frontal gyrus (IFG)	40 26 -6	34.52	4.20		
	Left transverse temporal gyrus / auditory cortex (AC)	-38 -28 12	34.31	4.20	2177	0.003
	Left superior temporal gyrus (STG)	-60 -20 -8	31.19	4.06		
("element" vs. "chunk" deviant) x (temporally-local vs. temporally-global)	Left superior parietal lobule (lSPL)	-26 -40 46	49.37	5.82	3073	0.003

Table 4.1. Source reconstruction results. Summary of significant clusters showing differences between conditions.

4.4.6 Dynamic causal modeling

To infer the most likely effective connectivity patterns underlying the observed ERP results, we used the six main cortical regions identified in the source reconstruction results as regions of interest (ROIs) to build a generative model of the ERP data. A fully interconnected model, fitted to each participants' ERP data, explained on average 71.03% of the ERP variance (SEM across participants 2.81%).

Bayesian model reduction was used to obtain connectivity and gain parameters of a range of reduced models, in which only a subset of parameters were allowed to be modulated by the two conditions (deviant vs. standard; interaction deviant element/chunk x temporally-local/global). Using this procedure, we identified a single winning model, in which "what" predictions (deviant vs. standard) modulated all types of connections (ascending, descending, lateral, and intrinsic), while the interaction between "what" and "when" predictions modulated three out of four types of connections (ascending, lateral, and intrinsic). The difference between the free-energy approximation to log-model evidence between the winning model and the next-best model (i.e., log Bayes factor) was 5.6615, corresponding to very strong evidence for the winning model (>99% probability). Therefore, the resulting Bayesian model average, implemented to integrate model parameter estimates from the entire model space, was mostly informed by the single winning model.

The posterior parameter estimates of the Bayesian model average are plotted in Figure 4.5A and reported in Table 4.2. The results revealed that deviant processing (as opposed to standard processing) significantly increased nearly all connectivity estimates (probability of increase >99.9% for all parameters), corresponding to an increase in

excitatory ascending connectivity and in inhibitory descending and intrinsic (gain) connectivity - with the exception of the intrinsic self-inhibition in the left SPL region, which was significantly decreased following deviant processing, as well as the bidirectional connectivity between the left SPL and right IFG, which was not affected by deviant processing.

The interaction between deviant type (deviant element vs. deviant chunk) and temporal predictability (temporally-global vs. temporally-local) modulated a more nuanced connectivity pattern. At the hierarchically lower level (between A1 and STG), deviants processed in a temporally congruent condition (i.e., deviant elements in the temporally-local condition, and deviant chunks in the temporally-global condition) decreased excitatory ascending connectivity from A1 to STG and inhibitory self-connectivity in A1. Conversely, at the hierarchically higher level (between STG and the fronto-parietal regions), deviants processed in a temporally congruent condition increased excitatory ascending connectivity from STG to SPL/IFC and inhibitory self-connectivity in the STG. Furthermore, deviants processed in a temporally congruent condition (1) increased lateral connectivity between the left and right STG, (2) decreased cross-hemispheric ascending connectivity between the STG regions and the fronto-parietal regions, and (3) increased self-inhibition in the left SPL region.

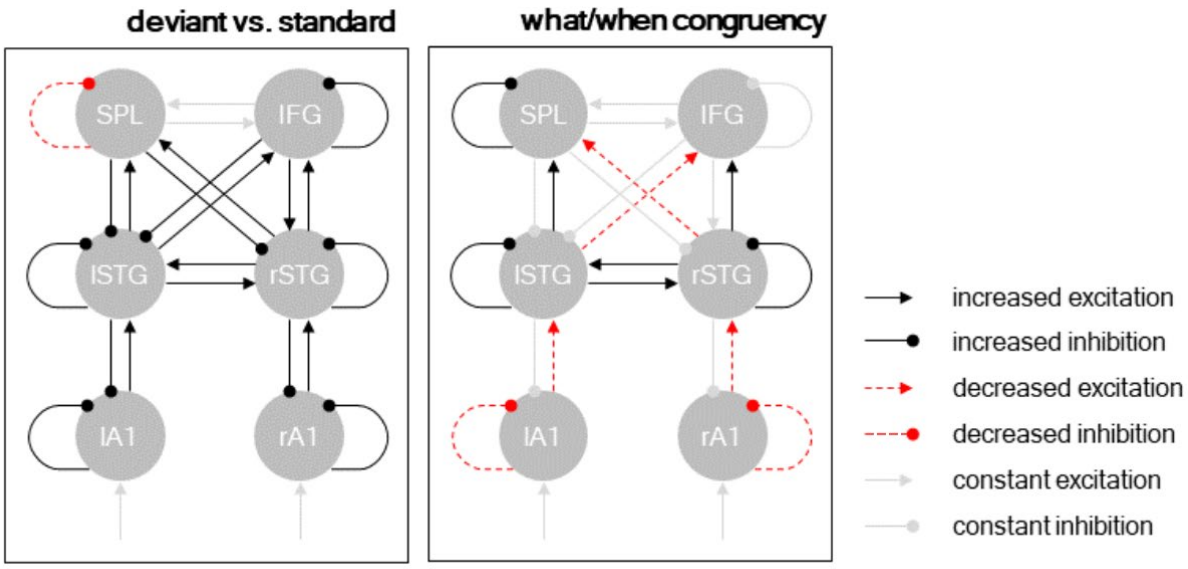

Figure 4.5 Dynamic causal modeling. Posterior model parameters. Separate panels show different condition-specific effects. Black arrows: excitatory connections; red arrows: inhibitory connections; solid lines: condition-specific increase; dashed lines: condition-specific decrease. Significant parameters ($p < 0.001$) shown in black/red, remaining connections (constant excitation/inhibition) shown in gray.

Connection type	**Connection label**	**Effect of content-based predictions (deviant vs. standard)**	**Effect of contextual interaction (congruent vs. incongruent)**
Intrinsic (gain)	lA1->lA1	50% self-inhibition increase	26% self-inhibition decrease
	rA1->rA1	9% self-inhibition increase	13% self-inhibition decrease
	lSTG->lSTG	44% self-inhibition increase	9% self-inhibition increase
	rSTG->rSTG	35% self-inhibition increase	33% self-inhibition increase
	lSPL->lSPL	42% self-inhibition decrease	12% self-inhibition increase
	rIFG->rIFG	11% self-inhibition increase	n.s.

