Selective enhancement of neural response consistency through self-organisation

Yujin Goto

The Graduate University for Advanced Studies, SOKENDAI

School of Life Science

Department of Physiological Sciences

Table of Contents

Summary

Perception is typically consistent: when the same sensory input is encountered again, subjective experience tends to be similar. Experimental neuroscience implicitly relies on a related assumption—that stimulus-evoked neural activity contains a consistent component that can be recovered from noisy recordings, for example, by averaging across trials. Yet such consistency is not guaranteed for a high-dimensional, recurrent, and intrinsically fluctuating system such as the brain. Neural activity varies across multiple scales, from local circuits to large-scale networks, and a growing literature argues that such variability is not merely a nuisance but can be functionally important, supporting flexible dynamics. The brain, therefore, faces a dual requirement: it must remain globally flexible while achieving strong input-conditioned consistency when needed. How this dual requirement is realised remains unclear.

In this thesis, I propose the selective consistency hypothesis: through experience, the brain selectively increases stimulus-conditioned neural consistency for a subset of inputs via synaptic self-organisation, and this selective increase supports more consistent perception for those inputs. This hypothesis is motivated by (i) evidence that response consistency is not fixed but can change with development and experience, and (ii) findings from perceptual learning paradigms in which perceptual selective consistency improves rapidly for complex stimuli that are difficult to explain through conventional feature-based accounts. The central behavioural model paradigm in this thesis is the noise-repetition detection (NRD) task, in which listeners judge whether concatenated white-noise segments are identical within a trial. A characteristic finding is that performance improves selectively for a repeatedly encountered exemplar, indicating experience-dependent gains in perceptual consistency for that specific stimulus.

To test the selective consistency hypothesis from complementary angles, I conducted (1) a computational simulation study and (2) a human EEG study during NRD performance.

In the simulation study, I examined whether a recurrent neural network can acquire selective consistency through local plasticity. Using an echo state network framework with weak Hebbian (Oja-type) plasticity in recurrent connectivity, I presented input time series analogous to NRD stimuli. The network developed higher selective consistency for repeatedly experienced inputs, while its responses to non-repeated inputs changed little. Importantly, this effect emerged without optimisation with respect to an explicit task objective, supporting the possibility that selective consistency can arise through self-organising dynamics. The magnitude of selective consistency acquisition depended on the network's baseline dynamical regime: it was maximised near criticality, at the boundary between overly ordered and overly chaotic dynamics (with the optimum slightly biased towards the chaotic side, where $\rho = 1.4$, $p<0.001$). In more ordered (e.g., $\rho \approx 0.9$) or more chaotic regimes (e.g., $\rho \approx$

1.9), selective consistency was not observed. This result motivates the idea that intrinsic network properties can constrain a system's capacity to acquire selective consistency.

In the EEG study, I tested whether neural and perceptual selective consistency develop together during NRD, and whether individual differences in learning relate to intrinsic neural properties. Following preregistered procedures, I collected behavioural and EEG data from 24 naïve participants (19 females, with an average age of 34.17 ± 9.02 years, range = 21–45) while they performed the NRD task. Participants listened to sounds composed of concatenated white-noise segments and judged on each trial whether the segments were identical. Behavioural results showed a significant learning effect: repeated-measures ANOVAs revealed a main effect of stimulus type on hit rate ($F_{1,23} = 5.42$, $p = 0.029$, $\eta p^2 = 0.191$) and on d' ($F_{1,23} = 12.19$, $p = 0.002$, $\eta p^2 = 0.346$), with no Type × Session interactions (hit rate: $F_{2,46} = 1.02$, $p = 0.367$; d': $F_{2,46} = 1.46$, $p = 0.242$). Perceptual reports of repetition were associated with higher within-trial neural consistency in sensory and parietal regions in broadband activity (theta, alpha, beta, with the most robust effect of beta: FDR corrected $q \ll 10^{-6}$). Conversely, perceptual reports of repetition were associated with lower within-trial consistency in delta-band activity in the parietal region (FDR corrected $q \leq 0.035$). These relationships strengthened for the repeatedly encountered stimulus as learning progressed (temporal and parietal beta: $q \ll 10^{-6}$; parietal delta: $q \ll 10^{-9}$). In addition, delta-band phase analyses revealed stimulus-specific inter-trial consistency in correct trials for the learnt stimulus, but only in sessions in which learning was successful. Because phase-based measures index cross-trial consistency in the timing of neural activity, this pattern is consistent with the view that learning is accompanied by increased stimulus-specific consistency in task-relevant computations at the network level.

Across the two projects, the findings converge on two conclusions. Firstly, selective-consistency acquisition varies substantially across systems and individuals. In the EEG experiment, learning performance distribution significantly deviated from normality (Shapiro–Wilk: $W = 0.85$, $p = 0.0026$), with high intraclass correlation across sessions, within-participant. Secondly, this variation is predictable from intrinsic properties of the underlying network: baseline dynamical regime in simulation and resting-state neural dynamics in humans. Given that disruptions of criticality have been linked to neurological and psychiatric conditions, the selective consistency framework suggests a principled way to reinterpret some impairments as constraints on the ability to acquire stimulus-specific consistency, rather than as simple increases in variability.

In summary, although the present work focused on a single paradigm (the NRD task), the results support the selective consistency hypothesis. Substantial challenges remain, including direct validation in vitro and formalising the mathematical mechanisms by which selective consistency can be acquired through self-organisation, as well as linking the framework more rigorously to existing computational theories. Nevertheless, the hypothesis is attractive in that it proposes a mechanism for

the brain's dual nature—overall flexibility/variability alongside stimulus-specific consistency—that emerges naturally as a dynamical property, without reliance on optimisation with respect to an explicit objective function. Rather, considering conventional optimisation-based accounts of brain function exploit dynamical properties, selective consistency may serve as a foundational substrate for them. Moreover, because dynamical properties are constrained by underlying structure, this framework may help bridge structural-level findings relevant to disease with higher-level functional abnormalities. Thus, despite many open questions, the selective consistency hypothesis proposed here may pave the way for a new direction in neuroscience.

Abbreviations

ADHD Attention Deficit Hyperactivity Disorder

ANN Artificial Neural Network

ASD Autism Spectrum Disorder

ASRS Adult ADHD Self-Report Scale

AQ Autism-spectrum Quotient

ANOVA Analysis of Variance

BNN Biological Neural Network

BOLD Blood Oxygen-Level-Dependent

cITPC Corrected ITPC

CLE Conditional Lyapunov Exponent

CR(R) Correct Rejection (Rate)

CSD Current Source Density

DNN Deep Neural Network

EEG Electroencephalogram (Electroencephalograph)

E-I Excitatory-Inhibitory

EMG Electromyography

EOG Electrooculography

ESN Echo State Network

ESP Echo State Property

FA(R) False Alarm (Rate)

FDR Benjamini–Hochberg False Discovery Rate

FEP Free Energy Principle

fMRI Functional Magnetic Resonance Imaging

FOI Frequency-band of interest

GLMM Generalised Linear Mixed Model

ICA Independent Component Analysis

ICC IntraClass Correlation

ITPC Inter-trial phase coherence

LOO Leave-One-Out

MBGD Minibatch-Based Gradient Descent

NRD Noise Repetition Detection (task)

PC Predictive coding

PSD Power-Spectrum Density

RefRN Referenced Repeated-Noise

RMSE Root Mean Squared Error

RN Repeated-Noise

RNN Recurrent Neural Network

RT Reaction time

ROI Region of interest

NRMSE Normalised Root Mean Squared Error

SE Spectral Exponent

SNN Spiking Neural Network

STDP Spike-Timing Dependent Plasticity

2AFC Two-Alternative Forced Choice

Symbols

In this thesis, I use **bold** style for vectors and matrices. *Italic* style is used for variables.

α	Hebbian learning rate
$C(s)$	selective consistency measure for a specific stimulus s
$C_{cost}(s)$	cost-like function, used for calculate $C(s)$
ch	channel index of EEG
$\boldsymbol{\theta}$	neural parameters (connectivity weight matrix...)
$\mathbb{E}$	expected value
ε	observation noise, stochastic fluctuation, and any other noise of the generative model
g	nonlinear projection function ($s \rightarrow u$, $u \rightarrow x$)
k	trial index
$\mathbf{J}$	Jacobian
λ	Lyapunov exponent
λ_c	conditional Lyapunov exponent
m	individual internal model
η	learning rate
p	probability
Φ	trajectory operator
s	sensory stimuli (e.g., flash, soundwave)
ses	session
S_{SC}	set of stimuli, which are already consistent
t	time index
τ	time, within a specific time window
T_{tr}	the last time point of the transition period for converging
T_{end}	the last time point of the stimulus
T_{SS}	time window after transition period. $[T_{tr}(s), T_{end}(s)]$
$\boldsymbol{u}$	sensory signal (e.g., visual, auditory signals from receptors. Input signal for the ANN)
$\boldsymbol{x}$	neural representation (activity at the middle layer of the ANN)
$\boldsymbol{W}$	weight matrix of the ANN's middle layer.
ω	time window
$\boldsymbol{W}^{\text{in}}$	a weight matrix connecting the input and the middle layer
$\boldsymbol{W}^{\text{out}}$	a weight matrix connecting the middle and the output layer
y	the output signal of the ANN (which can be used at the Decision-stage)
ρ	the spectral radius of the ANN

1. General Introduction

1.1. Consistency

In this thesis, I use consistency as a dynamical systems notion of reproducibility of trajectories rather than a synonym for low variability. Specifically, for a given stimulus input stream $\boldsymbol{u}(t)$, a neural system is consistent if repeated presentations of the same input drive its internal state trajectory into a similar region of state space, despite nuisance perturbations such as trial-to-trial noise, uncontrolled initial conditions, and contextual fluctuations. This definition is intentionally input-conditional: the brain may remain globally variable across time and states, while still exhibiting reliable, stimulus-conditioned responses when the input provides sufficient constraint. Throughout, I distinguish neural consistency (reproducibility of neural trajectories or representational features) from perceptual consistency (reproducibility of subjective perception), and I treat the former as a candidate mechanistic substrate for the latter.

Neural consistency—the property of producing similar internal neural states regardless of initial conditions or contextual fluctuations, and its contrast with variability—is essential for maintaining a stable representation of the environment and for controlling behaviour. Without such consistency, the brain would lose any sense of uniformity in the world; learning, adaptation, and any other meaningful cognitive operations would become impossible. Although many theories have been proposed regarding how the brain acquires information about the external world and learns to generate optimal actions, all of them, implicitly or explicitly, assume that the neural responses to a given input (observation) $\boldsymbol{u}$ are approximately consistent, aside from modulations arising from specific factors such as attention, neuromodulation, or behavioural context. Also, in experimental neuroscience, we assume the brain is a system that embeds consistent information representations for identical inputs while retaining sufficient flexibility to adjust its processing according to situational demands. On this basis, we usually record multiple trials of neural responses to the same input and identify consistent activity embedded in noisy neural responses. Neural activity patterns that reliably occur in response to the same stimulus are typically referred to as neural representations, and a substantial body of work has examined the brain regions, frequency bands, and even individual neurons that give rise to these consistent patterns. Understanding when and how neural systems exhibit neural consistency is, therefore, a foundational problem in neuroscience.

However, from the perspective of nonlinear dynamical systems theory, repeatedly producing similar internal states for the same input is not a trivial property for a high-dimensional, recurrent, nonlinear, and noisy system such as the brain. Theoretically, whether a complex dynamical system possesses consistency is determined by the properties of its connectivity matrix (Lukoševičius &

Jaeger, 2009; Sompolinsky et al., 1988). A connectivity matrix specifies the direction and strength of interactions among the system's components; in the brain, this corresponds to synaptic connections as well as anatomical, functional, and connectivity across regions. Consider a driven dynamical system given by

$$\frac{d\boldsymbol{x}(t)}{dt} = g\big(\boldsymbol{x}(t), \boldsymbol{u}(t)\big), \qquad (\text{eq. } 1)$$

where $\boldsymbol{x}(t)$ denotes the internal state, $\boldsymbol{u}(t)$ is an external input drive, and g is a nonlinear function determined by the system's dynamic property and structure (Uchida et al., 2008). In this framework, 'consistency' refers to the convergence of state-space trajectories across trials under the identical input $\boldsymbol{u}(t)$ (Uchida et al., 2004). Formally, for repeated trials indexed by k, consistency can be expressed as a convergence condition such as:

$$\lim_{t\to\infty} \left|\left|\boldsymbol{x}^{k}(t) - \boldsymbol{x}^{k+1}(t)\right|\right| = 0, \qquad (\text{eq. } 2)$$

for a fixed input $\boldsymbol{u}(t)$, meaning that different initial conditions of internal state and nuisance perturbations do not prevent dynamics from approaching the same attracting trajectories of state space (it can be written as an input-conditioned mapping $\boldsymbol{x}(t) = \Phi(\boldsymbol{u})$). In idealised settings—linear, feedforward, noise-free, and perfectly isolated from other influences—such convergence is expected. But real neural systems violate all of these conditions: neural activity is continuously perturbed by noise and ongoing dynamics; connectivity is strongly recurrent, so past states feed back into the present; and, crucially, the brain is never driven by "the stimulus alone". Even when the nominal sensory stimulus is identical, the total drive includes nonstationary sensory background, internal state fluctuations, and top-down influences (attention, expectation, neuromodulation), so the resulting input to the system is not exactly the same twice (Arieli et al., 1996). These considerations imply that consistency should be treated as a non-trivial dynamical achievement rather than an automatic baseline property of the brain (Sussillo & Abbott, 2009).

Importantly, a growing body of work also argues that neural variability is not merely nuisance noise to be eliminated, but can be functionally meaningful (Dinstein et al., 2015; Garrett et al., 2011; Stein et al., 2005; Terlau et al., 2025). In macro-level neuroscience using electroencephalography (EEG) and functional magnetic resonance imaging (fMRI), trial-to-trial variability in neural activity was traditionally dismissed as measurement noise (Arieli et al., 1996; Goris et al., 2014). However, accumulating evidence indicates that such variability carries functional meaning. For example, Garrett and colleagues demonstrated that moment-to-moment blood oxygen-level-dependent (BOLD) variability predicts cognitive performance better than mean activation, suggesting that variability reflects a system's processing capacity rather than noise (Garrett et al., 2011). In EEG, trial-to-trial variability decreases with perceptual learning and increases under conditions of uncertainty,

suggesting that variability tracks internal neural states relevant to decision making (Arazi et al., 2017). At a more mechanistic level, computational and theoretical studies have argued that variability enables efficient sampling of latent environmental structure and supports flexible switching between behavioural states (Berkes et al., 2011; Dinstein et al., 2015; Faisal et al., 2008; Orbán et al., 2016). Studies of spontaneous neural dynamics have likewise shown that variability is structured rather than random, correlating with behavioural variability, attention, and readiness to respond (Churchland et al., 2010; He, 2013). Together, these findings motivate a view in which the brain benefits from remaining globally variable and flexible, while still maintaining sufficient input-conditional reproducibility to form stable representations when needed.

Against this background, an increasingly influential perspective is that effective neural information processing requires an appropriate balance between consistency (stability, order) and variability (flexibility, exploration). An overly consistent system would lose the flexibility required to adjust its computations to contextual demands. In contrast, an overly variable system would fail to form stable representations and thus could not adapt to the environment. Related computational frameworks propose that neural systems may operate near the boundary between order and disorder (often described as the "edge of chaos" or criticality), where both reliable representations and flexible dynamics can coexist (Beggs, 2008, 2019; Beggs & Plenz, 2003; Chialvo, 2010; Chua et al., 2012; Kumar et al., 2017). Indeed, recent work has increasingly documented relationships among the balance, developmental stage, learning, psychiatric conditions, and ageing (Cocchi et al., 2017; de Arcangelis & Herrmann, 2010; Voytek et al., 2015; Wilkinson et al., 2024; Wilting & Priesemann, 2019). Clarifying how a fundamentally variable brain nevertheless produces input-conditional consistency is therefore a central question with implications for basic neuroscience, clinical research, and engineering (Arieli et al., 1996; Churchland et al., 2010). In the next section, I extend this problem by asking whether such reproducibility is acquired in a stimulus-dependent manner—namely, the selective consistency hypothesis.

1.2. Selective Consistency

Building on the definition above, I define “selective consistency” as the stimulus dependence of input-conditional consistency: through experience, only a subset of inputs (or input features) come to elicit strongly convergent trajectories, whereas other inputs continue to evoke weakly constrained, idiosyncratic dynamics. Selective consistency is therefore not the same as classical feature selectivity (a detector for a stimulus attribute), nor does it imply indiscriminate "collapse" of responses across different stimuli. Rather, it describes a learnt contraction of dynamics that can remain compatible with discrimination when different stimuli drive the system towards distinct attractor-like regions or distinct

stimulus-locked manifolds (Hopfield, 1982).

Integrating the background reviewed in the previous section yields a concrete hypothesis about how stimulus-dependent consistency emerges. The brain is a high-dimensional, noisy, and context-sensitive dynamical system; consequently, the consistency of neural representations for a given input should not be treated as an invariant, innate property. Instead, consistent responses are expected to be acquired for particular inputs repeatedly encountered via experience-dependent reorganisation of network dynamics. Empirically, this expectation is consistent with converging observations across development and learning: perceptual judgements often become more stable with exposure; developmental work has reported reductions in certain components of neural variability alongside more stable sensory representations; and some decoding studies suggest that neural representations can become more reliable across sessions (Jehee et al., 2012; Poort et al., 2015; Schoups et al., 2001). Collectively, these findings suggest that experience improves the consistency of neural representations in an input-dependent manner, thereby supporting stable perception.

In this thesis, I formalise this perspective as the selective consistency hypothesis: the proposition that the brain selectively acquires consistency only for inputs it has experienced, by reorganising network dynamics in a stimulus-specific manner while preserving overall variability and flexibility.

Non-Mathematical Definition

1. **Initial regime:** Early in development or before learning, neural trajectories can be weakly constrained, such that repeated presentations of the same input may yield variable representational states.
2. **Experience-dependent reorganisation:** With repeated exposure, neural connectivity and/or dynamics reorganise so that specific inputs drive the system into more reproducible (consistent) trajectories.
3. **Computational role:** This stimulus-specific increase in consistency can be partly dissociated from—and can provide a substrate for—downstream processes such as feature detection, decision formation, and higher-level learning rules.

Before developing the hypothesis further, it is useful to distinguish this framework from a classical line of work on feature selectivity in sensory systems. Numerous studies have demonstrated selective and reliable neural responses to particular stimulus features or categories (e.g., orientation selectivity in primary visual cortex (Hubel & Wiesel, 1962), categorical selectivity in higher-level areas (Haxby et al., 2001; Kanwisher et al., 1997), highly specific units in the medial temporal lobe, and even highly

specific cells, such as the so-called Jennifer Aniston neuron (Quiroga et al., 2005). Also, even for more explicitly dynamic stimuli, feature detector neurons which respond to specific combinations or queue of signals are well-known (Doupe & Solis, 1997; Solis et al., 2000; Suga et al., 1983). However, the crucial difference lies in where and how "consistency" is characterised. In classical feature-selectivity accounts, reliable responses are primarily attributed to specific detectors—individual neurons or small populations tuned to particular features. In contrast, selective consistency refers to a property of the network's global dynamics: a selective increase in the consistency of the system's trajectory in response to a given input pattern. On this view, selective consistency is not itself a task-specific computation; rather, it is a dynamical property that other computations can exploit. Accordingly, later sections (e.g., Section 4.3) discuss how selective consistency can be distinguished from changes in sensitivity and related to broader learning frameworks, such as predictive coding (Rao & Ballard, 1999). The key claim of this thesis is that selective consistency provides a principled precondition—input-dependent representational consistency—under which downstream computational theories of perception and learning can operate effectively.

Mathematical Definition

Based on the above considerations, selective consistency can be formalised physically as the property that differences in initial conditions decay over a transient period, yielding stable neural responses only for specific inputs $s \in S_{SC}$. Let the physical sensory stimulus (i.e, physical object, soundwave) be denoted by s, the corresponding sensory input signal by $\boldsymbol{u}(t;s)$, and the sensory representation on trial k by $\boldsymbol{x}^k(t;s)$. Let $\boldsymbol{\theta}$ represent internal model (i.e, synaptic connectivity), and ε denote stimulus-irreverent noise and intrinsic neural fluctuations. Under these definitions, the brain maps a stimulus s to a neural response following

$$\boldsymbol{x}^k(t;s) = \Phi\left(\boldsymbol{x}^k(t-1;s), \boldsymbol{u}(t;s), \boldsymbol{\theta}^k, \varepsilon^{\mathrm{k}}\right). \quad \text{(eq. 3)}$$

Here, the mathematical expression of selective consistency can be written by just adding the condition that the stimulus s is fixed in equation (2):

$$\lim_{t\to\infty} \left|\left|\boldsymbol{x}^k(t;s) - \boldsymbol{x}^{k+1}(t;s)\right|\right| \simeq 0. \quad \text{(eq. 4)}$$

In practice, neural responses are finite in duration, and the trial-to-trial differences cannot be exactly zero because of ε, so I define the following selective consistency measure:

$$C_{cost}(s) := \frac{1}{|T_{ss}(s)|} \int_{t\in T_{ss}(s)} \mathrm{Var}_k[\boldsymbol{x}^k(t;s)]\, dt \quad \text{(eq. 5)}$$

$$C(s) := \frac{1}{1+C_{cost}(s)} \in (0,1], \quad \text{(eq. 6)}$$

where $T_{ss}(s) = [T_{tr}(s), T_{end}(s)]$ is a time window of a stable state after the transition point $T_{tr}(s)$.

C takes values between 0 and 1, and larger values of $C(s)$ indicate higher selective consistency for the stimulus. Because the system is subject to noise ε, $C(s)$ will not reach 1 in practice, nor will it be exactly 0 for any stimulus.

1.3. Structure of This Thesis

The selective consistency hypothesis follows naturally from existing empirical and theoretical observations, yet it faces two major challenges.

Firstly, classical theories of nonlinear dynamical systems do not address how a system could acquire stimulus-specific consistency. In standard formulations, a system becomes globally more consistent only if its dynamics converge toward stable attractive trajectories. Applied to the brain, this would imply that a system that begins in a chaotic regime becomes progressively more consistent as it develops. However, such a scenario would lead to a loss of flexibility. As the brain matures, its dynamics would become increasingly rigid, eventually impairing its ability to adapt behaviour to changing environments. Thus, what is required is a mechanism that allows the system to maintain global preservation of flexibility while selectively increasing the consistency of responses only for stimuli that have been experienced. This mechanism represents a form of optimisation not captured by existing dynamical systems accounts.

Secondly, no study to date has provided a direct demonstration that neural and perceptual consistency are jointly acquired through experience. Although developmental studies and research on perceptual learning indirectly hint that such coupled changes must occur, they do not provide explicit evidence for the stimulus-specific emergence of consistency (Agus et al., 2010; Naik et al., 2023; Riggins & Scott, 2020). To validate the selective consistency hypothesis, a more direct, mechanistic investigation is required.

In the following two chapters, the selective consistency hypothesis is examined from both theoretical and empirical perspectives. Firstly, section 1.4 introduces the Noise Repetition Detection task, which serves as the experimental paradigm for this study in modelling selective consistency. In Chapter 2, I employ reservoir computing combined with Hebbian-like plasticity to demonstrate that selective consistency can emerge spontaneously through experience in a self-organising manner. In Chapter 3, I analyse human EEG data collected during the Noise Repetition Detection task to quantify neural consistency and to examine how this measure relates to experience and perceptual consistency.

Finally, in General Discussion, I integrate insights from these two projects and offer a comprehensive discussion of the principles underlying consistency formation in neural systems, as well as the relationship between selective consistency and existing theories of learning.

To maintain readability, I do not provide detailed explanations of specialised terminology from physics, mathematics, or biology in the main text. Where necessary, readers are referred to the Glossary via the embedded links.

1.4. Noise Repetition-Detection task

As the behavioural task for assessing selective consistency, I adapted an implicit unsupervised auditory perceptual learning—noise repetition detection (NRD) task—introduced by Agus and colleagues to study how the auditory system forms memories for arbitrary, complex sounds (Agus et al., 2010).

In a typical NRD task, listeners hear several-second-long white-noise excerpts and perform a two-alternative forced choice task to judge whether the sound contains a brief repeating segment or is pure noise. On "repeated-noise" (RN) trials, a short identical noise snippet (i.e., first and second 500 ms halves of 1-second stimuli) is seamlessly concatenated multiple times into a longer noise stream, whereas on "noise" (N) trials, the waveform is continuously regenerated so that no exact segment repeats within the trial. Crucially, small subsets of RN and N stimuli are secretly designated as "reference repeated noise" (RefRN) and "reference noise" (RefN): the same snippet recurs across many trials within a block, even though participants are never told of their existence and receive no explicit feedback (Fig. 1). Repetition-detection performance is then compared between RN trials, which repeat within a single trial, and RefRN trials, in which the same noise pattern recurs across trials. Basically, the behavioural results show improved detection accuracy only for RefRN (see 1.4.1 for details).

There are several reasons to employ the NRD task to test selective consistency. As a premise, the NRD task requires listeners to judge the identity of a given noise segment, so the resulting performance functionally reflects the consistency of perception for these segments. In this framework, the characteristic finding that performance remains low for RN but improves specifically for RefRN indicates the emergence of perceptual selective consistency for the RefRN waveform. Furthermore, because selective consistency is a dynamical construct, static stimuli are inherently unsuitable; this consideration already excludes many visual paradigms. Although dynamic visual stimuli such as optical flow or naturalistic videos could, in principle, be used, they are inappropriate if, consistent with the definition in Section 1.2, selective consistency is acquired through development and experience. Most such visual stimuli, despite their fine-grained variability, are sufficiently similar to things participants have encountered throughout their lives. In contrast, while listeners may have categorical experience with "white noise," each noise waveform is statistically independent, and its temporal evolution is in principle unpredictable from experience. For these reasons, the NRD task provides a

uniquely suitable paradigm for examining selective consistency in its pure form: it allows us to probe the consistency of perception and neural activity for a given temporal stimulus without contamination from prior familiarity, semantic structure, or predictable temporal statistics.

1.4.1. Previous Findings: Behavioural Aspects

Agus et al. and subsequent studies showed that listeners rapidly acquire robust memories for completely meaningless noise exemplars: detection of repetitions becomes almost perfect for RefRN (Agus et al., 2010), learning occurs without supervision or awareness even during sleep (Andrillon et al., 2017), and memory traces can persist for weeks (Agus et al., 2010). Also, some studies replicated these results using similar but different stimulus patterns: longer sample audio (Agus et al., 2010), shorter and separated by intermediate irrelevant audio segments (Andrillon et al., 2015; Ringer et al., 2022, 2023), random auditory pulse trains (Kang et al., 2017, 2018), random spectral pattern segments (Kang et al., 2021), tone clouds (Kumar et al., 2014; Agus & Pressnitzer, 2021), and visual and tactile pulse trains (Kang et al., 2018). All these studies showed performance benefits for RefRN—or stimuli playing a role like that—in terms of higher hit rate (HR), sensitivity (d') and faster reaction times compared to RN. Furthermore, this learning effect is robust even for patients with dyslexia (Agus et al., 2014). Taken together, these results indicate that the NRD task taps a form of unsupervised, largely implicit perceptual learning supported by idiosyncratic temporal features in the noise, rather than by any semantic or categorical structure.

Although this point has not been highlighted in subsequent studies, it is noteworthy that the distribution of learning performance for RefRN is bimodal: approximately two-thirds of the data are clustered around chance level, and the remaining data are scored almost perfectly in the first study (Agus et al., 2010). In the study, each participant completed multiple blocks, learning a different RefRN in each block. Interestingly, roughly half of the participants failed to exceed chance level in any block, whereas the other half achieved accuracy above 90% in at least one block. This pattern suggests that there are individual differences in the learning abilities and underlying neural substrates required for this task, and that even when the capacity for such learning is present, it does not necessarily manifest reliably on every occasion.

1.4.2. Previous Findings: Neuroscience Aspects

Subsequent studies assessing neural mechanisms of the NRD task reported increased event-related desynchronisation (Andrillon et al., 2015, 2017; Ringer et al., 2023) and inter-trial phase coherence (Luo et al., 2013; Andrillon et al., 2015, 2017) for RefRN. The first study measured neural activity

during the NRD task using magnetoencephalography (MEG) and reported that, as initially novel noise patterns are memorised, a reliable inter-trial phase coherence (ITPC) in low-frequency (3–8 Hz) responses appears in the auditory area (Luo et al., 2013). Moreover, they reported that the acquired patterns for different RefRN patterns were distinguishable. Andrillon et al. (2015) reported trials of learnt RefRN stimuli evoke reliable "memory-evoked potentials": early-latency ERP deflections with auditory topographies that resemble a standard N1–P2 complex from parietal, temporal, and occipital areas, in addition to consistent results regarding increased ITPC in the low frequency band (0.5–5 Hz). They also showed significantly higher stimulus-type decoding accuracy at the single-trial level, as assessed by logistic regression of the ERP, for RefRN than for RN (Andrillon et al., 2015). Those tendencies were confirmed even in the absence of task demands and attention, in an EEG study during human sleep (Andrillon et al., 2017) and in anaesthetised rats (Kang et al., 2021). A fMRI study revealed that the learnt patterns could be decoded using multi-voxel pattern analysis (Norman et al., 2006), with activity in the planum temporale and the hippocampus (Kumar et al., 2014). To sum up, previous findings regarding this task converge on strengthened phase alignment across trials in the temporal auditory area through learning on meaningless, featureless noise time series.

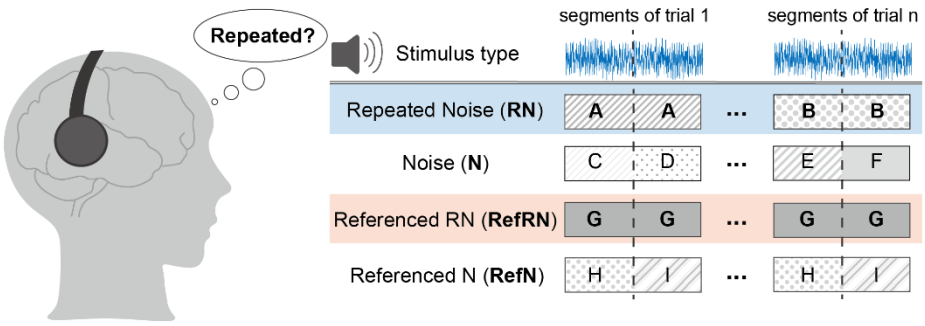

Figure 1. Abstract information on the Noise-repetition detection task.

Overview of the task. Typically, four types of white-noise stimuli are used, divided into two classes depending on whether the same noise segment is repeated (RN) or not (N). Participants report, in a 2AFC format, which class the stimulus belonged to—i.e., whether they perceived repetition in the sound. N and RN are each further subdivided into two stimulus types. RefN and RefRN use the same noise segment not only within a trial but also across trials, and are therefore learnable stimuli. The presence of these two stimulus types is usually not disclosed to participants.

1.4.3. Remaining Questions and Selective Consistency

In this study, I decompose the information-processing operations involved in the NRD task into two stages—Representation and Decision—and propose selective consistency as the mechanistic principle governing the Representation stage. By showing that this framework resolves the outstanding issues associated with NRD learning, I provide empirical and theoretical support for the selective consistency hypothesis.

Many studies have reported that, after only a handful of exposures to the RefRN waveform, neural activity becomes more consistent across trials (in terms of both amplitude and phase), and that this change correlates with behavioural performance. Although those findings are intriguing, the learning mechanisms underlying the NRD task have not yet been adequately discussed.

Firstly, despite the original study reporting significant variability in the learning effect, no study to date has assessed its neuroscientific basis (Agus et al., 2010). Given that roughly half of the participants failed to learn at all, it is plausible that this variability is strongly related to individual factors, such as age, musical experience, and neurobiological characteristics, including traits associated with neurodevelopmental disorders. Such individual differences may provide valuable clues for uncovering the underlying mechanisms. Moreover, even among the remaining half of the participants who were able to learn, learning did not occur consistently. Some participants learnt in some blocks but not in others. Thus, analysing those blocks separately is crucial for elucidating the neural mechanisms involved, yet no previous study has explicitly incorporated this aspect.

Secondly, despite its crucial role in understanding this task, one aspect that has been largely overlooked is the within-trial, across-segment comparison of neural activity. As participants do not, in fact, compare sounds across trials, but rather compare segments within a trial, any account of why repetition detection performance improves must be framed in terms of changes in within-trial neural activity (i.e., improved consistency across noise segments).

Lastly, there is the question of the mechanism that enables such rapid changes in behavioural and underlying neural activity. Because behavioural performance on this task approaches the ceiling after only a few exposures, the underlying neural changes—whether across or within trials—also must be realised through few-shot learning. Despite the scientific potential of such a computational mechanism, to my knowledge, no studies have successfully explained the NRD learning effect from a theoretical perspective.

All these remaining questions regarding the NRD learning effect converge to a need to decompose the information-processing involved in this task. At its core, the NRD task asks participants to judge the identity of a given sensory stimulus. This can be broadly divided into two stages: a sensory representation process, in which the external sensory stimulus is projected into an internal neural

activity pattern, and a comparison-based decision-making process, in which the representations are evaluated for identity (Gold & Shadlen, 2007). In the remaining parts of this thesis, these are referred to as the Representation stage and the Decision stage, respectively. Previous neurophysiological studies using the NRD paradigm have relied exclusively on across-trial comparisons and, as a result, have not been able to distinguish these two stages. Critically, evaluating the Representation stage requires within-trial analyses, because only trial-internal neural dynamics can reveal how a given noise segment is encoded before any decision comparison occurs. Separating these stages also enables a precise investigation of individual differences in learning and performance fluctuations across blocks. Moreover, it also helps consider the computational mechanisms—fundamentally different classes of learning mechanisms govern the Representation and Decision stages. By decomposing the NRD task into these two stages, the three unresolved issues outlined above converge on a single overarching question: whether and how NRD learning effects arise from changes in the Representation stage, the Decision stage, or even both.

The Representation stage corresponds to how each noise segment is encoded within neural activity. The sensory organs, transmission pathways (i.e., the inferior colliculus), and auditory cortical areas should be involved in this stage (De Martino et al., 2013; Nourski et al., 2014). The Decision stage, by contrast, aggregates a broader set of processes: comparing the currently represented noise segment with its short-term memory trace, judging whether they match, and generating a motor response. Modulatory factors such as task motivation and attention are also assumed to exert their effects at this stage. Therefore, brain-wide areas are involved at this stage, including the higher auditory cortex, regions of the frontal-parietal network (Ridderinkhof et al., 2004; Gold & Shadlen, 2007; Keuken et al., 2014), and the limbic system (Carter et al., 1998).

From this two-stage perspective, the existing literature suggests that NRD learning arises primarily within the Representation stage. In most domains, perceptual learning is explained by changes at the Decision stage (Dosher et al., 2013)—for example, through predictive coding (Rao & Ballard, 1999) or reinforcement learning (Niv, 2009). In predictive coding and related Bayesian frameworks, the brain uses priors constructed from experience to generate predictions about incoming inputs and the environment. Yet for white noise, the value at the next time point is determined entirely at random, making prediction impossible in principle. So, to explain this task in terms of prediction, one would effectively have to assume that the RefRN time series is memorised perfectly, which poses challenges given only a few exposures (Denham & Winkler, 2020). Reinforcement learning also seems less suitable as the dominant mechanism, because model updating requires feedback on decisions, whereas the NRD task provides neither correctness nor reward feedback. Indeed, selective and rapid changes in auditory cortical responses to RefRN have been observed even when stimuli are presented to anaesthetised rats, and learning effects are also evident in humans who were exposed to stimuli during

sleep and tested immediately upon awakening (Andrillon et al., 2017; Kang et al., 2021). These findings indicate that improved performance does not depend on decision-related processes such as strategic comparison, attention, or reward-based updating. Taken together, NRD learning is difficult to explain in terms of changes in the Decision stage and is more plausibly attributed to modifications within the Representation stage.

A representative update rule for the Representation stage is sensitivity sharpening at the level of individual neurons or neural networks (Gilbert et al., 2001; Jehee et al., 2012; Schoups et al., 2001). However, this framework is also not well-suited to explain the NRD learning mechanisms. Because behavioural performance on this task approaches the ceiling after only a few exposures, the underlying neural changes—whether across or within trials—also must be realised through few-shot learning. In the sensitivity framework, learning is considered to work by sharpening neurons in sensory areas that become selective to specific features, such as frequency preferences in primary auditory cortex (A1) and the inferior colliculus (De Martino et al., 2013), chunked spatiotemporal patterns (such as phonemes), or abstract concepts (such as music and human voice). Thus, in this scenario, explaining NRD learning would require neurons to become more sensitive to the stimulus's spectral properties or to compress its temporal structure so that RefRN can be distinguished from other RNs. However, all stimuli in the NRD task are white noise; their spectra are uniform and shared across all stimuli regardless of the condition. Moreover, from an information-theoretic standpoint, white noise is a maximally entropic, and therefore incompressible signal (Shannon, 1948). Consequently, to explain NRD learning through a sensitivity-sharpening mechanism, one would have to assume the existence—or rapid formation—of neurons or networks that respond selectively to an exact specific noise waveform (Masquelier, 2018). Given that RefRN exposure occurs only a handful of times, the feasibility of forming such particular detectors is extremely low.

In this thesis, through two complementary projects, I demonstrate that introducing selective consistency as a new update rule for the Representation stage resolves the limitations of existing accounts and provides empirical support for the selective-consistency framework. Consistency, in this context, is not the selectivity of a particular detector but the stability of dynamics at the level of the entire network. Theoretically, such stability is determined solely by the structure of recurrent connections (Sompolinsky et al., 1988). It therefore does not require optimisation driven by outcomes at the Decision stage—such as rewards, prediction errors, or correctness feedback. Consequently, if a mechanism exists for updating consistency, it should operate independently of stimulus predictability and remain effective even under NRD conditions, unlike other learning theories.

The computational implementation of achieving selective consistency is developed and demonstrated in Project 1 under the conceptualised NRD task setting. In concurrent Project 2, using human EEG experiments, I will provide evidence that NRD learning arises from changes in the

Representation stage. Furthermore, based on the implications of Project 1 and the behavioural and physiological data together, I will propose a mechanistic account of why individuals differ in their capacity to develop selective consistency (Fig. 2). Finally, in the general discussion, I will discuss the functional role and benefits of selective consistency, and its relationships with other existing learning frameworks (see Section 4.3).

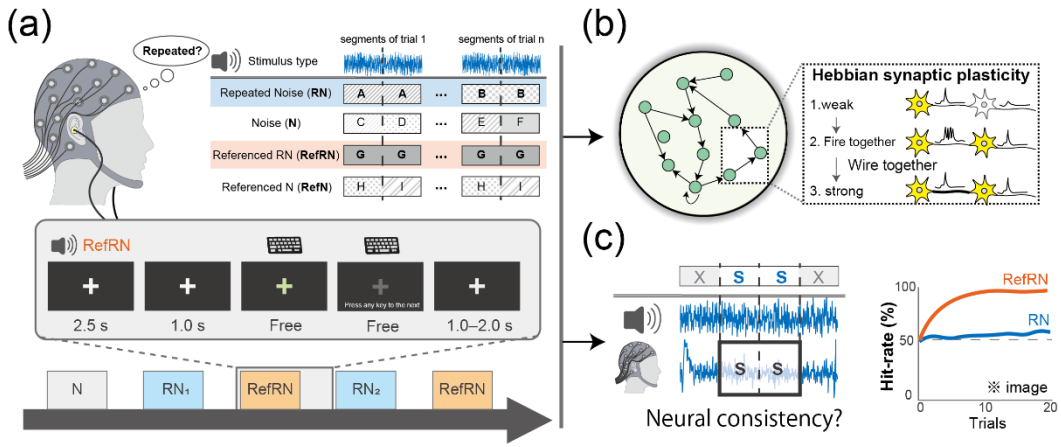

Figure 2. Structure of this thesis.

(a) Conceptualised behavioural paradigm of NRD task. (b) Project 1, simulation study. The aim is to explore the computational mechanism of selective consistency. (c) Project 2, human behavioural and EEG study. The aim is to investigate whether neural selective consistency emerges alongside perceptual selective consistency for RefRN stimuli. Detailed explanations for each will be provided in the following Chapters.

2. Project 1: Computational Mechanisms of Selective Consistency

2.1. Introduction

Project 1 is already published in PLoS Computational Biology as a peer-reviewed article (Goto & Kitajo, 2024).

2.1.1. Aim and Logic

In Project 1, I develop a minimal computational account of how selective consistency can emerge in the Representation stage of the NRD task. The goal is not to model repetition judgements at the Decision stage, but to demonstrate a more basic claim: repeated exposure to a particular exemplar can self-organise stimulus-conditioned convergence of recurrent dynamics, creating a reliable substrate that downstream readout or decision mechanisms could exploit. Accordingly, the model is not trained to detect repetitions, and no performance-driven optimisation (e.g., reward learning, error backpropagation, or supervised adjustment of recurrent weights) is applied.

NRD learning is difficult to explain within standard accounts because the RefRN advantage arises after only a handful of exposures without explicit feedback, even though the stimuli are spectrally similar and lack obvious compressible structure. Rather than assuming rapid formation of highly specific feature detectors, I test whether selective consistency can arise as a network-level dynamical property: convergence of stimulus-conditioned trajectories despite ongoing variability.

Moreover, because it is also suggested that individuals differ in their ability to achieve selective consistency, a candidate mechanism is required to explain why these differences arise.

To address these points keeping the mechanism as general as possible, I use a simple recurrent neural network as an abstract model of a local recurrent circuit that transforms sensory drive into time-evolving internal states. I then ask when a biologically plausible local rule—Hebbian-like plasticity—can modify the network's connectivity so that only repeatedly encountered inputs elicit consistent trajectories. Finally, I examine how the network's initial dynamical regime constrains this acquisition, focusing on criticality-related properties of the recurrent connectivity.

2.1.2. Working Hypotheses

H1. Self-organisation through Hebbian plasticity allows the network to acquire a selective consistency to a particular input

Given prior findings that NRD learning can progress during sleep and that neural activity becomes more consistent through development, it is natural to consider that the acquisition of selective consistency should proceed implicitly and unsupervised. In the nervous system, both unsupervised and supervised learning are realised through updates of synaptic connection strengths. The most common synaptic plasticity rule for unsupervised learning is the Hebbian rule (whereas supervised learning, such as backpropagation, is often described in terms of a different plasticity rule, the generalised delta rule).

H2. Network's recurrent connection property limits H1, and maximises it at the criticality

Individual differences in the ability to acquire selective consistency, as suggested by prior studies, should also be reproduced by this model. One important property related to consistency in an RNN is criticality. Criticality is determined by interactions among various properties of the connectivity matrix (such as the excitatory/inhibitory balance, connection density, and average connection strength). It is known that the brain achieves the highest performance in a critical regime that is neither too low nor too high, and that in early development and in disease, it deviates from this regime into subcritical or supercritical states. When the system becomes supercritical, it loses consistency, and when it becomes subcritical, it loses variability. Therefore, in this study, I hypothesised that if a network has a moderate initial level of criticality, plasticity can induce stimulus-selective changes that shift its conditional behaviour towards the subcritical side; if the network starts from the subcritical side, it cannot be variable; if the network starts from a supercritical regime, it cannot overcome its too strong variability.

2.2. Methods

In Project 1, I tested whether an RNN can acquire selective consistency via Hebbian plasticity and, if so, under what conditions. The representation stage models the process by which an external stimulus s is transformed into sensory afferent signal u, and by which u is further transformed into neural representations $\boldsymbol{x}$ in the sensory cortex. Even within the broad category of "sensory-cortex models", there exist diverse approaches, including DNNs incorporating laminar cortical architecture as characterised in the visual cortex, RNNs that assume within-layer horizontal connections, and neural mass models designed to capture more global activity. To avoid imposing specific anatomical assumptions about selective consistency, I adopt an RNN as an abstract model of an arbitrary local recurrent cortical circuit that generates time-evolving internal states in response to stimulus input. The simplest RNN has no explicit layers or clusters; it consists of a single network with a random recurrent connectivity matrix that allows excitatory and inhibitory connections. Anatomically, this can be regarded as an abstraction of within-layer horizontal connections, or more generally of recurrent connections within a cortical area. As will be shown later, the qualitative behaviour of an RNN depends on algebraic properties of its connectivity matrix (Sompolinsky et al., 1988). Accordingly, I demonstrate that the plasticity introduced into the RNN modifies the connectivity matrix and selectively enhances the consistency of stimulus-conditioned trajectories for particular inputs. I further show that the RNN's initial properties constrain the ability to acquire selective consistency.

2.2.1. Stimuli and Input Design

As my aim in this study was to simulate neural dynamics in the NRD task (Agus et al., 2010), the stimuli were generated in the same way (Fig. 3a). There were four stimulus types: N, RN, RefN, and RefRN. The N and RefN stimuli consisted of 1.0-s of white noise at a sample rate of 44 kHz. The RN and RefRN stimuli consisted of a repeated, identical noise segment concatenated twice. N and RN stimuli were generated anew for each trial. RefRN and RefN were generated in the same way as RN and N; however, their realisations were identical across all trials within each condition throughout a simulation run. The input signals were generated as follows. Firstly, each stimulus type was generated for 1.0-s (44,000 points). For RN and RefRN, the first and second segments (22,000 points) were set as the same time series. Secondly, all stimuli were filtered through an A-weighting filter, which is the most commonly used simple human auditory filter (Fig. S11) (Fletcher & Munson, 1933; Houser et al., 2017). Finally, sound signals were resampled to 2,000 Hz to reduce computational costs. For model training, five realisations per stimulus type (20 stimuli in total) were used as training data. The stimulus order was pseudorandom, except for the final presentation, which was set as one of N stimuli to avoid RefRN stimuli from being the last.

2.2.2. Network Model

Plastic Echo State Network

I used an echo state network (ESN), a reservoir computing model in which a fixed recurrent "reservoir" is driven by the input, with only a linear readout trained through learning (Fig. 3b)(Jaeger & Haas, 2004). This assumption aligns with the computational view adopted in this thesis: a local cortical circuit processes and maintains an input time series as a state trajectory, and downstream mechanisms read out that for information processing. Moreover, keeping the middle layer fixed supports real-time processing and enables fast, lightweight updates, and has therefore been highlighted as a useful abstraction of neural circuits under limited biological resources (Enel et al., 2016; Lukoševičius & Jaeger, 2009).

However, fixing the recurrent connectivity means the model's computational capability depends strongly on the reservoir's properties. To function effectively, ESN's reservoir is required to possess the echo state property (ESP) (Jaeger & Haas, 2004; Buehner & Young, 2006; Yildiz et al., 2012; Boedecker et al., 2012). In the ESN literature, the ESP can be summarised as follows: when the same input sequence $u(t)$ is applied, differences in initial conditions are faded, and the reservoir states $\boldsymbol{x}(t)$ becomes uniquely determined by the input history, $\boldsymbol{x}(t) = \Phi(u)$. In simple terms, the ESP describes asymptotic state convergence of the reservoir network under a driving input. This can be regarded as a special case of consistency as defined in this thesis. Thus, ESNs provide a natural

theoretical basis for discussing consistency in neural systems.

Although a general necessary and sufficient condition for the ESP is not fully known, one influential indicator is the spectral radius $\rho(\boldsymbol{W})$, the maximum absolute eigenvalue of the recurrent weight matrix $\boldsymbol{W}$ (Jaeger et al., 2007; Yildiz et al., 2012; Manjunath & Jaeger, 2013; Buehner & Young, 2006). When the activation function of each reservoir node is $f = tanh$, it is empirically known that ESNs satisfy the ESP for most time-series inputs when the spectral radius is below one. Conversely, as it exceeds one and moves further away, the system loses the ESP (consistency) (Rajan & Abbott, 2006). In this sense, the spectral radius corresponds to the edge of chaos (criticality) in the system dynamics (Bertschinger & Natschläger, 2004). Conveniently, because ρ is defined from the eigenvalues of the connectivity matrix $\boldsymbol{W}$, we can obtain an arbitrary spectral radius by scaling $\boldsymbol{W}$ (Lukoševičius, 2012). Therefore, although it is not a fully general criterion, I used the spectral radius as an approximate index of dynamical stability versus instability in this study.

To explore whether synaptic plasticity in the reservoir enables the system to achieve selective consistency for a learnt input, I incorporated a Hebbian plasticity rule (Oja, 1982) into the reservoir, deviating from the standard ESN (Jaeger & Haas, 2004). Alterations to the reservoir weight matrix affect the reservoir's consistency (ESP) via internal activity-dependent and unsupervised plasticity, rather than through processes such as backpropagation.

The primary simulation concept is as follows: if ESNs with plasticity exhibit greater consistency with Ref stimuli than with non-Ref stimuli after exposure to the stimulus set in the NRD task, then self-organising changes in the network may be a mechanism for acquiring selective consistency in the neural system.

Implementation

The ESN model has an input layer, a hidden recurrent layer (the reservoir), and an output layer (Lukoševičius & Jaeger, 2009).

The input signal, $u(t)$, is presented to the reservoir through the input weight matrix $\boldsymbol{W}^{in}$ from the input layer neuron following:

$$\boldsymbol{x}(t+1) = \boldsymbol{x}(t) + f\left(\boldsymbol{W}(t)\boldsymbol{x}(t) + \boldsymbol{W}^{in}u(t+1)\right) + \varepsilon, \qquad \text{(eq. 7)}$$

where $x_i(t), i = 1, \dots, T$ are the neural activations at time point t. ε replicates the internal Gaussian noise and represents fluctuations of each neuron. f is the activation function of the neurons and is defined as a hyperbolic tangent. $\boldsymbol{W} \in \mathbb{R}^{N \times N}$ is the synaptic weight matrix of the reservoir. In this study, I arranged it to be dynamic and thus replaced it with $\boldsymbol{W}(t)$. The weight matrix, denoted by

$\boldsymbol{W}^{in} \in \mathbb{R}^{N\times 1}$ connects an input neuron to the neurons in the reservoir and is generated as a set of uniformly distributed random numbers ranging from −1 to 1, fixed throughout the simulation.

The number of reservoir neurons was set to 500. The network was constructed as a sparse random network with a coupling density of $d = 0.1$. Non-zero elements of the weight matrix were defined as random numbers following a uniform distribution in the interval [−1,1].

To evaluate the impact of plasticity on reservoirs with varying degrees of consistency, I adjusted the spectral radius $\rho(\boldsymbol{W})$ of the reservoirs using the following modification:

$$\boldsymbol{W} = \boldsymbol{W}^{random}\frac{\rho^{desired}}{\rho(\boldsymbol{W}^{random})}. \qquad \text{(eq. 8)}$$

$\boldsymbol{W}^{random}$ is the weight matrix that was initiated randomly without regard for a spectral radius, and ρ^{desired} is the desired spectral radius (Lukoševičius, 2012; Yildiz et al., 2012). ρ^{desired} was generated in increments of 0.1 from 0.1 to 2.0 (see Table S1 for details).

To simulate the adaptation to the input signals, Oja's Hebbian rule was applied to the reservoir. This rule can be derived as a simple Hebbian plasticity rule that includes a forgetting factor to limit the explosion of the weight (Oja, 1982):

$$\Delta W_{ij} = \alpha x_i\left(x_j - x_i W_{ij}\right). \qquad \text{(eq. 9)}$$

The synaptic learning rate parameter α was set to 10^{-7}. Details of Oja's algorithm are available in the Oja's rule.

This study hypothesises that selective consistency is acquired through the RNN's plasticity. Consequently, in principle, an output layer is not necessarily required. However, it is not self-evident whether selective consistency is acquired across all neurons in the reservoir or is limited to a few elements. In the latter case, merely assessing the consistency of the reservoir's overall average or randomly selected neurons would not suffice to observe the phenomenon, and the optimisation process may help detect the consistent neurons. Thus, I considered the possibility of achieving selective consistency by combining weight optimisation in the output layer, as is common in reservoir computing, alongside plasticity.

The output layer has a neuron that has connections to reservoir neurons with a readout weight matrix $\boldsymbol{W}^{out} \in \mathbb{R}^{1\times N}$, and the output y of this model is calculated as a linear sum of the reservoir neurons' state:

$$y(t) = \boldsymbol{W}^{out}\boldsymbol{x}(t). \qquad \text{(eq. 10)}$$

The computational task assigned to the model was to make predictions one step ahead. This is based on the view of predictive coding, which assumes that the nervous system, especially the sensory

system, works in a predictive manner (Denham & Winkler, 2020; Friston, 2005; Rao & Ballard, 1999). Therefore, I regarded the output signal as the neural dynamics of the network, which can serve as a basis for repetition detection, but not for perception or decision itself.

To maximise the performance for the time-series prediction, $\boldsymbol{W}^{out}$ was optimised through the minibatch-based gradient descent (MBGD) (Bottou et al., 2018; Rumelhart et al., 1986). The gradient descent method is an iterative first-order optimisation algorithm used to find a local minimum/maximum of a given function. Hyperparameter η controls the rate of descent, typically set to 0.01–0.00001 and 0.01 in my simulation.

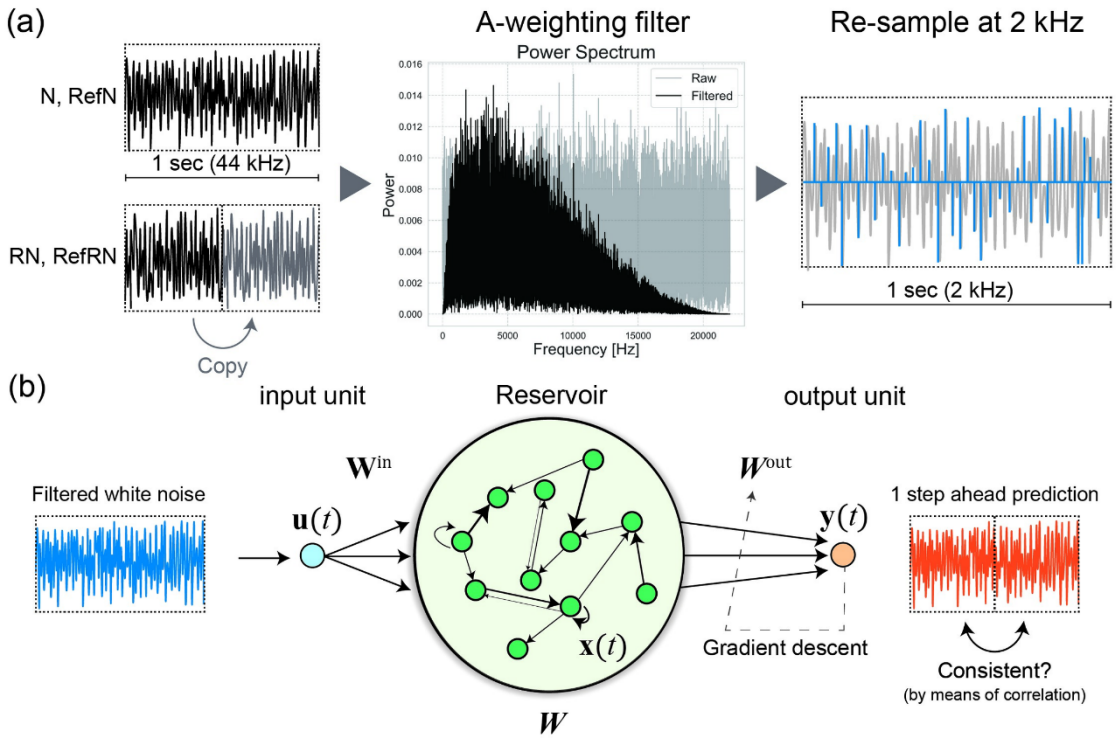

Figure 3. The model descriptions.

There are four stimulus types: N and RefN stimuli consist of 1.0-s of white noise, whereas RN and RefRN stimuli consist of 0.5-s of white noise repeated. The sampling frequency is 44 kHz. Subsequently, each time series is passed through an A-weighting filter, which reflects human auditory characteristics, peaking around 3,000 Hz and attenuating high frequencies. The middle figure shows the resulting power spectra of before (grey) and after (black) the A-weighting filter used in the simulation. After filter adaptation, each stimulus was resampled at 2,000 Hz to reduce computational costs. (b) An overview of the model. The resampled time series are presented to the neuron in the input layer as stimuli. $\boldsymbol{W}$, the reservoir weights matrix is dynamic and maintained by Oja's Hebbian plasticity rule. $\boldsymbol{W}^{out}$, the weights between the reservoir and neurons in the output layer are optimised using the gradient descent method. The model's output

target is one step ahead of the input time series. All parameter used in this study is listed in Table S1.

2.2.3. Evaluation

I evaluated changes in output time series, the degree of selective consistency, and prediction error for each trained reservoir across 200 test runs. In the 200 test runs, N and RN stimuli used are different time series from those in the training runs, while RefN and RefRN used are the same time series as in the training runs. During the test runs, I halted the Hebbian rule and the gradient descent optimisation of the output weights, and each trial was started with different initial values. This approach was chosen because, in a nonlinear network, due to the network's nonlinearity and ε—the internal Gaussian noise representing each neuron's fluctuations—the network's response exhibits different initial states and response trajectories across trials. The reproducibility of these responses across trials can serve as a measure of consistency.

Consistency

Selective consistency was evaluated by Pearson correlations between 200 time points from the first and second halves of each output time series—the first 100 ms of each of the 500-ms segments that compose the stimulus. I do not use the entire time series for both the first and second halves because if the network exhibits ESP, it will eventually show a consistent response over time, while the duration of the transient period T_{tr} is different. The fixed length of 200 time points, equal to one-fifth the length of the repetitive segment, is based on the average duration of the transient period observed in the networks evaluated in this study (see Fig. 4).

As Pearson's correlation coefficient is written as

$$\mathrm{corr}_{\mathrm{a,b}} = \frac{\mathrm{cov}(\mathrm{a},\mathrm{b})}{\sigma_{\mathrm{a}}\sigma_b} \qquad (\mathrm{eq.}\,11)$$

where cov is the covariance, $\sigma_{\mathrm{a}}, \sigma_{\mathrm{b}}$ are the standard deviation of each signal, this measure can be thought of as a special form of C in Eq.6, when available signals are only two.

Prediction Accuracy

I evaluated the accuracy of the time-series predictions learnt in the ESN. Prediction accuracy was assessed using the root mean square error (RMSE) and its normalisation (NRMSE) between the system output $y(t)$ and desired output $d(t)$. Let c index the four training/input conditions

$$c \in C_{all} = \{\text{RN Hebb}, \text{RefRN Hebb}, \text{RN static}, \text{RefRN static}\}$$

and let $k = 1, \ldots, K$ index trials, $t = 1, \ldots, T$ time points. For a given spectral radius ρ, define the time-resolved RMSE for condition c as

$$\mathrm{RMSE}_{\rho,c}(t) = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(d_{n,c}(t) - y_{n,c}(t)\right)^2}. \qquad (\text{eq. } 12)$$

Also, using a ρ-specific normalisation constant as the mean RMSE across time and across the four conditions C_{all}:

$$\overline{\mathrm{RMSE}}_{\rho} = \frac{1}{4T}\sum_{c \in C_{all}}\sum_{t=1}^{T}\mathrm{RMSE}_{\rho,c}(t), \qquad (\text{eq. } 13)$$

the normalised RMSE is simply

$$\mathrm{NRMSE}_{\rho,c}(t) = \frac{\mathrm{RMSE}_{\rho,c}(t)}{\overline{\mathrm{RMSE}}_{\rho}}. \qquad (\text{eq. } 14)$$

The definition of NRMSE is slightly different from the general normalisation of RMSE, which is done by simply dividing RMSE by the mean of $y(t)$. This is because the network's output signal tends to increase with larger spectral radii; thus, the normalisation scale must be adjusted for each spectral radius to avoid this bias. In addition, since the output time series is a white-noise-like signal with zero mean, it was difficult to normalise using a simple mean, so I used RMSE for normalisation. In this equation, normalisation is performed by dividing by the average RMSE across the four conditions: Hebbian vs. non-Hebbian and RefRN vs. RN stimuli.

Statistics

The statistical significance of selective consistency and prediction error was tested using a nonparametric rank-order test based on surrogate data (Lancaster et al., 2018). Firstly, the difference between the evaluated selective consistency and the prediction error of the Hebbian and non-Hebbian paired distributions for each spectral radius and stimulus type was calculated. Next, I randomly shuffled the two pairwise distributions and evaluated the difference between them to obtain a surrogate difference value. This was done 5,000 times to obtain a surrogate distribution of differences. The statistical significance level of the real data is determined by the percentile rank (p) of the difference in the original data relative to this surrogate difference distribution. The results of statistical tests were adjusted for multiple comparisons using the Bonferroni method, considering the number of spectral

radius levels and the combinations of stimulus types. The notation for percentile ranks (PR) used here is as follows: (*; PR < 5%, $p < 0.05$, **; PR < 1%, $p < 0.01$, ***; PR < 0.5%, $p < 0.005$, ****; PR < 0.1%, $p < 0.001$).

2.3. Results

2.3.1. Output Signals and Spectral Radius

I firstly investigated whether the network's output signal exhibited the correct behaviour as a predictive signal for a white noise time series. I plotted three representative, randomly selected time series of the responses (Fig. 4). The behaviour of the output signals was noise-like in both the Hebbian and non-Hebbian networks. I confirmed that the plasticity during the training session changed slightly but not significantly in spectral radii (Table S2). Additionally, I found that even within the same network, responses differed across distinct trials at the beginning of the output time series, a period I considered transient. This is attributed to the network's sensitivity to the initial state. This tendency became stronger as the spectral radius increased, a characteristic typical of reservoir computing. The system's consistency is defined by the length of the transient period, and initial state differences strongly influence the system's outputs. If the system is highly consistent, its response will settle onto the same trajectory after a short transient period. Therefore, we can say that both networks, with and without plasticity, show consistency across a range of relatively low spectral radii.

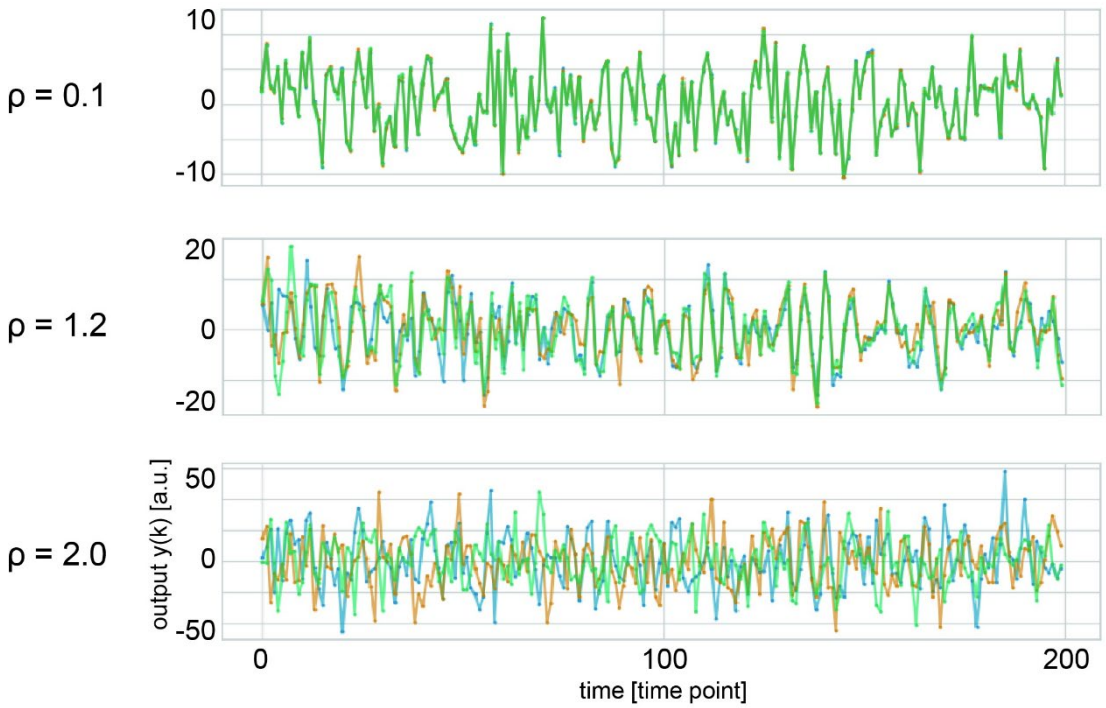

Figure 4. Representative examples of output time–series data.

These are outputs from RefRN for Hebbian networks, although a similar tendency was observed across different network and stimulus types. The results for three distinct spectral radii ($\rho = 0.1, 1.2, 2.0$) are plotted separately. Each graph plots the overlaid output values on the vertical axis against the first 200 time points on the horizontal axis. The different colours correspond to three different output trials. Time is not plotted in its entirety; instead, the first 200 points are magnified and plotted. When ρ is low, the lines converge following the transient period, indicating the identical response trajectory regardless of variations in the initial states. As ρ increases, the network no longer has ESP and behaves completely differently across distinct trials.

*2.3.2. Selective Consistency for RefRN (**H1**)*

The inter-segment correlation analysis revealed that the plastic model showed selective consistency for the RefRN stimulus, whereas RN consistency did not change. To evaluate the selective consistency for RefRN and RN, I compared the correlation between the first and second segments of output time series for RN and RefRN stimulus of both plastic and non-plastic models (Fig. 5a) and time series of five randomly selected nodes of the reservoir (Fig. 5b). Significant differences in consistency were observed between the non-plastic and plastic models for RefRN. In contrast, no

differences in RN were observed between the two models. Additionally, these tendencies were confirmed at each node of the recurrent network, indicating that selective consistency was acquired at the reservoir level, not the output layer. Also, direct comparison between RN and RefRN of the Hebbian network showed significant selective consistency for RefRN (Fig. S1). These indicate that the plasticity induced in the reservoir made the network exhibit selective consistency only for repeatedly presented input signals, while retaining the original property for other signals.

As supplementary control, the comparisons of the Pearson's correlation of both RefN and N conditions are shown in Figure S2. Since the segments in these two conditions consist of different time series, the correlation coefficients were consistently near zero, regardless of whether plasticity was present or the spectral radius was used, indicating no significant difference between the conditions. Also, inter-trial evaluation is shown in Figure S3 (very weak, but significant selective consistency for RefRN).

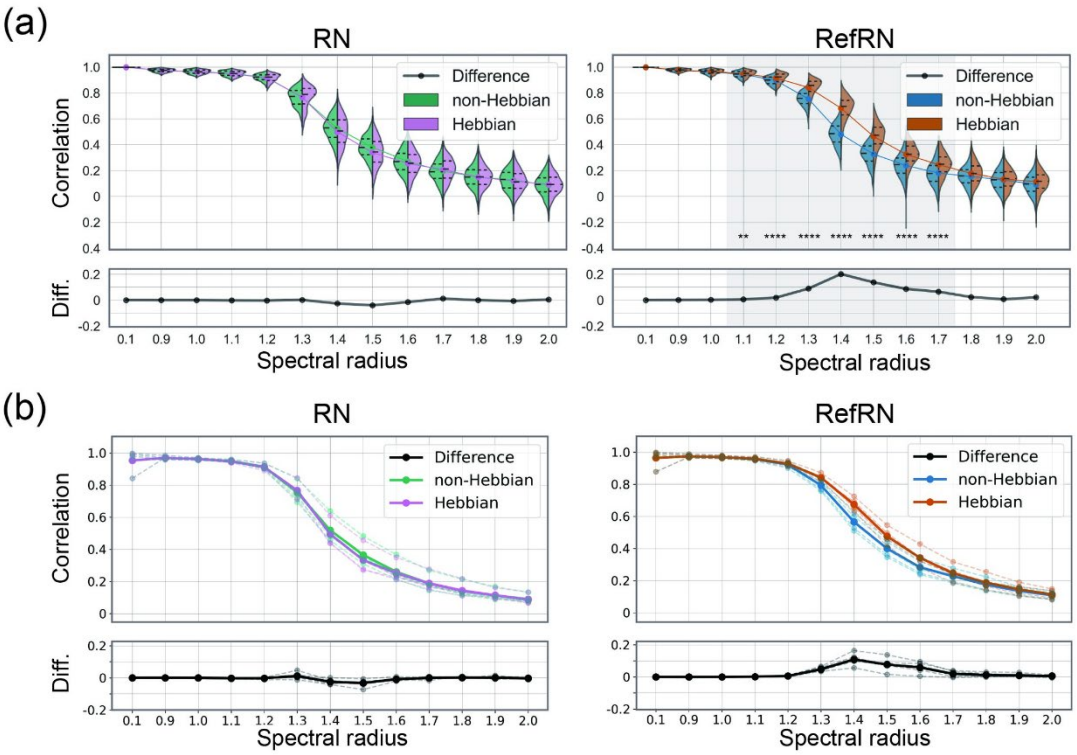

Figure 5. The evaluation of selective consistency.

The consistency was evaluated by correlation between the first and second segment time series for each test run with repeated noise (RN; left) and referenced repeated noise (RefRN; right). (a) The evaluation is at the output neuron level. The violin plots show the probability density

distributions and interquartile ranges of Hebbian (right; magenta and brown) and non–Hebbian (left; green and cyan), respectively (****; PR < 0.01%, p < 0.001). The coloured line plots connect the mean values for each condition. The black lines in the bottom windows show the difference between Hebbian and non–Hebbian models. The horizontal axis represents the spectral radius of the evaluated networks. (b) The evaluation of the five randomly selected reservoir neurons. Each dotted line represents five distinct neurons. The solid lines represent the mean value for these five neurons.

*2.3.3. Dependence on the Dynamical Regime (**H2**)*

Furthermore, the prominence of selective consistency varied depending on the spectral radius. Correlation decreased near the spectral radius exceeding 1, regardless of the presence of plasticity. Generally, reservoir consistency decreases as the spectral radius increases. Selective consistency was not acquired in less complex or more complex reservoirs with a smaller or larger spectral radius than the near-critical dynamical regime around 1.4 (Fig.5, Fig.6).

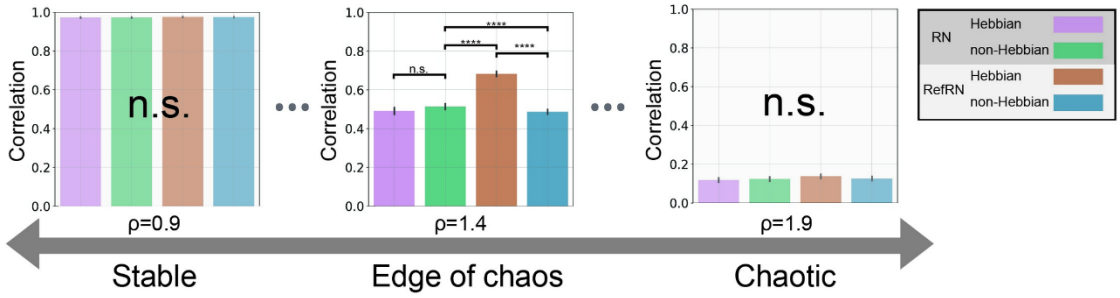

Figure 6. The edge of chaos and selective consistency.

Each histogram bin shows the averaged inter–segment correlation for four conditions (magenta; RN of Hebbian network, green; RN of non–Hebbian network, brown; RefRN of Hebbian network, cyan; RefRN of non–Hebbian network). The histograms represent results from networks with spectral radii of 0.9, 1.4, and 1.9, from left to right, which correspond to stable, edge of chaos, and chaotic, respectively (****; PR < 0.01%, p < 0.001). The error bars represent 95% confidence intervals. Notably, the edge of chaos region was chosen for its strong observation of selective consistency (see Fig. 5). Statistical significance was tested using a nonparametric rank–order test based on the surrogate data model. It was found that in networks that are either stable or, conversely, chaotic, there was no difference between conditions, and differences were observed only in the edge of chaos region.

2.3.4. Dissociating Selective Consistency from Decision

Because the reservoir's response depends on the spectral radius and input scale, optimising readout weights to achieve prediction accuracy may not be crucial for achieving selective consistency. To test this, I also conducted the same analysis for the reservoir without optimising readout weights $\boldsymbol{W}^{out}$. Figure 7 shows the selective consistency without optimising readout weights. Similar to the results with the optimisation process, significant differences in consistency were observed between the non-plastic and plastic models for RefRN, whereas no differences were observed for RN.

The prediction error for each stimulus type was not affected by the presence or absence of plasticity, spectral radius or stimulus type. As we verified above, changes to the output layer's connections during optimisation do not affect selective consistency. However, it remains unclear how the model's temporal prediction accuracy varies with different spectral radii and with the presence or absence of plasticity. Therefore, I finally examined the impact of changes in the accuracy of the time-series prediction. Figure 8 shows the prediction errors for RN and RefRN for varying spectral radii with and without plasticity. Although higher spectral radii resulted in higher prediction errors for both RN and RefRN, there were no differences in prediction error between RN and RefRN, regardless of the spectral radii or plasticity (Fig. 8a). This indicates that the prediction accuracy for repeatedly exposed stimuli (RefRN) is not selectively improved by introducing plasticity. Additionally, comparisons of NRMSE across spectral radii did not reveal prominent selective minimisation or maximisation of prediction error for RefRN (Fig. 8b).

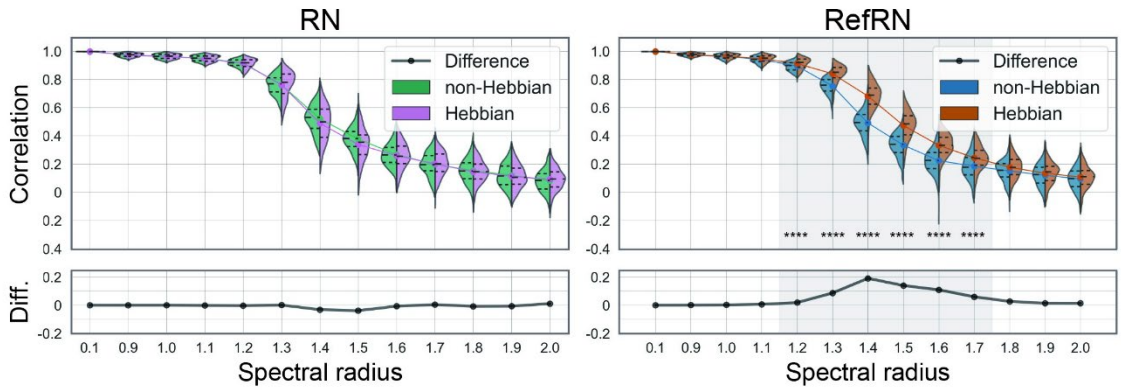

Figure 7. The evaluation of selective consistency without optimising the readout weights. The figure styles are the same as Figure 5a. The consistency was evaluated by the correlation between the first and second segment time series for each test run for repeated noise (RN; left) and referenced repeated noise (RefRN; right). The violin plots show the probability density

distributions and interquartile ranges of the Hebbian (right; magenta and brown) and non–Hebbian (left; green and cyan) models, respectively (****; PR < 0.01%, p < 0.001). The coloured line plots connect the mean values for each condition. The black lines in the bottom windows show the difference between Hebbian and non–Hebbian models. The horizontal axis represents the spectral radius of the evaluated networks.