Extrinsic (ascending)	lA1->lSTG	21% excitation increase	16% excitation decrease
	rA1->rSTG	13% excitation increase	10% excitation decrease
	lSTG->lSPL	33% excitation increase	20% excitation increase
	rSTG->rIFG	34% excitation increase	33% excitation increase
	lSTG->rIFG	41% excitation increase	13% excitation decrease
	rSTG->lSPL	34% excitation increase	10% excitation decrease
Extrinsic (descending)	lSTG->lA1	140% inhibition increase	n.s.
	rSTG->rA1	151% inhibition increase	n.s.
	lSPL->lSTG	111% inhibition increase	n.s.
	rIFG->rSTG	10% inhibition increase	n.s.
	rIFG->lSTG	24% inhibition increase	n.s.
	lSPL->rSTG	166% inhibition increase	n.s.
Extrinsic (lateral)	lSTG->rSTG	88% excitation increase	60% excitation increase
	rSTG->lSTG	13% excitation increase	20% excitation increase
	lSPL->rIFG	n.s.	n.s.
	rIFG->lSPL	n.s.	n.s.

Table 4.2. Dynamic causal modeling results. Summary of significant condition-specific effects on connectivity estimates.

4.5 Discussion

In the present study, we found that “when” predictions modulate MMR to violations of “what” predictions in a contextually specific fashion, such that more local “when” predictions modulated responses to single deviant elements, while more global “when” predictions modulated responses to deviant chunks, indicating a congruence effect in the processing of “what” and “when” predictions at different contextual levels in the auditory

system. The authors interpret this as levels of processing hierarchy, however it is worth noting that alternate interpretations of this effect could rely on simple contextual pairings such as position effects within the sequence. While “what” and “when” kinds of predictions showed interactive effects for both levels, both interaction effects (e.g. chunk-rate/deviant chunk and element-rate/deviant element) were associated with similar spatiotemporal patterns of EEG evoked activity modulations, and linked in the DCM analysis to a widespread connectivity increase at relatively late stages of cortical processing (between the STG and the fronto-parietal network). These findings suggest that the integration of “what” and “when” predictions, while contextually specific, is mediated by a shared and distributed cortical network.

Deviant responses to “what” prediction violations within melodic sequences and tone contours are well documented, having been used to explore a variety of phenomena in the auditory system (see Yu et al., 2015 for a partial review). Deviant tones within familiar musical scales have been found to elicit higher MMR amplitudes compared to those of unfamiliar scales are tones presented without a scale structure (Brattico et al., 2001), as well as higher deviant responses to out-of-scale notes in unfamiliar melodies (Brattico et al., 2006). Deviant responses to manipulated musical characteristics within melodic sequences (e.g. timing, pitch, transposition, melodic contour) have similarly been demonstrated in musician and non-musician groups (Tervaniemi et al., 2014; Vuust et al., 2011). In predictive coding frameworks, such evoked responses can be understood in the context of prediction error, wherein bottom-up error signaling triggers the adjustment of higher-level models of the stimulus train formed as a result of perceptual learning during repeated stimulus presentation (Garrido et al., 2009). Such hierarchical relationships

have been quantified using DCM (Auksztulewicz and Friston, 2016), and are consistent with our analysis of the evoked responses observed herein. The resultant model shows increased connectivity throughout the network, consistent with increased error signaling (ascending connections), predictive template updates (descending connections), and gain connectivity evident in a decrease in gain following predictions errors. Our source reconstruction was equally consistent with existing literature revealing bilateral activity in the primary auditory cortex (A1) and higher-order auditory regions in the superior temporal gyrus (STG), as well as the right inferior frontal gyrus (IFG) (Garrido et al., 2008; Giroud et al., 2020).

Turning to “when” predictions, the results of our frequency domain analysis show that the EEG spectrum largely follows that of the stimulus spectrum. However, ITPC peaks at the pair-tone rate of ‘fully predictable’ and ‘temporally local’ sequences are significantly larger than neighboring frequencies, which is not the case in the stimulus spectrum, indicating that ITPC peaks do not just follow the stimulus spectrum but also reflect the neural processing of sequence structures at higher levels (e.g. chunking (Kotz et al., 2018)) . Indeed, our behavioral results show faster reaction times in temporally predictable conditions, supporting the notion of neural entrainment to stimulus periodicity, results which mirror previous behavioral studies (Morillon et al., 2016). We found that the EEG-based ITPC response at tone-rate is stronger near central electrodes, with results consistent with existing EEG studies (Ding et al., 2017). Additionally, the chunk-rate effect is predominantly present in the right hemisphere, suggesting that the contextual structure of non-linguistic sequences can be entrained by parallel neural activity in different regions at distinct time scales - consistent with existing research (Giroud et al., 2020).

Interestingly, the ITPC differences between conditions (temporally-local vs. global) emerged during the experiment in chunk-rate peaks, but not in telement-rate peaks, suggesting that rapid learning could modulate neural entrainment to auditory sequences with different regularities at the chunk-rate level. Similarly, a previous study (Moser et al., 2021) found significant differences in non-linguistic triplet-rate ITPC peaks between structured and random conditions, occurring during early exposure. This ITPC difference suggests a fine shift in sequence encoding, with different regularities from single elements to integrated chunks. Notably, we also found correlations between the ITPC difference conditions and the congruence effect of ERP amplitude, indicating a mutual network between neural entrainment and prediction.

In addition to their dissociable main effects on neural activity, “what” and “when” predictions modulated element-evoked response amplitude interactively and in a contextually specific manner, such that faster “when” predictions amplified MMRs to less complex “what” prediction violations (single elements), while slower “when” predictions amplified MMRs to more complex “when” prediction violations (chunks). These findings extend the result of previous studies, which showed that “when” predictions modulate MMR amplitude (Jalewa et al., 2021; Lumaca et al., 2019; Takegata and Morotomi, 1999; Todd et al., 2018; Yabe et al., 1997), by showing that these modulatory effects are specific with respect to the complexity of “what” predictions. Dynamic causal modeling of our ERP data showed partially dissociable connectivity patterns between the main effect of "what" predictions (i.e., all deviants vs. all standards), which increased recurrent connectivity throughout the network (Auksztulewicz and Friston, 2015; Fitzgerald et al., 2021; Garrido et al., 2008), and "what"/"when" interactive effects, which had a more nuanced pattern of

effects on neural activity. Specifically, congruent “what” and “when” predictions decreased recurrent connectivity at lower parts of the network (between A1 and the STG), while at the same time increasing recurrent connectivity at higher parts of the network (between STG and the fronto-parietal regions). Previous DCM work has shown similar dissociations between processing deviants based on violations of relatively simple predictions vs. complex contextual information, indicating the higher-order regions as sensitive to complex prediction violations (Fitzgerald et al., 2021). Additionally, in the current results, the main effect of “what” predictions and the contextually specific integration of “what” and “when” predictions had opposing effects on the neural gain estimates for the left SPL region, which displayed decreased self-inhibition (increased gain) following deviant processing but increased self-inhibition (decreased gain) following prediction integration. These results mirror our source reconstruction, wherein deviants presented in congruent temporal conditions were associated with decreased left parietal activity, and imply the left parietal cortex - recently shown to mediate the integration of “what” and “when” information in speech processing (Orpella et al., 2020) - in the more general process of integrating “what” and “when” predictions also for non-speech stimuli. It is worth noting that while “when” predictions did not elicit a significant main effect on ERP amplitude, it is possible this finding may have resulted from design constraints, as all conditions contained only “what” (repetition detection) tasks, suggestive of previous studies on the role of attention in parallel temporal and mnemonic predictive processing (Lakatos et al., 2013; Wollman and Morillon, 2018).