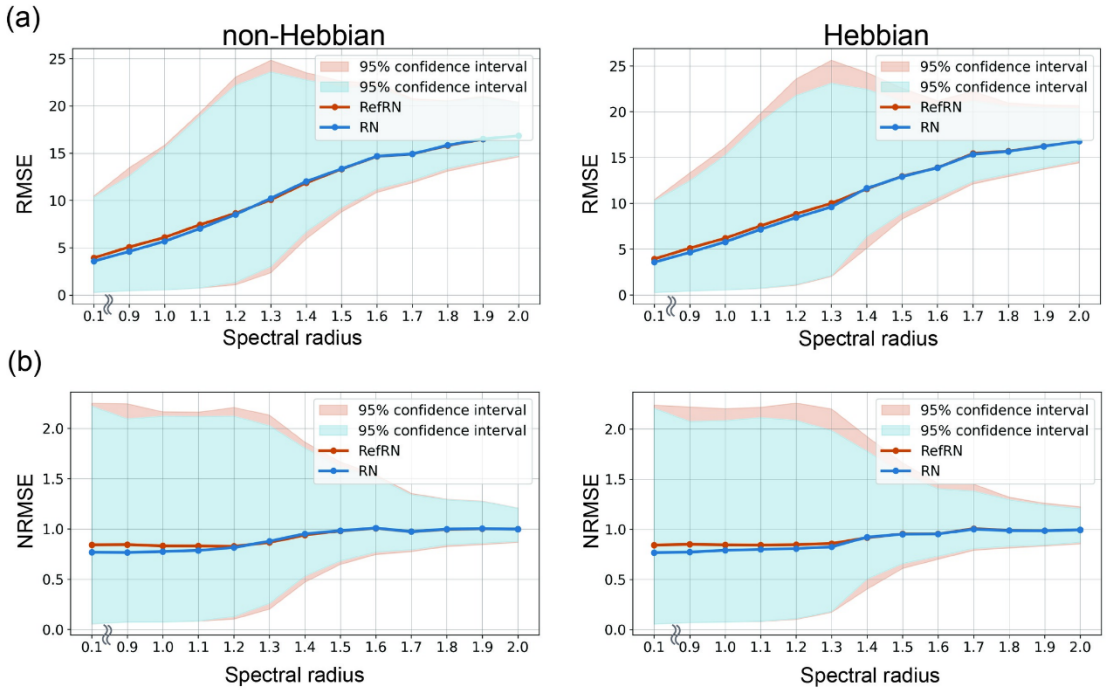

Figure 8. Prediction error with varying spectral radii.

(a) Root mean squared error (RMSE) series and (b) normalised root mean square error (NRMSE) series. The plots show the average RMSE or NRMSE between network predictions and the correct future time series. Non–plastic and plastic models are shown on the left and right, respectively. The results of RN and RefRN are represented by cyan and brown lines, respectively. The 95% confidence intervals are depicted as light cyan (RN) and pink (RefRN) filled areas.

2.4. Discussion

2.4.1. Summary of Main Findings

In the model, repeated exposure to one specific repeated-noise stimulus (RefRN) caused the

recurrent network to develop stimulus-selective convergence of activity. After learning, correlations between the early parts of the first and second segments of RefRN trials were significantly higher in the plastic than in the non-plastic network. In contrast, no such effect was observed with other noise stimuli (RN, RefN, N). Notably, this "selective consistency" was a population property: convergence was reducible in the output unit and a small subset of neurons. When I varied the spectral radius, selective consistency was maximal in a mildly supercritical regime, slightly beyond the critical boundary. In more subcritical networks, dynamics became too stable, compressing differences between stimuli. In contrast, strongly supercritical networks exhibited large trial-to-trial variability and little convergence regardless of the stimulus types. Crucially, selective consistency emerged without any optimisation algorithms, indicating that local plasticity alone can self-organise stimulus-specific dynamical stability.

2.4.2. Mechanistic Interpretation

The stimulus-conditional dynamics of a reservoir for a particular input depend strictly on the internal connectivity matrix and intrinsic noise. In reservoir computing, as adopted in the present study, the internal connectivity matrix is not updated to optimise performance. In most neural-network approaches, by contrast, the network architecture itself is optimised using rules such as backpropagation. In my modelling framework, however, the RNN was assumed to be an activity-generating and changing network whose dynamics evolve largely independently of task demands. At the same time, a separate external mechanism reads out its activity (representations) to implement task-specific decisions (here, repetition detection). The key advantage is that it allows a clearer separation of components, because the sensory network is not described as a model tailored to a particular task; in this respect, it is a better match to the brain (Enel et al., 2016).

Accordingly, we can attribute the observed change in system behaviour—identified here as acquisition of selective consistency—directly to updates of the connectivity matrix induced by the plasticity rule. Unlike standard reservoir computing, the present model allowed the internal connectivity matrix to change slightly via input-dependent plasticity, even though task requirements did not drive this change. The fact that networks without plasticity (as expected) showed no change in consistency before versus after simulation further confirms that selective consistency acquisition was realised by plasticity.

A notable aspect of the results is that consistency increased selectively only for the stimulus that had been repeatedly experienced, while the network's algebraic properties and behaviour for other inputs changed little. Similarity for stimuli other than the learnt RefRN (and RN consistency) did not change over the course of the simulation. If similarity for RN had increased alongside RefRN, this

might appear desirable at first glance, but would also imply that the system had lost context-dependent flexibility. If similarity had also increased for N and RefN, the system would no longer exhibit stimulus-specific dynamics; instead, it would produce broadly similar responses regardless of input, rendering it uninformative. Conversely, if RefRN consistency increased while responses to other stimuli became more dissimilar, the premise of task-independent dynamics—one of the key benefits of a reservoir—would be undermined. Instead, I found no such trade-offs: responses to these other stimuli remained unchanged before and after the simulation, and the spectral radius also remained stable. In other words, the system acquired stimulus-dependent consistency while preserving its overall flexibility.

Nevertheless, the specific algebraic changes that give rise to this selective effect remain unclear. Because addressing this question fully goes beyond what can be established in the present study, I develop this discussion further in Section 4.4.1.

2.4.3. Relation to the Information Processing

From an information-processing perspective, acquiring selective consistency can be regarded as an adaptive dynamical systems property that shifts stimulus-dependent system behaviour towards criticality, where information-processing capacity is maximised. selective consistency acquisition capacity was maximised in a regime that was slightly more chaotic than the critical point typically considered optimal for information processing. Post hoc, this can be explained by noting that systems with larger baseline fluctuations yield a larger change when stabilised. Functionally, however, this suggests a "physical filter" mechanism: for stimuli that are reproducible and therefore important to the system, selective consistency increases information-processing capacity, whereas for other stimuli, the system maintains flexibility and remains relatively unaffected.

Indeed, although output optimisation was not directly important for the acquisition of selective consistency, this does not mean that selective consistency is irrelevant to information processing; rather, it can play an important role. As we argued earlier, the optimisation process for the output weight does not play a crucial role in achieving selective consistency. However, the effect of selective consistency on the optimisation process warrants further discussion. In general, the computational performance of the ESN model depends on its consistency, and the network must have an ESP. This indicates that improvements in the network's computational performance (e.g., prediction accuracy) depend on the level of consistency. Therefore, in a situation where the network shows selective consistency for only some input patterns, the optimisation process is more affected by those reproducible patterns than by those with no reproducibility. This seems plausible for adapting to the environment and for efficient information processing, because the agent is not significantly affected

by non-repeated noises that are likely to be ignorable. Thus, selective consistency may help the system reconcile the impact of statistically reproducible and non-reproducible information on model optimisation based on its experience, thereby enabling efficient information processing. This linking will be discussed in Section 4.3.

2.4.4. Limitations

To avoid redundancy, I refrain from detailing project-specific limitations here and instead discuss them jointly for the two projects in Section 4.1.4.

2.4.5. Predictions for EEG/Behaviour and Motivation for Project 2

Although it is difficult to map the modelling results directly onto the EEG experiment, the present work can provide important insights into individual differences in the capacity to acquire selective consistency. As noted above, previous findings have suggested that there are individual differences—at least at the perceptual level—in the ability to acquire selective consistency (Agus et al., 2010). Furthermore, Project 1 showed that selective consistency acquisition depends on network properties. Specifically, the capacity to acquire selective consistency (the difference between consistency for RefRN and consistency for RN) was maximised when the spectral radius lay in an intermediate regime (slightly supercritical relative to criticality).

Because the brain's connectivity matrix is unknown, it is not feasible to estimate the spectral radius directly in the brain; however, from a criticality perspective, it is possible to use alternative indices (Beggs & Plenz, 2003; Hesse & Gross, 2014). Mathematically, the spectral radius is determined by the eigenvalues of the connectivity matrix and thus varies with complex factors such as the density and strength of connections and the balance between excitatory and inhibitory connections. Considering that the connection matrix in the biological brain consists of synaptic connections, structurally, spectral radius-like quantities in the brain are expected to reflect the functions of subplates during development (De Carlos & O'Leary, 1992; Friauf et al., 1990; Kanold & Luhmann, 2010), synaptic pruning, and subsequent synaptic plasticity and dendric spine morphoplasticity driven by memory and learning (Hayashi-Takagi et al., 2015; Holtmaat & Svoboda, 2009; Roberts et al., 2010), along with the resulting E-I balance (Poil et al., 2012). The appropriate expression of genes and molecules, as well as glial cells that influence these factors, is also important (Chung et al., 2013; Hirai et al., 2005; Hori et al., 2020; Kim & Kandler, 2003). Functionally, the E-I balance can change over

time, independent of structural changes, due to influences such as attention (Harris & Thiele, 2011). Thus, even without directly calculating the spectral radius, it is possible to conduct discussions within a consistent framework by employing criticality-related alternative metrics that correlate with the aforementioned biological factors (Colombo et al., 2019; Gao et al., 2017). Therefore, if the individual differences in perceptual selective consistency suggested by prior studies are confirmed, it may be possible to explain them in terms of criticality-related indices as a trait-like property of an individual's brain.

This line of reasoning motivates hypotheses ***H5***–***H6*** in Project 2. Therefore, the relationship between selective consistency acquisition and spectral radius will be further discussed in Section 4.1.3, along with the results from Project 2.

3. Project 2: Neural Selective Consistency in Human EEG

3.1. Introduction

3.1.1. Aim and Logic

In Project 2, I conducted a human EEG and behavioural experiment to test whether neural selective consistency is acquired alongside improvements in perceptual selective consistency during the NRD task, linking behavioural improvements in repetition detection to changes in neural consistency. To that end, I evaluated the relationships between task performance and both within-trial (across noise segments) and across-trial neural similarity measures (for RN and RefRN, which turn into consistency measures).

This approach addresses three gaps in the existing NRD literature. Firstly, although prior neuroscience studies reported increased ITPC for RefRN, they did not test whether these effects track learning success, correctness, or individual differences. Secondly, because NRD decisions depend on comparing noise segments within a trial, an adequate neural account requires explicit within-trial analyses. Finally, motivated by the results for ***H2*** and by broader work linking criticality-related dynamics to information-processing capacity (Cocchi et al., 2017; Del Papa et al., 2017, 2017; Hesse & Gross, 2014; Wilting & Priesemann, 2019), I examine whether individual differences in the acquisition of selective consistency relate to individual brain criticality, as evaluated with resting-state EEG.

3.1.2. Hypotheses

H3. Repetition detection is more likely when within-trial neural similarity is high

A correct response in the NRD task is a correct judgement of whether the noise segments are identical or different within a trial. Therefore, if selective consistency supports this task, within-trial neural similarity for RN and RefRN stimuli (for these stimuli, similarity equals consistency) should be higher in trials in which repetition is correctly detected, and learning should increase within-trial consistency for RefRN. A related question is whether incorrect trials for N and RefN stimuli—during which participants experienced illusory repetition—show higher within-trial neural activity similarity than trials in which this did not occur. If all of these predictions are met, I can summarise the results as follows: decisions in the NRD task depend on the similarity between neural activity patterns corresponding to the noise segments, and learning effects for RefRN arise because experience-dependent neural selective consistency (high similarity) is acquired.

Under this framework, selective consistency is assumed to arise at the level of the sensory cortex and therefore requires a discussion separate from the function that makes decisions using similar or dissimilar neural activity as input. For this reason, I primarily evaluate ***H3*** using electrodes that reflect sensory cortical activity.

Notably, there are three possible forms of increase in consistency. Firstly, the similarity observed at the single-trial level may increase with learning. Secondly, single-trial similarity may remain unchanged, while the frequency of similar dynamics may increase. Thirdly, both may occur. In all cases, the selective consistency formulation is satisfied.

H4. ITPC for RefRN increases preferentially in sessions where learning is successful

If selective consistency is correct, across-trial neural consistency (ITPC) shows a *stimulus* × *learning* interaction: the RefRN–RN difference in ITPC is larger in sessions with behavioural learning than in sessions without learning, and may further depend on correctness. Previous work investigating the neural mechanisms of the NRD task reported elevated low-frequency ITPC over temporal and parietal regions during RefRN listening (Luo et al., 2013; Andrillon et al., 2015). ITPC indexes the across-trial similarity in brain activity at the phase level. Therefore, these results suggest selective consistency for RefRN. However, even when learning occurred, there were still some missed trials, and the sound waves of RNs differed across trials; it is unclear whether higher ITPC reflects the cognitive processes of repetition-detection and learning. Although Andrillon and colleagues reported a correlation between ITPC and a behavioural measurement of accuracy and reaction time (Andrillon et al., 2015), rigorous evaluation requires splitting the dataset by correct versus incorrect trials and assessing how the presence or absence of learning affects the results.

H5. Individuals vary substantially in their ability to acquire neural and perceptual selective consistency

To understand the neural basis of the NRD task, not only session- and trial-level variability but also individual differences are crucial. I position ***H5*** as a prerequisite hypothesis for H6. Although Agus and colleagues discussed individual differences in NRD learning in the original paper, subsequent studies have not explicitly addressed this. In Project 2, I first test whether I can reproduce these individual differences at the behavioural level (***H5-1***). If I can, I then test whether corresponding differences are also present in the neural selective consistency examined in ***H3*** and ***H4*** ***(H5-2)***.

If ***H5*** is satisfied, the working hypothesis that selective consistency can account for learning in the NRD task will be largely supported.

H6-1: People with relatively subcritical brains exhibit higher consistency

H6-2 Selective consistency shows an inverted-U relationship between spectral exponent and is maximised in a slightly supercritical regime

Finally, if ***H5*** is supported, I attempt to explain the ability to acquire neural/perceptual consistency in terms of global properties of brain activity that are not task-dependent. Specifically, I use the degree of criticality, as indicated by the results supporting ***H2***, to account for individual differences in overall consistency and selective consistency acquisition ability. In ***H2***, more subcritical systems showed robust consistency regardless of stimulus type, and near the critical regime, the capacity to acquire selective consistency showed an inverted-U shape, maximising slightly on the supercritical side. Therefore, in Project 2, I test whether a similar relationship holds between overall performance (***H6-1***) and learning effects ***(H6-2)*** and a criticality measure estimated from each individual's resting-state EEG.

I also examine its relationship with developmental-disorder trait measures, which have recently been discussed in the context of criticality. Disruptions of E–I balance, and abnormalities in connection density can alter the algebraic properties of the network and thereby change the degree of criticality (Poil et al., 2012; Vogels et al., 2011). Indeed, studies of criticality and related indices in individuals with neurodevelopmental or psychiatric conditions associated with such abnormalities have reported distributions that differ from those in healthy individuals (Bruining et al., 2020; Markicevic et al., 2020; Rubenstein & Merzenich, 2003; Shew & Plenz, 2013; Wilting & Priesemann, 2019). Because this study targets healthy adults, I cannot examine clinical relationships directly; however, as a supplementary analysis, I examine the relationships among individual differences in developmental

traits, degree of criticality, and selective consistency.

3.2. Methods

3.2.1. NRD Task Procedure

Participants

Twenty-six typically developed healthy adults with no history of neurological or psychiatric disorders participated in this study. Data from two participants were excluded for the following reasons: one due to high developmental disorder traits as determined by screening, and one due to a recording problem. The final sample size of 24 (19 females, with an average age of 34.17 ± 9.02 years, range = 21–45) was determined based on previous studies (Agus et al., 2010; Andrillon et al., 2015; Kang et al., 2018) and our pre-registered a priori power analysis, which used 12 participants (Goto et al., 2024). All participants were right-handed and had normal vision and hearing. Written informed consent was obtained from all participants after a full explanation of the procedure. The ethics committee of the National Institutes of Natural Sciences approved the study protocol.

After providing informed consent, participants completed a pre-experiment questionnaire alone that collected basic demographic information, handedness, dominant ear, musical experience, and scores on the Autism-Spectrum Quotient (AQ) and the Adult ADHD Self-Report Scale (ASRS) (Kessler et al., 2005; Baron-Cohen et al., 2001). Upon completing all experimental sessions, participants filled out a post-experiment questionnaire. This included: (1) a 1–10 rating of their overall confidence throughout the task, (2) a binary Yes/No question regarding their awareness of the presence of RefRN stimuli, and (3) an open-ended comment section about the experiment.

Stimuli (Fig. 9)

As I introduced in Section 1.4, in NRD tasks, participants are asked to discriminate between white-noise stimuli (N) and repeated-noise stimuli (RN), which contain repeated white-noise segments within trials but not across trials (Agus et al., 2010). If participants can consistently recognise RN as containing repetition, this indicates perceptual consistency; however, performance typically remains near chance level. Unbeknownst to participants, a randomly selected instance of an RN stimulus—referred to as Referenced RN (RefRN)—was presented multiple times throughout the experimental session. Previous studies showed improved repetition-detection performance in RefRN trials compared to RN trials (Agus et al., 2010; Andrillon et al., 2015; Luo et al., 2013), indicating that perceptual selective consistency for the noise segment is acquired in RefRN despite the absence of acoustic differences between RN and RefRN stimuli.

In Project 2, I developed a variant of the NRD paradigm adapted for an EEG study. As in the original paradigm, there were four stimulus types, and each segment consisted of a 0.5s segment of

white noise. The two main modifications were (i) adding task-irrelevant noise segments before and after the segments of interest, and (ii) using three repeated segments. With these changes, an RN or RefRN trial became a 2.5s stimulus, in which the central 1.5s consisted of three repetitions of the same noise segment; N and RefN became a 2.5s plain white noise.

When comparing EEG responses to repeated stimuli within a trial, it is necessary to consider onset-locked and offset-locked ERPs. During behavioural tasks, EEG typically shows a large response immediately after stimulus onset and offset (Hillyard & Picton, 1978; Parker et al., 1982). These activities are particularly prominent in sensory cortices and reflect responses locked to the events of stimulus onset and offset. Therefore, using the original NRD stimulus structure would contaminate the first segment with onset-related activity, and the last segment with offset-related activity, especially in low-frequency components. To deal with this problem, I added task-irrelevant noise segments before and after the segments of interest. These additional segments always used different noise patterns on every trial, irrespective of stimulus type.

To evaluate stimulus-related responses, such as ERPs, multiple trials—usually tens to hundreds—are required to obtain precise results. However, in this study, I compare EEG activity across time windows corresponding to auditory noise segments within a trial. With only two segments, we cannot determine whether low similarity across EEG segments reflects a genuine physiological tendency or an inevitable measurement noise. In addition, because there is a temporal lag between stimulus onset and its EEG reflection, it is unclear which time point within an analysis window corresponds to the onset of an auditory noise segment, further reducing the precision of segment comparisons. To address these issues and improve the reliability of the consistency measure, I used three noise segments: providing three combinations of across-segment comparison.

Training Session

Before the first experimental session, participants underwent a brief training session. After the task explanation, participants were initially presented with demonstration sounds composed of ten 0.5-second noise segments, forming 5-second RN or N stimuli. Once a participant successfully detected RN in three consecutive trials, the number of segments was gradually reduced to 8, 6, 5, and finally 4. Participants were not exposed to the three-segment condition used in the actual experimental sessions. Upon completing the four-segment condition, participants proceeded to the main experiment. The duration of each participant's training session was recorded.

Experimental Session

In the experimental sessions, participants were asked to distinguish between Ns and RNs for each trial, which consisted of a 2.5-second auditory stimulus followed by a 1–2-second temporal interval. After the interval, participants responded immediately in a 2AFC manner using their right hand on a

mini keyboard, pressing '1' when they perceived repetition and '2' when they did not. After the button press, the screen transitioned to a 'Rest' phase, allowing participants to blink. The subsequent trial began upon pressing a button again, with an inter-stimulus interval jittered between 1 and 2 seconds.

The experiment included three sessions, each consisting of 120 trials: 30 Ns, 30 RNs, 30 times RefN, and 30 times RefRN. Thus, participants were exposed to three different RefRNs across sessions—each RefRN varied across sessions but was identical across participants. Within each session, trials were pseudo-randomised in blocks of 20 (5 trials per condition) to avoid biased stimulus presentation. The stimulus order regarding sessions was also randomised across participants. Randomisation was performed by applying MATLAB's '*randperm*' function 10 times. To keep their attention and alertness, a forced resting period of at least 1 minute was inserted between sessions.

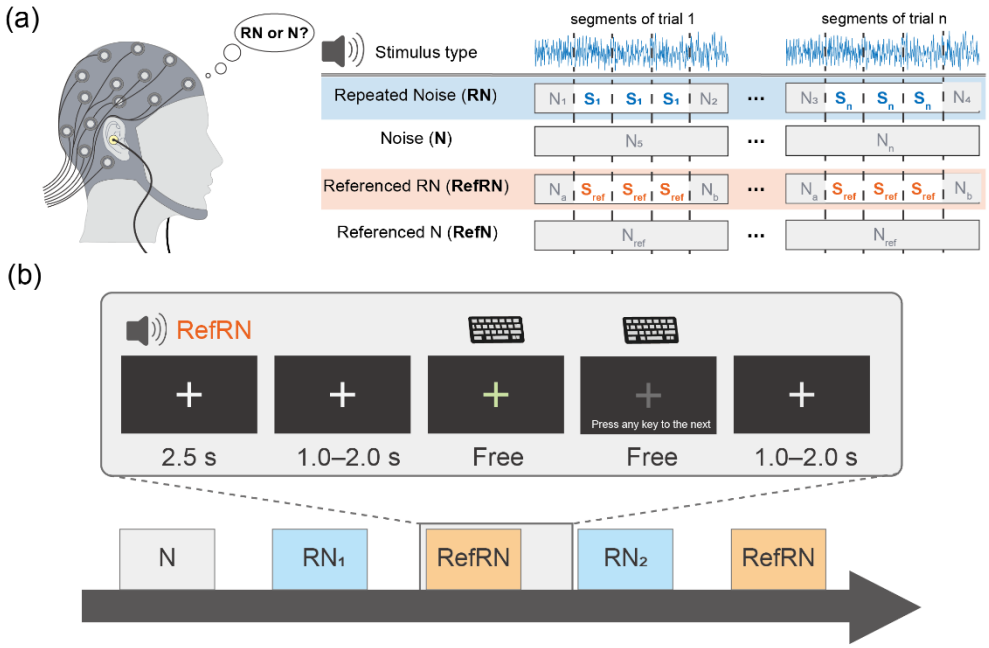

Figure 9. NRD experimental procedure.

(a) Twenty-four typically developing healthy adults participated in an EEG-adapted version of the NRD task. As in the standard paradigm, four stimulus types were used, defined by the presence or absence of within-trial repetition and of across-trial repetition. Each noise segment comprised three repetitions of a 0.5-s segment. To prevent contamination by onset- and offset-related ERPs, 0.5-s "washing" segments unrelated to the task were appended both before and after the stimulus. All washing segments were generated randomly, both within and across trials. The experiment consisted of three sessions; each included 30 trials for each stimulus type (120 per session). (b) Example trial sequence. Each stimulus was presented for 2.5-s. After a 1-

second wait, participants provided their response. Following the response, a brief pause was provided, and the next trial began after a 1–2-s inter-trial interval (ITI).

Behavioural Data Analysis

Repetition-detection performance was analysed using measures from signal detection theory. To compute d′ for RN and RefRN, I used their respective HR, together with a common false-alarm rate (FAR) derived from responses to N and RefN.

In addition to performance, I defined a categorical label for each trial indicating whether the participant perceived a repetition. To examine the relationship between neural activity and perception, accuracy-based labelling alone is insufficient given the nature of the stimuli. For RN and RefRN, a correct response corresponds to repetition being perceived, whereas for N and RefN, an incorrect response corresponds to repetition being perceived (illusory repetitive perception). Accordingly, I labelled correct and incorrect trials for RN and RefRN as "*Perceived*" and "*Not perceived*" respectively, and reversed this mapping for N and RefN.

Reaction time was measured as the interval between the onset of the response screen and the button press. In a 2AFC task, reaction time is typically used as a proxy for response confidence (Johnson, 1939; Pleskac & Busemeyer, 2010). Therefore, an ideal design would instruct participants to press the button as soon as they have made a decision after stimulus presentation. However, in this study, I introduced a minimum 1s jitter between stimulus offset and the response period to avoid overlap between motor-related EEG (and EMG noise) and the brain activity during repeated-sound listening, which was the target of analysis. This may weaken the usefulness of reaction time as an index of confidence.

Apparatus

The experiment was conducted in a shielded room to reduce external magnetic noise, using Psychtoolbox-3 (Kleiner et al., 2007) under conditions designed to minimise EMG contamination and noise from the experimental equipment as far as possible. To avoid acoustic and electrical noise, Auditory stimuli were presented via EEG-compatible air-tube earphones (ER-2 Tubephones, Etymotic Research, Inc.).

3.2.2. Recording Method for Neural Activity

EEG is one of the most suitable measurement methods for an initial empirical test of selective consistency. Methods for evaluating neural activity range from invasive to non-invasive, and from

micro- to macro-scale. In this study, I define selective consistency not as implemented by specific detectors, but as a property of the response trajectories of local neural circuits. Therefore, what is required is not cell-level recording but macro-scale observation at the circuit level.

Given the complexity of the task, I targeted humans in this initial study. Moreover, based on ***H5***–***H6***, the ability to acquire selective consistency may be associated with specific developmental/neural conditions, so a sample of typically developing healthy adults is preferable. These considerations point to non-invasive methods in humans. Candidate techniques include EEG, MEG, and fMRI; however, fMRI has low temporal resolution and is not well-suited to separating neural activity patterns evoked by 0.5-s sensory stimuli that lack clear distinguishing features. For these reasons, I used an EEG available in my laboratory. EEG approximately reflects the summed activity of neuronal populations near each electrode and provides high temporal resolution, allowing me to capture differences in activity patterns even when the active regions are similar. Although the spatial resolution differs, this situation is analogous to that in Project 1: the RNN corresponds to a neuronal population, and the signals from the output nodes correspond to EEG signals recorded at each electrode.

Recording

The EEG data were recorded at a sampling rate of 2000 Hz using an actiCHamp 64-channel system (BP-100-2115, Brain Products GmbH). Sixty-four electrodes were positioned according to the international 10–10 system. The recording reference was the average of the two earlobes, and the ground electrode was placed on the mastoid process. Each electrode was electrically coupled to the scalp using conductive gel, and recording commenced only after the impedance at each electrode had stabilised below 5 kΩ.

All experiments were conducted in a soundproofed, electrically shielded dark room. Participants were seated with their heads supported by a chin rest at a viewing distance of 60 cm from a monitor (ZOWIE XL2546, BenQ; 1920 × 1080 resolution; 100 Hz refresh rate; display size, 54.6 × 30.3 cm).

To monitor ocular artefacts, vertical and horizontal electrooculograms (EOG) were also recorded. For the EOG recordings, conductive paste was applied to the electrodes, which were then secured to the face using medical tape. After completion of the three experimental sessions, participants underwent a 3-min eyes-closed resting-state recording.

Pre-processing (Fig. 10a)

EEG data were preprocessed in MNE-Python (Larson et al., 2025). Raw EEG data were resampled to 1,000 Hz, band-pass filtered between 1 and 70 Hz, and notch-filtered at 60 Hz and its harmonics

(120, 180, and 240 Hz). Task-related epochs were extracted from −1.5 to 3.0 s relative to stimulus onset for each condition and baseline-corrected using the pre-stimulus interval (−1.5 to 0 s). Independent component analysis was then performed using FastICA (Hyvarinen, 1999) with 62 components. Components associated with ocular and muscle artefacts were automatically identified using the ica.find_bads_eog and ica.find_bads_muscle functions and were subsequently removed. An automated bad-channel detection step was also applied, and any identified channels were interpolated. Finally, noisy epochs were rejected using amplitude criteria (EEG: 100 µV; EOG: 200 µV) within the −1.0 to 0.2 s interval, together with a flatline criterion for EEG channels (1 µV).

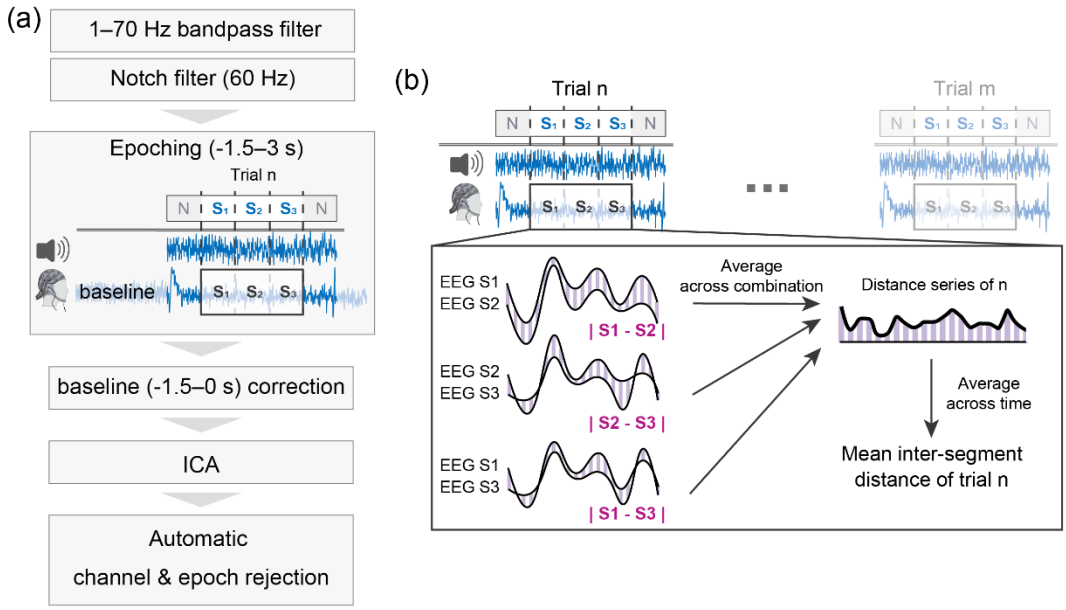

Figure 10. EEG analysis pipeline.

(a) EEG pre-processing pipeline. The recorded EEG data were band-pass filtered between 1 and 70 Hz and notch-filtered at 60 Hz, then segmented into epochs from −1.5 to 3 s relative to stimulus onset. After baseline correction using the pre-stimulus time window, noise was removed using ICA and an automatic noise-detection method. (b) Computation of the L1 distance. For each frequency band, channel, and trial, I computed the absolute error between the time series of the three segments. The resulting L1-distance time series was then averaged over time to enable trial-wise comparisons, yielding a single representative value for each trial.

L1 distance: Within-trial Consistency Measure (Fig. 10b)

To quantify within-trial similarity of the EEG signal, I computed an L1-based difference measure between three EEG segments of the same trial, separately for each condition (N, RN, RefN, and

RefRN). For a given frequency band, epochs were again narrow-band filtered in small steps across the band (using repeated filtering at $f \pm 0.1$ (delta and theta) or $f \pm 1.0$ (alpha and beta) Hz within the frequency band, see Region and frequency of interest selection). From each trial, I extracted 0.5-s windows and computed the pointwise absolute difference between pairs of windows separated by 0.5-s and 1.0-s (specifically, comparing windows starting at 0.5-s vs 1.0-s, 1.0-s vs 1.5-s, and 0.5-s vs 1.5-s). These absolute differences were accumulated across the frequency samples and then averaged across both frequency and the three within-trial comparisons, yielding a channel × time-series (within the 0.5-s window) L1-difference profile for each session and trial.

Let $\mathrm{EEG}^{l}_{\mathrm{ses},k,ch}(\omega;\tau)$ denote the band-limited EEG segment of time window ω for session ses, trial k, and channel ch after narrow-band filtering around a frequency sample f_l. For a fixed window length of $\tau \in T_\omega$ (here, $T_\omega = 0.5$-s), we can quantify across-segment dissimilarity time-series $d^{l}_{ses,k,ch}(\omega_1,\omega_2;\tau)$ by the pointwise absolute difference between two segments:

$$d^{l}_{ses,k,ch}(\omega_1,\omega_2;\tau) = \left|\mathrm{EEG}^{l}_{\mathrm{ses},k,ch}(\omega_1;\tau) - \mathrm{EEG}^{l}_{ses,k,ch}(\omega_2;\tau)\right|. \qquad \text{(eq. 15)}$$

Below, for the sake of simplicity, I will not explicitly write ses, k, ch. Under this paradigm, as we have three combinations of t_0 for two time windows to compare: $\mathrm{W}_{\mathrm{comb}} = \{(0.5,1.0),(1.0,1.5),(0.5,1.5)\}$ s for segment comparisons, we can obtain the EEG L1 metric for a stimulus s, $\mathrm{D}_{\mathrm{L1}}(\tau;s)$ of a channel by averaging d^{l} across all frequency samples $\{f_l\}_{l=1}^{L}$ within the frequency band of interest and across the three pairs as follows:

$$D_{\mathrm{L1}}(\tau;s) = \frac{1}{3L}\sum_{l=1}^{L}\sum_{(\omega_1,\omega_2)\in \mathrm{W}_{\mathrm{comb}}} d^{l}(\omega_1,\omega_2;\tau), \qquad where\ \tau \in [0,T_\omega]\ . \qquad \text{(eq. 16)}$$

This yields an L1-distance time series (channel × τ) for each trial. Note that Eq. 16 is a sort of dissimilarity measure; lower values indicate greater within-trial similarity of neural activity.

If we also average $\mathrm{D}_{\mathrm{L1}}(\tau;s)$ over $\tau \in [0,T_\omega]$, we can obtain a proxy of C_{cost} in Eq. 5 as:

$$C^{L1}_{cost}(s) = \frac{1}{T_\omega}\sum_{\tau=0}^{T_\omega} D_{L1}(\tau;s)\,. \qquad \text{(eq. 17)}$$

If the EEG segments are perfectly consistent, Eq. 16 equals zero. Conversely, we can define $C^{L1}_{cost}(s)$ as a surrogate index of $C_{cost}(s)$ in a discrete-time form with a small number of samples. In principle, it is also possible to use the variance across the three segments, like the original $C_{cost}(s)$, but with few samples the variance has large estimation error, and its distribution becomes strongly skewed, leading to low precision. For this reason, I use D_{L1} in this study.

cITPC: Across-trial Consistency Measure

Consistency across trials was quantified using inter-trial phase coherence. For each condition (N, RN, RefN, and RefRN), epochs were narrow-band filtered in small steps across the delta-theta frequency range of interest (implemented as repeated filtering at $f \pm 0.1$ Hz within the frequency bands of interest, see Region and frequency of interest selection). The analytic signal was obtained via the Hilbert transform. Instantaneous phase was extracted at each channel and time point, and ITPC was computed as the magnitude of the circular mean of unit phase vectors across trials, then averaged across the sampled frequencies within the band.

To minimise sample-size bias, I corrected ITPC estimates for the number of trials. Because ITPC is the mean of unit vectors, its finite-sample estimate does not become exactly zero even when phases are random, and it tends to take larger values with smaller sample sizes (M. X. Cohen, 2014). This issue is particularly important in this study because I compute ITPC separately by stimulus type and by correctness, and, as we will see in Section 3.3, some participants produced extremely few incorrect trials for RefRNs. To address this, I implemented a bias-correction method for ITPC, cITPC, which I later found to be equivalent to the *pairwise phase consistency (PPC)* transformation proposed by Aydore and colleagues (Aydore et al., 2013).

The standard across-trial phase-locking measure called ITPC or PLV is determined as:

$$\mathrm{ITPC}(t,f,ch) = \left| \frac{1}{K} \sum_{k=1}^{K} e^{i\theta}(t,f,ch,n) \right|. \qquad \text{(eq. 18)}$$

Because the uncorrected ITPC depends on trial number (K) and therefore introduces different biases across conditions, I used the following expression, derived by the mathematical procedure described in the Glossary (Corrected ITPC), as the bias-corrected ITPC measure in this study.

$$\mathrm{cITPC}(t,f,ch) = \frac{K}{K-1}\left(\mathrm{ITPC}^2(t,f,ch) - \frac{1}{K}\right). \qquad \text{(eq. 19)}$$

This correction reduces the sample-size bias of ITPC. It adjusts its expected value to zero under the null hypothesis (H_0: ITPC = 0), and it is mathematically equivalent to PPC (Vinck et al., 2010). Note that whereas standard ITPC ranges in $[0,1]$, cITPC takes values in the range $-\frac{1}{K-1} \leq PPC \leq 1$ (Fig. S4). Consequently, statistically detectable differences between two conditions do not by themselves warrant the interpretation that either condition shows meaningful phase locking; in particular, when the cITPC values in both conditions remain close to 0, I do not interpret the effect as evidence of phase alignment. Finally, cITPC was averaged within each ROI by averaging channel-wise cITPC values.

Spectral Exponent: Criticality Measure

As a criticality-related measure, I initially used the spectral exponent (SE), which is the simplest proxy of criticality (Beggs & Plenz, 2003; Beggs, 2008; de Arcangelis & Herrmann, 2010; He, 2014a; Donoghue et al., 2020), although it is not a sufficient enough condition of criticality (Bédard et al., 2006). Data from the parietal region (see ROI selection) was analysed after preprocessed resting-state EEG data (rereference, filtering, ICA, noise extraction). For each channel, I estimated the power spectral density (PSD) using Welch's method (Hann window; segment length $2 \times f_s$; overlap $f_s/2$, where f_s is sampling rate 1,000 Hz). The PSD was restricted to $0 < f < 70$ Hz. To stabilise the linear fit on log–log axes, the spectrum was re-sampled on a log-frequency grid by linear interpolation in $(\log_{10} f, \log_{10} P)$ space. The spectral exponent was then obtained by ordinary least squares regression of $\log_{10} P(f)$ on $\log_{10} f$:

$$\log_{10} P(f) = \beta \log_{10} f + b, \qquad \text{(eq. 20)}$$

where β is the fitted slope. Note that under the common parameterisation $P(f) \propto 1/f^{\alpha}$, the relationship is $\alpha = -\beta$.

Region and Frequency of Interest Selection

The regions of interest (ROIs) were defined based on previous studies. Prior work investigating the neural mechanisms of the NRD task has reported that ITPC specific to RefRN is observed over parietal sites (Andrillon et al., 2015; Luo et al., 2013). In contrast, studies examining neural activity when the stimuli are presented to anaesthetised mice (Kang et al., 2021), as well as fMRI BOLD "decoding" studies (Kumar et al., 2014), have reported contributions from the auditory cortex. Therefore, in the present study, I regarded the auditory cortex as the locus of the Representation stage and the parietal region as the locus of the Decision stage involved in repetition detection, and defined these as ROIs. Additionally, because decision processes often involve frontal areas, I included a frontal region. As a result, three ROIs were defined. For the specific electrode composition, see Figure S5b–d. To validate the defined ROIs, I checked the conventional ITPC topographical map for the obtained data and confirmed their similarity to those reported in previous studies (Figure S5a) (Andrillon et al., 2015).

The frequency of interests (FOIs) differed across analyses. Previous ITPC studies have reported involvement of the delta band (Andrillon et al., 2015). Accordingly, I defined the main FOI for ITPC as delta (0.5–4 Hz).

The L1-distance analysis was conducted exploratorily in the delta ([0.5,4) Hz), theta ([4,8) Hz), alpha ([8,13) Hz), and beta ([13,30) Hz) bands. Although a mouse ECoG study has suggested

involvement of the beta band, it should not be directly applied to humans, given species differences (Kang et al., 2021). And the RefRN decoding study used fMRI, which is not frequency-dependent (Kumar et al., 2014). Moreover, beta is a plausible candidate for the frequency bands dominant in sensory processing, but it is also necessary to examine its coupling with the low-frequency components observed in ITPC (Andrillon et al., 2015, 2017; Luo et al., 2013). Therefore, in this study, I assumed that high-frequency L1 reflects Representation-related components, whereas low-frequency L1 is assumed to reflect Decision-related components. Then I performed independent analyses across a broad range of frequency bands.

3.2.3. Statistics

Behavioural Measures

The learning effects of RefRN were assessed by comparing behavioural performance (HR, d', and RT) between RefRN and RN. Because a different RefRN exemplar was used in each session, learning was treated as session-specific. For each behavioural outcome, I firstly summarised performance at the session level for each condition.

In addition to whole-session performance, a learning-focused index was computed from the last 20 RefRN trials in each session (HR and d′) to capture late-session asymptotic performance, which was used for subsequent learner stratification (Agus et al., 2010). For group comparison between those obtained "well-learners" and "poor-learners", Welch's t-test, χ^2 test, and Fisher's exact test were used (Welch, 1947).

Overall group-level effects were tested using repeated-measures analyses with Session (1–3) as a within-participant factor. For HR_{Ref} and d'_{Ref}, a repeated-measures ANOVA with within-participant factors Session and Type (RN vs RefRN), and for CRR, an analogous Session × Type model contrasting N vs RefN was used. For reaction times, a repeated-measures ANOVA with within-participant factors Session and Condition (N, RN, RefN, RefRN) was used. Significant omnibus effects were followed by pre-specified paired comparisons (RN vs RefRN for learning; N vs RefN for CRR) and, for RT, post-hoc paired comparisons across the four conditions. The effect sizes were reported as partial η^2 for ANOVA. Also, to quantify within-participant consistency of behavioural performance across sessions, the main effect of the Session in the ANOVA and the intraclass correlation coefficient (case 3) were assessed (Shrout & Fleiss, 1979).

Before paired comparisons, the normality of the paired differences was assessed using the Shapiro–Wilk test (Shapiro & Wilk, 1965). If supported, paired-samples t-tests were used; otherwise, Wilcoxon signed-rank tests were applied. For post-hoc paired comparisons, p-values were controlled using the Benjamini–Hochberg false discovery rate (FDR) procedure within each comparison family

(planned comparisons vs. RT post-hoc comparisons) (Benjamini et al., 2001; Benjamini & Hochberg, 1995). Effect sizes were reported as Cohen's *d*z for paired tests.

L1-based Consistency Measure

To test whether within-trial neural pattern similarity (indexed by the L1 distance; smaller values indicate greater similarity) predicts the subjective perception of repetition on a trial-by-trial basis, I fitted a binomial generalised linear mixed models (GLMM) with a logit link (Breslow & Clayton, 1993), using a binary dependent variable of perceptual outcome: *Perceived* (1: corresponding to Hit and false alarm (FA) trials) vs *Not perceived* (0: corresponding to Miss and CR trials). The main predictor was the within-trial L1 distance (*L1*), computed for each frequency band of interest (FOI: delta, theta, alpha, and beta) and each region of interest (ROI: temporal, frontal, and parietal). To evaluate the interaction effect between L1 and learning, a session-wise learning metric was also included as a fixed effect. *Learning* was quantified using two alternative indices: standardised HR_{Ref} and d'_{Ref} (z-scored across all participant × session observations). Subject and session identity were included as random intercepts to account for within-subject dependence across trials.

I fitted two model variants. $GLMM_{all}$ included all stimulus conditions, with condition entered as a nuisance fixed effect (sum-coded), and tested the overall association between L1 distance and perception across conditions. $GLMM_{RefRN}$ was fitted to RefRN trials only to assess whether the effects observed in the pooled data were preserved when restricting the analysis to the trained stimulus. Because HR_{Ref} and d'_{Ref} are derived from behavioural performance on RefRN and are therefore not statistically independent of the perceptual outcome, I do not interpret the main effect of the learning index (it should always be significant).

For inference, defined primary terms as (i) the main effect of *L1* and (ii) the interaction *L1* × *learning*. The multiple comparisons were controlled using Benjamini-Hochberg FDR across the full analysis grid within each GLMM family ($GLMM_{all}$, $GLMM_{RefRN}$), pooling across all FOIs and ROIs. Additionally, to minimise dependence on a particular learning index, results were considered robust only when they were significant after FDR correction for both metrics within the same FOI × ROI cell.

ITPC-based Consistency Measure

To identify time-resolved differences in cITPC, I used a two-sided, time-resolved cluster-based Monte Carlo permutation test (Maris & Oostenveld, 2007). For each comparison, a *t*-statistic was computed at each time point, clusters were formed by thresholding $|t|$ at the 97.5th percentile of the permutation-derived $|t|$ distribution, and cluster mass was defined as the sum of *t*-values within a

contiguous suprathreshold segment. Family-wise error was controlled by comparing the observed cluster masses against a null distribution of the maximum cluster mass across permutations (add-one correction), and clusters were considered significant at the cluster level ($p_{cluster} < 0.05$). In interpreting the results, I focused on post-stimulus clusters. The permutation was run 10,000 times to obtain the permuted distribution.

To avoid reporting condition differences that arose solely from fluctuations around zero, I additionally required that at least one of the two conditions exhibited a significant positive deviation from zero within the same time interval. Specifically, I performed one-sided (>0) one-sample cluster-based permutation tests against a zero null for each condition separately, using sign-flip permutations and the maximum positive cluster mass to construct the null distribution. The between-condition cluster was retained only if it overlapped with at least one significant positive cluster from these one-sample tests; to control family-wise error across the two one-sample gate tests, I applied a Bonferroni correction (α/2 for each). Unless stated otherwise, I focused on post-stimulus clusters for interpretation. Permutations were run 10,000 times to obtain stable null distributions.

Unless stated otherwise, statistical tests were two-tailed with α = 0.05, with FDR control applied as specified above for families of post-hoc comparisons and for fixed-effect terms in mixed-effects models.

Spectral Exponent

To examine whether individual differences in the SE predict task performance, I conducted two statistical analyses: (i) Pearson's correlation between the spectral exponent and mean performance across the entire task, and (ii) evaluation of a non-linear relationship between the spectral exponent and the learning effect using a quadratic regression model. The mean performance d'_{mean} and the learning effect d'_{diff} were defined as:

$$d'_{\text{mean}} = \frac{d'_{\text{RefRN}} + d'_{RN}}{2}, \qquad d'_{diff} = d'_{RefRN} - d'_{RN}. \qquad \text{(eq. 21)}$$

For interpretability, the model was centred at $SE = -1.0$, therefore $x_c = SE + 1$, and the following equation was estimated by ordinary least squares (OLS):

$$\mathrm{d}'_{\text{diff}} = \beta_0 + \beta_1 x_c + \beta_2 x_c^2. \qquad \text{(eq. 22)}$$

A nested model comparison between a linear model assessed support for non-linearity ($\beta_2 = 0$) and the quadratic model, testing whether adding the quadratic term improved model fit (partial F-test). To verify an inverted-U relationship, I conducted one-sided tests, following the Lind–Mehlum U-test procedure (Lind & Mehlum, 2010), to determine whether the derivative was positive at the left

endpoint (minimum observed SE) and negative at the right endpoint (maximum observed SE). The vertex (SE yielding the maximal learning effect) was computed from the estimated quadratic coefficients as $SE^* = -1 - \beta_1/(2\beta_2)$, and I confirmed whether it lay within the observed range.

For robustness, I performed bootstrap resampling of participants with replacement (20,000 iterations) to evaluate the distributions of the quadratic coefficient and the vertex estimate. In addition, leave-one-out (LOO) analyses were conducted to assess the influence of individual participants on the conclusions. As influence diagnostics for the quadratic regression model, Cook's distance, leverage, and studentised residuals were computed (Cook, 1977).

3.3. Results

3.3.1. Demographics

Twenty-four healthy, typically developed human participants (19 females, with an average age of 34.17 ± 9.02 years, range = 21–45) completed three NRD sessions (360 trials total; 30 trials × 4 stimulus type × 3 sessions). Demographic and questionnaire summary statistics are reported in Table 1.

Table 1. Demographics.

For non-binary variables, the Shapiro-Wilk test was used to assess normality. P-values from the Shapiro-Wilk test were then BH-FDR corrected. Several continuous variables deviated from normality (Shapiro–Wilk): age (W = 0.85, p = 0.002, q=0.006), musical training (W = 0.82, p = 0.001, q=0.006), overall confidence rating (W = 0.87, p = 0.005, q=0.011), practice time (W = 0.85, p = 0.002, q=0.006), ASRS total score (W = 0.85, p = 0.002, q=0.006), and an AQ sub score (communication: W=0.894, p=0.016, q=0.029), whereas AQ total score did not show a clear deviation (W = 0.95, p = 0.261, q=0.287).

Characteristics	***n***	**Mean**	**SD**	**Min**	**Max**	**Shapiro-Wilk**		
						stats	***p***	***q***
Gender								
Men	5							
Women	19							
Age		34.17	9.02	21	45	0.850	0.002	0.006
Musical training	14	4.83	5.82	0	20	0.818	0.001	0.006

Dominant ear								
Right	15							
Left	7							
Both	2							
Confidence		3.46	2.36	1	8	0.869	0.005	0.011
Practice time		364.67	190.94	147	890	0.846	0.002	0.006
AQ score		19.00	6.88	9	32	0.949	0.261	0.287
Social skill		4.63	2.75	1	9	0.919	0.055	0.086
Attention switching		4.00	1.64	1	7	0.958	0.392	0.392
Attention to detail		4.08	2.15	1	9	0.938	0.151	0.207
Communication		2.79	2.34	0	8	0.894	0.016	0.029
Imagination		3.50	1.64	1	8	0.942	0.184	0.225
ASRS score		2.75	2.42	9	15	0.851	0.002	0.006

Note. N=24

3.3.2. Perceptual Selective Consistency for RefRN

Learning-related Improvement for RefRN (HR and d')

The learning effect in repetition-detection performance for RefRN stimuli was replicated as previously reported (Agus et al., 2010; Luo et al., 2013; Andrillon et al., 2015). Trial-course analysis showed rapid improvement in both measures for RefRN, whereas RN performance remained comparatively stable across the experiment (Fig. 12a, b). Consistent with this pattern, a repeated-measures ANOVA (Session × Type) revealed a main effect of stimulus type (RN vs RefRN) for both HR and d', with no Session × Type interaction (Table 2). Note that the Ref stimuli examples differ across trials. Specifically, HR was higher for RefRN than RN (Type: $F_{1,23}$= 5.42, p = 0.029, ηp^2 = 0.191; interaction: $F_{2,46}$= 1.02, p = 0.367), and d' was also higher for RefRN than RN (Type: $F_{1,23}$= 12.19, p = 0.002, ηp^2 = 0.346; interaction: $F_{2,46}$= 1.46, p = 0.242). These ANOVA results were mirrored by paired comparisons averaged across sessions (Fig. 12c, Table S3): RefRN outperformed RN in HR (RN = 0.70 ± 0.18; RefRN = 0.76 ± 0.20; t = 2.328, BH-FDR q = 0.044; Cohen's dz = 0.475) and d' (RN = 0.87 ± 0.99; RefRN = 1.19 ± 1.17; t = 3.492, q = 0.006; Cohen's dz = 0.713).

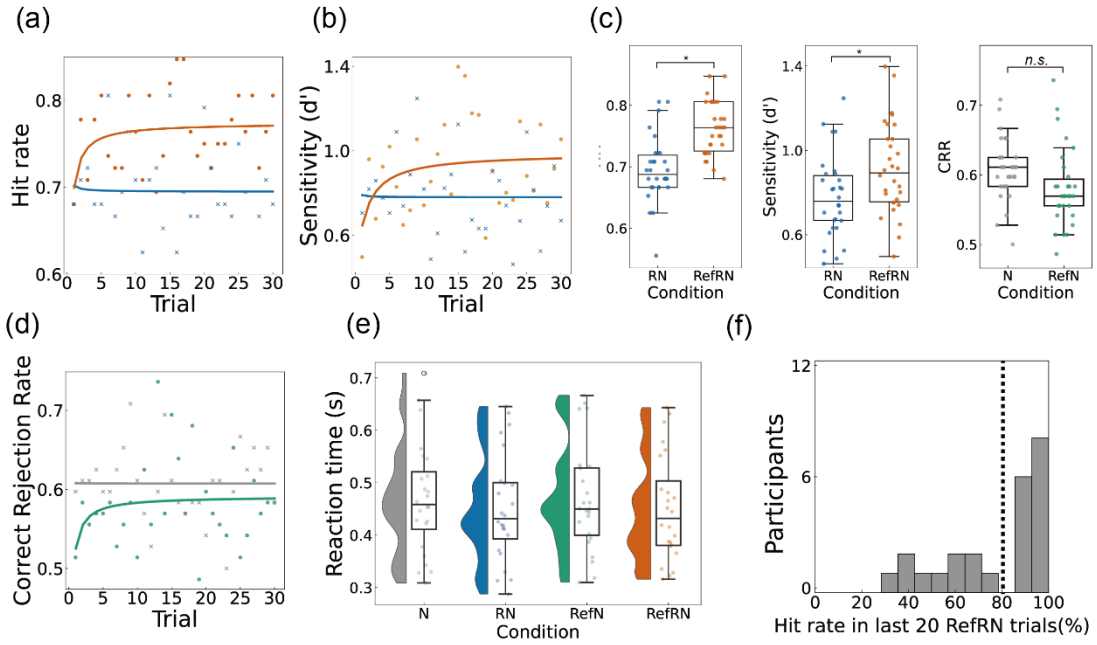

Figure 12. Behavioural results.

N, RN, RefN, and RefRN are shown in grey, blue, green, and orange, respectively. (a) Trial course of hit rate. (b) Trial course of sensitivity. (c) Paired comparisons of HR (left), d' (middle), and CRR (right). *: FDR-corrected $q < 0.05$. (d) Trial course of CRR. (e) Reaction times for each condition. Diamonds indicate outliers. (f) Participant number distribution of the learning effect. The dashed line indicates the threshold separating "well learners" and "poor learners" (HR = 0.8).

Table 2. Repeated-measures ANOVA for main behavioural performance measures across sessions and stimulus type.

Source	*df1*	*df2*	*F*	*p-value*	η^2
HR across sessions and stimulus					
Session	2	46	0.096	0.909	0.004
Type (RN / RefRN)	1	23	5.422	0.029	0.191
Session × Type	2	46	1.024	0.368	0.043
d' across sessions and stimulus					
Session	2	46	1.500	0.234	0.061

Type (RN / RefRN)	1	23	12.19	0.002	0.346
Session × Type	2	46	1.463	0.242	0.060
CRR across sessions and stimulus					
Session	2	46	0.907	0.411	0.038
Type (N / RefN)	1	23	1.241	0.277	0.051
Session × Type	2	46	0.101	0.904	0.004
RT across sessions and stimulus					
Session	2	46	24.895	4.71×10^{-8}	0.520
Type (N / RN / RefN / RefRN)	3	69	6.461	0.001	0.219
Session × Type	6	138	0.433	0.856	0.018

No Clear Learning-related Change for RefN

In contrast, the correct rejection rate (CRR) did not show a reliable RefN-specific improvement (Fig. 12d). The repeated-measures ANOVA indicated no main effect of Type (N vs RefN) on CRR (Type: $F_{1,23} = 1.24$, $p = 0.277$, $\eta p^2 = 0.051$) and no interaction with session (Table 2). Consistently, the paired comparison showed a non-significant tendency for RefN to have slightly lower CRR than N (N = 0.61 ± 0.26; RefN = 0.58 ± 0.27; $t = -1.114$, $q = 0.277$; Fig. 12c, Table S3). These results indicate that, though RefN is also a learnable ("referenced") stimulus, data showed no clear learning effect for this non-repeated, relatively long noise stimulus.

Reaction Time: Repetition-related but not RefRN-specific Speed-up

RT, used as a proxy for confidence, did not show a learning-related speeding specific to RefRN (RN = 0.45 ± 0.10, RefRN = 0.45 ± 0.10, $t = 0.348$, BH-FDR $q = 0.731$), nor a difference between N and RefN (N = 0.47 ± 0.11, RefN = 0.47 ± 0.10, Wilcoxon $W = 105.5$, $q = 0.594$) (Fig. 12e, Table S3).

However, RT exhibited systematic differences across stimulus types overall. The ANOVA showed a significant main effect of Session ($F_{2,46} = 24.89$, $p = 4.71\times10^{-8}$, $\eta p^2 = 0.520$) and Type (N/RN/RefN/RefRN) ($F_{3,69} = 6.46$, $p = 0.000642$, $\eta p^2 = 0.219$), with no Session × Type interaction (Table 2). Pairwise tests further indicated a consistent ordering of RN ≈ RefRN < N ≈ RefN: RT was longer for N than RN (BH-FDR $q = 0.017$) and longer for N than RefRN ($q = 0.010$), and similarly longer for RefN than RN ($q = 0.010$) and longer for RefN than RefRN ($q = 0.010$), but the effect size was small, on the order of 0.01-s (Table S3).

Within-participant Consistency Across Sessions

Learning-related performance was highly consistent within individuals across sessions. Consistent with no significant main effects of Session in ANOVA (Table 2) for HR ($F_{2,46}$= 0.096, p = 0.909, ηp^2 = 0.004), d' ($F_{2,46}$= 1.500, p = 0.234, ηp^2 = 0.061), and CRR ($F_{2,46}$= 0.907, p = 0.411, ηp^2 = 0.038), intraclass correlation coefficients (ICC(3,k), k = 3 sessions) indicated strong reliability for both HR (RN: 0.84; RefRN: 0.75) and d′ (RN: 0.91; RefRN: 0.83, Table S4). Thus, the learning effect for three different RefRN examples is consistent within an individual across sessions.

Individual Differences, Group Definition (H5-1)

Supporting ***H5-1***, I replicated the individual differences and bimodal-like distribution in learning performance reported by Agus et al. in their first work on NRD (Agus et al., 2010). Despite a clear and consistent within-participant learning effect for RefRN, learning performance showed pronounced individual differences and deviated from normality (Fig. 12f, Shapiro–Wilk: W = 0.85, p = 0.0026). Thus, following the plotting method of Agus et al. (2010), sessions/participants were labelled as "well-learnt" vs "poor-learnt" (data ratio: Participant level, 14/10; Session level, 43/29) based on the bimodal distribution of HR in the last 20 RefRN trials, using a threshold of HR = 0.8 (Fig. 12f).

As expected, the groups differed strongly on behavioural indices (Table S5). Well-learners showed a higher overall HR also for RN (p = 0.001, q = 0.008), not only for RefRN (the measure used to separate these groups). They also showed higher overall d′ for RefRN (p = 0.001, q = 0.007), and even when focusing on the last 20 RefRN trials, d′ remained substantially higher in well-learners (1.98 ± 1.01 vs 0.24 ± 1.11; p = 0.001, q = 0.007), corroborating the idea that grouping by HR captured stable sensitivity differences rather than noise. In contrast, FAR, CRR, and RT did not show reliable group differences (all $q \geq 0.618$).

Also, group comparisons (Table S5; FDR-corrected across variables) indicated that demographic and trait variables did not significantly differ between groups after correction. Age did not differ (Welch's t: p = 0.883, q = 1.000), and there were no reliable differences in musical training, dominant ear, or dominant hand (all q > 0.811). The sex ratio showed a trend (Fisher's exact: p = 0.053, q = 0.174), but the ratio is hugely biased. AQ total score did not differ (p = 1.000, q = 1.000), and ASRS total score showed only a non-significant tendency (p = 0.077, q = 0.221). It should be noted that the research design was not intended to detect between-group differences in these measures, and that the use of FDR correction across numerous test items, so these results do not conclude that no differences exist.

Practice time to clear training was shorter in well-learners at the uncorrected level (*Well:* 281 ± 104-s, *Poor:* 481 ± 227-s, Welch's t, $p = 0.024$), but this did not survive FDR correction ($q = 0.091$). Overall confidence did not significantly differ between groups ($p = 0.817$, $q = 0.988$).

3.3.3. Neural Selective Consistency for RefRN

Within-trial Neural Pattern Similarity and Repetitive Perception (H3, H5-2)

In $GLMM_{all}$, the L1 distance across segments was a robust predictor of repetitive perception. For all ROIs, theta, alpha, and beta L1 effect was significantly negative, meaning that smaller L1 (more similar neural activity) was associated with a higher probability of repetitive perception (Theta: temporal and frontal ROIs ($\beta \approx -0.10$ to -0.02; $q \leq 0.022$ across metrics/ROIs), Alpha: temporal/frontal/parietal ($\beta \approx -0.10$ to -0.03; all $q \ll 0.001$), and Beta: temporal/frontal/parietal ($\beta \approx -0.14$ to -0.03; $q \leq 0.049$ across metrics/ROIs)). All data are summarised in Table S6 and Table S7. When data are restricted to trials of Well-learners, this tendency becomes even stronger. As it is the most visually clear, representative FOI of the alpha band of Well-learners' data is presented in Figure 13.

In contrast, the effect of delta L1 was positive ($\beta > 0$) across all ROIs, meaning that larger L1 (less similar neural activity) was associated with a higher probability of *Perceived* ($\beta \approx +0.04$ to $+0.09$; all $q \ll 10^{-9}$ across metrics/ROIs).

When the analysis was restricted to RefRN trials (Table 3), robust L1 effects remained detectable but were no longer ubiquitous across FOI×ROI. Beta L1 remained a strong negative predictor of perception in temporal and parietal ROIs (temporal: $\beta \approx -0.24$ to -0.51; parietal: $\beta \approx -0.09$ to -0.18; all $q \leq 0.0136$, and many $q \ll 10^{-6}$). Theta showed a negative L1 effect only in the frontal ROI ($\beta \approx -0.07$ to -0.10; $q \leq 0.010$). The positive L1 effect of delta activity remained in parietal ROI ($\beta \approx +0.04$; $q \leq 0.035$).

Next, I tested whether session-wise learning modulates the L1–*Perceived* relationship via the interaction term *L1 × learning*. In $GLMM_{all}$, the interaction and its sign were robust across all FOIs and ROIs, except for temporal delta activity. Theta, alpha, and beta activities interact with the learning effect negatively, meaning that as learning increases, the negative L1 slope becomes more negative (i.e., *Perceived* becomes more strongly associated with a smaller L1). Delta interaction was positive in frontal and parietal ROIs, indicating that as learning increases, *Perceived* becomes more strongly associated with larger L1. In $GLMM_{RefRN}$, the learning modulation effects that met the strict robustness criterion were observed mainly in beta in temporal and parietal ROIs ($\beta \approx -0.21$ to -0.53; $q \ll 10^{-6}$) and delta in the parietal ROI ($\beta \approx +0.11$ to $+0.15$; $q \ll 10^{-9}$).

Overall, these results are consistent with the idea that the association between perceived repetition and neural similarity is frequency-dependent: perceived trials are linked to greater neural similarity,

especially beta in temporal and parietal areas, but to greater delta dissimilarity in parietal cortex, and these relations can become stronger with learning.

		Percived	*Perceived* × learning interaction	
FOI	ROI	β (SE)	HR β (SE)	d' β (SE)
Delta	Temoral	0.014 (0.028)	-0.007 (0.030)	0.079 (0.037)
	Parietal	0.037 (0.015) *	0.107 (0.016) ***	0.151 (0.019) ***
	Frontal	-0.027 (0.014)	0.015 (0.011)	0.012 (0.012)
Theta	Temoral	-0.053 (0.043)	-0.059 (0.048)	-0.073 (0.055)
	Parietal	-0.041 (0.021)	-0.065 (0.023)	-0.021 (0.026)
	Frontal	-0.102 (0.023) ***	-0.012 (0.023)	-0.004 (0.025)
Alpha	Temoral	-0.078 (0.027)	-0.022 (0.028)	-0.060 (0.037)
	Parietal	-0.038 (0.011)	-0.020 (0.011)	-0.009 (0.017)
	Frontal	-0.061 (0.016)	-0.014 (0.017)	0.003 (0.021)
Beta	Temoral	-0.510 (0.053) ***	-0.291 (0.052) ***	-0.526 (0.059) ***
	Parietal	-0.178 (0.032) ***	-0.208 (0.029) ***	-0.253 (0.035) ***
	Frontal	-0.070 (0.033)	-0.105 (0.030)	-0.003 (0.033)

Table 3. Results of the generalised linear mixed model for RefRN trials.

Coefficient estimates and standard errors for $GLMM_{RefRN}$ across all FOI × ROI combinations. The β for *Perceived* corresponds to the value from the d′ model. *: $q < 0.05$, **: $q < 0.005$, ***: $q < 0.0005$ for both measures after global FDR correction.

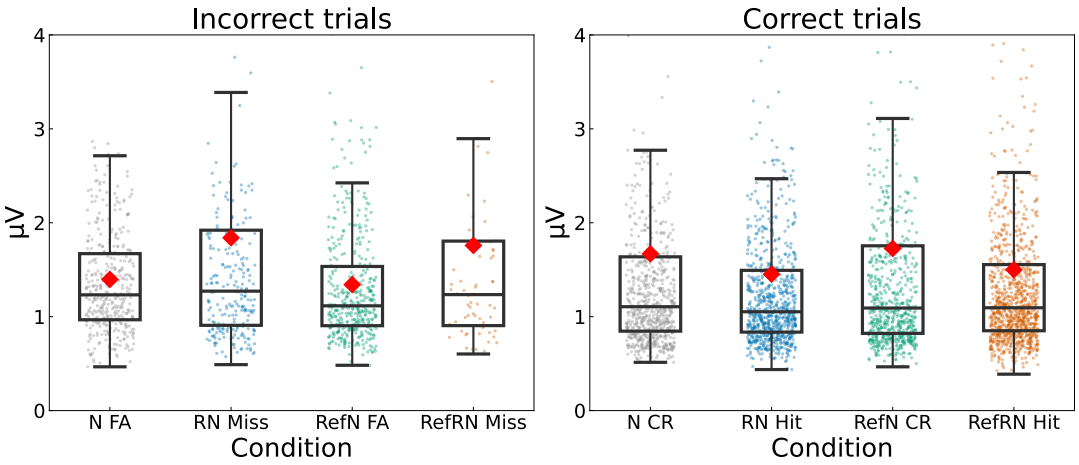

Figure 13. A typical example of L1 distance comparisons.

The well-learners' temporal ROI L1 distances of each condition are presented separately according to the correctness of their response. The left panel shows incorrect trials and the right panel shows correct trials, plotted separately for each of the four conditions. Thus, false alarms in N and RefN, as well as hits in RN and RefRN, correspond to *Perceive* trials in which participants reported hearing repetition. Dots represent all pooled trials from all participants. The black boxplots for each condition show Q1–Q3, with the horizontal line indicating the median; whiskers extend to 1.5×IQR. The red diamonds represent the mean values.

Across-trial Neural Similarity and Learning Dependence (H4, H5-2)

Scrutiny of cITPC during repetitive segments showed that an increase in cITPC during RefRN was not consistently observed, but occurred only when the stimulus was well learnt, and repetition was perceived. When I compared the cITPC time courses during the repetitive segments, the delta phase was not aligned across trials for RN hit trials, RefRN hit trials in poor learners, or RefRN miss trials in well learners; a significant increase was observed only for RefRN hit trials in well learners. However, the ROI comparison did not reveal region-specific effects.

The central finding regarding the cITPC analyses was that learning selectively shaped phase-consistency effects in the RefRN condition. In the full dataset, RefRN trials showed higher cITPC than N and RN trials in line with previous studies (Andrillon et al., 2015, 2017)(Fig. S9, N: 0.7–2.4-s, $p_{cluster} = 0.001$; RN: 0.8–2.2-s, $p_{cluster} = 0.001$). Although there were no qualitative differences across the ROIs, when trials were stratified by correctness and learning, a similar RefRN-related high cITPC cluster was present only for the Hit trials in the well-learnt subset (Fig. 14a, b). In contrast, no significant RefRN-related high cITPC cluster was detected in the poor-learnt subset, either Hit or Miss trials (Fig. 14a, c). RN trials did not show significant improvement in cITPC regardless of the correctness (Fig. 14a, b, c). All of these comparisons indicate that the high ITPC during RefRN listening reported in previous studies is observed only under a very limited set of conditions: when the stimulus is RefRN, learning has occurred, and the repetition is perceived.

For N and RefN, cITPC showed no phase alignment during listening, regardless of whether the response was a CR or a FA (Fig. 14d).

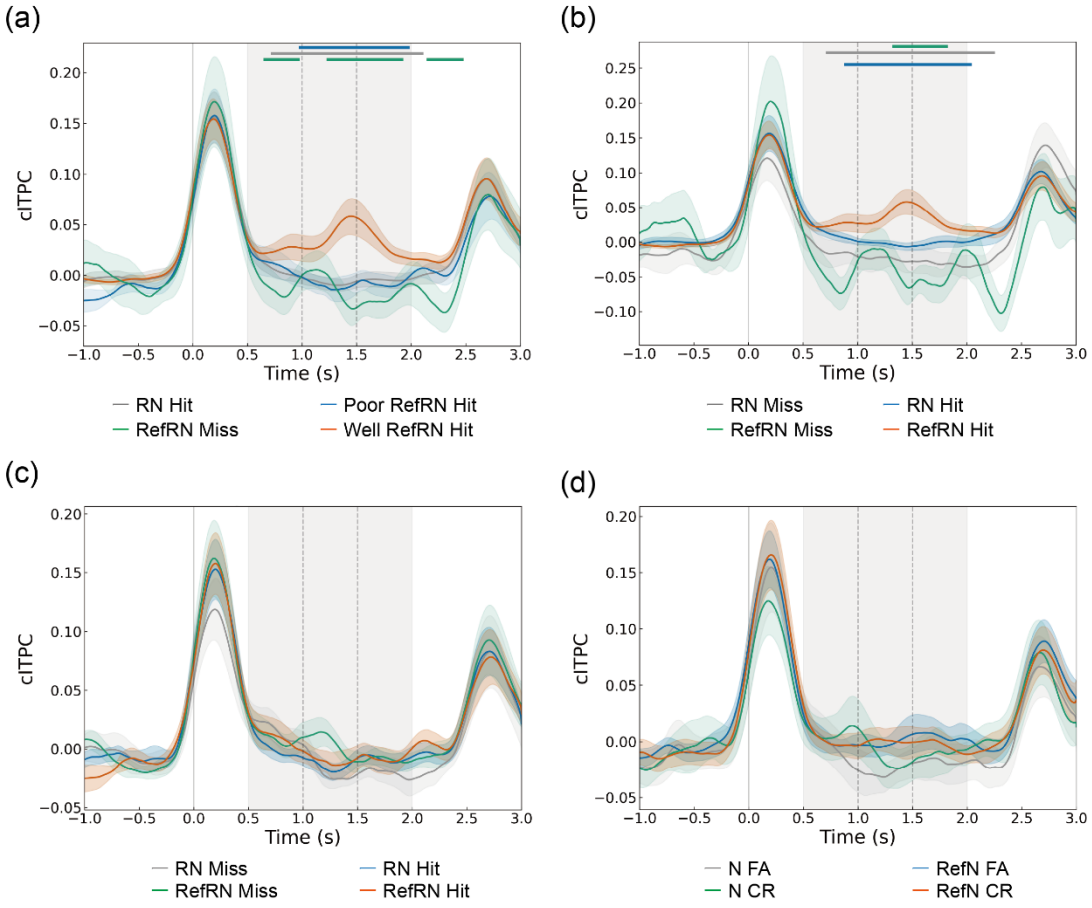

Figure 14. Main results of cITPCs comparisons from all ROIs.

(a) Condition-wise cITPC time courses for RN and RefRN stimuli. Coloured solid lines indicate: grey, Hit trials of RN; green, Miss trials of RefRN; blue, Hit trials of RefRN from "Poor-learnt"; orange, Hit trials of RefRN from "well-learnt". Shaded areas indicate the 95% confidence interval for each condition. Time intervals showing significant clusters in the cluster-based Monte Carlo permutation test for well learners' RefRN hits are marked by bars at the top ($p_{cluster} < 0.05$). The colour of each bar matches the corresponding condition (grey: RN Hit; green: RefRN Miss; blue: poor-learner Hit). All significant clusters are as follows: *Well* RefRN Hit > RN Hit: 0.7–2.1-s, $p_{cluster} = 0.001$; *Well* RefRN Hit > *Poor* RefRN Hit: 1.0–2.0-s, $p_{cluster} = 0.001$; *Well* RefRN Hit > RefRN Miss: three clusters in 0.6–2.5-s, all $p_{cluster} < 0.04$. (b) Condition-wise cITPC time courses for RN and RefRN stimuli from "well-learnt" sessions. Grey: RN Miss; blue: RN Hit; green: RefRN Miss; orange: RefRN Hit. Time intervals showing significant clusters in the cluster-based Monte Carlo permutation test for RefRN hits are marked by bars at the top ($p_{cluster} < 0.05$). The colour of each bar matches the corresponding condition (grey: RN Miss; green: RefRN Miss; blue: RN Hit). All significant

clusters are as follows: *Well* RefRN Hit > *Well* RN Miss: 0.7–2.3-s, $p_{cluster} = 0.001$; *Well* RefRN Hit > *Well* RefRN Miss: 1.3–1.8-s, $p_{cluster} = 0.003$; *Well* RefRN Hit > *Well* RN Hit: 0.9–2.0-s, all $p_{cluster} = 0.001$. (c) Same style plot as b, but for "Poor-learnt" session data. There were no significant differences across conditions. (d) Condition-wise cITPC time courses for N and RefN stimuli. Grey: N FA; blue: N CR; green: RefN FA; orange: RefN CR. There were no significant differences across conditions.

3.3.4. Resting-state Criticality Proxy Predicts Selective Consistency Acquisition (H6)

Examining the relationship between SE and overall repetition sensitivity, average d′ showed a significant negative correlation with SE (Fig. 15a, Pearson's $r = -0.415$, $p = 0.043$). That is, participants with larger SE values (i.e., a shallower slope, indicating supercritical) tended to show lower overall performance, averaged across RefRN and RN.