Previous studies have shown that the processing of musical information requires predictive mechanisms for timing of content of auditory events, and that such predictions

can have modulatory effects at different cortical levels when presented within the framework of melodic expectation (Di Liberto et al., 2020; Royal et al., 2016). Musical stimuli presents us with an intriguing opportunity to investigate predictive coding mechanisms, as the statistical regularities within musical frameworks are well defined and intrinsically learned. In particular, such structures allow us to disassociate “what” and “when” predictions while keeping other elements of a stimulus stream intact across manipulations and trials. Studies have demonstrated an early right anterior negativity (ERAN) in contexts where musical syntax has been violated, as opposed to the comparatively low-level acoustic diavations that elicit a MMN response (see Koelsch et al., 2019 for review). Because the presence of musical syntax violations require knowledge acquired through long-term repeated exposure to music, long-term memory recall is involved in establishing those regularities. The role of memory in syntactical prediction violation is an avenue ripe for further investigation, and future studies may wish to extend our paradigm to further probe the observed late-series ITPC pair-rate differences in that context. Furthermore, since “what” and “when” predictions are also ubiquitous in other stimulus domains - most prominently in speech perception - future research should test whether similar contextual specificity of “what” and “when” predictions as observed here also governs speech processing.

Chapter 5. Summary and Conclusions

1. The results of Chapter 2 show that in both awake humans monitored with EEG and anesthetized rats fitted with ECoG arrays, the acoustic frequency of recent

tokens could be decoded from neural activity evoked by pure tones as well as that evoked by neutral frozen noise burst stimuli presented during silent-state auditory sensory memory retention. This finding demonstrates that memory contents can be decoded in different species and different states using homologous methods, suggesting that the mechanisms of sensory memory encoding are evolutionarily conserved across species.

2. The results of Chapter 3 show that mnemonic and predictive representations of auditory stimuli can be simultaneously decoded from neural activity measured under passive listening in anesthetized rats. Memory and prediction decoding is observed at overlapping latencies, but based on largely uncorrelated data features, suggesting partly dissociable underlying mechanisms. Predictive representations are dynamically updated over the course of stimulation, suggesting a gradual formation of prediction, even under anesthesia.

3. The results of Chapter 4 show that temporal predictions interactively modulate neural activity evoked by mispredicted stimulus contents in a contextually congruent manner, such that local (vs. global) time-based predictions modulated content-based predictions of sequence elements (vs. chunks). These modulations were shared between contextual levels in terms of the spatiotemporal distribution of neural activity, suggesting that the brain integrates different predictions with a high degree of contextual specificity, but in a shared and distributed cortical network.

The above experiments investigated the neural mechanisms responsible for the brain's ability to process complex sequences of auditory sensory information, allowing it to learn statistical regularities, form predictive models based on those regularities, and encode mnemonic representations of sensory events used in storage and error correction. Future research could expand on the findings outlined above in 1. by applying similar methods in different attentional/conscious states (e.g. asleep or unconscious humans and awake rats), or refine the stimulus paradigms to further differentiate between passive sensory memory and active working memory processes across species. As the results discussed in 2. presented streams of stimulus tokens that were interrupted by bursts of white noise, researchers of the auditory "filling-in" phenomenon may find an analog in our paradigm, as modifications to the paradigm could provide an interesting platform to explore physically masked but perceptually and behaviorally perceived stimulus tokens.

As the findings discussed in 2. also demonstrate a correlation between statistical learning and relative decoding strength of predictable tokens within a sequence, this technique could potentially be employed to probe the effect of such models on syntactic predictability during the presentation of naturalistic stimuli such as music and language. Indeed, the frameworks proposed by Dehaene, et. al (2015) offer an interesting lens through which these results might be interpreted. To briefly recap, the framework proposes a taxonomy five underlying neural mechanisms corresponding to sequence processing at five increasing layers of abstraction. The first type of sequence processing relies on timing and transition information to parse steams based on temporal regularities, while in the second layer chunking occurs based also on the content of the stream, where

several tokens are grouped into a single (e.g. a “chunk”). The third layer of abstraction comes in the form of ordinal knowledge, where the relative position of tokens within a sequence are processed. In the fourth layer, so-called “algebraic patterns” are processed, wherein relationships across chunked tokens (e.g. AAB and XXY containing patterns of two identical items followed by a third item which is different). Finally, in the fifth layer, “nested tree structures” rely on symbolic rules such grammar and semantic meaning.

In the context of this framework, my use of streamed triplets the stimulus design of 3. can allow for several types of neural representations at several of the proposed hierarchical levels. As all stimulus tokens were presented in repeated triplet streams, they would at the most basic level require chunking into those triplets. In order for the brain to form the predictions that were observed in my decoding analysis, algebraic patterns would also need to be extracted, as the form of each triplet would necessitate the processing of a given triplet form. Although our experiment was not designed to answer this question, it would be interesting in a follow-up analysis to investigate if different algebraic patterns result in different learning effects (e.g. an XXY pattern as opposed to a XYZ pattern). However, this may be difficult in my design as there was no control for how the triplets might be chunked, given they were presented in streams with a fixed ISI – a XXY triplet may well be chunked as XYX or YXX. Similarly, ordinal knowledge would be difficult to quantify, given that relative position of chunked elements may be in an ambiguous order, particularly after omissions/bursts when the stream of triplets would resume at an arbitrary starting point. This line of inquiry seems particularly well-suited to human experimentation where studies might be designed specifically to establish behavioral metrics into how such streams and predictions are processed within the

framework proposed by Dehaene et al (2015). The observed topographical differences in mnemonic and predictive decodability of predictable tokens may also lend important insights into the segregation of prediction-specific neuronal populations within the same hierarchically-organized cortical pathways. Extensions of 2. are currently being drafted to employ similar paradigms in human fMRI to enhance our understanding of topographical specificity of mnemonic and predictive representations and single-unit measurements in rodent prefrontal and sensory cortices to further investigate the presence of laminar-based hierarchies implied by the predictive coding framework.

The findings outlined in 3. correspond to “what” and “when” predictions that are also ubiquitous in speech perception. As with results presented in 2. (above), the Dehaene framework could be applied to several aspects of my findings, which may provide fertile ground for extensions and alternate interpretations. For example, in my interpretation, all stages of the proposed framework are engaged by the stimuli, owing to the nature of its design. On the simplest layer of the taxonomy, each tone within the sequence is either presented at a fixed timing interval or has that interval manipulated, allowing me to probe transition and timing knowledge. The loud-soft tones further become chunked into pairs, while ordinal knowledge is implicit in the understanding of scale contour, and algebraic patterns would be present in the repeated pattern of 7 tones which make up either an ascending or descending scale. Finally, nested tree structures are present in my interpretation of error detection on global scales, as an understanding of the scale as a whole would be required in order for experimental manipulations at the penultimate tone to elicit a global “what” error when the trajectory of the scale has been disrupted (e.g. “this is a full/correct scale” in the context of nested structures as opposed

to “a scale consists of 7 tones where each tone is high/lower than the next” in the context algebraic patterns). An alternate interpretation of what I have categorized as “hierarchal” effects could be one of ordinal effects – the observed effect may be a result of errors that always occur at a given position within scales, as global and local “what” manipulations always occur at either the final or penultimate position within a scale. Although my analyses imply this is not the case, further analysis or a refinement of experimental design may be needed to fully discount this alternate interpretation. As my paradigm was designed to map onto layers of hierarchy in language processing (e.g. chunked pairs and scales are analogous to words and sentences), future research lines could investigate similar phenomena in language processing, and explore the extent to which contextual specificity of “what” and “when” predictions also govern speech processing.

Chapter 6. References

Alaerts, J., Luts, H., Hofmann, M., Wouters, J., 2009. Cortical auditory steady-state responses to low modulation rates. Int. J. Audiol. 48, 582–593. https://doi.org/10.1080/14992020902894558

Alain, C., Woods, D.L., Knight, R.T., 1998. A distributed cortical network for auditory sensory memory in humans. Brain Res. 812, 23–37. https://doi.org/10.1016/S0006-8993(98)00851-8

Arnal, L.H., Giraud, A.-L., 2012. Cortical oscillations and sensory predictions. Trends Cogn. Sci. 16, 390–398. https://doi.org/10.1016/j.tics.2012.05.003

Astikainen, P., Stefanics, G., Nokia, M., Lipponen, A., Cong, F., Penttonen, M., Ruusuvirta, T., 2011. Memory-Based Mismatch Response to Frequency Changes in Rats. PLoS ONE 6. https://doi.org/10.1371/journal.pone.0024208

Auksztulewicz, R., Friston, K., 2016. Repetition suppression and its contextual determinants in predictive coding. Cortex, Special Issue:Repetition suppression-an integrative view 80, 125–140. https://doi.org/10.1016/j.cortex.2015.11.024

Auksztulewicz, R., Friston, K., 2015. Attentional Enhancement of Auditory Mismatch Responses: a DCM/MEG Study. Cereb. Cortex 25, 4273–4283. https://doi.org/10.1093/cercor/bhu323

Auksztulewicz, R., Myers, N.E., Schnupp, J.W., Nobre, A.C., 2019. Rhythmic Temporal Expectation Boosts Neural Activity by Increasing Neural Gain. J. Neurosci. 39, 9806–9817. https://doi.org/10.1523/JNEUROSCI.0925-19.2019

Auksztulewicz, R., Schwiedrzik, C.M., Thesen, T., Doyle, W., Devinsky, O., Nobre, A.C., Schroeder, C.E., Friston, K.J., Melloni, L., 2018. Not All Predictions Are Equal: “What” and “When” Predictions Modulate Activity in Auditory Cortex through Different

Mechanisms. J. Neurosci. 38, 8680–8693. https://doi.org/10.1523/JNEUROSCI.0369-18.2018
Ball, T., Kern, M., Mutschler, I., Aertsen, A., Schulze-Bonhage, A., 2009. Signal quality of simultaneously recorded invasive and non-invasive EEG. NeuroImage 46, 708–716. https://doi.org/10.1016/j.neuroimage.2009.02.028
Barron, H.C., Auksztulewicz, R., Friston, K., 2020. Prediction and memory: A predictive coding account. Prog. Neurobiol. 192, 101821. https://doi.org/10.1016/j.pneurobio.2020.101821
Bastos, A.M., Usrey, W.M., Adams, R.A., Mangun, G.R., Fries, P., Friston, K.J., 2012. Canonical microcircuits for predictive coding. Neuron 76, 695–711. https://doi.org/10.1016/j.neuron.2012.10.038
Baumgarten, T.J., Maniscalco, B., Lee, J.L., Flounders, M.W., Abry, P., He, B.J., 2021. Neural integration underlying naturalistic prediction flexibly adapts to varying sensory input rate. Nat. Commun. 12, 2643. https://doi.org/10.1038/s41467-021-22632-z
Belardinelli, P., Ortiz, E., Braun, C., 2012. Source Activity Correlation Effects on LCMV Beamformers in a Realistic Measurement Environment. Comput. Math. Methods Med. 2012, e190513. https://doi.org/10.1155/2012/190513
Bellmund, J.L.S., Polti, I., Doeller, C.F., 2020. Sequence Memory in the Hippocampal-Entorhinal Region. J. Cogn. Neurosci. 32, 2056–2070. https://doi.org/10.1162/jocn_a_01592
Benjamini, Y., Hochberg, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B Methodol. 57, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x
Bianco, R., Harrison, P.M., Hu, M., Bolger, C., Picken, S., Pearce, M.T., Chait, M., 2020. Long-term implicit memory for sequential auditory patterns in humans. eLife 9, e56073. https://doi.org/10.7554/eLife.56073
Bigelow, J., Rossi, B., Poremba, A., 2014. Neural correlates of short-term memory in primate auditory cortex. Front. Neurosci. 8. https://doi.org/10.3389/fnins.2014.00250
Bobadilla-Suarez, S., Ahlheim, C., Mehrotra, A., Panos, A., Love, B., 2019. Measures of Neural Similarity. Comput. Brain Behav. https://doi.org/10.1007/s42113-019-00068-5
Brattico, E., Näääätäänen, R., Tervaniemi, M., 2001. Context Effects on Pitch Perception in Musicians and Nonmusicians: Evidence from Event-Related-Potential Recordings. Music Percept. 19, 199–222. https://doi.org/10.1525/mp.2001.19.2.199
Brattico, E., Tervaniemi, M., Näätänen, R., Peretz, I., 2006. Musical scale properties are automatically processed in the human auditory cortex. Brain Res. 1117, 162–174. https://doi.org/10.1016/j.brainres.2006.08.023
Britton, B., Blumstein, S.E., Myers, E.B., Grindrod, C., 2009. The role of spectral and durational properties on hemispheric asymmetries in vowel perception. Neuropsychologia 47, 1096–1106. https://doi.org/10.1016/j.neuropsychologia.2008.12.033
Cappotto, D., Auksztulewicz, R., Kang, H., Poeppel, D., Melloni, L., Schnupp, J., 2021. Decoding the Content of Auditory Sensory Memory Across Species. Cereb. Cortex N. Y. N 1991. https://doi.org/10.1093/cercor/bhab002
Capsius, B., Leppelsack, H.-J., 1996. Influence of urethane anesthesia on neural processing in the auditory cortex analogue of a songbird. Hear. Res. 96, 59–70. https://doi.org/10.1016/0378-5955(96)00038-X
Carbajal, G.V., Malmierca, M.S., 2018. The Neuronal Basis of Predictive Coding Along the Auditory Pathway: From the Subcortical Roots to Cortical Deviance Detection. Trends Hear. 22, 2331216518784822. https://doi.org/10.1177/2331216518784822
Casado-Román, L., Carbajal, G.V., Pérez-González, D., Malmierca, M.S., 2020. Prediction error signaling explains neuronal mismatch responses in the medial prefrontal cortex. PLOS Biol. 18, e3001019. https://doi.org/10.1371/journal.pbio.3001019