Also, the learning effect (d′(RefRN) − d′(RN)) and SE showed a quadratic nonlinear relationship as the model in Project 1 expected (Fig. 15b). The quadratic model showed a moderate fit ($R^2 = 0.206$). The overall model F-test did not reach significance ($F(2, 21) = 2.73$, $p = 0.088$). However, in the model, the quadratic term had a significant negative coefficient ($\beta_2 = -6.886$, $p = 0.029$). Compared with the linear model, adding the quadratic term significantly improved model fit (partial F-test: $F = 5.459$, $p = 0.0295$). The observed SE range was [−1.370, −0.617]. Testing the sign of the slope at these endpoints showed a positive slope at the left endpoint (slope = 4.658, one-sided $p = 0.020$) and a negative slope at the right endpoint (slope = −5.701, one-sided $p = 0.018$). These results support an inverted-U relationship in which the learning effect increases and then decreases within the observed range. The vertex estimated from the quadratic regression was $SE^* = -1.031$, indicating that the learning effect tended to be maximal around $SE \approx -1$.

In the bootstrap analysis, 14,652 resamples (73.3%) satisfied the inverted-U conditions and had a vertex within the observed range. Restricting to resamples meeting these criteria, the 95% interval for the quadratic coefficient was [−13.40, −2.24], and the 95% interval for the vertex SE* was [−1.135, −0.771]. In addition, the LOO analysis showed that in 23 of 24 cases (95.8%) the quadratic term remained negative and significant ($p < 0.05$), and the significance of adding the quadratic term as well as the endpoint conditions of the U-test were likewise preserved in 23 cases. However, influence diagnostics identified one participant with a high Cook's distance (4.62) and high leverage (0.787). Excluding this participant in the LOO condition resulted in the vertex moving outside the observed range, and the inverted-U conditions were not satisfied.

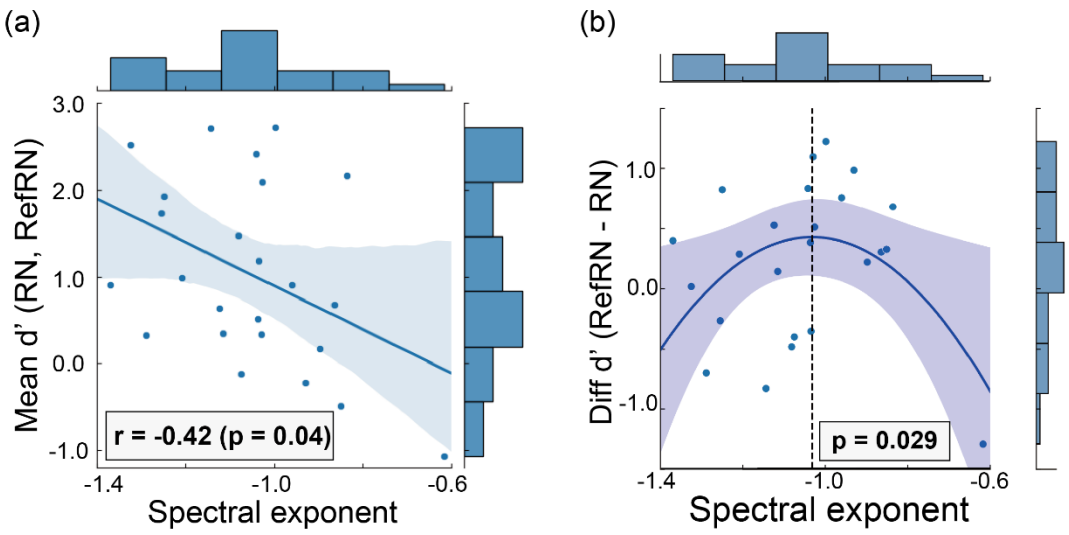

Figure 15. Relationships between spectral exponent and performance

(a) Spearman's correlation between the spectral exponent and mean task performance. Each dot represents an individual; the solid line shows the fitted linear regression line, and the shaded area denotes its 95% confidence interval. (b) Quadratic regression of learning performance on the spectral exponent. Adding the quadratic term significantly improved the fit over the linear model (partial F-test: $p = 0.029$). The Lind–Mehlum U-test satisfied the endpoint conditions, with a positive slope at the lower bound of the observed range ($pL = 0.020$) and a negative slope at the upper bound ($pR = 0.018$), supporting an inverted-U relationship. The estimated vertex is indicated by the dashed line (SE* = −1.03). Marginal histograms show the distributions of each variable.

As an exploratory analysis, I also examined associations with self-report developmental disorder trait scores. ASRS scores showed a positive correlation with overall task sensitivity (mean d′; $r = 0.55$, $p = 0.005$). Also, it was negatively correlated with SE ($r = −0.52$, $p = 0.01$), suggesting that participants with smaller SE values tended to have higher ASRS scores (Fig. S10). In contrast, the ASD main and sub scores did not show significance, even without multiple-comparison correction.

3.4. Discussion

3.4.1. Summary of Main Findings

This study aimed to determine whether neural selective consistency is acquired alongside perceptual selective consistency in the NRD task (***H3***–***H4***) and to explain individual differences in this acquisition (***H5***) in terms of criticality (***H6***).

In the within-trial analyses, I found that, regardless of stimulus type, trials in which participants perceived repetition showed higher similarity, especially in the beta band between the corresponding EEG segments (***H3***). In contrast, the delta band showed the opposite direction. In addition, these relationships become stronger as learning progresses. Also, cITPC analyses revealed that the previously reported "high ITPC for RefRNs" in the delta band was found only in sessions in which learning was successful, and only in hit trials (***H4***).

Performance across the three sessions was consistent within individuals, but comparisons across individuals revealed a bimodal distribution, with some participants showing learning and others not (***H5-1***). I also observed a clear difference in neural selective consistency corresponding to this behavioural difference in perceptual selective consistency (***H5-2***). Finally, the spectral exponent of resting-state EEG—used as a proxy for criticality—explained individual differences in the ability to acquire selective consistency, suggesting that this ability is grounded in the dynamical properties of each individual's neural network (***H6***).

3.4.2. What They Learnt in RefRN?

In this study, I assumed that learning in the NRD task reflects perceptual selective consistency, rather than changes in strategy, "memorisation" of particular waveforms, or increased sensitivity to a typical noise waveform. The behavioural analyses showed that both the HR and d' for RefRNs increased, indicating that at least some form of learning occurred. Moreover, because the ANOVA showed no session effect for either RN or RefRN, the results are not readily explained by task familiarisation or a change in strategy. In addition, because reaction time did not differ between RN and RefRN, it is also unlikely that participants became more sensitive to the waveform of RefRN per se rather than to repetition. Taken together, these findings suggest that RefRN learning can reasonably be interpreted as perceptual selective consistency.

It is also important that the CRR for RefN did not increase, despite RefN being a Ref stimulus. To obtain a unique representation for a given sensory input, not only stimulus identity but also discriminability is required. Consistency guarantees a mapping from external stimuli to internal representations, but without injectivity it cannot be regarded as a complete internal model. Whether

injectivity is acquired would be reflected in the correct-rejection rate for RefN; however, because RefN has twice the duration of the waveform that would need to be "remembered" in RefRN, it is unclear whether the present results indicate a failure to acquire discriminability for noise stimuli. I discuss this issue in detail in the General Discussion.

3.4.3. Two-stage Model of Repetition-detection and Neural Correlates

The within-trial and across-trial evaluations of neural selective consistency obtained in this study can be explained consistently using the Representation–Decision stage model of the NRD task proposed in the General Introduction.

Within-trial similarity of the beta band mainly reflected the Representation stage. The L1 analysis results showed the intuitive pattern that consistency was higher in hit trials for both RN and RefRN, in which participants correctly perceived repetition for these stimuli. At the same time, within-trial neural similarity was also higher in FA trials for N and RefN ("illusory repetition"). When L1 was evaluated across all stimulus types, theta, alpha, and beta bands showed similar trends across all ROIs. These results indicate that, regardless of the actual sound, when sensory representations were similar, perception also tended to be similar. Moreover, in the RefRN-specific analysis, the beta band—being the only band in which L1 decreased with learning—also matched the ECoG band observed in the auditory cortex of anaesthetised mice during NRD stimulus listening (Kang et al., 2021). In addition, beta oscillations have been widely discussed in relation to working memory (Schmidt et al., 2019). Beta activity in the temporal cortex has also been shown, in studies of auditory illusions, to better account for subjectively perceived sounds than for the physical acoustic input (Vinnik et al., 2012).

Multi-frequency computations may implement decision-stage processing for repetition detection in the parietal cortex. In the present study, the learning-related decrease in beta-band L1 was observed not only in the temporal ROI but also in the parietal ROI. Previous work has likewise reported that a large-scale cortical network, including the parietal region in the beta band, predicts auditory perception (Hipp et al., 2011). In addition, the L1 analysis of parietal EEG showed a distinctive pattern: only in this area, across all stimuli, *Perceive* trials exhibited greater between-segment dissimilarity in delta-band activity, with a further relationship for this distance to increase with RefRN learning. Coordinated delta–beta activity across temporo-parietal regions is important for the precision of temporal information processing (Arnal et al., 2015). Furthermore, there is strong evidence that the parietal cortex is a dominant locus of evidence accumulation and decision-making in perceptual 2AFC tasks (Gold & Shadlen, 2007; Hanks et al., 2006; Keuken et al., 2014; O'Connell et al., 2012; Zhou & Freedman, 2019). Taken together, these findings support an interpretation in which representations formed in the auditory cortex propagate to the parietal cortex, where they are compared, and a decision

of perceptual identity is made once a sufficient level of confidence is reached.

The delta cITPC findings (mainly observed in the parietal area) further support this representational–decisional distinction. Previous studies have consistently reported high ITPC for RefRNs (Andrillon et al., 2015, 2017; Luo et al., 2013), but I found that cITPC was confined to the RefRN × learning × hit-trial interaction. This restriction suggests that cITPC did not directly reflect learning effects or repetition perception, as had often been assumed. If cITPC reflected learning, it should have been observed even in miss trials within high-learning sessions; if it reflected repetition perception, it should also have been observed in RN hit trials and in N/RefN false-alarm trials. Instead, the results are more consistent with the interpretation that cITPC reflects the across-trial consistency of the timing of decision-related processing, rather than decision processing itself. When across-segment consistency in sensory signals is low, even if the same decision "repetition" is made for the same RefRN stimulus, the timing of that decision may vary across trials because the judgement is exploratory, and the timing of accumulation to bound differs across trials (O'Connell et al., 2012). Because phase-based analyses are sensitive to temporal misalignment, such trial-to-trial timing variability would yield low cITPC values (van Diepen & Mazaheri, 2018), even if similar information processing were occurring. After selective consistency is acquired, confidence may increase, the temporal reproducibility of decision-related processing may improve, and cITPC may therefore become higher.

In summary, the present data can be interpreted as follows. The presented stimulus is represented in the auditory-cortical broadband activity, primarily in the beta band. The represented signal is then compared in parietal decision-related regions. When evidence for identity accumulates beyond a certain threshold, a decision of perceptual identity is made. However, representations of noise are inconsistent, so they cannot reliably produce the same signal even for identical inputs at first. After several exposures to RefRN, once selective consistency is acquired, evidence accumulation in the parietal cortex accelerates, leading to consistent perceptual decisions. Consequently, when trial-to-trial comparisons are restricted to correct trials for learnt RefRN stimuli, neural timing becomes aligned and is observed as condition-specific cITPC.

3.4.4. Individual Variability and Criticality

The behavioural results showed clear individual differences in the ability to acquire perceptual selective consistency, with a bimodal distribution between participants who learnt easily and those who did not. The ICC and ANOVA results further indicated that this tendency was consistent within individuals. However, there were no clear demographic differences between these groups; only ADHD traits were suggested to differ.

The spectral exponent was proposed as a potential explanation for individual differences in the ability to acquire selective consistency. The spectral exponent computed from resting-state EEG showed a tendency for participants with steeper SEs to perform better overall in the NRD task, and that the learning effects for RefRNs are maximised around an SE value close to -1.0, which is typically taken to indicate criticality. These were consistent with the non-linear relationship I observed in Project 1 for ***H2***.

In general, in the critical regime around a spectral exponent of -1.0, the balance between flexibility and stability is thought to be optimised, leading to maximal computational efficiency (Beggs, 2008; Hesse & Gross, 2014; Shew & Plenz, 2013). Similar patterns are often observed in performance on other learning tasks (de Arcangelis & Herrmann, 2010; Del Papa et al., 2017). In addition, atypical exponents have been discussed in relation to specific conditions (Robertson et al., 2019). This relationship will be further discussed in Sections 4.1.3, 4.2, and 4.4.4. Consistent with this, in my data, ADHD traits also relate to NRD learning ability (with a significant correlation in the overall analysis and a tendency in the "well-learner" vs. "poor-learner" group analysis, but this did not survive correction for multiple comparisons). Therefore, to discuss the present pattern in more detail, it may be necessary to conduct similar analyses not only in typically developing healthy adults but also in clinical populations.

3.4.5. Limitations

To avoid redundancy, I refrain from detailing project-specific limitations here and instead discuss them jointly for the two projects in Section 4.1.4.

4. General Discussion

4.1. NRD Task Through The Lens of Selective Consistency

In this thesis, I focused on the NRD task—an unusual form of learning that has been difficult to explain within conventional learning theories—to test the selective consistency hypothesis (Agus et al., 2010; Denham & Winkler, 2020). Across the two projects, I examined six working hypotheses (***H1–H6***) derived from viewing this task through the selective consistency framework (Goto et al., 2024; Goto & Kitajo, 2024). Although the strength of evidence varied across hypotheses, they were broadly supported, demonstrating the usefulness of the selective consistency hypothesis for explaining the perceptual learning mechanisms observed in the NRD task.

4.1.1. Summary of Main Findings

This thesis investigated how experience can selectively stabilise neural dynamics for specific stimuli in the NRD paradigm, combining a recurrent network simulation (Project 1) and human EEG (Project 2). Across these two projects, I asked whether "selective consistency"—stimulus-specific convergence of neural/perceptual dynamics' trajectories—can emerge from local plasticity in recurrent networks, and whether such convergence is observable in human cortical dynamics during rapid auditory implicit perceptual learning. Through these, by decomposing the neural information-processing underlying the NRD task into Representation and Decision stages, and by treating selective consistency as the mechanism that modifies neural activity within the Representation stage, I show that this framework resolves the key outstanding issues associated with NRD learning.

Project 1: Simulation Study

In Project 1, I constructed a plastic recurrent neural network model inspired by the Representation stage of the NRD task, and examined how repeated exposure to a RefRN-like repeating noise pattern shapes the network's internal dynamics. The simulations showed that a high correlation between within-trial segments emerged selectively for the RefRN input, driven by weak Hebbian-like plasticity (***H1***). This convergence was expressed at the level of the entire middle layer as an unsupervised population property, rather than at the level of the output unit or a small subset of neurons. Systematically varying the spectral radius revealed that this stimulus-selective convergence was maximised not in the most stable, subcritical regime, nor in strongly chaotic dynamics, but in a slightly more complex, mildly supercritical regime near the edge of chaos (***H2***). Importantly, the emergence of this selective consistency did not depend on explicit optimisation of task performance or readout weights, indicating that plasticity alone can self-organise stimulus-specific consistency.

Project 2: EEG Study

In Project 2, I asked whether analogous stimulus-selective convergence can be observed in human EEG during the NRD task. Firstly, using GLMMs on EEG data, I quantified the similarity between within-trial segments of frequency-dependent activity. In auditory regions, beta L1 distances between segments were smaller when participants perceived repetition, and this relationship was strengthened by learning (***H3***). In contrast, parietal delta activity showed larger inter-segment distances when they felt repetition, and this also correlated with learning. In addition, I showed that the previously reported high delta ITPC in parietal regions for RefRN is not a general property of RefRN trials but occurs only when participants both learn and perceive repetition correctly (***H4***). Altogether, these findings suggest

that parietal delta activity reflected decision-related or evaluative processes that diverged across segments rather than converged. Finally, individual learning ability in the NRD task (***H5***) was related to individual resting spectral exponent, and selective consistency learning ability was maximal in participants whose cortical activity operated near criticality (***H6***).

4.1.2. Selective Consistency as The Mechanism of The NRD Learning Effect

Results from these two projects support the hypothesis that neural selective consistency achieved through self-organisation to repeatedly exposed RefRN noise segments yields perceptual selective consistency—i.e., an NRD learning effect.

Firstly, Project 1 demonstrated that the selective consistency hypothesis can explain the most puzzling aspect of the NRD task: the rapid learning of white-noise patterns. As discussed in the General Introduction, it is difficult to explain the learning effect for white noise, as the spectra are uniform — theoretically incompressible — and shared across all stimuli regardless of condition (Shannon, 1948). Typically, auditory sensory neurons exhibit selectivity for stimulus frequency (Kayser et al., 2007; R. Kandel et al., 2021), allowing the brain to distinguish stimulus patterns. However, the uniformity of the white noise spectra makes it impossible to distinguish among the stimulus patterns in this case. Also, it severely limits the ability to "learn" or "remember" the pattern using conventional, well-established learning algorithms (Denham & Winkler, 2020).

Selective consistency, in contrast, is acquired in a self-organising manner without any of these requirements. In Project 1, the model was not explicitly trained to detect repetitions, nor was any optimisation, such as error backpropagation, applied to the recurrent-layer connectivity. Nevertheless, through Hebbian-like plasticity, the recurrent network self-organised to increase the consistency of its responses selectively for RefRN. As the network's conditional dynamics are automatically converged, it is possible to detect the identity of white noise regardless of the network's "performance"—e.g., prediction accuracy in a predictive coding framework (Friston et al., 2006; Friston & Kiebel, 2009b; Rao & Ballard, 1999). Indeed, perceptual psychology has widely embraced the view that, rather than the physical similarity of stimuli per se, the representational geometry of the evoked neural activity patterns better predicts participants' similarity judgements and generalisation (Jozwik et al., 2022; Kilian-Hütten et al., 2011; Kriegeskorte & Kievit, 2013).

Thus, although the precise computational strategy by which the brain detects repetitions at the Decision stage remains unknown, the acquisition of selective consistency provides the necessary substrate that enables any downstream mechanism—whatever its form—to perform repetition detection. For example, within a predictive-coding framework, the brain is assumed to generate predictions about future inputs (in this case, the next step of white noise) and to update its internal

model by minimising prediction error. As noted earlier, however, forecasting the temporal evolution of white noise is challenging in principle, and therefore model updates based solely on prediction error cannot account for learning in the NRD task. In contrast, when the dynamics of the intermediate (recurrent) layer stabilise, the model produces consistent errors. That is, even if the predictions themselves are always incorrect, the pattern of error is reproduced in the same way each time. In the context of this highly specialised task—detecting repetition—such consistent errors are sufficient to achieve the computational goal.

In Project 2, repeated exposure to RefRN likewise suggested a selective enhancement of neural representations for noise segments. Firstly, I confirmed a direct correspondence between perceptual consistency and neural representational consistency: perception tended to be similar when activity in the temporal ROI—likely reflecting population dynamics in sensory cortex—was similar. In contrast, because repeated perception was accompanied by more divergent activity across segments, I also identified the parietal ROI as a region likely related to the decision processes underlying repetition perception. These correspondences were strengthened with learning. Moreover, in the parietal ROI, beta-band similarity—of the same band observed in the temporal ROI—was also evident, specifically in repetition-perception trials. It indicates that the neural representation in the temporal area is referenced at the parietal region.

Taken together, these findings align well with my hypothesis that RefRN learning can be explained by a two-layer model comprising (i) stabilised representations acquired through selective consistency and (ii) decision processes that detect repetition using those representations. A key advantage of the SC-based account is that it does not require any specific neuronal population to be tuned as a feature detector; instead, "memory" is realised through the dynamical properties of network activity. Because the NRD task employs atypical stimuli, it has been difficult to explain within existing learning models, but the present study resolves this difficulty by focusing on dynamical systems properties.

Whether the selective consistency hypothesis is useful beyond the NRD task—i.e., in a more generalised account—goes somewhat beyond the scope of the present work and is discussed from Section 4.2 onwards.

4.1.3. Requirements for The Ability to Acquire Selective Consistency

It is also important that, in the course of this work, both projects identified individual differences in selective consistency acquisition capacity. In Project 1, I showed that selective consistency acquisition capacity varies with the spectral radius of the recurrent network. The spectral radius is a quantity for which having $\rho < 1.0$ provides a degree of assurance of ESP: if ρ is too small, the system loses flexibility, whereas if it is too large, the system loses stability. My results indicated that

selective consistency acquisition capacity was maximised around $\rho \approx 1.4$, i.e., slightly on the chaotic side of criticality. In the complex-systems literature, information-processing capacity is often argued to peak near the critical point—around $\rho = 1.0$ in the present framework. Thus, my finding suggests that the brain may possess an experience-dependent mechanism that tunes its dynamics so that information-processing capacity is maximised only for particular, experienced inputs, while processing of unfamiliar inputs remains relatively inefficient. The possible advantages of such a mechanism are discussed in detail in Sections 4.2 and 4.3.

In Project 2, analysis of the spectral exponent in resting-state EEG likewise revealed a relationship in which selective consistency acquisition capacity was maximised at an intermediate spectral exponent—used as a criticality proxy. Here, the vertex aligned with a canonical critical-point range; however, because participants were restricted to typically developing healthy adults, variability in spectral exponent was limited, and the estimated vertex location is therefore less reliable. If I had more participants with higher spectral exponents, I would have a clearer tendency. Nonetheless, both sets of results suggest that selective consistency acquisition capacity depends on network properties, giving rise to individual differences.

In summary, the results relevant to ***H2***, ***H5*** and ***H6*** imply that—if, as will be discussed in Section 4.2, selective consistency reflects a general brain function rather than a phenomenon specific to the NRD task—then structural constraints of neural networks, such as E/I balance, modulate adaptive capacity to the environment. This is particularly intriguing in light of recent discussions on biased criticality—its shifts with age, its alteration in developmental disorders, and its disruption in various neurological and psychiatric conditions. If criticality deviates from the optimal regime, selective consistency is compromised, leading to chronically low representation precision. Such a reduction in precision could in turn contribute to a wide range of functional deficits. In the future, this framework may be extended to account for perceptual and behavioural abnormalities observed in specific neurological or psychiatric disorders.

4.1.4. Limitations

Across the two projects, I obtained several intriguing findings relevant to the selective consistency hypothesis. However, each approach has its own limitations, and important challenges remain to be addressed before this hypothesis can be firmly established.

Project 1

A Project 1–specific limitation concerns variation in model choice. In this study, I focused on an

ESN as the recurrent model, introduced Oja's Hebbian rule as the plasticity mechanism, and used a hyperbolic tangent as the neuronal activation function (Jaeger et al., 2007; Oja, 1982). However, there are several possible neural network models and algorithms for reproducing plasticity in the nervous system beyond Hebbian learning, such as spike-timing-dependent plasticity (STDP) in spiking neural networks (SNNs) (Maass, 1997; Markram et al., 1997; Taherkhani et al., 2020). SNNs employ integrate-and-fire units that, like biological neurons, emit spikes and transmit information when the membrane potential—modulated by synaptic inputs—crosses a threshold. They also incorporate excitatory and inhibitory neurons, enabling manipulation of network dynamics via the E/I balance. Moreover, STDP can be regarded as a more refined form of the Hebbian rule used here (D'amour & Froemke, 2015; Taherkhani et al., 2020). In this sense, SNN-based models with STDP are more biologically plausible than the model adopted in the present research.

Because I did not test alternative models, it remains unclear to what extent the present conclusions generalise when the model is changed. However, at the same time, extending the same line of argument to other models poses a significant challenge: it is not straightforward to define an explicit criterion corresponding to the "spectral exponent" in this study—a parameter that controls the complexity of system dynamics. For the reservoir with a hyperbolic tangent used here, an established threshold for the echo state property (ESP) is known at $\rho = 1$. Although research has examined general ESP conditions and metrics applicable across different networks, to my knowledge, no established methodology is yet universally applicable (Buehner & Young, 2006; Yildiz et al., 2012; Manjunath & Jaeger, 2013). Consequently, in other networks, it becomes difficult to connect criticality to individual differences in selective consistency acquisition capacity in the same way. For a similar reason, I did not investigate STDP, which requires an SNN framework.

Therefore, a conservative summary of my findings is that at least one class of RNN can acquire selective consistency for a given noise stimulus without requiring an explicit optimisation algorithm. To elevate this account to one that can be applied more directly to the brain, it will be necessary either to reproduce the results using models with higher biological plausibility that address the issues above or to obtain direct empirical evidence from the biological nervous system. This point is discussed in Section 4.4.4.

Project 2

Although this study was preregistered, some of the findings still require replication to establish converging evidence (Open Science Collaboration, 2015). Except for ITPC, which had already been highlighted in previous work, I aimed to explore whether learning effects can be accounted for by neural selective consistency. For this reason, I conducted a power analysis based on behavioural

performance and ITPC values and preregistered the required sample size (Goto et al., 2024). Therefore, the results for the L1 measure and the spectral exponent are exploratory and require replication in follow-up studies. In particular, the robustness analyses further indicated that the inverted-U relationship between selective consistency and SE was not uniform and could be influenced by specific participants. In the bootstrap analysis, the inverted-U conditions were satisfied in approximately 70% of resamples, and the LOO analysis preserved the main conclusion in 23 of 24 cases. Therefore, although the present result does not hold only due to a small number of outliers, the uncertainty inherent in non-linear estimation with the current sample size (N = 24) means that replication in an independent sample is required. Relatedly, if data could be obtained from clinical groups expected to show more atypical SE, this would allow for a more rigorous discussion.

Using the EEG imposed several analytical constraints. The most crucial problem concerns spatial resolution. Because the EEG has low spatial resolution, I cannot discuss in detail which brain areas or layers the observed selective consistency occurred in. In particular, although the L1 measure differed between temporal and frontal/parietal electrodes, cITPC showed no clear differences between them. I also attempted to increase spatial resolution using CSD, but the results did not change. Spatial considerations, therefore, remain an issue for future work when addressing the functional role of selective consistency. However, selective consistency, by definition, is not necessarily restricted to a single region and may propagate across layers and areas, so this limitation may not be decisive. I discuss this point in the Section 4.4.5.

Finally, the timing of the noise segment's arrival and the transition period until EEG responses align are unknown. Because there is a time lag between sound emission and processing through the brainstem and auditory cortex, and subsequent expression in EEG, the exact onset of the corresponding “EEG segment” cannot be determined in principle. This issue arises in all experimental systems, but it matters when comparing brain activity across multiple within-trial time windows. Ideally, consistency should be evaluated while accounting for the transition period, but here I used the average over the time window. Indeed, in Project 1, I set the analysis window data-driven to align with the transient period. In biological systems, if the precise time at which a signal arrives at the recorded region were known (rather than merely the stimulus onset time), a more detailed discussion that maps more directly onto the theory and the model should be possible. One possible way to tackle this point will be discussed in Section 4.4.4. With more precise timing, a more detailed analysis might have been possible. For example, methods such as dynamic time warping might address this issue (Berndt, 1994). Still, because this study is an initial investigation of selective consistency, I did not adopt such approaches to avoid increasing analytical complexity.

As an NRD Paradigm Study

Although the primary aim of this thesis was not to elucidate the neural mechanisms of the NRD task, it is still necessary—if the present work is to be positioned as NRD research—to state the remaining open questions.

The NRD task is an unusually specialised form of learning: learning a stimulus that contains no explicit statistical features distinguishing one signal from another. To date, there has been no compelling theoretical or neuroscientific mechanism that can readily account for this ability (Denham & Winkler, 2020; Masquelier, 2018). The present study introduced the concept of selective consistency. In Project 1, I showed that, given plasticity and experience, it is theoretically possible to make response trajectories consistent even for noise stimuli, and proposed the hypothesis that NRD can be explained not by conventional learning mechanisms but as a special form of learning driven by changes in consistency. In Project 2, I further demonstrated empirically that similarity in neural activity—represented in high-frequency activity across a broad network including sensory cortex and parietal regions—explains similarity in perception regardless of the physical sound, and that this relationship becomes stronger with learning. Moreover, because low-frequency parietal activity showed responses specific to the combination of learning and successful repetition detection, I suggested that this region implements repetition detection by leveraging neural representations that have become consistent.

Crucially, however, even if all of these claims are correct, it remains unclear how "low-frequency processing in the parietal cortex" detects identity in white noise. The parietal cortex is widely known to support evidence accumulation in perceptual decision-making. In this scheme, evidence is typically accumulated from signals provided by feature-selective "detectors", and a decision is made once the accumulated evidence exceeds a threshold. Yet, as discussed above, white noise does not contain discriminable features that could be used to detect. It is also known that the brain does not process time-varying sensory inputs in a raw form, but compresses them into chunks. However, such chunking is, in principle, not feasible for white noise (Shannon, 1948). Thus, while the present study offers an account of how learning progresses, the neural implementation of noise-repetition detection remains unresolved.

In theory, recurrent neural networks like the brain can maintain a memory of recent history through their intrinsic recurrence, such that the system dynamics carry information about past inputs over a finite timescale. If an identical input recurs within the time window over which this memory is retained, then—given neuronal populations that respond strongly to such recurrences—noise repetition detection might be achievable purely through dynamics, without relying on explicit stimulus features. If selective consistency emerges only for particular stimuli, such a mechanism could yield stimulus-selective detection performance. This account would require a long memory capacity, which is

typically achieved in a supercritical regime (though excessive supercriticality compromises stability). This may relate to my observation that individuals with higher than ideal complexity—indexed by the spectral radius in Project 1 and by ADHD trait in Project 2—performed better. This possibility requires further investigation.

A further limitation of the NRD task, not specific to this study, is that learning progresses unusually fast. Typically, learning curves develop gradually, allowing evaluation by splitting data into pre- and post-learning phases or by fitting a linear approximation. In NRD, however, when learning is successful, accuracy can reach almost 100% after fewer than five exposures (Agus et al., 2010). As a result, it is difficult to compare neural changes within RefRN as learning progresses, and analyses are largely limited to contrasts with RN. Ideally, the selective consistency indices used here would show a gradual strengthening with learning. In this study, I observed a correlation in the GLMM results: greater learning effects were associated with higher consistency for RefRNs in the L1 measure, but a more detailed account may be required.

Finally, although the concept of selective consistency and the spectral exponent results explained individual differences in learning effects to some extent, I did not explain why sometimes differences arise across sessions. The ICC results indicated consistent performance within participants overall, but the data include individuals with substantial variability. In addition, previous work reported RefRN waveform-specific effects that cannot be explained by acoustic features (waveforms that are easier or harder to learn) (Agus et al., 2010). Neither the present study nor subsequent work has tested this direction, so its reproducibility is unknown; however, if it is reproducible, it will require a neuroscientific explanation. Under the current formulation of selective consistency, differences in ease of adaptation as a function of stimulus features do not follow, and this remains a potential issue.

As a Selective Consistency Study

The present work also has several limitations with respect to its central aim: testing the selective consistency hypothesis. Firstly, my validation was confined to a specific experimental paradigm—the NRD task —and thus cannot be straightforwardly generalised to more conventional perceptual paradigms. However, what makes this difficult is that selective consistency acquisition occurs automatically in daily experience. In many perceptual tasks, although the exact time-varying stimulus sequence may be novel, participants have already accumulated extensive experience with similar stimuli or with the constituent features derived from their decomposition. Consequently, in typical adult participants, their brains may already be adapted, making it uncertain whether any measurable change in neural consistency can be induced under laboratory conditions. If, as argued in the General Introduction, selective consistency acts early in development as a foundational mechanism supporting

broad information processing, then observations will be required from before this adaptation occurs—from the early developing brain, in vitro preparations, or neural tissue in which such functions are impaired. This point will be further discussed in Section 4.4.2.

A second issue concerns the correspondence between the selective consistency measure originally defined in Eq. 6 and the indices used in each project. Given the nature of NRD stimuli, identical inputs occur only a few times within a trial. In my definition, selective consistency refers to an asymptotic reduction in the variance across input-evoked dynamics; however, because only two or three trajectories can be obtained, I used correlations and waveform distances as proxy measures. Although these proxies can be related mathematically to variance, it may be desirable to use experimental settings that allow sampling more trajectories to improve estimation precision. Importantly, however, one must also note that selective consistency acquisition would continue during such sampling.

Finally, in Project 1, I observed that plasticity updates the system's connectivity matrix and that selective consistency emerges as a consequence, but the specific mathematical mechanism remains unknown. The stimulus-conditioned dynamics of a reservoir depend strictly on its internal connectivity matrix and the environmental noise applied to the system. Because environmental noise was fixed throughout the simulations and changes in connectivity were driven solely by plasticity, we can conclude that selective consistency emerged because the connectivity matrix changed to satisfy an unknown mathematical property. However, I did not identify the condition in the present study. I will speculate it as far as I can in Section 4.4.1.

If these issues can be resolved, the selective consistency framework could be extended in multiple directions, including rigorous links to other learning theories, the development of brain-like computing, and mechanistic accounts of neurological and psychiatric disorders. In the following section, I therefore provide a careful discussion and agenda for possible future work, while acknowledging that parts of this discussion necessarily go beyond what can be strictly concluded from the present results.

4.2. Selective Consistency as a General Property of The Brain

To generate behaviour optimal for a given environment, the brain must observe the external world and represent it as an internal state. Many theories of brain information processing have been proposed, but a common premise across them is that the brain reads out the state of the world and encodes it in neural activity (Rao & Ballard, 1999). In an environment that contains unlimited information, sensory organs sample only the information needed; signals are transformed into electrical activity according to the transduction rules of each modality, then conveyed to the brain. Processing in the brain is also hierarchical across regions. Depending on the region, diverse computations are performed, including whether and how signals contribute to semantic or conscious processing, feature decomposition, and

integrative processing. However, all such functions presuppose consistency in the signals they receive from upstream—namely, that the same information is conveyed in the same signal (or at least with the same constituent components).

Consistency, in this sense, is the principle that identical inputs give rise to identical responses, and this thesis has examined brain function from this perspective. Consistency has been extensively discussed in physics, for example, in studies of lasers (Uchida et al., 2004, 2007, 2008). A key difference between such physical systems and the brain is that, whereas in physical systems the response to an input equals the system's behaviour, the brain exhibits a two-layer structure: neural activity and the resulting perception or behaviour. In this thesis, I distinguished these layers and referred to the neural response to identical sensory stimuli as neural consistency and the resulting perceptual outcome as perceptual consistency. Concretely, in the NRD task, I treated correct perception of repetition in listening to stimuli composed of repeated instances of the same white-noise segment as an index of perceptual consistency, and adopted a logical framework in which neural consistency is posited as the mechanistic basis for this perception.

At the same time, it is also essential that the brain exhibits variability: it can change its responses to identical inputs depending on context, and it can engage in rich spontaneous processing even in the absence of external input (Dinstein et al., 2015; Faisal et al., 2008; Stein et al., 2005). One major trend in contemporary neuroscience concerns the brain's variability and flexibility. Traditionally, much of the trial-to-trial variability in neural activity measured under the same condition was treated as noise. However, over the past two decades, it has become increasingly clear that response variability to an identical stimulus is not merely measurement noise, but can reflect pre-stimulus ongoing activity and cortical state, as well as fluctuations in top-down control such as attention. Indeed, single-trial responses can be described as the sum of a reproducible evoked component and ongoing activity present before stimulus onset (Arieli et al., 1996), and stimulus presentation can broadly suppress (quench) cortical variability, thereby stabilising network state (Arazi et al., 2017, 2017; Churchland et al., 2010; Daniel & Dinstein, 2021; Wolff et al., 2019). Correlated fluctuations in population activity also influence encoding and readout (Averbeck et al., 2006), and attentional improvements in behaviour can be explained by reductions in correlated variability (M. R. Cohen & Maunsell, 2009). In addition, theories have been proposed in which spontaneous activity reflects the statistical structure of an internal model shaped by experience, and in which variability itself encodes inferential uncertainty (Orbán et al., 2016). Collectively, these perspectives increasingly support the view that neural variability underpins information processing. If one were to assume an excessively consistent brain, information processing would ignore context and memory entirely, making it impossible to adapt to the environment.

Therefore, the brain must maintain variable and flexible dynamics overall while reducing

variability and producing consistent behaviour when necessary (Buonomano & Maass, 2009). The well-known phenomenon of neural variability quenching—widespread suppression of response variability following stimulus presentation—is often cited as direct evidence for this dual requirement (Arazi et al., 2017; Daniel & Dinstein, 2021; Wolff et al., 2019). How the brain implements this duality remains an active topic of debate.

In this thesis, I therefore proposed the selective consistency hypothesis, namely that the degree of consistency varies stimulus-dependently. Many studies report that neural activity in some developmental stages exhibits high variability (Caras & Sanes, 2019; Montez et al., n.d.; Naik et al., 2023). Responses in cultured neuronal networks are likewise highly unstable (Maeda et al., 1995; Wagenaar et al., 2005). In other words, the brain cannot, from the outset, reliably produce variability-quenching and neural consistency in response to external inputs. Moreover, even in adults, perceptual consistency is low for complex stimuli that have not been previously experienced, and it can be enhanced through experience (Agus et al., 2010; Kang et al., 2018). Taken together, these observations led us to propose that experience-dependent acquisition of neural selective consistency occurs in a stimulus-specific manner, while overall flexibility is preserved—thereby realising the dual requirement of consistency and variability.

Here, I focused on the NRD task because perceptual selective consistency can be acquired in this paradigm, yet it has been difficult to explain the neural mechanisms using existing learning theories alone. As shown in Section 4.1, results indicate that the selective consistency hypothesis provides a good account of NRD learning, at least at the level of explanation required for this task.