Cheung, S.W., Nagarajan, S.S., Bedenbaugh, P.H., Schreiner, C.E., Wang, X., Wong, A., 2001. Auditory cortical neuron response differences under isoflurane versus pentobarbital anesthesia. Hear. Res. 156, 115–127. https://doi.org/10.1016/S0378-5955(01)00272-6
Constantinidis, C., Procyk, E., 2004. The primate working memory networks. Cogn. Affect. Behav. Neurosci. 4, 444–465.
Costa-Faidella, J., Grimm, S., Slabu, L., Díaz-Santaella, F., Escera, C., 2011. Multiple time scales of adaptation in the auditory system as revealed by human evoked potentials: Adaptation in the human auditory system. Psychophysiology 48, 774–783. https://doi.org/10.1111/j.1469-8986.2010.01144.x
Daniel, T.A., Katz, J.S., Robinson, J.L., 2016. Delayed match-to-sample in working memory: A BrainMap meta-analysis. Biol. Psychol. 120, 10–20. https://doi.org/10.1016/j.biopsycho.2016.07.015
de Cheveigné, A., Parra, L.C., 2014. Joint decorrelation, a versatile tool for multichannel data analysis. NeuroImage 98, 487–505. https://doi.org/10.1016/j.neuroimage.2014.05.068
de Cheveigné, A., Simon, J.Z., 2008. Denoising based on spatial filtering. J. Neurosci. Methods 171, 331–339. https://doi.org/10.1016/j.jneumeth.2008.03.015
De Maesschalck, R., Jouan-Rimbaud, D., Massart, D.L., 2000. The Mahalanobis distance. Chemom. Intell. Lab. Syst. 50, 1–18. https://doi.org/10.1016/S0169-7439(99)00047-7
Dehaene, S., Meyniel, F., Wacongne, C., Wang, L., Pallier, C., 2015. The Neural Representation of Sequences: From Transition Probabilities to Algebraic Patterns and Linguistic Trees. Neuron 88, 2–19. https://doi.org/10.1016/j.neuron.2015.09.019
Denham, S.L., Winkler, I., 2020. Predictive coding in auditory perception: challenges and unresolved questions. Eur. J. Neurosci. 51, 1151–1160. https://doi.org/10.1111/ejn.13802
Di Liberto, G.M., Pelofi, C., Bianco, R., Patel, P., Mehta, A.D., Herrero, J.L., de Cheveigné, A., Shamma, S., Mesgarani, N., 2020. Cortical encoding of melodic expectations in human temporal cortex. eLife 9, e51784. https://doi.org/10.7554/eLife.51784
Ding, N., Melloni, L., Yang, A., Wang, Y., Zhang, W., Poeppel, D., 2017. Characterizing Neural Entrainment to Hierarchical Linguistic Units using Electroencephalography (EEG). Front. Hum. Neurosci. 11, 481. https://doi.org/10.3389/fnhum.2017.00481
Ding, N., Melloni, L., Zhang, H., Tian, X., Poeppel, D., 2016. Cortical tracking of hierarchical linguistic structures in connected speech. Nat. Neurosci. 19, 158–164. https://doi.org/10.1038/nn.4186
Ding, N., Simon, J.Z., 2013. Power and phase properties of oscillatory neural responses in the presence of background activity. J. Comput. Neurosci. 34, 337–343. https://doi.org/10.1007/s10827-012-0424-6
Doelling, K.B., Assaneo, M.F., 2021. Neural oscillations are a start toward understanding brain activity rather than the end. PLOS Biol. 19, e3001234. https://doi.org/10.1371/journal.pbio.3001234
Fairhall, A.L., Lewen, G.D., Bialek, W., de Ruyter Van Steveninck, R.R., 2001. Efficiency and ambiguity in an adaptive neural code. Nature 412, 787–792. https://doi.org/10.1038/35090500
Fitzgerald, K., Auksztulewicz, R., Provost, A., Paton, B., Howard, Z., Todd, J., 2021. Hierarchical Learning of Statistical Regularities over Multiple Timescales of Sound Sequence Processing: A Dynamic Causal Modeling Study. J. Cogn. Neurosci. 33, 1549–1562. https://doi.org/10.1162/jocn_a_01735
Fries, P., 2005. A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. Trends Cogn. Sci. 9, 474–480. https://doi.org/10.1016/j.tics.2005.08.011
Friston, K., 2005. A theory of cortical responses. Philos. Trans. R. Soc. B Biol. Sci. 360, 815–836. https://doi.org/10.1098/rstb.2005.1622