Importantly, the above argument places no constraints on stimulus properties or task demands and is, therefore, in principle, applicable to a wide range of situations. Throughout this thesis, I decomposed the NRD task into two stages: (i) representation of the stimulus (noise) in the sensory system and (ii) a decision process implementing repetition detection. And I treated selective consistency as a form of learning operating at the representation stage. This separation allows the selective consistency account to be distinguished from task-specific considerations about repetition detection. Indeed, changes in neural activity to NRD white-noise stimuli have been reported even during sleep or under anaesthesia, further suggesting that these stages can be treated independently (Andrillon et al., 2017; Kang et al., 2021). Although using noise stimuli retains some task-specificity at the representation level, this choice primarily served to illustrate the selective consistency framework clearly. In general, virtually any perceptual experiment can be decomposed—at a coarse level—into representations and downstream decisions based on those representations. However, as discussed in Section 4.1.4, when experiments target the mature brain using other, more familiar sensory stimuli, extensive prior experience is likely to reduce the observable change in consistency.

One of the most intriguing findings was that the capacity to acquire selective consistency depends on intrinsic network properties, and is maximised in networks whose characteristics lie slightly outside the regime traditionally considered optimal for information processing. Both the simulation and EEG results revealed a non-linear relationship between selective consistency acquisition capacity and parameters that govern the complexity of system dynamics, such as the spectral radius and the spectral exponent. Specifically, the relationship was well captured by a quadratic curve with a peak around the vicinity of criticality—a regime in which dynamics are neither overly stable nor overly flexible (Beggs, 2008; Beggs & Plenz, 2003). Notably, Project 1 showed that selective consistency was maximised in a regime slightly more flexible, complex, and variable than criticality—supercriticality —which is typically assumed to be too flexible and to achieve no more than stable information processing. In Project 2, the estimated peak coincided with criticality; yet, given that subtracting one outlier showing low performance moved the vertex into a more supercritical regime, and stronger ADHD traits were associated with better task performance, it remains possible that, with a larger sample size, the vertex would likewise shift towards a somewhat higher-complexity regime.

Selective consistency, therefore, suggests a possible resolution to an apparent tension: the brain may operate in a regime more flexible than criticality, while still achieving consistent information processing when needed. Criticality is often regarded as desirable because if dynamics are too stable and fixed, context- and memory-dependent processing becomes difficult, whereas if dynamics are too flexible, information becomes conflated and consistent responses are lost (Hengen & Shew, 2025; Zimmern, 2020). This has motivated extensive discussion of the idea that the brain self-organises towards criticality (self-organised criticality) (Bak et al., 1987; Beggs, 2008; Dorogovtsev et al., 2008; Plenz et al., 2021). Selective consistency, however, as indicated by my simulations, could allow the system to operate in a regime even more biased towards complexity than criticality, while selectively shifting stimulus-specific dynamics towards criticality for learnt inputs. In other words, the brain could maintain a higher degree of flexibility than is often assumed—remaining strongly influenced by prior history, immediate context, other concurrent inputs, broader context, and attentional state—yet increase consistency only when a "familiar" stimulus is encountered, by virtue of its dynamical properties.

If such a mechanism indeed operates, then for stimuli with low consistency, canonical forms of learning and information processing (will be discussed in Section 4.3) would not function effectively, whereas they would become effective specifically for stimuli that have acquired selective consistency. In this sense, selective consistency may act as a physical filter that separates signals that should be used for information processing (e.g., learning) from those that should not.

Finally, under this view, dysfunction of selective consistency—as a key physical filter supporting information processing—should impact a wide range of downstream computations that depend on it.

For selective consistency to operate, at least two conditions are required: the system must start from an initial state near the critical point, and plasticity must function appropriately. If either condition is compromised, selective consistency will not be acquired. In developmental disorders and some neurological or psychiatric conditions, the estimated neural dynamics have been reported to deviate substantially from criticality (Alamian et al., 2022; Arviv et al., 2016; Bruining et al., 2020; Yin et al., 2022). Impairments of plasticity are also widely recognised to be implicated in many disorders. However, the mechanisms by which such structural alterations translate into specific functional symptoms remain incompletely understood. Consistent with this, in Project 1, selective consistency failed to emerge when the system was extremely far from criticality. More broadly, the selective-consistency framework may provide a way to bridge these levels of description.

4.3. Connections to Other Learning Theories

4.3.1. Selectivity of Sensory Neurons

When discussing stimulus-specific neural responses, it is necessary to consider neuronal selectivity in the sensory cortex, one of the canonical topics in neuroscience. Traditionally, it has been shown that neurons in the visual cortex respond selectively to particular orientations, and that such tuning can be sharpened through learning (Hubel & Wiesel, 1962; Jehee et al., 2012; Schoups et al., 2001). Beyond static stimuli, there are also neurons that respond selectively to dynamic stimuli (e.g., speech or musical sounds) (Doupe & Solis, 1997; Mesgarani et al., 2014). At higher levels of processing, some neurons respond selectively to information about specific individuals, such as names, face images, and voices (Quiroga et al., 2005). Taken together, these findings consistently indicate that neurons can be tuned as “detectors” for particular stimuli or concepts, and that stable engagement of such neurons could support perceptual consistency across trials.

In contrast, selective consistency is fundamentally a dynamical systems property and does not explicitly implement functions such as signal detection or other forms of information processing; therefore, it differentiates from neuronal selectivity. As shown in Project 1, it emerges as a self-organised change in trajectories. Precisely because of this, however, selective consistency has the advantage of remaining applicable to cases in which learning and the differentiation of “detectors” are difficult, such as the NRD task. The view that the system as a whole can preserve memory as a dynamical system (Buonomano & Maass, 2009), without requiring dedicated detectors, is conceptually parsimonious because it avoids the need to assign a one-to-one responsible unit for conceivable concepts.

At the same time, considering detector-like neurons highlights an additional requirement: for such neurons to detect their preferred features, they must receive consistent input each time. From this

perspective, one can also organise the account as follows: selective consistency stabilises signal representations, and then specific neurons or networks come to respond selectively to these stabilised representations, thereby supporting stable perception. This is closely analogous to the role played by the output neurons in Project 1. If the system does not possess consistent dynamics, the learning rule, such as backpropagation, cannot achieve its computational goal, as I will discuss in the next section.

4.3.2. Bayesian Brain Hypothesis, Predictive Coding, and Free Energy Principle

Here, I will connect selective consistency and Bayesian inference–based frameworks—currently the most widely accepted computational theories of perception and perceptual learning. For readers who need a description of those theories, see Predictive Coding and Free Energy Principle before reading the following statements.

Selective consistency implies that, even under the same internal model m, the variability of neural responses $\boldsymbol{x}$ differs across stimuli s, and that this variability changes with experience. Under the standard PC observation equation,

$$\boldsymbol{x}(t) = g(s(t); m) + \boldsymbol{\varepsilon_x}(t), \quad \boldsymbol{\varepsilon_x}(\mathrm{t}) \backsim \mathrm{N}(0, \boldsymbol{\sigma_x^2}) \qquad (\mathrm{eq.}\,23)$$

the observation noise $\boldsymbol{\varepsilon_x}(t)$ is assumed to be stimulus-independent and determined solely by neural activity. The selective-consistency framework, however, posits that the error term contains a stimulus-dependent component.

Accordingly, if we decompose the error as $\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon_x} + \boldsymbol{\varepsilon_{x;s}}$, assuming linear additivity of stimulus-common part $\boldsymbol{\varepsilon_x}(t)$ and stimulus-dependent part $\boldsymbol{\varepsilon_{x;s}}(t)$, the observation equation can be rewritten as

$$\boldsymbol{x}(\mathrm{t}) = g(s(t); m) + \boldsymbol{\varepsilon_x}(t) + \boldsymbol{\varepsilon_{x;s}}(t), \quad \boldsymbol{\varepsilon_x}(t) \backsim \mathrm{N}\big(0, \boldsymbol{\sigma_x^2}(t)\big), \boldsymbol{\varepsilon_s}(t) \backsim \mathrm{N}\left(0, \boldsymbol{\sigma_{x;s}^2}(t)\right). \qquad (\mathrm{eq.}\,24)$$

In this case, the precision Π is given by

$$\Pi_{\varepsilon}(t; s) = \left(\boldsymbol{\sigma_x^2}(t) + \boldsymbol{\sigma_{x;s}^2}(t)\right)^{-1}. \qquad (\mathrm{eq.}\,25)$$

Alternatively, the stimulus-dependent variability can be interpreted as a heteroscedastic scaling of the noise. In this formulation, the observation equation becomes

$$\boldsymbol{x}(t) = g(s(t); m) + \boldsymbol{\sigma_x}(t)\boldsymbol{\sigma_{x;s}}(t)\boldsymbol{\eta}(t), \qquad \boldsymbol{\eta}(t) \backsim \mathrm{N}(0, \boldsymbol{I}) \qquad (\mathrm{eq.}\,26)$$

and the corresponding precision becomes

$$\Pi_{\varepsilon}(t; s) = \left(\boldsymbol{\sigma_x^2}(t)\boldsymbol{\sigma_{x;s}^2}(t)\right)^{-1}, \qquad (\mathrm{eq.}\,27)$$

that is, the inverse squared scale (gain) of the state- and stimulus-dependent standard error. Or finally, a more realistic formulation includes a global baseline noise term $\boldsymbol{\varepsilon}_0(t)$, yielding the full observation equation and precision

$$\boldsymbol{x}(t) = g(s(t);m) + \boldsymbol{\varepsilon}_0(t) + \boldsymbol{\sigma}_{\boldsymbol{x}}(t)\sigma_{x;s}(t)\boldsymbol{\eta}(t), \qquad where \;\; \boldsymbol{\varepsilon}_0(t) \backsim \mathrm{N}(0, \boldsymbol{\sigma}_0^2), \boldsymbol{\eta}(t) \backsim \mathrm{N}(0, \boldsymbol{I}),$$

$$\Pi_\varepsilon(t;s) = \left(\boldsymbol{\sigma}_0^2 + \boldsymbol{\sigma}_{\boldsymbol{x}}^2(t)\boldsymbol{\sigma}_{x;s}^2(t)\right)^{-1}. \qquad \text{(eq. 28)}$$

In conventional discussions of the PC and FEP, the precision Π_ε—defined as the inverse of the variance $\boldsymbol{\sigma}_{\boldsymbol{x}}^2(t)$—plays a central role in scaling the prediction error $\boldsymbol{\epsilon}_u$. Through this precision weighting, synaptic gain is adjusted, thereby determining how strongly prediction errors contribute to model updating (Feldman & Friston, 2010; Friston et al., 2006). Precision itself has been argued to be controlled by functional factors such as attention, as well as by neuromodulatory systems involving dopamine, NMDA receptors, acetylcholine, and noradrenaline (Feldman & Friston, 2010; Parr & Friston, 2017; Owens et al., 2018). Under my formulation, these conventional sources of precision modulation are captured by the term $\boldsymbol{\sigma}_{\boldsymbol{x}}^2(t)$. Importantly, all of these mechanisms are context-dependent rather than stimulus-dependent.

In contrast, all formulations proposed here (Eqs. 25, 27, and 28) explicitly incorporate a stimulus-dependent variance $\boldsymbol{\sigma}_{x;s}^2(t)$, which has not been considered in standard PC and FEP accounts. Because selective consistency is itself defined in terms of the across-trial variance of neural responses to a given stimulus $\mathrm{Var}_k[\boldsymbol{x}^k(t;s)]$. Once selective consistency is defined in this dynamical sense, a reduction in trial-to-trial dispersion follows as a statistical consequence. Specifically, for a fixed stimulus s, stronger trajectory contraction implies that the empirical conditional distribution of neural states becomes tighter, so that $\mathrm{Var}_k[\boldsymbol{x}^k(t;s)]$ decreases as selective consistency increases. Therefore, it is natural to model this tightening as a reduction in the likelihood variance parameter,

$$\mathrm{Var}_k[\boldsymbol{x}^k(t;s)] \approx \boldsymbol{\sigma}_{x;s}^2(t). \qquad \text{(eq. 29)}$$

Consequently, the acquisition of selective consistency for a stimulus *s* corresponds, in predictive-coding terms, to a reduction in $\boldsymbol{\sigma}_{\boldsymbol{u};s}^2(t)$ for that stimulus. By equations 25, 27, and 28, this is mathematically equivalent to an increase in precision $\Pi_\varepsilon(t;s)$. In other words, selective consistency functions as a stimulus-specific adjustment of prediction-error gain grounded in the system's intrinsic dynamics.

This mapping also clarifies implications for perception. In Bayesian formulations,

$$\mathrm{p}(s|\boldsymbol{x}) \;\propto\; \mathrm{p}(\boldsymbol{x}|s)\mathrm{p}(s), \qquad \text{(eq. 30)}$$

perception corresponds to inferring the environmental state s from neural activity $\boldsymbol{x}$. From Eq. 29, if selective consistency has not been acquired for a stimulus, the conditional distribution

$p(\boldsymbol{x}|s)$ exhibits high variance, meaning that the observation $\boldsymbol{x}$ provides weak and unreliable information about s. In that regime, posterior inference becomes dominated by prior fluctuations and nuisance variability, yielding unstable perception even when the nominal external input is unchanged. Conversely, selectively reducing $\boldsymbol{\sigma}^2_{\boldsymbol{u};s}(t)$ makes the likelihood sharper, stabilising inference for those stimuli.

In summary, neural consistency directly governs perceptual stability through the likelihood's variance and is therefore essential for perceptual consistency. Up to this point, this conclusion aligns with existing Bayesian accounts of perception. The critical extension proposed here is that the brain may initially operate with high likelihood variance, and that this variance can be selectively reduced in a stimulus-dependent manner—not as a consequence of stimulus complexity, contextual modulation, or attention, as traditionally assumed, but as a consequence of the intrinsic dynamical properties of the brain as a nonlinear system (Feldman & Friston, 2010; Parr & Friston, 2017; Owens et al., 2018). This implies that perceptual stability—and, under Bayesian formulations, the learning rate—can be naturally enhanced through experience in a stimulus-specific manner. Under this interpretation, selective consistency acts as a physical filter early in development, determining which inputs acquire consistent neural representations and thereby differentiating stimuli that should be learnt from those that need not be.

4.4. Open Questions and Future Directions

The selective consistency framework reveals a range of intriguing possibilities. However, rigorous validation will require addressing several outstanding challenges that were beyond the scope of the present work due to constraints in time, resources, and expertise. Below, I provide an overview of these issues.

4.4.1. How Is It Possible to Achieve Conditional Trajectory Convergence through Plasticity

As noted in the Limitations, it is worth investigating what specific changes in the connectivity matrix enable the system to exhibit stimulus-dependent consistency while preserving overall flexibility. The present thesis reports empirically—both in simulations and experiments—that selective consistency can be acquired. This is a sufficient starting point, but more rigorous conditions need to be derived mathematically. In this regard, dynamical systems theory and algebraic approaches are likely to be useful (Strogatz, 2024).

From the perspective of non-linear dynamical systems, selective consistency can be understood,

in a straightforward way, as convergence of input-conditioned trajectories in an input-driven system. That is, whether differences in initial state and intrinsic fluctuations persist, or are they gradually quenched over time when the same input drives the system. Let $\boldsymbol{x}_t \in \mathbb{R}^N$ denotes the latent state of the n neural populations and the input signal $\boldsymbol{u}_t(s) \in \mathbb{R}^M$ induced by stimulus s. The state update is given by Eq. (3) introduced in the General Introduction

$$\boldsymbol{x}_t^k(s) = \Phi\left(\boldsymbol{x}_{t-1}^k(s), \boldsymbol{u}(s), \boldsymbol{\theta}^k, \varepsilon^{\mathrm{k}}\right) \qquad \text{(eq. 31)}$$

Here, $\boldsymbol{\theta}^k$ denotes parameters of the connectivity matrix, such as synaptic connections. And k is the trial index. Introducing the state difference between trial k and trial l, $\boldsymbol{\delta x}_t(s) = \boldsymbol{x}_t^k - \boldsymbol{x}_t^l(s)$, the dynamical definition of consistency is to track how this difference evolves over time under the same input $\boldsymbol{u}_{\mathrm{t}}(s)$.

Linearising the dynamics around the stimulus-conditioned trajectory yields

$$\boldsymbol{\delta x}_{t+1}(s) \approx \mathbf{J}_{\mathrm{t}}(s)\boldsymbol{\delta x}_t(s), \qquad \mathbf{J}_t(s) = \frac{\partial \Phi}{\partial \boldsymbol{x}}\Big|_{(\boldsymbol{x}_t(s),\ \boldsymbol{u}_t(s))}. \qquad \text{(eq. 32)}$$

The Jacobian $\mathbf{J}_t(s) \in \mathbb{R}^{N\times N}$ is a local linear map that determines, at a given time and in the neighbourhood of a given state, how much an infinitesimal deviation is amplified or attenuated at the next time step (Strogatz, 2024; W.Hirsch et al., 2017). In general, $\mathbf{J}_t(s)$ depends on the state $\boldsymbol{x}_{\mathrm{t}}(s)$ and the input $\boldsymbol{u}_{\mathrm{t}}(s)$, and therefore varies over time. Consequently, even for neural systems with the same connectivity $\boldsymbol{\theta}$, different stimuli can drive the system through different regions of state space, leading to stimulus-dependent statistics of $\mathbf{J}_t$.

A concept that summarises this average rate of expansion of deviations under conditioning on the same input is the maximal conditional Lyapunov exponent (CLE) (Eckmann & Ruelle, 1985; Strogatz, 2024; W.Hirsch et al., 2017), written in the form

$$\lambda_c = \lim_{T\to\infty} \frac{1}{T} \log \left\| \prod_{t=0}^{T-1} \mathbf{J}_{\mathrm{t}} \right\|, \qquad \text{(eq. 33)}$$

where $\lambda_c < 0$ implies that, under an identical input, inter-trial differences decay exponentially on average, and responses from different trials are likely to converge onto the same bundle of trajectories. Conversely, if $\lambda_c(s) > 0$, such differences tend to be amplified, so trial-to-trial variability is more likely to persist even for the same stimulus. In the literature on drive–response synchronisation and generalised synchronisation, a negative conditional exponent in the response system has repeatedly been used as a criterion for the response to stably follow an identical drive (Kocarev & Parlitz, 1996; Pecora & Carroll, 1990; Rulkov et al., 1995).

Accordingly, selective consistency can be expressed as stimulus-dependent variation in $\lambda_c(s)$, such that repeated experience shifts $\lambda_c(s)$ towards more negative values for the experienced stimulus.

Within this framework, plasticity corresponds to updates of $\boldsymbol{\theta}$, but these updates need not proceed uniformly across stimuli. Repeatedly presented stimuli repeatedly visit a relatively restricted region of state space (a particular trajectory bundle), so updates accumulate preferentially in the neighbourhood of that region. As a result, changes in connectivity necessarily act in a biased manner on the dynamics of the regions traversed by that stimulus. In contrast, for stimuli that are not repeatedly exposed, or for stimuli whose input sequences vary across trials such that the same trajectory bundle is unlikely to be revisited, updates are less likely to concentrate in a specific region, making comparable reorganisation less likely. This asymmetry naturally yields stimulus selectivity in the change of its dynamics.

An important caveat, however, is that although updates can be expected to accumulate more readily in a particular region, it is not self-evident that they will necessarily reduce $\mathbf{J}_t(s)$ in that region. Depending on the direction of the updates, it is also possible that $\mathbf{J}_t(s)$ becomes larger specifically for the experienced stimulus, driving the system towards greater "selective inconsistency". Therefore, for trial-to-trial variability to actually decrease for repeated stimuli (i.e., for $\lambda_c(s)$ to shift towards negative values), some mechanism is required that selectively biases connectivity changes towards suppressing transverse inconsistency. This non-triviality is why, at least for the present author, expressing selective consistency acquisition mathematically is difficult.

Nevertheless, several plausible mechanisms can be considered that could bias updates in the direction of reducing $\mathbf{J}_t(s)$. As a starting point, $\mathbf{J}_t(s)$ is a matrix (with dimensionality determined by the number of neurons). Each element describes local expansion or contraction along a particular direction, and the CLE is given by the log average of these local stretch factors. Therefore, achieving selective consistency ($\lambda_c(s) < 0$) does not require every direction to shift towards contraction. If contraction-inducing changes dominate while expansion-inducing changes are constrained, the net exponent $\lambda_c(s)$ can become negative.

A key ingredient here is homeostasis in synaptic plasticity. In general, rules that "strengthen or weaken connections between temporally correlated neurons" can lead to divergence (weights grow without bound as updates accumulate). Consequently, regularisation or forgetting terms are typically introduced to maintain the overall energy balance of the system. This is a widely shared experimental and theoretical view, and Oja's Hebbian rule used in this thesis is one such example. Although there are multiple ways to implement homeostasis, they all act to suppress runaway excitation and excessive weight growth (Chen et al., 2013; Oja, 1982).

Let the local amplification factor under the linearised dynamics $\delta\boldsymbol{x}$ be denoted by

$$g_t(s) = \frac{\|\delta\boldsymbol{x}_{t+1}\|}{\|\delta\boldsymbol{x}_t\|} \approx \frac{\|\mathbf{J}_t\delta\boldsymbol{x}_t\|}{\|\delta\boldsymbol{x}_t\|}. \qquad \text{(eq. 34)}$$

Intuitively, the maximal CLE can be expressed as the long-time average of log $g_t(s)$ (see Eq. 33). Under this view, error amplification dominates when large contributions with $g_t(s) > 1$ pushes the average upward, shifting $\lambda_c(s)$ towards positive values; conversely, error contraction dominates when contributions with $g_t(s) < 1$ pulls the average downward, making $\lambda_c(s)$ negative. For g_t to become large, overall activity at time *t* + *1* needs to be large relative to time *t*, which is more likely when recurrent gain is strong. However, when a homeostatic plasticity rule, such as Oja's Hebbian rule

$$\Delta W_{ij} = \alpha x_i \left(x_j - x_i W_{ij}\right) \quad \text{(eq. 35)}$$

where i and j represent neurons, is present, synaptic weights are adjusted not only by the Hebbian term $\alpha x_i x_j$ but also by the normalisation term $-\alpha x_i^2 W_{ij}$. Crucially, the magnitude of the normalisation term scales with postsynaptic activity x_i increases. In other words, states with stronger activity—and thus a greater tendency for recurrent influences to persist—receive stronger suppressive updates. As a result, when the system is biased towards inconsistency ($\lambda > 0$, with large and frequent positive log g_t), changes that would further increase g_t (and hence amplify across-trial differences) are selectively suppressed, whereas changes that decrease g_t are not equivalently constrained. Through such asymmetric updating, the contribution of $g_t < 1$ components may increase over time, potentially driving $\lambda_c(s)$ below zero.

By contrast, in systems that are already weakly non-linear ($\lambda < 0$), the state depends more strongly on the current input and is less influenced by recurrent history. In this regime, synaptic updates are less shaped by the temporal structure of the stimulus. Because recurrent activity is not strongly amplified, the normalisation term does not selectively suppress changes along particular directions, and the net change in λ may remain minimal (cancel each other).

Also, if the system possesses strong non-linearity ($\lambda \gg 0$), the same input no longer visits a restricted region of state space, so updates do not accumulate primarily in the neighbourhood of that region.

From this perspective, Project 1's pattern—SC acquisition in the supercritical regime (corresponding to $\lambda > 0$) and little change in the subcritical regime—may be interpretable. In summary, consistency corresponds to having a negative maximal conditional Lyapunov exponent, and selectivity arises because plasticity accumulates only in limited regions of state space that are repeatedly visited conditioned by the stimulus. Plasticity can, in principle, scale the Jacobian in directions that either amplify or suppress deviations. However, when homeostatic plasticity is present, in a mildly supercritical regime ($\lambda > 0$), changes that would increase difference amplification are selectively constrained. Consequently, contraction-promoting changes are expressed more strongly, shifting $\lambda_c(s)$ towards 0 (and potentially below).

That said, this account remains speculative rather than proven, and further verification will be required.

4.4.2. Consistency and Discrimination

Consistency is the property that enables the brain to treat the same stimulus as the same, without being driven by initial states or background noise. Put differently, it may be viewed as the presence of an attractive conditional trajectory that overcomes subtle perturbations. However, this also entails a risk: even if stimuli differ in subtle but meaningful ways, increased consistency could pull them into the same trajectory. If so, does increasing consistency inevitably reduce the ability to discriminate between similar stimuli?

A distinct representation requires two properties: the same object should be represented consistently, and it should be distinguishable from other objects. In other words, to form appropriate representations of external concepts, the system must acquire discriminability alongside consistency (DiCarlo & Cox, 2007; O'Reilly & McClelland, 1994; Yassa & Stark, 2011). Because the present study did not test this experimentally, whether discriminability is preserved or co-acquired as selective consistency develops remains an important open question. So far, the lack of improvement in CRR for RefN denies any improvement in discrimination for experienced stimuli, at least within the data range of this study.

Experimentally, within an NRD framework, one approach would be to train participants on multiple RefRN time series and then present mixture stimuli (*MixedN*) composed of combinations of those learnt sequences. Perceptually, discriminability could be assessed by testing whether CRR for *MixedN* are higher than for standard N. Neural activity could also be recorded and analysed using the same metrics as in this thesis—for example, testing whether between-segment neural distances for *MixedN* exceed the mean distance observed for N. Another candidate paradigm would be to parametrically randomise a learnt RefRN time series and estimate the proportion of randomisation at which perceptual and neural selective consistency disappear, and whether surprise-related neural responses emerge at that point.

If the system can maintain attractors that are both consistent and discriminable, it is also critical to ask how many such attractors can be supported. From a purely dynamical systems perspective—considering the arrangement of attractors in state space—in principle, an unbounded storage capacity (Amit et al., 1985). Empirically, however, constraints should appear: for instance, if participants learn far more than two or three RefRN patterns, performance may eventually plateau or decline, earlier patterns may be forgotten, or interference may arise. Even if learning remains accurate, it would be informative to examine how neural activity differs among RefRN1, RefRN2, and RefRN10. Pursuing these questions could contribute to understanding the brain's remarkable memory capacity.

Building on the formulation in Section 4.4.1, the question of how a similar stimulus $\boldsymbol{u}_2$ is treated

after selective consistency has been acquired for a stimulus $\boldsymbol{u}_1$ can be described in the same framework. If learning expands the basin of the attracting trajectory bundle formed to $\boldsymbol{u}_1$, $\boldsymbol{u}_2$may more readily fall within that basin, and responses to $\boldsymbol{u}_2$ may be pulled into $\boldsymbol{u}_1$'s basin—yielding assimilation. In the context of generalised synchronisation, even if a functional relationship from drive to response, $\boldsymbol{x} = \Phi(\boldsymbol{u})$, can be established, there is no guarantee that the mapping is one-to-one; similar but distinct inputs may be mapped to similar responses (Boccaletti et al., 2002).

However, acquiring consistency for $\boldsymbol{u}_1$ does not necessarily entail uniformly stronger convergence in all directions of state space (Moreno-Bote et al., 2014; Strogatz, 2024). If trial-to-trial differences under the same input—arising mainly along directions associated with initial-state errors or intrinsic fluctuations (in Project 1, the imposed fluctuational noise)—are selectively suppressed, while directions along which input differences project are not suppressed, then consistency can increase without forcing $\boldsymbol{u}_2$ to be assimilated into $\boldsymbol{u}_1$. Alternatively, if the system can acquire multiple stable trajectories (i.e., a bifurcation) along directions carrying input differences, $\boldsymbol{u}_2$ could coexist as a distinct representation rather than being absorbed into $\boldsymbol{u}_1$. Whether such mechanisms exist in the brain, and if so, how they are achieved, remains unknown. Nonetheless, at least in principle, acquiring an attractive trajectory for $\boldsymbol{u}_1$ does not immediately imply a loss of discriminability with respect to other stimuli (Averbeck et al., 2006).

4.4.3. Consistency and Multistability

Speaking of bifurcation, a related but distinct topic is multistable perception (Leopold & Logothetis, 1999; Sterzer et al., 2009). Identical sensory stimuli do not necessarily fall neatly into just binary patterns—being perceived identically or failing to be recognised as identical. For many stimuli, perception maintains a one-to-one mapping to the sensory input; however, phenomena known as multistable perception demonstrate a one-to-many relationship, in which the same sensory stimulus can be stably associated with multiple perceptions (Sterzer et al., 2009). This has been extensively studied in the context of the idea that perception is generated in accordance with predictions rather than the physical stimulus itself, as well as in relation to perceptual stability and switching (Brascamp et al., 2018; Hohwy et al., 2008). Like the present work, multistable perception has been argued to be well explained from a dynamical systems perspective, in terms of bifurcations, hysteresis, and catastrophes (Moreno-Bote et al., 2007; Pisarchik et al., 2014). These theories can all be described as phase-transition-like phenomena in state space driven by changes in hidden parameters.

Research on multistable perception typically presupposes that multiple mutually exclusive, consistent trajectories have already been acquired (Pisarchik et al., 2014). In light of the selective consistency framework—where the issue is the acquisition of the trajectories themselves—future work

might instead observe how multistable perception emerges during early development (Brown & Miracle, 2003; Shimojo et al., 1986), or experimentally create multistability by manipulating brain activity to control how trajectories are acquired for a novel stimulus using brain-stimulation methods (Carmel et al., 2010).

4.4.4. Biological Experiments Regarding Self-organised Selective Consistency and Its Constraints

As noted in the Limitations, validation using in vitro and invasive experiments is also essential. Project 1 relied on simulations and Project 2 on macroscopic non-invasive measurements; therefore, the argument needs to be complemented by microscopic, invasive recordings that bridge these two levels.

As developed in the Discussion, my results suggest that the capacity to acquire selective consistency depends on the structural properties of the system's connectivity. Because synaptic connections realise the connectivity matrix in the biological brain, it is structurally expected to be influenced by the functions of subplate circuits during development (De Carlos & O'Leary, 1992; Friauf et al., 1990; Kanold & Luhmann, 2010), synaptic pruning, and subsequent synaptic plasticity and dendric spine morphoplasticity (Hayashi-Takagi et al., 2015; Holtmaat & Svoboda, 2009; Paolicelli et al., 2011; Roberts et al., 2010), along with the resulting E-I balance (Poil et al., 2012; Vogels et al., 2011). The appropriate expression of genes and molecules, and glial cells that influence these factors, is also important (Chung et al., 2013; Hirai et al., 2005; Hori et al., 2020; Kim & Kandler, 2003; Ullian et al., 2001).

Dysfunctions in these elements have been extensively discussed in relation to neurological and psychiatric disorders and neurodevelopmental conditions (Marín, 2012; Rubenstein & Merzenich, 2003; Sekar et al., 2016; Sellgren et al., 2019; Yizhar et al., 2011). However, the logical gap between structural abnormalities and higher-order functional symptoms, such as atypical cognition, has not been fully resolved (Sellgren et al., 2019; Yizhar et al., 2011). A dynamical systems perspective—including SC—may help bridge this gap, because dynamical systems characterise the responses that emerge under structural constraints. Since information processing is grounded in system responses, this framework may provide a viable link between structure and function.

As a concrete experimental paradigm, for example, use optogenetics to activate a set of neurons in the sensory cortex in a fixed temporal sequence, and then quantify trial-to-trial consistency of the evoked neural activity (Carrillo-Reid et al., 2016; Marshel et al., 2019). Driving the same neuronal ensemble in the same order is formally analogous to repeated presentation of a particular sensory stimulus, as used in the present work. To minimise confounds from additional inputs from other

cortical areas, the barrel cortex in the mouse somatosensory system would be a promising target (Petersen, 2007). Because each barrel corresponds to an individual whisker and forms a relatively independent barrel-shaped cluster, it may allow a biologically grounded analogue of the reservoir used in Project 1, enabling a more direct mapping between model-based and physiological arguments.

By comparing such experiments across combinations such as developmental versus adult stages, neurodevelopmental-disorder models versus wild-type, or schizophrenia models versus wild-type, and evaluating consistency in each case, it may be possible to identify specific biological constraints that were inaccessible in the present study.

Assessing dynamical systems properties—including selective consistency—in such target neural systems could therefore serve as a bridge between "hardware" and "software" levels of explanation (Breakspear, 2017).

4.4.5. Cascading Selective Consistency

The selective consistency evaluated in this thesis was assessed in settings that primarily targeted the sensory cortex: the model was intended to represent sensory areas, and in the EEG data, the within-trial similarity was observed mainly in the temporal ROI, plausibly reflecting auditory cortical activity given that the sensory input was auditory. Although the cortex can be subdivided into layers, scalp EEG has limited spatial resolution and is influenced broadly by activity spanning roughly layers 2/3 to layer 5 rather than providing layer-specific information (Murakami & Okada, 2006). In general, sensory inputs first reach the cortex via thalamic relays (Sherman, 2007), and within the cortex, they are received primarily in layer 4 (Miller et al., 2001). Information processed in layer 4 is then propagated via layers 1–3 and 5–6 to many other targets, including other cortical regions (e.g., higher-order sensory and association cortices, sensorimotor areas) and subcortical structures (e.g., the hippocampus), giving rise to perceptual experience and behaviour. Given this architecture, it is unlikely that selective consistency acquisition occurs only within a specific layer of a sensory cortex.

For any neuronal population—defined in various senses, such as a cortical layer or area—the output of upstream populations serves as its input. This is the foundation of the hierarchical structure assumed in many computational models (Bastos et al., 2012; Friston, 2008; Friston & Kiebel, 2009a). Even in the present model, while motivated by sensory cortex, the architecture was simply an input–intermediate–output system. If we consider the brain as a whole, it is more appropriate to think in terms of a chain of similar systems (i.e., a deep recurrent neural network), in which the input to RNN_{n+1} is the output of RNN_n (Felleman & Van Essen, 1991). If the output of RNN_n is not consistent, then for the same stimulus s, the "input" $\boldsymbol{u}_t(s)$ to RNN_{n+1} would itself vary across trials. This was precisely the motivation for separating representation and decision in this thesis and assuming that acquiring

consistency at the representation stage must occur first. Because the only constraint imposed in my model was plasticity, the same logic could in principle apply to many parts of the nervous system. From this perspective, it is natural to hypothesise that selective consistency acquisition does not occur in a single locus, but emerges in a cascading manner across information-processing stages.

Although indirect, my results are consistent with such a cascading-SC hypothesis. Signals strong enough to be detectable in scalp EEG are unlikely to originate from highly localised activity confined to a specific layer in a single region (Murakami & Okada, 2006). Indeed, within-trial neural similarity during repetition perception (and, likewise, consistency for actually repeated stimuli) was observed not only in the temporal ROI—presumably reflecting auditory cortex—but also in the parietal ROI (and frontal ROI for the pooled GLMM).

Testing the cascading-SC hypothesis is difficult with EEG because of its limited spatial resolution. Future work will therefore require approaches such as wide-field calcium imaging and optogenetic perturbation for invasive observation and intervention, or non-invasive yet large-scale measurements such as MEG.

If the cascading acquisition of selective consistency is empirically supported, it could motivate a more unified complementary view of brain adaptation. In addition to conventional goal-directed "backward" learning theories—where downstream rewards or prediction errors drive upstream structural updates—the brain may also adapt through a dynamical "forward" process: self-organised increases in response selective consistency at upstream stages induce downstream selective consistency.

4.5. Significance of The Present Work

This thesis is organised as a three-layer structure: under the overarching theme of selective consistency, I adopted the NRD task as a model paradigm and developed two NRD projects grounded in the selective consistency hypothesis. Accordingly, the contributions of this work can be discussed at each layer; however, because the present section concerns the overall significance of the thesis, I focus here on the implications for selective-consistency research rather than the NRD-specific discussion (detailed in Section 4.1).

The central significance of the selective consistency hypothesis is that it provides a parsimonious framework for explaining the brain's dual requirement—overall flexibility alongside stimulus-conditioned consistency—in terms of self-organised changes in dynamical properties. Although contemporary neuroscience has increasingly recognised that neural variability reflects information processing, much work has nevertheless focused on how such variability is suppressed, or how

information processing can proceed despite it. However, the selective consistency hypothesis reframes the explanatory burden: rather than asking only where and how consistent information processing occurs, it motivates asking under what dynamical conditions exposed inputs yield consistent neural trajectories, and how experience changes those conditions in an input-specific manner. This perspective also clarifies why "more consistency" is not the goal: what matters is input-conditioned consistency that can be selectively expressed without collapsing overall variability.

It is also important that this objective can be achieved through self-organisation (or, at least, that it is possible). Although the results of Project 1 have limitations for direct translation to the brain, they nevertheless showed that a homeostatic Hebbian rule in a recurrent neural network can give rise to selective consistency. Because the mechanism does not rely on supervised tuning, it is attractive in that it requires relatively less speculation about how the proposed algorithm might be implemented in the real brain. Moreover, by formulating the process as self-organisation, I was able to make its constraints explicit: the acquisition of selective consistency depended strongly on the baseline dynamical regime, with an optimum near criticality. This indicates that the capacity to acquire selective consistency can vary substantially, and that this variation is plausibly constrained by intrinsic dynamical properties of the system. Because dynamical properties are constrained by structure, the framework also provides a route to connect circuit-level or resting-state markers to behavioural competence in a mechanistically interpretable way.

Another important, and more speculative, implication concerns the cascading form of selective consistency discussed in Section 4.4.5. Because cortical processing is hierarchical, inconsistent outputs from one stage imply inconsistent inputs to the next, motivating the idea that selective consistency may emerge in a cascading manner across populations. The present EEG findings are at least compatible with this picture, in that within-trial similarity during repetition perception was observed not only in temporal but also in parietal regions. Although I did not test this directly, expanding the simulation model into a deep recurrent neural network would allow this hypothesis to be evaluated.

If such cascading selective consistency is empirically supported in future work, it would motivate a complementary view of brain adaptation: beyond conventional "backward" learning—where downstream errors or rewards drive upstream updates—the brain may also adapt through a "forward" dynamical process, in which self-organised increases in selective consistency at upstream stages induce downstream selective consistency. A view in which whole-brain, cascading changes in dynamical properties directly shape information processing could provide a more unified framework, given that dynamical properties are constrained by underlying structure. That is, if structural determinants such as synapses or glial cells are compromised, as in some disorders, this could hinder changes in dynamical properties (e.g., selective consistency), thereby resulting in deficits in

information processing.

References

Agus, T. R., Carrión-Castillo, A., Pressnitzer, D., & Ramus, F. (2014). Perceptual learning of acoustic noise by individuals with dyslexia. *Journal of Speech, Language, and Hearing Research*, *57*(3), 1069–1077. https://doi.org/10.1044/1092-4388(2013/13-0020)

Agus, T. R., & Pressnitzer, D. (2021). Repetition detection and rapid auditory learning for stochastic tone clouds. *The Journal of the Acoustical Society of America*, *150*(3), 1735–1749. https://doi.org/10.1121/10.0005935

Agus, T. R., Thorpe, S. J., & Pressnitzer, D. (2010). Rapid formation of robust auditory memories: Insights from noise. *Neuron*, *66*(4), 610–618. https://doi.org/10.1016/j.neuron.2010.04.014

Alamian, G., Lajnef, T., Pascarella, A., Lina, J.-M., Knight, L., Walters, J., Singh, K. D., & Jerbi, K. (2022). Altered brain criticality in schizophrenia: New insights from magnetoencephalography. *Frontiers in Neural Circuits*, *16*, 630621. https://doi.org/10.3389/fncir.2022.630621

Amit, D. J., Gutfreund, H., & Sompolinsky, H. (1985). Storing infinite numbers of patterns in a spin-glass model of neural networks. *Physical Review Letters*, *55*(14), 1530–1533. https://doi.org/10.1103/PhysRevLett.55.1530

Andrillon, T., Kouider, S., Agus, T., & Pressnitzer, D. (2015). Perceptual learning of acoustic noise generates memory-evoked potentials. *Current Biology*, *25*(21), 2823–2829. https://doi.org/10.1016/j.cub.2015.09.027

Andrillon, T., Pressnitzer, D., Léger, D., & Kouider, S. (2017). Formation and suppression of acoustic memories during human sleep. *Nature Communications*, *8*(1), 179. https://doi.org/10.1038/s41467-017-00071-z

Arazi, A., Censor, N., & Dinstein, I. (2017). Neural Variability Quenching Predicts Individual

Perceptual Abilities. *The Journal of Neuroscience*, *37*(1), 97–109. https://doi.org/10.1523/JNEUROSCI.1671-16.2016

Arieli, A., Sterkin, A., Grinvald, A., & Aertsen, A. (1996). Dynamics of Ongoing Activity: Explanation of the Large Variability in Evoked Cortical Responses. *Science*, *273*(5283), 1868–1871. https://doi.org/10.1126/science.273.5283.1868

Arnal, L. H., Doelling, K. B., & Poeppel, D. (2015). Delta-beta coupled oscillations underlie temporal prediction accuracy. *Cerebral Cortex (New York, N.Y.: 1991)*, *25*(9), 3077–3085. https://doi.org/10.1093/cercor/bhu103

Arviv, O., Medvedovsky, M., Sheintuch, L., Goldstein, A., & Shriki, O. (2016). Deviations from critical dynamics in interictal epileptiform activity. *Journal of Neuroscience*, *36*(48), 12276–12292. https://doi.org/10.1523/JNEUROSCI.0809-16.2016

Averbeck, B. B., Latham, P. E., & Pouget, A. (2006). Neural correlations, population coding and computation. *Nature Reviews Neuroscience*, *7*(5), 358–366. https://doi.org/10.1038/nrn1888

Aydore, S., Pantazis, D., & Leahy, R. M. (2013). A note on the phase locking value and its properties. *NeuroImage*, *74*, 231–244. https://doi.org/10.1016/j.neuroimage.2013.02.008

Bak, P., Tang, C., & Wiesenfeld, K. (1987). Self-organized criticality: An explanation of the 1/*f* noise. *Physical Review Letters*, *59*(4), 381–384. https://doi.org/10.1103/PhysRevLett.59.381

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The autism-spectrum quotient (AQ): Evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders*, *31*(1), 5–17. https://doi.org/10.1023/a:1005653411471

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*(4), 695–711. https://doi.org/10.1016/j.neuron.2012.10.038

Bédard, C., Kröger, H., & Destexhe, A. (2006). Does the 1 / f frequency scaling of brain signals reflect self-organized critical states? *Physical Review Letters*, *97*(11), 118102. https://doi.org/10.1103/PhysRevLett.97.118102

Beggs, J. M. (2008). The criticality hypothesis: How local cortical networks might optimize information processing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *366*(1864), 329–343. https://doi.org/10.1098/rsta.2007.2092

Beggs, J. M. (2019). The critically tuned cortex. *Neuron*, *104*(4), 623–624. https://doi.org/10.1016/j.neuron.2019.10.039

Beggs, J. M., & Plenz, D. (2003). Neuronal Avalanches in Neocortical Circuits. *The Journal of Neuroscience*, *23*(35), 11167–11177. https://doi.org/10.1523/JNEUROSCI.23-35-11167.2003

Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N., & Golani, I. (2001). Controlling the false discovery rate in behavior genetics research. *Behavioural Brain Research*, *125*(1–2), 279–284. https://doi.org/10.1016/s0166-4328(01)00297-2

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Berkes, P., Orbán, G., Lengyel, M., & Fiser, J. (2011). Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science (New York, N.Y.)*, *331*(6013), 83–87. https://doi.org/10.1126/science.1195870

Berndt, D. J. (1994). Using dynamic time warping to find patterns in time series. *Proceedings of AAAI Workshop on Knowledge Discovery in Databases, 1994*, 359–370.

Bertschinger, N., & Natschläger, T. (2004). Real-Time Computation at the Edge of Chaos in Recurrent

Neural Networks. *Neural Computation*, *16*(7), 1413–1436. https://doi.org/10.1162/089976604323057443

Boccaletti, S., Kurths, J., Osipov, G., Valladares, D. L., & Zhou, C. S. (2002). The synchronization of chaotic systems. *Physics Reports*, *366*(1), 1–101. https://doi.org/10.1016/S0370-1573(02)00137-0

Boedecker, J., Obst, O., Lizier, J. T., Mayer, N. M., & Asada, M. (2012). Information processing in echo state networks at the edge of chaos. *Theory in Biosciences*, *131*(3), 205–213. https://doi.org/10.1007/s12064-011-0146-8

Bottou, L., Curtis, F. E., & Nocedal, J. (2018). Optimization methods for large-scale machine learning. *SIAM Review*, *60*(2), 223–311. https://doi.org/10.1137/16M1080173

Brascamp, J., Sterzer, P., Blake, R., & Knapen, T. (2018). Multistable Perception and the Role of the Frontoparietal Cortex in Perceptual Inference. *Annual Review of Psychology*, *69*(Volume 69, 2018), 77–103. https://doi.org/10.1146/annurev-psych-010417-085944

Breakspear, M. (2017). Dynamic models of large-scale brain activity. *Nature Neuroscience*, *20*(3), 340–352. https://doi.org/10.1038/nn.4497

Breslow, N. E., & Clayton, D. G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, *88*(421), 9–25. https://doi.org/10.1080/01621459.1993.10594284

Brown, A. M., & Miracle, J. A. (2003). Early binocular vision in human infants: Limitations on the generality of the Superposition Hypothesis. *Vision Research*, *43*(14), 1563–1574. https://doi.org/10.1016/S0042-6989(03)00177-9

Bruining, H., Hardstone, R., Juarez-Martinez, E. L., Sprengers, J., Avramiea, A.-E., Simpraga, S., Houtman, S. J., Poil, S.-S., Dallares, E., Palva, S., Oranje, B., Matias Palva, J., Mansvelder, H. D., & Linkenkaer-Hansen, K. (2020). Measurement of excitation-inhibition ratio in autism

spectrum disorder using critical brain dynamics. *Scientific Reports*, *10*(1), 9195. https://doi.org/10.1038/s41598-020-65500-4

Buehner, M., & Young, P. (2006). A tighter bound for the echo state property. *IEEE Transactions on Neural Networks*, *17*(3), 820–824. https://doi.org/10.1109/TNN.2006.872357

Buonomano, D. V., & Maass, W. (2009). State-dependent computations: Spatiotemporal processing in cortical networks. *Nature Reviews Neuroscience*, *10*(2), 113–125. https://doi.org/10.1038/nrn2558

Caras, M. L., & Sanes, D. H. (2019). Neural Variability Limits Adolescent Skill Learning. *The Journal of Neuroscience*, *39*(15), 2889–2902. https://doi.org/10.1523/JNEUROSCI.2878-18.2019

Carmel, D., Walsh, V., Lavie, N., & Rees, G. (2010). Right parietal TMS shortens dominance durations in binocular rivalry. *Current Biology*, *20*(18), R799-800. https://doi.org/10.1016/j.cub.2010.07.036

Carrillo-Reid, L., Yang, W., Bando, Y., Peterka, D. S., & Yuste, R. (2016). Imprinting and recalling cortical ensembles. *Science*, *353*(6300), 691–694. https://doi.org/10.1126/science.aaf7560

Carter, C. S., Braver, T. S., Barch, D. M., Botvinick, M. M., Noll, D., & Cohen, J. D. (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science*, *280*(5364), 747–749. https://doi.org/10.1126/science.280.5364.747

Chen, J.-Y., Lonjers, P., Lee, C., Chistiakova, M., Volgushev, M., & Bazhenov, M. (2013). Heterosynaptic plasticity prevents runaway synaptic dynamics. *Journal of Neuroscience*, *33*(40), 15915–15929. https://doi.org/10.1523/JNEUROSCI.5088-12.2013

Chialvo, D. R. (2010). Emergent complex neural dynamics. *Nature Physics*, *6*(10), 744–750. https://doi.org/10.1038/nphys1803

Chua, L., Sbitnev, V., & Kim, H. (2012). NEURONS ARE POISED NEAR THE EDGE OF CHAOS. *International Journal of Bifurcation and Chaos*, *22*(04), 1250098.

https://doi.org/10.1142/S0218127412500988

Chung, W.-S., Clarke, L. E., Wang, G. X., Stafford, B. K., Sher, A., Chakraborty, C., Joung, J., Foo, L. C., Thompson, A., Chen, C., Smith, S. J., & Barres, B. A. (2013). Astrocytes mediate synapse elimination through MEGF10 and MERTK pathways. *Nature*, *504*(7480), 394–400. https://doi.org/10.1038/nature12776

Churchland, M. M., Yu, B. M., Cunningham, J. P., Sugrue, L. P., Cohen, M. R., Corrado, G. S., Newsome, W. T., Clark, A. M., Hosseini, P., Scott, B. B., Bradley, D. C., Smith, M. A., Kohn, A., Movshon, J. A., Armstrong, K. M., Moore, T., Chang, S. W., Snyder, L. H., Lisberger, S. G., … Shenoy, K. V. (2010). Stimulus onset quenches neural variability: A widespread cortical phenomenon. *Nature Neuroscience*, *13*(3), 369–378. https://doi.org/10.1038/nn.2501

Cocchi, L., Gollo, L. L., Zalesky, A., & Breakspear, M. (2017). Criticality in the brain: A synthesis of neurobiology, models and cognition. *Progress in Neurobiology*, *158*, 132–152. https://doi.org/10.1016/j.pneurobio.2017.07.002

Cohen, M. R., & Maunsell, J. H. R. (2009). Attention improves performance primarily by reducing interneuronal correlations. *Nature Neuroscience*, *12*(12), 1594–1600. https://doi.org/10.1038/nn.2439

Cohen, M. X. (2014). *Analyzing Neural Time Series Data: Theory and Practice*. The MIT Press. https://doi.org/10.7551/mitpress/9609.001.0001

Colombo, M. A., Napolitani, M., Boly, M., Gosseries, O., Casarotto, S., Rosanova, M., Brichant, J.-F., Boveroux, P., Rex, S., Laureys, S., Massimini, M., Chieregato, A., & Sarasso, S. (2019). The spectral exponent of the resting EEG indexes the presence of consciousness during unresponsiveness induced by propofol, xenon, and ketamine. *NeuroImage*, *189*, 631–644. https://doi.org/10.1016/j.neuroimage.2019.01.024

Cook, R. D. (1977). Detection of Influential Observation in Linear Regression. *Technometrics*, *19*(1),

15–18. https://doi.org/10.1080/00401706.1977.10489493

D'amour, J. A., & Froemke, R. C. (2015). Inhibitory and Excitatory Spike-Timing-Dependent Plasticity in the Auditory Cortex. *Neuron*, *86*(2), 514–528. https://doi.org/10.1016/j.neuron.2015.03.014

Daniel, E., & Dinstein, I. (2021). Individual magnitudes of neural variability quenching are associated with motion perception abilities. *Journal of Neurophysiology*, *125*(4), 1111–1120. https://doi.org/10.1152/jn.00355.2020

de Arcangelis, L., & Herrmann, H. J. (2010). Learning as a phenomenon occurring in a critical state. *Proceedings of the National Academy of Sciences*, *107*(9), 3977–3981. https://doi.org/10.1073/pnas.0912289107

De Carlos, J. A., & O'Leary, D. D. (1992). Growth and targeting of subplate axons and establishment of major cortical pathways. *The Journal of Neuroscience*, *12*(4), 1194–1211. https://doi.org/10.1523/JNEUROSCI.12-04-01194.1992

De Martino, F., Moerel, M., van de Moortele, P.-F., Ugurbil, K., Goebel, R., Yacoub, E., & Formisano, E. (2013). Spatial organization of frequency preference and selectivity in the human inferior colliculus. *Nature Communications*, *4*(1), 1386. https://doi.org/10.1038/ncomms2379

Del Papa, B., Priesemann, V., & Triesch, J. (2017). Criticality meets learning: Criticality signatures in a self-organizing recurrent neural network. *PLOS ONE*, *12*(5), e0178683. https://doi.org/10.1371/journal.pone.0178683

Denham, S. L., & Winkler, I. (2020). Predictive coding in auditory perception: Challenges and unresolved questions. *European Journal of Neuroscience*, *51*(5), 1151–1160. https://doi.org/10.1111/ejn.13802

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, *11*(8), 333–341. https://doi.org/10.1016/j.tics.2007.06.010

Dinstein, I., Heeger, D. J., & Behrmann, M. (2015). Neural variability: Friend or foe? *Trends in Cognitive Sciences*, *19*(6), 322–328. https://doi.org/10.1016/j.tics.2015.04.005

Donoghue, T., Haller, M., Peterson, E. J., Varma, P., Sebastian, P., Gao, R., Noto, T., Lara, A. H., Wallis, J. D., Knight, R. T., Shestyuk, A., & Voytek, B. (2020). Parameterizing neural power spectra into periodic and aperiodic components. *Nature Neuroscience*, *23*(12), 1655–1665. https://doi.org/10.1038/s41593-020-00744-x

Dorogovtsev, S. N., Goltsev, A. V., & Mendes, J. F. F. (2008). Critical phenomena in complex networks. *Reviews of Modern Physics*, *80*(4), 1275–1335. https://doi.org/10.1103/RevModPhys.80.1275

Dosher, B. A., Jeter, P., Liu, J., & Lu, Z.-L. (2013). An integrated reweighting theory of perceptual learning. *Proceedings of the National Academy of Sciences*, *110*(33), 13678–13683. https://doi.org/10.1073/pnas.1312552110

Doupe, A. J., & Solis, M. M. (1997). Song- and order-selective neurons develop in the songbird anterior forebrain during vocal learning. *Journal of Neurobiology*, *33*(5), 694–709. https://doi.org/10.1002/(SICI)1097-4695(19971105)33:5%253C694::AID-NEU13%253E3.0.CO;2-9

Eckmann, J.-P., & Ruelle, D. (1985). Ergodic theory of chaos and strange attractors. *Reviews of Modern Physics*, *57*(3), 617–656. https://doi.org/10.1103/RevModPhys.57.617

Enel, P., Procyk, E., Quilodran, R., & Dominey, P. F. (2016). Reservoir Computing Properties of Neural Dynamics in Prefrontal Cortex. *PLOS Computational Biology*, *12*(6), e1004967. https://doi.org/10.1371/journal.pcbi.1004967

Faisal, A. A., Selen, L. P. J., & Wolpert, D. M. (2008). Noise in the nervous system. *Nature Reviews Neuroscience*, *9*(4), 292–303. https://doi.org/10.1038/nrn2258

Feldman, H., & Friston, K. J. (2010). Attention, Uncertainty, and Free-Energy. *Frontiers in Human*

Neuroscience, *4*, 215. https://doi.org/10.3389/fnhum.2010.00215

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, *1*(1), 1–47. https://doi.org/10.1093/cercor/1.1.1-a

Fletcher, H., & Munson, W. A. (1933). Loudness, its definition, measurement and calculation. *The Journal of the Acoustical Society of America*, *5*(2), 82–108. https://doi.org/10.1121/1.1915637

Friauf, E., McConnell, S. K., & Shatz, C. J. (1990). Functional synaptic circuits in the subplate during fetal and early postnatal development of cat visual cortex. *The Journal of Neuroscience*, *10*(8), 2601–2613. https://doi.org/10.1523/JNEUROSCI.10-08-02601.1990

Friedman, N., Ito, S., Brinkman, B. A. W., Shimono, M., DeVille, R. E. L., Dahmen, K. A., Beggs, J. M., & Butler, T. C. (2012). Universal critical dynamics in high resolution neuronal avalanche data. *Physical Review Letters*, *108*(20), 208102. https://doi.org/10.1103/PhysRevLett.108.208102

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622

Friston, K. (2008). Hierarchical models in the brain. *PLOS Computational Biology*, *4*(11), e1000211. https://doi.org/10.1371/journal.pcbi.1000211

Friston, K., & Kiebel, S. (2009a). Cortical circuits for perceptual inference. *Neural Networks, Cortical Microcircuits*, *22*(8), 1093–1104. https://doi.org/10.1016/j.neunet.2009.07.023

Friston, K., & Kiebel, S. (2009b). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1211–1221. https://doi.org/10.1098/rstb.2008.0300

Friston, K., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology-Paris, Theoretical and Computational Neuroscience: Understanding Brain*

Functions, *100*(1), 70–87. https://doi.org/10.1016/j.jphysparis.2006.10.001

Gao, R., Peterson, E. J., & Voytek, B. (2017). Inferring synaptic excitation/inhibition balance from field potentials. *NeuroImage*, *158*, 70–78. https://doi.org/10.1016/j.neuroimage.2017.06.078

Garrett, D. D., Kovacevic, N., McIntosh, A. R., & Grady, C. L. (2011). The Importance of Being Variable. *The Journal of Neuroscience*, *31*(12), 4496–4503. https://doi.org/10.1523/JNEUROSCI.5641-10.2011

Gilbert, C. D., Sigman, M., & Crist, R. E. (2001). The neural basis of perceptual learning. *Neuron*, *31*(5), 681–697. https://doi.org/10.1016/s0896-6273(01)00424-x

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, *30*(Volume 30, 2007), 535–574. https://doi.org/10.1146/annurev.neuro.29.051605.113038

Goris, R. L. T., Movshon, J. A., & Simoncelli, E. P. (2014). Partitioning neuronal variability. *Nature Neuroscience*, *17*(6), 858–865. https://doi.org/10.1038/nn.3711

Goto, Y., Hagihara, M., & Kitajo, K. (2024). EEG dynamics related to noise repetition-detection performance. *OSF*. https://doi.org/10.17605/OSF.IO/TYAB6

Goto, Y., & Kitajo, K. (2024). Selective consistency of recurrent neural networks induced by plasticity as a mechanism of unsupervised perceptual learning. *PLOS Computational Biology*, *20*(9), e1012378. https://doi.org/10.1371/journal.pcbi.1012378

Hanks, T. D., Ditterich, J., & Shadlen, M. N. (2006). Microstimulation of macaque area LIP affects decision-making in a motion discrimination task. *Nature Neuroscience*, *9*(5), 682–689. https://doi.org/10.1038/nn1683

Harris, K. D., & Thiele, A. (2011). Cortical state and attention. *Nature Reviews Neuroscience*, *12*(9), 509–523. https://doi.org/10.1038/nrn3084

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed

and overlapping representations of faces and objects in ventral temporal cortex. *Science (New York, N.Y.)*, *293*(5539), 2425–2430. https://doi.org/10.1126/science.1063736

Hayashi-Takagi, A., Yagishita, S., Nakamura, M., Shirai, F., Wu, Y. I., Loshbaugh, A. L., Kuhlman, B., Hahn, K. M., & Kasai, H. (2015). Labelling and optical erasure of synaptic memory traces in the motor cortex. *Nature*, *525*(7569), 333–338. https://doi.org/10.1038/nature15257

He, B. J. (2013). Spontaneous and task-evoked brain activity negatively interact. *The Journal of Neuroscience*, *33*(11), 4672–4682. https://doi.org/10.1523/JNEUROSCI.2922-12.2013

He, B. J. (2014a). Scale-free brain activity: Past, present, and future. *Trends in Cognitive Sciences*, *18*(9), 480–487. https://doi.org/10.1016/j.tics.2014.04.003

He, B. J. (2014b). Scale-free brain activity: Past, present, and future. *Trends in Cognitive Sciences*, *18*(9), 480–487. https://doi.org/10.1016/j.tics.2014.04.003

Hengen, K. B., & Shew, W. L. (2025). Is criticality a unified setpoint of brain function? *Neuron*, *113*(16), 2582-2598.e2. https://doi.org/10.1016/j.neuron.2025.05.020

Hesse, J., & Gross, T. (2014). Self-organized criticality as a fundamental property of neural systems. *Frontiers in Systems Neuroscience*, *8*, 166. https://doi.org/10.3389/fnsys.2014.00166

Hillyard, S. A., & Picton, T. W. (1978). ON and OFF components in the auditory evoked potential. *Perception & Psychophysics*, *24*(5), 391–398. https://doi.org/10.3758/BF03199736

Hipp, J. F., Engel, A. K., & Siegel, M. (2011). Oscillatory synchronization in large-scale cortical networks predicts perception. *Neuron*, *69*(2), 387–396. https://doi.org/10.1016/j.neuron.2010.12.027

Hirai, H., Pang, Z., Bao, D., Miyazaki, T., Li, L., Miura, E., Parris, J., Rong, Y., Watanabe, M., Yuzaki, M., & Morgan, J. I. (2005). Cbln1 is essential for synaptic integrity and plasticity in the cerebellum. *Nature Neuroscience*, *8*(11), 1534–1541. https://doi.org/10.1038/nn1576

Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: An

epistemological review. *Cognition*, *108*(3), 687–701. https://doi.org/10.1016/j.cognition.2008.05.010

Holtmaat, A., & Svoboda, K. (2009). Experience-dependent structural synaptic plasticity in the mammalian brain. *Nature Reviews Neuroscience*, *10*(9), 647–658. https://doi.org/10.1038/nrn2699

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*(8), 2554–2558. https://doi.org/10.1073/pnas.79.8.2554

Hori, K., Yamashiro, K., Nagai, T., Shan, W., Egusa, S. F., Shimaoka, K., Kuniishi, H., Sekiguchi, M., Go, Y., Tatsumoto, S., Yamada, M., Shiraishi, R., Kanno, K., Miyashita, S., Sakamoto, A., Abe, M., Sakimura, K., Sone, M., Sohya, K., … Hoshino, M. (2020). AUTS2 Regulation of Synapses for Proper Synaptic Inputs and Social Communication. *iScience*, *23*(6), 101183. https://doi.org/10.1016/j.isci.2020.101183

Houser, D. S., Yost, W., Burkard, R., Finneran, J. J., Reichmuth, C., & Mulsow, J. (2017). A review of the history, development and application of auditory weighting functions in humans and marine mammals. *The Journal of the Acoustical Society of America*, *141*(3), 1371. https://doi.org/10.1121/1.4976086

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*(1), 106–154. https://doi.org/10.1113/jphysiol.1962.sp006837

Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, *10*(3), 626–634. https://doi.org/10.1109/72.761722

Jaeger, H., & Haas, H. (2004). Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication. *Science*, *304*(5667), 78–80.

https://doi.org/10.1126/science.1091277

Jaeger, H., Maass, W., & Principe, J. (2007). Special issue on echo state networks and liquid state machines. *Neural Networks*, *20*(3), 287–289. https://doi.org/10.1016/j.neunet.2007.04.001

Jehee, J. F. M., Ling, S., Swisher, J. D., van Bergen, R. S., & Tong, F. (2012). Perceptual learning selectively refines orientation representations in early visual cortex. *The Journal of Neuroscience*, *32*(47), 16747–16753a. https://doi.org/10.1523/JNEUROSCI.6112-11.2012

Johnson, D. M. (1939). *Confidence and speed in the two-category judgment*. [s.n.]. https://cir.nii.ac.jp/crid/1970867909835657510

Jozwik, K. M., O'Keeffe, J., Storrs, K. R., Guo, W., Golan, T., & Kriegeskorte, N. (2022). Face dissimilarity judgments are predicted by representational distance in morphable and image-computable models. *Proceedings of the National Academy of Sciences*, *119*(27), e2115047119. https://doi.org/10.1073/pnas.2115047119

Kang, H., Agus, T. R., & Pressnitzer, D. (2017). Auditory memory for random time patterns. *The Journal of the Acoustical Society of America*, *142*(4), 2219–2232. https://doi.org/10.1121/1.5007730

Kang, H., Auksztulewicz, R., An, H., Abi Chacra, N., Sutter, M. L., & Schnupp, J. W. H. (2021). Neural correlates of auditory pattern learning in the auditory cortex. *Frontiers in Neuroscience*, *15*, 610978. https://doi.org/10.3389/fnins.2021.610978

Kang, H., Lancelin, D., & Pressnitzer, D. (2018). Memory for random time patterns in audition, touch, and vision. *Neuroscience*, *389*, 118–132. https://doi.org/10.1016/j.neuroscience.2018.03.017

Kanold, P. O., & Luhmann, H. J. (2010). The subplate and early cortical circuits. *Annual Review of Neuroscience*, *33*, 23–48. https://doi.org/10.1146/annurev-neuro-060909-153244

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience*, *17*(11),

4302–4311. https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997

Kayser, C., Petkov, C. I., & Logothetis, N. K. (2007). Tuning to sound frequency in auditory field potentials. *Journal of Neurophysiology*, *98*(3), 1806–1809. https://doi.org/10.1152/jn.00358.2007

Kessler, R. C., Adler, L., Ames, M., Demler, O., Faraone, S., Hiripi, E., Howes, M. J., Jin, R., Secnik, K., Spencer, T., Ustun, T. B., & Walters, E. E. (2005). The World Health Organization adult ADHD self-report scale (ASRS): A short screening scale for use in the general population. *Psychological Medicine*, *35*(2), 245–256. https://doi.org/10.1017/S0033291704002892

Keuken, M. C., Müller-Axt, C., Langner, R., Eickhoff, S. B., Forstmann, B. U., & Neumann, J. (2014). Brain networks of perceptual decision-making: An fMRI ALE meta-analysis. *Frontiers in Human Neuroscience*, *8*, 445. https://doi.org/10.3389/fnhum.2014.00445

Kilian-Hütten, N., Valente, G., Vroomen, J., & Formisano, E. (2011). Auditory cortex encodes the perceptual interpretation of ambiguous sound. *The Journal of Neuroscience*, *31*(5), 1715–1720. https://doi.org/10.1523/JNEUROSCI.4572-10.2011

Kim, G., & Kandler, K. (2003). Elimination and strengthening of glycinergic/GABAergic connections during tonotopic map formation. *Nature Neuroscience*, *6*(3), 282–290. https://doi.org/10.1038/nn1015

Kleiner, M., Brainard, D. H., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3. *Perception*, *36*, 1–16. https://doi.org/10.1068/v070821

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–719. https://doi.org/10.1016/j.tins.2004.10.007

Kocarev, L., & Parlitz, U. (1996). Generalized synchronization, predictability, and equivalence of unidirectionally coupled dynamical systems. *Physical Review Letters*, *76*(11), 1816–1819.

https://doi.org/10.1103/PhysRevLett.76.1816

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*(8), 401–412. https://doi.org/10.1016/j.tics.2013.06.007

Kumar, S., Bonnici, H. M., Teki, S., Agus, T. R., Pressnitzer, D., Maguire, E. A., & Griffiths, T. D. (2014). Representations of specific acoustic patterns in the auditory cortex and hippocampus. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1791), 20141000. https://doi.org/10.1098/rspb.2014.1000

Kumar, S., Strachan, J. P., & Williams, R. S. (2017). Chaotic dynamics in nanoscale NbO2 Mott memristors for analogue computing. *Nature*, *548*(7667), 318–321. https://doi.org/10.1038/nature23307

Lancaster, G., Iatsenko, D., Pidde, A., Ticcinelli, V., & Stefanovska, A. (2018). Surrogate data for hypothesis testing of physical systems. *Physics Reports, Surrogate Data for Hypothesis Testing of Physical Systems*, *748*, 1–60. https://doi.org/10.1016/j.physrep.2018.06.001

Larson, E., Gramfort, A., Engemann, D. A., Leppakangas, J., Brodbeck, C., Jas, M., Brooks, T. L., Sassenhagen, J., McCloy, D., Luessi, M., King, J.-R., Höchenberger, R., Brunner, C., Goj, R., Favelier, G., van Vliet, M., Wronkiewicz, M., Appelhoff, S., Rockhill, A., … user27182. (2025). *MNE-Python* [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.17675410

Leopold, D. A., & Logothetis, N. K. (1999). Multistable phenomena: Changing views in perception. *Trends in Cognitive Sciences*, *3*(7), 254–264. https://doi.org/10.1016/s1364-6613(99)01332-7

Lind, J. T., & Mehlum, H. (2010). With or without U? The appropriate test for a U-shaped relationship. *Oxford Bulletin of Economics and Statistics*, *72*(1), 109–118. https://doi.org/10.1111/j.1468-

0084.2009.00569.x

Lukoševičius, M. (2012). A practical guide to applying echo state networks. In G. Montavon, G. B. Orr, & K.-R. Müller (Eds), *Neural Networks: Tricks of the Trade: Second Edition* (pp. 659–686). Springer. https://doi.org/10.1007/978-3-642-35289-8_36

Lukoševičius, M., & Jaeger, H. (2009). Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, *3*(3), 127–149. https://doi.org/10.1016/j.cosrev.2009.03.005

Luo, H., Tian, X., Song, K., Zhou, K., & Poeppel, D. (2013). Neural response phase tracks how listeners learn new acoustic representations. *Current Biology*, *23*(11), 968–974. https://doi.org/10.1016/j.cub.2013.04.031

Maass, W. (1997). Networks of spiking neurons: The third generation of neural network models. *Neural Networks*, *10*(9), 1659–1671. https://doi.org/10.1016/S0893-6080(97)00011-7

Maeda, E., Robinson, H. P., & Kawana, A. (1995). The mechanisms of generation and propagation of synchronized bursting in developing networks of cortical neurons. *Journal of Neuroscience*, *15*(10), 6834–6845. https://doi.org/10.1523/JNEUROSCI.15-10-06834.1995

Manjunath, G., & Jaeger, H. (2013). Echo State Property Linked to an Input: Exploring a Fundamental Characteristic of Recurrent Neural Networks. *Neural Computation*, *25*(3), 671–696. https://doi.org/10.1162/NECO_a_00411

Marín, O. (2012). Interneuron dysfunction in psychiatric disorders. *Nature Reviews. Neuroscience*, *13*(2), 107–120. https://doi.org/10.1038/nrn3155

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Markicevic, M., Fulcher, B. D., Lewis, C., Helmchen, F., Rudin, M., Zerbi, V., & Wenderoth, N. (2020). Cortical Excitation:Inhibition Imbalance Causes Abnormal Brain Network Dynamics as

Observed in Neurodevelopmental Disorders. *Cerebral Cortex (New York, N.Y.: 1991)*, *30*(9), 4922–4937. https://doi.org/10.1093/cercor/bhaa084

Markram, H., Lübke, J., Frotscher, M., & Sakmann, B. (1997). Regulation of Synaptic Efficacy by Coincidence of Postsynaptic APs and EPSPs. *Science*, *275*(5297), 213–215. https://doi.org/10.1126/science.275.5297.213

Marshel, J. H., Kim, Y. S., Machado, T. A., Quirin, S., Benson, B., Kadmon, J., Raja, C., Chibukhchyan, A., Ramakrishnan, C., Inoue, M., Shane, J. C., McKnight, D. J., Yoshizawa, S., Kato, H. E., Ganguli, S., & Deisseroth, K. (2019). Cortical layer-specific critical dynamics triggering perception. *Science*, *365*(6453), 558. https://doi.org/10.1126/science.aaw5202

Masquelier, T. (2018). STDP allows close-to-optimal spatiotemporal spike pattern detection by single coincidence detector neurons. *Neuroscience*, *389*, 133–140. https://doi.org/10.1016/j.neuroscience.2017.06.032

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic Feature Encoding in Human Superior Temporal Gyrus. *Science*, *343*(6174), 1006–1010. https://doi.org/10.1126/science.1245994

Miller, K. D., Pinto, D. J., & Simons, D. J. (2001). Processing in layer 4 of the neocortical circuit: New insights from visual and somatosensory cortex. *Current Opinion in Neurobiology*, *11*(4), 488–497. https://doi.org/10.1016/S0959-4388(00)00239-7

Montez, D. F., Calabro, F. J., & Luna, B. (n.d.). The expression of established cognitive brain states stabilizes with working memory development. *eLife*, *6*, e25606. https://doi.org/10.7554/eLife.25606

Moreno-Bote, R., Beck, J., Kanitscheider, I., Pitkow, X., Latham, P., & Pouget, A. (2014). Information-limiting correlations. *Nature Neuroscience*, *17*(10), 1410–1417. https://doi.org/10.1038/nn.3807

Moreno-Bote, R., Rinzel, J., & Rubin, N. (2007). Noise-induced alternations in an attractor network model of perceptual bistability. *Journal of Neurophysiology*, *98*(3), 1125–1139. https://doi.org/10.1152/jn.00116.2007

Murakami, S., & Okada, Y. (2006). Contributions of principal neocortical neurons to magnetoencephalography and electroencephalography signals. *The Journal of Physiology*, *575*(Pt 3), 925–936. https://doi.org/10.1113/jphysiol.2006.105379

Naik, S., Adibpour, P., Dubois, J., Dehaene-Lambertz, G., & Battaglia, D. (2023). Event-related variability is modulated by task and development. *NeuroImage*, *276*, 120208. https://doi.org/10.1016/j.neuroimage.2023.120208

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology, Special Issue: Dynamic Decision Making*, *53*(3), 139–154. https://doi.org/10.1016/j.jmp.2008.12.005

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430. https://doi.org/10.1016/j.tics.2006.07.005

Nourski, K. V., Steinschneider, M., McMurray, B., Kovach, C. K., Oya, H., Kawasaki, H., & Howard, M. A. (2014). Functional organization of human auditory cortex: Investigation of response latencies through direct recordings. *NeuroImage*, *101*, 598–609. https://doi.org/10.1016/j.neuroimage.2014.07.004

O'Connell, R. G., Dockree, P. M., & Kelly, S. P. (2012). A supramodal accumulation-to-bound signal that determines perceptual decisions in humans. *Nature Neuroscience*, *15*(12), 1729–1735. https://doi.org/10.1038/nn.3248

Oja, E. (1982). Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, *15*(3), 267–273. https://doi.org/10.1007/BF00275687

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*,

349(6251), aac4716. https://doi.org/10.1126/science.aac4716

Orbán, G., Berkes, P., Fiser, J., & Lengyel, M. (2016). Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron*, *92*(2), 530–543. https://doi.org/10.1016/j.neuron.2016.09.038

O'Reilly, R. C., & McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and recall: Avoiding a trade-off. *Hippocampus*, *4*(6), 661–682. https://doi.org/10.1002/hipo.450040605

Owens, A. P., Allen, M., Ondobaka, S., & Friston, K. J. (2018). Interoceptive inference: From computational neuroscience to clinic. *Neuroscience and Biobehavioral Reviews*, *90*, 174–183. https://doi.org/10.1016/j.neubiorev.2018.04.017

Paolicelli, R. C., Bolasco, G., Pagani, F., Maggi, L., Scianni, M., Panzanelli, P., Giustetto, M., Ferreira, T. A., Guiducci, E., Dumas, L., Ragozzino, D., & Gross, C. T. (2011). Synaptic Pruning by Microglia Is Necessary for Normal Brain Development. *Science*, *333*(6048), 1456–1458. https://doi.org/10.1126/science.1202529

Parker, D. M., Salzen, E. A., & Lishman, J. R. (1982). Visual-evoked responses elicited by the onset and offset of sinusoidal gratings: Latency, waveform, and topographic characteristics. *Investigative Ophthalmology & Visual Science*, *22*(5), 675–680.