Friston, K., Buzsáki, G., 2016. The Functional Anatomy of Time: What and When in the Brain. Trends Cogn. Sci. 20, 500–511. https://doi.org/10.1016/j.tics.2016.05.001
Friston, K., Kilner, J., Harrison, L., 2006. A free energy principle for the brain. J. Physiol. Paris 100, 70–87. https://doi.org/10.1016/j.jphysparis.2006.10.001
Friston, K., Penny, W., 2011. Post hoc Bayesian model selection. NeuroImage 56, 2089–2099. https://doi.org/10.1016/j.neuroimage.2011.03.062
Gaese, B.H., Ostwald, J., 2001. Anesthesia Changes Frequency Tuning of Neurons in the Rat Primary Auditory Cortex. J. Neurophysiol. 86, 1062–1066. https://doi.org/10.1152/jn.2001.86.2.1062
Garrido, M.I., Friston, K.J., Kiebel, S.J., Stephan, K.E., Baldeweg, T., Kilner, J.M., 2008. The functional anatomy of the MMN: A DCM study of the roving paradigm. NeuroImage 42, 936–944. https://doi.org/10.1016/j.neuroimage.2008.05.018
Garrido, M.I., Kilner, J.M., Stephan, K.E., Friston, K.J., 2009. The mismatch negativity: A review of underlying mechanisms. Clin. Neurophysiol. 120, 453–463. https://doi.org/10.1016/j.clinph.2008.11.029
Gavornik, J.P., Bear, M.F., 2014. Learned spatiotemporal sequence recognition and prediction in primary visual cortex. Nat. Neurosci. 17, 732–737. https://doi.org/10.1038/nn.3683
Giroud, J., Trébuchon, A., Schön, D., Marquis, P., Liegeois-Chauvel, C., Poeppel, D., Morillon, B., 2020. Asymmetric sampling in human auditory cortex reveals spectral processing hierarchy. PLOS Biol. 18, e3000207. https://doi.org/10.1371/journal.pbio.3000207
Grootswagers, T., Wardle, S.G., Carlson, T.A., 2017. Decoding Dynamic Brain Patterns from Evoked Responses: A Tutorial on Multivariate Pattern Analysis Applied to Time Series Neuroimaging Data. J. Cogn. Neurosci. 29, 677–697. https://doi.org/10.1162/jocn_a_01068
Haegens, S., Zion Golumbic, E., 2018. Rhythmic facilitation of sensory processing: A critical review. Neurosci. Biobehav. Rev. 86, 150–165. https://doi.org/10.1016/j.neubiorev.2017.12.002
Heilbron, M., Chait, M., 2018. Great Expectations: Is there Evidence for Predictive Coding in Auditory Cortex? Neuroscience, Sensory Sequence Processing in the Brain 389, 54–73. https://doi.org/10.1016/j.neuroscience.2017.07.061
Henin, S., Turk-Browne, N.B., Friedman, D., Liu, A., Dugan, P., Flinker, A., Doyle, W., Devinsky, O., Melloni, L., 2021. Learning hierarchical sequence representations across human cortex and hippocampus. Sci. Adv. 7, eabc4530. https://doi.org/10.1126/sciadv.abc4530
HiJee, K., Ryszard, A., Hong, C.C., Drew, C., Gurusamy, R.V., Hendrik, S.J.W., 2021. Memory Transfer of Random Time Patterns Across Modalities. bioRxiv 2020.11.24.395368. https://doi.org/10.1101/2020.11.24.395368
Hsu, Y.-F., Hämäläinen, J.A., Waszak, F., 2013. Temporal expectation and spectral expectation operate in distinct fashion on neuronal populations. Neuropsychologia 51, 2548–2555. https://doi.org/10.1016/j.neuropsychologia.2013.09.018
Huang, Y., Matysiak, A., Heil, P., König, R., Brosch, M., 2016. Persistent neural activity in auditory cortex is related to auditory working memory in humans and nonhuman primates. eLife 5. https://doi.org/10.7554/eLife.15441
Ille, N., Berg, P., Scherg, M., 2002. Artifact correction of the ongoing EEG using spatial filters based on artifact and brain signal topographies. J. Clin. Neurophysiol. Off. Publ. Am. Electroencephalogr. Soc. 19, 113–124. https://doi.org/10.1097/00004691-200203000-00002
Jalewa, J., Todd, J., Michie, P.T., Hodgson, D.M., Harms, L., 2021. Do rat auditory event related potentials exhibit human mismatch negativity attributes related to predictive coding? Hear. Res., Stimulus-specific adaptation, MMN and predicting coding 399, 107992. https://doi.org/10.1016/j.heares.2020.107992

Kamiński, J., Rutishauser, U., 2019. Between persistently active and activity-silent frameworks: novel vistas on the cellular basis of working memory. Ann. N. Y. Acad. Sci. https://doi.org/10.1111/nyas.14213
Kawahara, H., 2006. STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. Acoust Sci Technol 27349. https://doi.org/10.1250/ast.27.349
Kilner, J.M., Kiebel, S.J., Friston, K.J., 2005. Applications of random field theory to electrophysiology. Neurosci. Lett. 374, 174–178. https://doi.org/10.1016/j.neulet.2004.10.052
Koelsch, S., Vuust, P., Friston, K., 2019. Predictive Processes and the Peculiar Case of Music. Trends Cogn. Sci. 23, 63–77. https://doi.org/10.1016/j.tics.2018.10.006
Kotz, S.A., Ravignani, A., Fitch, W.T., 2018. The Evolution of Rhythm Processing. Trends Cogn. Sci., Special Issue: Time in the Brain 22, 896–910. https://doi.org/10.1016/j.tics.2018.08.002
Kotz, S.A., Schwartze, M., 2010. Cortical speech processing unplugged: a timely subcortico-cortical framework. Trends Cogn. Sci. 14, 392–399. https://doi.org/10.1016/j.tics.2010.06.005
Kriegeskorte, N., Mur, M., Bandettini, P., 2008. Representational similarity analysis - connecting the branches of systems neuroscience. Front. Syst. Neurosci. 2.
Lakatos, P., Musacchia, G., O'Connel, M.N., Falchier, A.Y., Javitt, D.C., Schroeder, C.E., 2013. The Spectrotemporal Filter Mechanism of Auditory Selective Attention. Neuron 77, 750–761. https://doi.org/10.1016/j.neuron.2012.11.034
Ledoit, O., Wolf, M., 2004. A well-conditioned estimator for large-dimensional covariance matrices. J. Multivar. Anal. 88, 365–411. https://doi.org/10.1016/S0047-259X(03)00096-4
Libby, A., Buschman, T.J., 2021. Rotational dynamics reduce interference between sensory and memory representations. Nat. Neurosci. 1–12. https://doi.org/10.1038/s41593-021-00821-9
Little, S., Bonaiuto, J., Meyer, S.S., Lopez, J., Bestmann, S., Barnes, G., 2018. Quantifying the performance of MEG source reconstruction using resting state data. NeuroImage 181, 453–460. https://doi.org/10.1016/j.neuroimage.2018.07.030
Litvak, V., Friston, K., 2008. Electromagnetic source reconstruction for group studies. NeuroImage 42, 1490–1498. https://doi.org/10.1016/j.neuroimage.2008.06.022
Luft, C.D.B., Meeson, A., Welchman, A.E., Kourtzi, Z., 2015. Decoding the future from past experience: learning shapes predictions in early visual cortex. J. Neurophysiol. 113, 3159–3171. https://doi.org/10.1152/jn.00753.2014
Lumaca, M., Trusbak Haumann, N., Brattico, E., Grube, M., Vuust, P., 2019. Weighting of neural prediction error by rhythmic complexity: A predictive coding account using mismatch negativity. Eur. J. Neurosci. 49, 1597–1609. https://doi.org/10.1111/ejn.14329
Luo, D., Li, K., An, H., Schnupp, J.W., Auksztulewicz, R., 2021. Learning boosts the decoding of sound sequences in rat auditory cortex. Curr. Res. Neurobiol. 2, 100019. https://doi.org/10.1016/j.crneur.2021.100019
Mahajan, Y., Peter, V., Sharma, M., 2017. Effect of EEG Referencing Methods on Auditory Mismatch Negativity. Front. Neurosci. 11, 560. https://doi.org/10.3389/fnins.2017.00560
Malmierca, M.S., Niño-Aguillón, B.E., Nieto-Diego, J., Porteros, Á., Pérez-González, D., Escera, C., 2019. Pattern-sensitive neurons reveal encoding of complex auditory regularities in the rat inferior colliculus. NeuroImage 184, 889–900. https://doi.org/10.1016/j.neuroimage.2018.10.012
Mishra, J., Gazzaley, A., 2016. Cross-species Approaches to Cognitive Neuroplasticity Research. NeuroImage 131, 4–12. https://doi.org/10.1016/j.neuroimage.2015.09.002

Mongillo, G., Barak, O., Tsodyks, M., 2008. Synaptic theory of working memory. Science 319, 1543–1546. https://doi.org/10.1126/science.1150769
Morillon, B., Schroeder, C.E., Wyart, V., Arnal, L.H., 2016. Temporal Prediction in lieu of Periodic Stimulation. J. Neurosci. 36, 2342–2347. https://doi.org/10.1523/JNEUROSCI.0836-15.2016
Morlet, D., Fischer, C., 2014. MMN and Novelty P3 in Coma and Other Altered States of Consciousness: A Review. Brain Topogr. 27, 467–479. https://doi.org/10.1007/s10548-013-0335-5
Murray, J.D., Bernacchia, A., Roy, N.A., Constantinidis, C., Romo, R., Wang, X.-J., 2017. Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. Proc. Natl. Acad. Sci. 114, 394–399. https://doi.org/10.1073/pnas.1619449114
Musacchia, G., Large, E., Schroeder, C.E., 2014. Thalamocortical mechanisms for integrating musical tone and rhythm. Hear. Res. 308, 50–59. https://doi.org/10.1016/j.heares.2013.09.017
Myers, N.E., Rohenkohl, G., Wyart, V., Woolrich, M.W., Nobre, A.C., Stokes, M.G., 2015. Testing sensory evidence against mnemonic templates. eLife 4, e09000. https://doi.org/10.7554/eLife.09000
Näätänen, R., Jacobsen, T., Winkler, I., 2005. Memory-based or afferent processes in mismatch negativity (MMN): a review of the evidence. Psychophysiology 42, 25–32. https://doi.org/10.1111/j.1469-8986.2005.00256.x
Nakamura, T., Michie, P.T., Fulham, W.R., Todd, J., Budd, T.W., Schall, U., Hunter, M., Hodgson, D.M., 2011. Epidural Auditory Event-Related Potentials in the Rat to Frequency and duration Deviants: Evidence of Mismatch Negativity? Front. Psychol. 2. https://doi.org/10.3389/fpsyg.2011.00367
Natan, R.G., Rao, W., Geffen, M.N., 2017. Cortical Interneurons Differentially Shape Frequency Tuning following Adaptation. Cell Rep. 21, 878–890. https://doi.org/10.1016/j.celrep.2017.10.012
Nees, M.A., 2016. Have We Forgotten Auditory Sensory Memory? Retention Intervals in Studies of Nonverbal Auditory Working Memory. Front. Psychol. 7. https://doi.org/10.3389/fpsyg.2016.01892
Nemrodov, D., Niemeier, M., Patel, A., Nestor, A., 2018. The Neural Dynamics of Facial Identity Processing: Insights from EEG-Based Pattern Analysis and Image Reconstruction. eNeuro 5. https://doi.org/10.1523/ENEURO.0358-17.2018
Nieto-Diego, J., Malmierca, M.S., 2016. Topographic Distribution of Stimulus-Specific Adaptation across Auditory Cortical Fields in the Anesthetized Rat. PLoS Biol. 14, e1002397. https://doi.org/10.1371/journal.pbio.1002397
O'Connor, K., Ison, J.R., 1991. Echoic memory in the rat: effects of inspection time, retention interval, and the spectral composition of masking noise. J. Exp. Psychol. Anim. Behav. Process. 17, 377–385. https://doi.org/10.1037//0097-7403.17.4.377
Orpella, J., Ripollés, P., Ruzzoli, M., Amengual, J.L., Callejas, A., Martinez-Alvarez, A., Soto-Faraco, S., de Diego-Balaguer, R., 2020. Integrating when and what information in the left parietal lobe allows language rule generalization. PLoS Biol. 18, e3000895. https://doi.org/10.1371/journal.pbio.3000895
Parras, G.G., Nieto-Diego, J., Carbajal, G.V., Valdés-Baizabal, C., Escera, C., Malmierca, M.S., 2017. Neurons along the auditory pathway exhibit a hierarchical organization of prediction error. Nat. Commun. 8, 2148. https://doi.org/10.1038/s41467-017-02038-6
Pasternak, T., Greenlee, M.W., 2005. Working memory in primate sensory systems. Nat. Rev. Neurosci. 6, 97–107. https://doi.org/10.1038/nrn1603

Poeppel, D., 2003. The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time.' Speech Commun., The Nature of Speech Perception 41, 245–255. https://doi.org/10.1016/S0167-6393(02)00107-3
Polley, D. B., Read, H. L., Storace, D. A., & Merzenich, M. M. (2007). Multiparametric auditory receptive field organization across five cortical fields in the albino rat. Journal of neurophysiology, 97(5), 3621-3638., n.d.
Rosch, R.E., Auksztulewicz, R., Leung, P.D., Friston, K.J., Baldeweg, T., 2019. Selective Prefrontal Disinhibition in a Roving Auditory Oddball Paradigm Under N-Methyl-D-Aspartate Receptor Blockade. Biol. Psychiatry Cogn. Neurosci. Neuroimaging 4, 140–150. https://doi.org/10.1016/j.bpsc.2018.07.003
Royal, I., Vuvan, D.T., Zendel, B.R., Robitaille, N., Schönwiesner, M., Peretz, I., 2016. Activation in the Right Inferior Parietal Lobule Reflects the Representation of Musical Structure beyond Simple Pitch Discrimination. PLoS ONE 11, e0155291. https://doi.org/10.1371/journal.pone.0155291
Rubin, J., Ulanovsky, N., Nelken, I., Tishby, N., 2016. The Representation of Prediction Error in Auditory Cortex. PLOS Comput. Biol. 12, e1005058. https://doi.org/10.1371/journal.pcbi.1005058
Rust, N.C., Palmer, S.E., 2021. Remembering the Past to See the Future. Annu. Rev. Vis. Sci. 7, 349–365. https://doi.org/10.1146/annurev-vision-093019-112249
Ruusuvirta, T., Penttonen, M., Korhonen, T., 1998. Auditory cortical event-related potentials to pitch deviances in rats. Neurosci. Lett. 248, 45–48. https://doi.org/10.1016/S0304-3940(98)00330-9
Salisbury, D.F., 2012. Finding the missing stimulus mismatch negativity (MMN): emitted MMN to violations of an auditory gestalt. Psychophysiology 49, 544–548. https://doi.org/10.1111/j.1469-8986.2011.01336.x
Schönwiesner, M., Rübsamen, R., von Cramon, D.Y., 2005. Spectral and temporal processing in the human auditory cortex--revisited. Ann. N. Y. Acad. Sci. 1060, 89–92. https://doi.org/10.1196/annals.1360.051
Schroeder, C.E., Lakatos, P., 2009. Low-frequency neuronal oscillations as instruments of sensory selection. Trends Neurosci. 32, 9–18. https://doi.org/10.1016/j.tins.2008.09.012
Schröger, E., Bendixen, A., Denham, S.L., Mill, R.W., Bőhm, T.M., Winkler, I., 2014. Predictive Regularity Representations in Violation Detection and Auditory Stream Segregation: From Conceptual to Computational Models. Brain Topogr. 27, 565–577. https://doi.org/10.1007/s10548-013-0334-6
Schumacher, J.W., Schneider, D.M., Woolley, S.M.N., 2011. Anesthetic state modulates excitability but not spectral tuning or neural discrimination in single auditory midbrain neurons. J. Neurophysiol. 106, 500–514. https://doi.org/10.1152/jn.01072.2010
Spaak, E., Watanabe, K., Funahashi, S., Stokes, M.G., 2017. Stable and Dynamic Coding for Working Memory in Primate Prefrontal Cortex. J. Neurosci. 37, 6503–6516. https://doi.org/10.1523/JNEUROSCI.3364-16.2017
Spector, F., 2011. Echoic Memory, in: Kreutzer, J.S., DeLuca, J., Caplan, B. (Eds.), Encyclopedia of Clinical Neuropsychology. Springer New York, New York, NY, pp. 923–924. https://doi.org/10.1007/978-0-387-79948-3_1121
Spitzer, B., Blankenburg, F., 2012. Supramodal Parametric Working Memory Processing in Humans. J. Neurosci. 32, 3287–3295. https://doi.org/10.1523/JNEUROSCI.5280-11.2012
Stokes, M.G., 2015. 'Activity-silent' working memory in prefrontal cortex: a dynamic coding framework. Trends Cogn. Sci. 19, 394–405. https://doi.org/10.1016/j.tics.2015.05.004
Takegata, R., Morotomi, T., 1999. Integrated neural representation of sound and temporal features in human auditory sensory memory: an event-related potential study. Neurosci. Lett. 274, 207–210. https://doi.org/10.1016/S0304-3940(99)00711-9

Tark, K.-J., Curtis, C.E., 2009. Persistent neural activity in the human frontal cortex when maintaining space that is off the map. Nat. Neurosci. 12, 1463–1468. https://doi.org/10.1038/nn.2406
Tervaniemi, M., Huotilainen, M., Brattico, E., 2014. Melodic multi-feature paradigm reveals auditory profiles in music-sound encoding. Front. Hum. Neurosci. 8, 496. https://doi.org/10.3389/fnhum.2014.00496
Tivadar, R.I., Knight, R.T., Tzovara, A., 2021. Automatic Sensory Predictions: A Review of Predictive Mechanisms in the Brain and Their Link to Conscious Processing. Front. Hum. Neurosci. 15.
Todd, J., Petherbridge, A., Speirs, B., Provost, A., Paton, B., 2018. Time as context: The influence of hierarchical patterning on sensory inference. Schizophr. Res., Mismatch Negativity 191, 123–131. https://doi.org/10.1016/j.schres.2017.03.033
Todorovic, A., Auksztulewicz, R., 2021. Dissociable neural effects of temporal expectations due to passage of time and contextual probability. Hear. Res. 399, 107871. https://doi.org/10.1016/j.heares.2019.107871
Trübutschek, D., Marti, S., Ueberschär, H., Dehaene, S., 2018. Probing the limits of activity-silent non-conscious working memory. bioRxiv 379537. https://doi.org/10.1101/379537
Ulanovsky, N., Las, L., Farkas, D., Nelken, I., 2004. Multiple Time Scales of Adaptation in Auditory Cortex Neurons. J. Neurosci. 24, 10440–10453. https://doi.org/10.1523/JNEUROSCI.1905-04.2004
van Ede, F., Chekroud, S.R., Stokes, M.G., Nobre, A.C., 2018. Decoding the influence of anticipatory states on visual perception in the presence of temporal distractors. Nat. Commun. 9, 1–12. https://doi.org/10.1038/s41467-018-03960-z
Vuust, P., Brattico, E., Glerean, E., Seppänen, M., Pakarinen, S., Tervaniemi, M., Näätänen, R., 2011. New fast mismatch negativity paradigm for determining the neural prerequisites for musical ability. Cortex 47, 1091–1098. https://doi.org/10.1016/j.cortex.2011.04.026
Wacongne, C., Changeux, J.-P., Dehaene, S., 2012. A Neuronal Model of Predictive Coding Accounting for the Mismatch Negativity. J. Neurosci. 32, 3665–3678. https://doi.org/10.1523/JNEUROSCI.5003-11.2012
Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., Diedrichsen, J., 2016. Reliability of dissimilarity measures for multi-voxel pattern analysis. NeuroImage 137, 188–200. https://doi.org/10.1016/j.neuroimage.2015.12.012
Wang, X.-J., 2001. Synaptic reverberation underlying mnemonic persistent activity. Trends Neurosci. 24, 455–463. https://doi.org/10.1016/S0166-2236(00)01868-3
Winkler, I., Reinikainen, K., Näätänen, R., 1993. Event-related brain potentials reflect traces of echoic memory in humans. Percept. Psychophys. 53, 443–449. https://doi.org/10.3758/BF03206788
Wipf, D., Nagarajan, S., 2009. A unified Bayesian framework for MEG/EEG source imaging. NeuroImage 44, 947–966. https://doi.org/10.1016/j.neuroimage.2008.02.059
Wolff, M., Kandemir, G., Stokes, M., Akyurek, E., 2019. Impulse responses reveal unimodal and bimodal access to visual and auditory working memory. https://doi.org/10.1101/623835
Wolff, M.J., Ding, J., Myers, N.E., Stokes, M.G., 2015. Revealing hidden states in visual working memory using electroencephalography. Front. Syst. Neurosci. 9. https://doi.org/10.3389/fnsys.2015.00123
Wolff, M.J., Jochim, J., Akyürek, E.G., Stokes, M.G., 2017. Dynamic hidden states underlying working-memory-guided behavior. Nat. Neurosci. 20, 864–871. https://doi.org/10.1038/nn.4546
Wollman, I., Morillon, B., 2018. Organizational principles of multidimensional predictions in human auditory attention. Sci. Rep. 8, 13466. https://doi.org/10.1038/s41598-018-31878-5

Yabe, H., Tervaniemi, M., Reinikainen, K., Näätänen, R., 1997. Temporal window of integration revealed by MMN to sound omission. NeuroReport 8, 1971–1974.
Yu, L., Hu, J., Shi, C., Zhou, L., Tian, M., Zhang, J., Xu, J., 2021. The causal role of auditory cortex in auditory working memory. eLife 10, e64457. https://doi.org/10.7554/eLife.64457
Zurita, P., Villa, A.E.P., de Ribaupierre, Y., de Ribaupierre, F., Rouiller, E.M., 1994. Changes of single unit activity in the cat's auditory thalamus and cortex associated to different anesthetic conditions. Neurosci. Res. 19, 303–316. https://doi.org/10.1016/0168-0102(94)90043-4