Parr, T., & Friston, K. J. (2017). Uncertainty, epistemics and active inference. *Journal of The Royal Society Interface*, *14*(136), 20170376. https://doi.org/10.1098/rsif.2017.0376

Pecora, L. M., & Carroll, T. L. (1990). Synchronization in chaotic systems. *Physical Review Letters*, *64*(8), 821–824. https://doi.org/10.1103/PhysRevLett.64.821

Petersen, C. C. H. (2007). The functional organization of the barrel cortex. *Neuron*, *56*(2), 339–355. https://doi.org/10.1016/j.neuron.2007.09.017

Pisarchik, A. N., Jaimes-Reátegui, R., Magallón-García, C. D., & Castillo-Morales, C. O. (2014). Critical slowing down and noise-induced intermittency in bistable perception: Bifurcation

analysis. *Biological Cybernetics*, *108*(4), 397–404. https://doi.org/10.1007/s00422-014-0607-5

Plenz, D., Ribeiro, T. L., Miller, S. R., Kells, P. A., Vakili, A., & Capek, E. L. (2021). Self-Organized Criticality in the Brain. *Frontiers in Physics*, *9*, 639389. https://doi.org/10.3389/fphy.2021.639389

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, *117*(3), 864–901. https://doi.org/10.1037/a0019737

Poil, S.-S., Hardstone, R., Mansvelder, H. D., & Linkenkaer-Hansen, K. (2012). Critical-state dynamics of avalanches and oscillations jointly emerge from balanced excitation/inhibition in neuronal networks. *The Journal of Neuroscience*, *32*(29), 9817–9823. https://doi.org/10.1523/JNEUROSCI.5990-11.2012

Poort, J., Khan, A. G., Pachitariu, M., Nemri, A., Orsolic, I., Krupic, J., Bauza, M., Sahani, M., Keller, G. B., Mrsic-Flogel, T. D., & Hofer, S. B. (2015). Learning enhances sensory and multiple non-sensory representations in primary visual cortex. *Neuron*, *86*(6), 1478–1490. https://doi.org/10.1016/j.neuron.2015.05.037

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, *435*(7045), 1102–1107. https://doi.org/10.1038/nature03687

R. Kandel, E., D. Koester, J., H. Mack, S., & A. Siegelbaum, S. (2021). *Principles of neural science* (6th edn). McGraw Hill. https://doi.org/10.1036/9780071390118

Rajan, K., & Abbott, L. F. (2006). Eigenvalue spectra of random matrices for neural networks. *Physical Review Letters*, *97*(18), 188104. https://doi.org/10.1103/PhysRevLett.97.188104

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional

interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87. https://doi.org/10.1038/4580

Ridderinkhof, K. R., Ullsperger, M., Crone, E. A., & Nieuwenhuis, S. (2004). The role of the medial frontal cortex in cognitive control. *Science*, *306*(5695), 443–447. https://doi.org/10.1126/science.1100301

Riggins, T., & Scott, L. S. (2020). P300 development from infancy to adolescence. *Psychophysiology*, *57*(7), e13346. https://doi.org/10.1111/psyp.13346

Ringer, H., Schröger, E., & Grimm, S. (2022). Perceptual learning and recognition of random acoustic patterns. *Auditory Perception & Cognition*, *5*(3–4), 259–281. https://doi.org/10.1080/25742442.2022.2082827

Ringer, H., Schröger, E., & Grimm, S. (2023). Within- and between-subject consistency of perceptual segmentation in periodic noise: A combined behavioral tapping and EEG study. *Psychophysiology*, *60*(2), e14174. https://doi.org/10.1111/psyp.14174

Roberts, T. F., Tschida, K. A., Klein, M. E., & Mooney, R. (2010). Rapid spine stabilization and synaptic enhancement at the onset of behavioural learning. *Nature*, *463*(7283), 948–952. https://doi.org/10.1038/nature08759

Robertson, M. M., Furlong, S., Voytek, B., Donoghue, T., Boettiger, C. A., & Sheridan, M. A. (2019). EEG power spectral slope differs by ADHD status and stimulant medication exposure in early childhood. *Journal of Neurophysiology*, *122*(6), 2427–2437. https://doi.org/10.1152/jn.00388.2019

Rubenstein, J. L. R., & Merzenich, M. M. (2003). Model of autism: Increased ratio of excitation/inhibition in key neural systems. *Genes, Brain, and Behavior*, *2*(5), 255–267. https://doi.org/10.1034/j.1601-183x.2003.00037.x

Rulkov, N. F., Sushchik, M. M., Tsimring, L. S., & Abarbanel, H. D. I. (1995). Generalized

synchronization of chaos in directionally coupled chaotic systems. *Physical Review E*, *51*(2), 980–994. https://doi.org/10.1103/PhysRevE.51.980

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

Schmidt, R., Ruiz, M. H., Kilavik, B. E., Lundqvist, M., Starr, P. A., & Aron, A. R. (2019). Beta oscillations in working memory, executive control of movement and thought, and sensorimotor function. *Journal of Neuroscience*, *39*(42), 8231–8238. https://doi.org/10.1523/JNEUROSCI.1163-19.2019

Schoups, A., Vogels, R., Qian, N., & Orban, G. (2001). Practising orientation identification improves orientation coding in V1 neurons. *Nature*, *412*(6846), 549–553. https://doi.org/10.1038/35087601

Sejnowski, T. J. (1999). The Book of Hebb. *Neuron*, *24*(4), 773–776. https://doi.org/10.1016/S0896-6273(00)81025-9

Sekar, A., Bialas, A. R., de Rivera, H., Davis, A., Hammond, T. R., Kamitaki, N., Tooley, K., Presumey, J., Baum, M., Van Doren, V., Genovese, G., Rose, S. A., Handsaker, R. E., Daly, M. J., Carroll, M. C., Stevens, B., & McCarroll, S. A. (2016). Schizophrenia risk from complex variation of complement component 4. *Nature*, *530*(7589), 177–183. https://doi.org/10.1038/nature16549

Sellgren, C. M., Gracias, J., Watmuff, B., Biag, J. D., Thanos, J. M., Whittredge, P. B., Fu, T., Worringer, K., Brown, H. E., Wang, J., Kaykas, A., Karmacharya, R., Goold, C. P., Sheridan, S. D., & Perlis, R. H. (2019). Increased synapse elimination by microglia in schizophrenia patient-derived models of synaptic pruning. *Nature Neuroscience*, *22*(3), 374–385. https://doi.org/10.1038/s41593-018-0334-7

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples)†. *Biometrika*, *52*(3–4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591

Sherman, S. M. (2007). The thalamus is more than just a relay. *Current Opinion in Neurobiology*, *17*(4), 417–422. https://doi.org/10.1016/j.conb.2007.07.003

Shew, W. L., & Plenz, D. (2013). The Functional Benefits of Criticality in the Cortex. *The Neuroscientist*, *19*(1), 88–100. https://doi.org/10.1177/1073858412445487

Shimojo, S., Bauer, J., O'Connell, K. M., & Held, R. (1986). Pre-stereoptic binocular vision in infants. *Vision Research*, *26*(3), 501–510. https://doi.org/10.1016/0042-6989(86)90193-8

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Solis, M. M., Brainard, M. S., Hessler, N. A., & Doupe, A. J. (2000). Song selectivity and sensorimotor signals in vocal learning and production. *Proceedings of the National Academy of Sciences*, *97*(22), 11836–11842. https://doi.org/10.1073/pnas.97.22.11836

Sompolinsky, H., Crisanti, A., & Sommers, H. J. (1988). Chaos in random neural networks. *Physical Review Letters*, *61*(3), 259–262. https://doi.org/10.1103/PhysRevLett.61.259

Stein, R. B., Gossen, E. R., & Jones, K. E. (2005). Neuronal variability: Noise or part of the signal? *Nature Reviews Neuroscience*, *6*(5), 389–397. https://doi.org/10.1038/nrn1668

Sterzer, P., Kleinschmidt, A., & Rees, G. (2009). The neural bases of multistable perception. *Trends in Cognitive Sciences*, *13*(7), 310–318. https://doi.org/10.1016/j.tics.2009.04.006

Strogatz, S. H. (2024). *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering* (3rd edn). Chapman and Hall/CRC. https://doi.org/10.1201/b17597

Suga, N., O'Neill, W. E., Kujirai, K., & Manabe, T. (1983). Specificity of combination-sensitive neurons for processing of complex biosonar signals in auditory cortex of the mustached bat.

Journal of Neurophysiology, *49*(6), 1573–1626. https://doi.org/10.1152/jn.1983.49.6.1573

Sussillo, D., & Abbott, L. F. (2009). Generating coherent patterns of activity from chaotic neural networks. *Neuron*, *63*(4), 544–557. https://doi.org/10.1016/j.neuron.2009.07.018

Taherkhani, A., Belatreche, A., Li, Y., Cosma, G., Maguire, L. P., & McGinnity, T. M. (2020). A review of learning in biologically plausible spiking neural networks. *Neural Networks*, *122*, 253–272. https://doi.org/10.1016/j.neunet.2019.09.036

Terlau, J., Martini, J., & Helfrich, R. F. (2025). Structure in noise: Recurrent connectivity shapes neural variability to balance perceptual and cognitive demands in the human brain. *Neuron*, *0*(0). https://doi.org/10.1016/j.neuron.2025.10.015

Uchida, A., McAllister, R., & Roy, R. (2004). Consistency of Nonlinear System Response to Complex Drive Signals. *Physical Review Letters*, *93*(24), 244102. https://doi.org/10.1103/PhysRevLett.93.244102

Uchida, A., Yoshimori, S., McALLISTER, R., & Roy, R. (2007). Consistency in Lasers. *The Review of Laser Engineering*, *35*(1), 38–42. https://doi.org/10.2184/lsj.35.1_38

Uchida, A., Yoshimura, K., Davis, P., Yoshimori, S., & Roy, R. (2008). Local conditional Lyapunov exponent characterization of consistency of dynamical response of the driven Lorenz system. *Physical Review E*, *78*(3), 036203. https://doi.org/10.1103/PhysRevE.78.036203

Ullian, E. M., Sapperstein, S. K., Christopherson, K. S., & Barres, B. A. (2001). Control of Synapse Number by Glia. *Science*, *291*(5504), 657–661. https://doi.org/10.1126/science.291.5504.657

van Diepen, R. M., & Mazaheri, A. (2018). The Caveats of observing Inter-Trial Phase-Coherence in Cognitive Neuroscience. *Scientific Reports*, *8*(1), 2990. https://doi.org/10.1038/s41598-018-20423-z

Vinck, M., van Wingerden, M., Womelsdorf, T., Fries, P., & Pennartz, C. M. A. (2010). The pairwise phase consistency: A bias-free measure of rhythmic neuronal synchronization. *NeuroImage*,

51(1), 112–122. https://doi.org/10.1016/j.neuroimage.2010.01.073

Vinnik, E., Itskov, P. M., & Balaban, E. (2012). β- And γ-band EEG power predicts illusory auditory continuity perception. *Journal of Neurophysiology*, *108*(10), 2717–2724. https://doi.org/10.1152/jn.00196.2012

Vogels, T. P., Sprekeler, H., Zenke, F., Clopath, C., & Gerstner, W. (2011). Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks. *Science (New York, N.Y.)*, *334*(6062), 1569–1573. https://doi.org/10.1126/science.1211095

von Helmholtz, H. (1924). *Helmholtz's treatise on physiological optics, Vol. 1, Trans. From the 3rd German ed* (pp. xxi, 482). Optical Society of America. https://doi.org/10.1037/13536-000

Voytek, B., Kramer, M. A., Case, J., Lepage, K. Q., Tempesta, Z. R., Knight, R. T., & Gazzaley, A. (2015). Age-related changes in 1/f neural electrophysiological noise. *The Journal of Neuroscience*, *35*(38), 13257–13265. https://doi.org/10.1523/JNEUROSCI.2332-14.2015

Wagenaar, D. A., Madhavan, R., Pine, J., & Potter, S. M. (2005). Controlling bursting in cortical cultures with closed-loop multi-electrode stimulation. *Journal of Neuroscience*, *25*(3), 680–688. https://doi.org/10.1523/JNEUROSCI.4209-04.2005

Welch, B. L. (1947). The generalisation of student's problems when several different population variances are involved. *Biometrika*, *34*(1–2), 28–35. https://doi.org/10.1093/biomet/34.1-2.28

W.Hirsch, M., Smale, S., & L. Devaney, R. (2017). *Differential Equations, Dynamical Systems & Linear Algebra* (3rd edn). ELSEVIER INC. https://doi.org/10.1016/s0079-8169(08)x6044-1

Wilkinson, C. L., Yankowitz, L. D., Chao, J. Y., Gutiérrez, R., Rhoades, J. L., Shinnar, S., Purdon, P. L., & Nelson, C. A. (2024). Developmental trajectories of EEG aperiodic and periodic components in children 2–44 months of age. *Nature Communications*, *15*(1), 5788. https://doi.org/10.1038/s41467-024-50204-4

Wilting, J., & Priesemann, V. (2019). 25 years of criticality in neuroscience—Established results, open

controversies, novel concepts. *Current Opinion in Neurobiology*, *58*, 105–111. https://doi.org/10.1016/j.conb.2019.08.002

Wolff, A., Yao, L., Gomez-Pilar, J., Shoaran, M., Jiang, N., & Northoff, G. (2019). Neural variability quenching during decision-making: Neural individuality and its prestimulus complexity. *NeuroImage*, *192*, 1–14. https://doi.org/10.1016/j.neuroimage.2019.02.070

Yassa, M. A., & Stark, C. E. L. (2011). Pattern separation in the hippocampus. *Trends in Neurosciences*, *34*(10), 515–525. https://doi.org/10.1016/j.tins.2011.06.006

Yildiz, I. B., Jaeger, H., & Kiebel, S. J. (2012). Re-visiting the echo state property. *Neural Networks*, *35*, 1–9. https://doi.org/10.1016/j.neunet.2012.07.005

Yin, W., Li, T., Mucha, P. J., Cohen, J. R., Zhu, H., Zhu, Z., & Lin, W. (2022). Altered neural flexibility in children with attention-deficit/hyperactivity disorder. *Molecular Psychiatry*, *27*(11), 4673–4679. https://doi.org/10.1038/s41380-022-01706-4

Yizhar, O., Fenno, L. E., Prigge, M., Schneider, F., Davidson, T. J., O'Shea, D. J., Sohal, V. S., Goshen, I., Finkelstein, J., Paz, J. T., Stehfest, K., Fudim, R., Ramakrishnan, C., Huguenard, J. R., Hegemann, P., & Deisseroth, K. (2011). Neocortical excitation/inhibition balance in information processing and social dysfunction. *Nature*, *477*(7363), 171–178. https://doi.org/10.1038/nature10360

Zhou, Y., & Freedman, D. J. (2019). Posterior parietal cortex plays a causal role in perceptual and categorical decisions. *Science (New York, N.Y.)*, *365*(6449), 180–185. https://doi.org/10.1126/science.aaw8347

Zimmern, V. (2020). Why brain criticality is clinically relevant: A scoping review. *Frontiers in Neural Circuits*, *14*, 54. https://doi.org/10.3389/fncir.2020.00054

Appendix & Supporting information

6.1. Data Availability

The simulation code and its description of Project 1 are available on the GitHub repository (https://github.com/Yujingoto/Goto_etal_PLoSCB2024).

The preregistered methods, demographic and behavioural data, and EEG data used for Project 2 are available on the OSF repository (https://osf.io/tyab6/overview).

6.2. Supplementary Tables

Table S1. Experimental setting. For all simulations, I chose the following common parameters for the network constructions, MBGD, and Hebbian learning in this table.

Parameter	Variable	Value
Input dimension	N_{u}	1
Reservoir dimension	N_x	5×10^2
Reservoir connection density	d	10^{-1}
Internal noise level in the reservoir	ε	$10^{-1} \times N(0,1)$
Spectral radii	ρ	[0.1,0.9,1.0,1.1,1.2,1.3,1.4,1.5,1.6,1.7,1.8,1.9,2.0]
Output dimension	N_{y}	1
MBGD learning rate	η	10^{-1}
Hebbian learning rate	α	10^{-7}
Number of each stimulus type for a training dataset	-	5
Number of tests	-	200
Time points for transition	-	300

Table S2. The spectral radii of the Hebbian and non-Hebbian networks after the training session averaged over 200 runs.

Data shows plasticity changes the spectral radii little.

Network	Spectral radius (ρ)												
Initial	0.1	0.9	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
non-Hebbian	0.1	0.9	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
Hebbian	**0.999**	**0.900**	**1.000**	**1.102**	**1.204**	**1.308**	**1.411**	**1.511**	**1.610**	**1.709**	**1.809**	**1.908**	**2.007**

Table S3. Paired comparison of main behavioural measures.

Contrast	Value 1	Value 2	Mean difference	test	statistic	p (q) value	Effect size
d' (RefRN-RN)	1.19±1.17	0.87±0.99	0.334	paired t	3.492	0.002(0.006)	0.713
HR (RefRN-RN)	0.76±0.20	0.70±0.18	0.067	paired t	2.328	0.029(0.044)	0.475
CRR (RefN-N)	0.58±0.27	0.61±0.26	-0.026	paired t	-1.114	0.277(0.277)	-0.227
RT (N-RefN)	0.47±0.11	0.47±0.10	0.0003	wilcoxon	105.500	0.495(0.594)	-0.166
RT (N-RN)	0.47±0.11	0.45±0.10	0.018	paired_t	2.754	0.113(0.017)	0.562
RT (N-RefRN)	0.47±0.11	0.45±0.10	0.020	paired_t	3.320	0.003(0.010)	0.678
RT (RefN-RN)	0.47±0.10	0.45±0.10	0.018	paired_t	3.104	0.005(0.010)	0.634
RT (RefN-RefRN)	0.47±0.10	0.45±0.10	0.020	paired_t	3.283	0.003(0.010)	0.670
RT (RN-RefRN)	0.45±0.10	0.45±0.10	0.002	paired_t	0.348	0.731(0.731)	0.071

Table S4. Results of ICC for HR and d' (3,k)

Condition	HR	CRR	d'	RT
N		0.923		0.944
RN	0.844		0.906	0.942
RefN		0.899		0.947
RefRN	0.754		0.835	0.933

Table S5. Group differences between "Well-learners" and "Poor-learners".

Well-learners and poor-learners were classified using a threshold of 0.8 for the hit rate in the last 20 RefRN trials. Group differences were then examined for each behavioural measure and demographic characteristic. Although these variables were not strictly independent, statistical tests were conducted for all possible combinations, and the resulting p-values were corrected for multiple comparisons using the false discovery rate (FDR) procedure.

variable	Well learners (n=14)	Poor learners (n=10)	test	p	q (FDR)
Age (years)	33.9 ± 9.2	34.5 ± 9.3	Welch's t	0.883	1.000
Sex	F: 9; M: 5	F: 10	Fisher's exact	0.053	0.174
Musical training experience (years)	0: 6; 6: 2; 7: 2; 2: 1; 3: 1; 13: 1; 14: 1	0: 4; 3: 2; 15: 1; 7: 1; 10: 1; 20: 1	Chi-square	0.471	0.811
Dominant ear	R: 8; L: 5; B: 1	R: 7; L: 2; B: 1	Chi-square	0.703	0.898
Dominant hand	R: 14	R: 10	Chi-square	1.000	1.000
Practice time (s)	281 ± 104	481 ± 227	Welch's t	0.024	0.091
Overall confidence	3.36 ± 2.21	3.6 ± 2.68	Welch's t	0.817	0.988
AQ total	19.0 ± 6.0	19.0 ± 8.3	Welch's t	1.000	1.000
ASRS	3.43 ± 2.77	1.80 ± 1.48	Welch's t	0.077	0.221
HR: RN (overall)	0.79 ± 0.14	0.57 ± 0.15	Welch's t	0.001	0.008
d': RN (overall)	1.244 ± 0.955	0.351 ± 0.808	Welch's t	0.022	0.091
d': RefRN (overall)	1.801 ± 0.961	0.330 ± 0.870	Welch's t	0.001	0.007
FAR: N (overall)	0.351 ± 0.283	0.452 ± 0.238	Welch's t	0.353	0.677
FAR: RefN (overall)	0.392 ± 0.270	0.454 ± 0.268	Welch's t	0.585	0.841
CRR: N (overall)	0.649 ± 0.283	0.548 ± 0.238	Welch's t	0.353	0.677
CRR: RefN (overall)	0.608 ± 0.270	0.546 ± 0.268	Welch's t	0.585	0.841
RT: N (overall, s)	0.491 ± 0.103	0.439 ± 0.106	Welch's t	0.242	0.618
RT: RefN (overall, s)	0.489 ± 0.101	0.442 ± 0.103	Welch's t	0.275	0.633
RT: RN (overall, s)	0.461 ± 0.098	0.439 ± 0.114	Welch's t	0.632	0.855
RT: RefRN (overall, s)	0.464 ± 0.104	0.435 ± 0.095	Welch's t	0.494	0.811
d' in last 20 RefRN trials (overall)	1.982 ± 1.009	0.239 ± 1.112	Welch's t	0.001	0.007

Table S6. L1 main effect predicting Perceived (only effects significant for both HR_{ref} and d'_{ref} after global FDR).

Each value is written as a coefficient (SE), *q*-value after FDR. Conditions not shown here did not survive the global FDR correction.

Model	FOI	ROI	HR_{ref}	d'_{ref}
$GLMM_{all}$	Theta	temporal	-0.100 (0.018), *q*=7.09e-08	-0.088 (0.018), *q*=6.1e-07
$GLMM_{all}$	Theta	frontal	-0.023 (0.010), *q*=0.0216	-0.041 (0.010), *q*=4.41e-05
$GLMM_{all}$	Alpha	temporal	-0.104 (0.011), *q*=3.92e-19	-0.074 (0.011), *q*=2.09e-12
$GLMM_{all}$	Alpha	frontal	-0.046 (0.007), *q*=7.11e-11	-0.033 (0.007), *q*=3.38e-06
$GLMM_{all}$	Alpha	parietal	-0.041 (0.005), *q*=2.71e-16	-0.032 (0.005), *q*=1.57e-10
$GLMM_{all}$	Beta	temporal	-0.142 (0.023), *q*=1.59e-09	-0.238 (0.023), *q*=3.65e-09
$GLMM_{all}$	Beta	frontal	-0.058 (0.015), *q*=0.000126	-0.091 (0.015), *q*=2.35e-09
$GLMM_{all}$	Beta	parietal	-0.028 (0.014), *q*=0.0493	-0.042 (0.014), *q*=0.00365
$GLMM_{all}$	Delta	temporal	0.094 (0.012), *q*=6.56e-15	0.094 (0.012), *q*=9.45e-15
$GLMM_{all}$	Delta	frontal	0.062 (0.006), *q*=5.35e-27	0.066 (0.006), *q*=6.13e-30
$GLMM_{all}$	Delta	parietal	0.039 (0.006), *q*=1.69e-11	0.074 (0.006), *q*=1.32e-29
$GLMM_{RefRN}$	Theta	frontal	-0.069 (0.023), *q*=0.00949	-0.102 (0.023), *q*=3.36e-05
$GLMM_{RefRN}$	Beta	temporal	-0.239 (0.055), *q*=6.95e-05	-0.510 (0.053), *q*=4.75e-20
$GLMM_{RefRN}$	Beta	parietal	-0.091 (0.033), *q*=0.0136	-0.178 (0.032), *q*=1.7e-07
$GLMM_{RefRN}$	Delta	parietal	0.044 (0.016), *q*=0.0138	0.037 (0.015), *q*=0.035

Table S7. Learning modulation of the L1 effect (only effects significant for both HR_{ref} and d'_{ref} after global FDR)

Each value is written as a coefficient (SE), *q*-value after FDR. Conditions not shown here did not survive the global FDR correction.

Model	FOI	ROI	HR_{ref}	d'_{ref}
$GLMM_{all}$	Theta	temporal	-0.059 (0.021), q=0.00476	-0.151 (0.019), q=1.15e-15
$GLMM_{all}$	Theta	frontal	-0.053 (0.010), q=1.35e-07	-0.073 (0.009), q=1.7e-14
$GLMM_{all}$	Theta	parietal	-0.035 (0.010), q=0.000514	-0.064 (0.009), q=1.24e-11
$GLMM_{all}$	Alpha	temporal	-0.049 (0.012), q=0.000107	-0.088 (0.011), q=6.56e-15
$GLMM_{all}$	Alpha	frontal	-0.046 (0.007), q=1.11e-09	-0.059 (0.007), q=2.37e-18
$GLMM_{all}$	Alpha	parietal	-0.021 (0.005), q=6.93e-05	-0.031 (0.005), q=3.71e-09
$GLMM_{all}$	Beta	temporal	-0.083 (0.023), q=0.000326	-0.255 (0.023), q=8.89e-29
$GLMM_{all}$	Beta	frontal	-0.125 (0.014), q=4.97e-19	-0.131 (0.014), q=1.16e-20
$GLMM_{all}$	Beta	parietal	-0.085 (0.013), q=2.53e-10	-0.125 (0.014), q=3.92e-19
$GLMM_{all}$	Delta	frontal	0.033 (0.005), q=8.45e-13	0.040 (0.005), q=2.15e-17
$GLMM_{all}$	Delta	parietal	0.078 (0.006), q=6.12e-33	0.083 (0.006), q=5.06e-44
$GLMM_{RefRN}$	Beta	temporal	-0.291 (0.052), q=1.7e-07	-0.526 (0.059), q=8.56e-18
$GLMM_{RefRN}$	Beta	parietal	-0.208 (0.029), q=1.05e-11	-0.253 (0.035), q=7.5e-12
$GLMM_{RefRN}$	Delta	parietal	0.107 (0.016), q=2.53e-10	0.151 (0.019), q=1.06e-14

6.3. Supplementary Figures

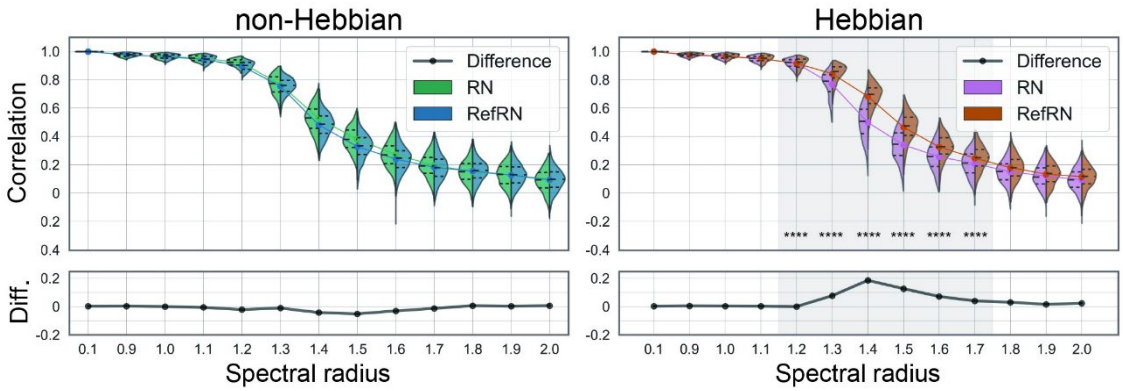

Figure S1. The evaluation of selective consistency for RN and RefRN of non-Hebbian and Hebbian networks.

The figure style is the same as Figure 5a, but in a different style of comparison. Colours for each condition are as follows: non-Hebbian RN; green, non-Hebbian RefRN; cyan, Hebbian RN; magenta, and Hebbian RefRN; brown.

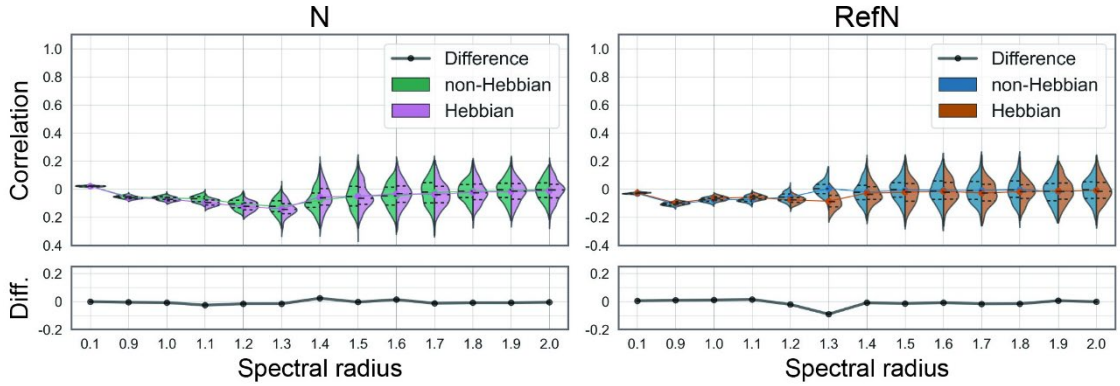

Figure S2. The evaluation of the between-segment similarity for RefN and N stimuli.

The figure style is the same as Figure 4a. The similarity was evaluated by the correlation between the first and second segment time series for each test run for repeated noise (N; left) and referenced repeated noise (RefN; right). The violin plots show probability density distributions and interquartile ranges of Hebbian (right side; magenta and brown) and non-Hebbian (left side; green and cyan) models, respectively. The coloured line plots connect the mean values for each condition. The black lines in the bottom windows show the difference between Hebbian and non-Hebbian models. The horizontal axis represents the spectral radius of the evaluated networks.

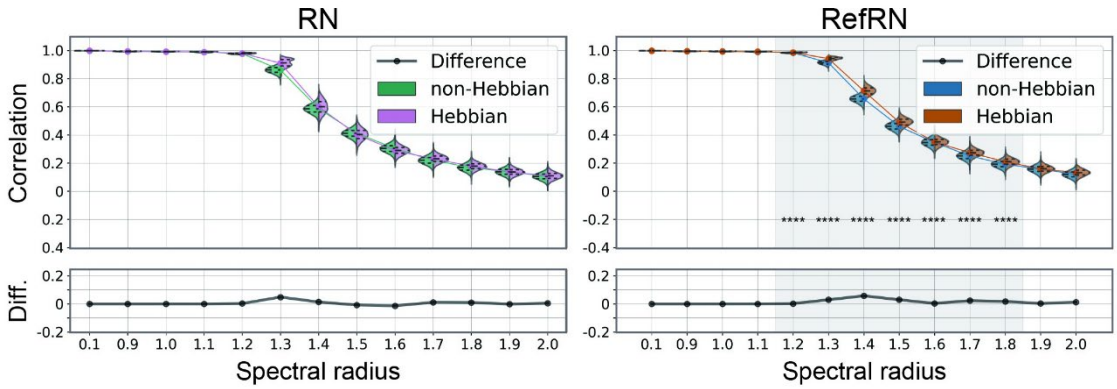

Figure S3. The inter-trial level selective consistency for RefRN and RN.

The consistency was evaluated by the mean of the correlation between all time series. The violin plots show probability density distributions and interquartile ranges of Hebbian (right side; magenta and brown) and non-Hebbian (left side; green and cyan) models, respectively (****; PR < 0.01%, p < 0.001). The coloured line plots connect the mean values for each condition. The black lines in the bottom windows show the difference between Hebbian and non-Hebbian models. The horizontal axis represents the spectral radius of the evaluated networks. Weaker but significant differences between conditions can be seen in the same way as the inter-segment level comparison.

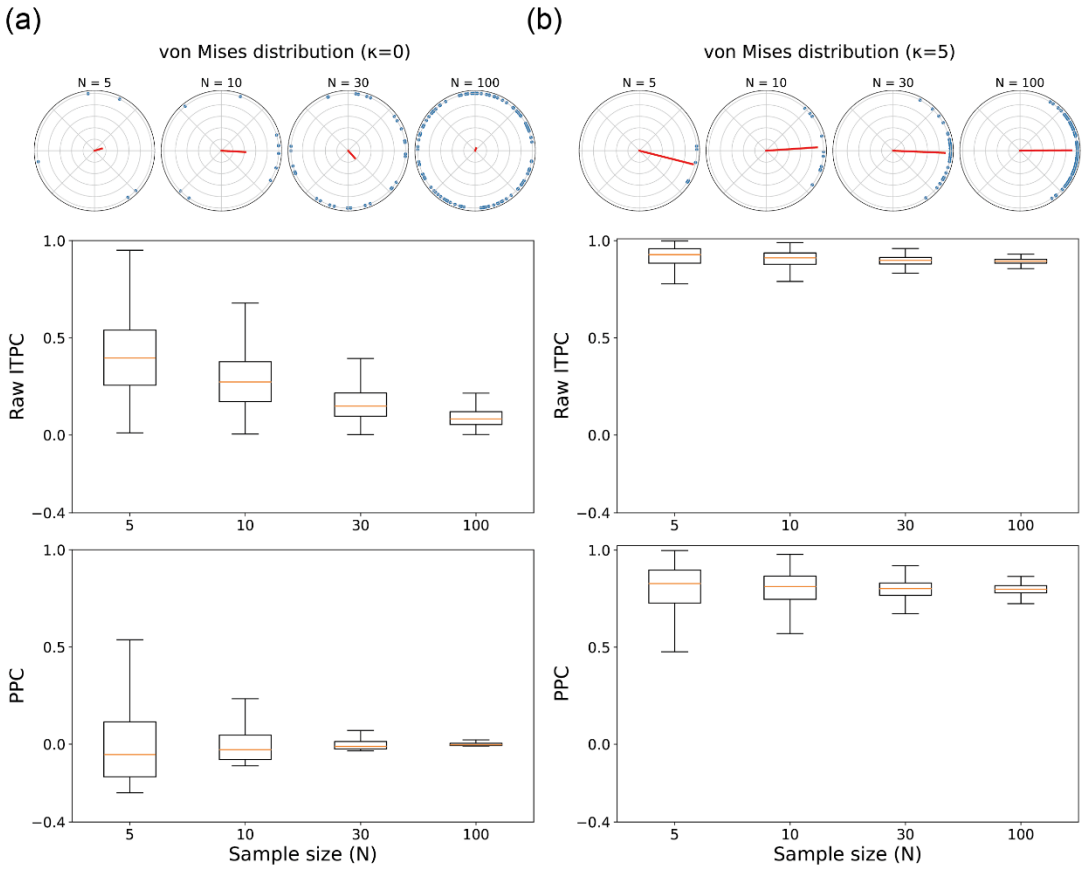

Figure S4. Evaluation of ITPC correction method.

Comparison of ITPC and cITPC computed from simulated datasets with varying sample sizes, generated from von Mises phase distributions with different concentration parameters (κ). The top row shows the phase distributions generated for each sample size. Each blue dot represents an independent random sample, and the red lines represent the evaluated mean vector of these samples. The middle row shows results for conventional ITPC, and the bottom row shows results for cITPC. The boxplots summarise results across sample sizes on the x-axis; thus, horizontal shifts reflect sample-size-dependent bias in each metric. The median is shown by the orange line. (a) Comparison under a fully random phase distribution ($\kappa = 0$). (b) Comparison under a strong phase-locking condition ($\kappa = 5$).

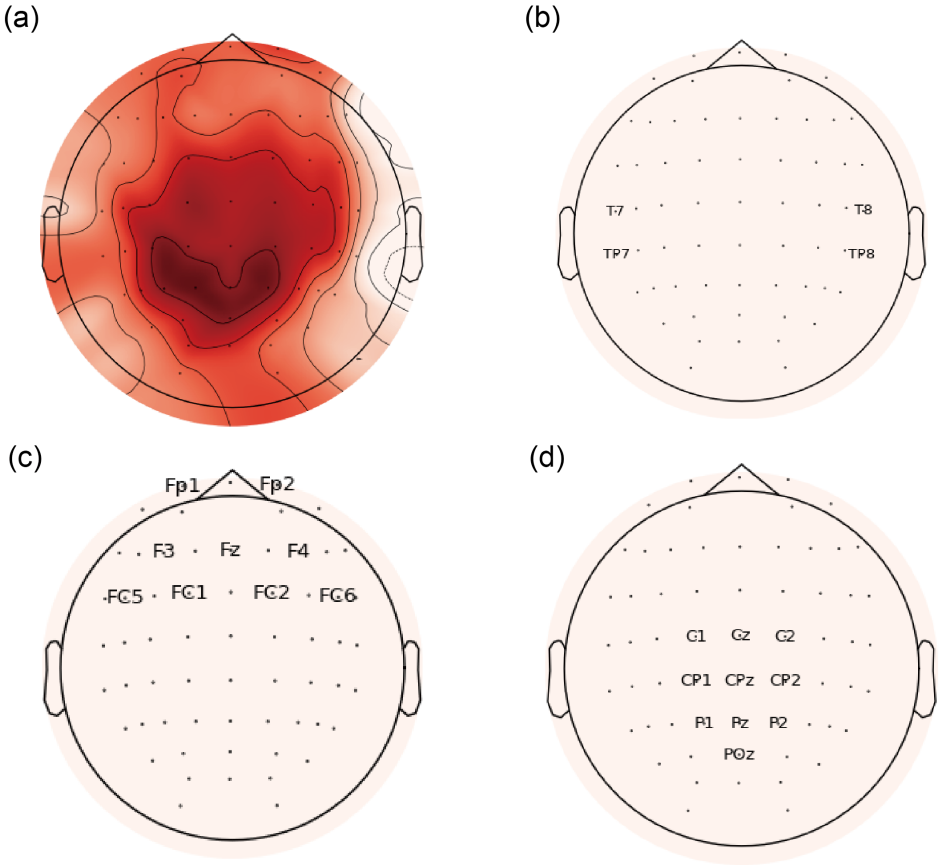

Figure S5. Topographical map of ROIs.

(a) The ITPC topographical map of the differences between RefRN and RN. Colour indicates big differences. (b) Channels of temporal ROI. (c) Channels of frontal ROI. (d) Channels of parietal ROI.

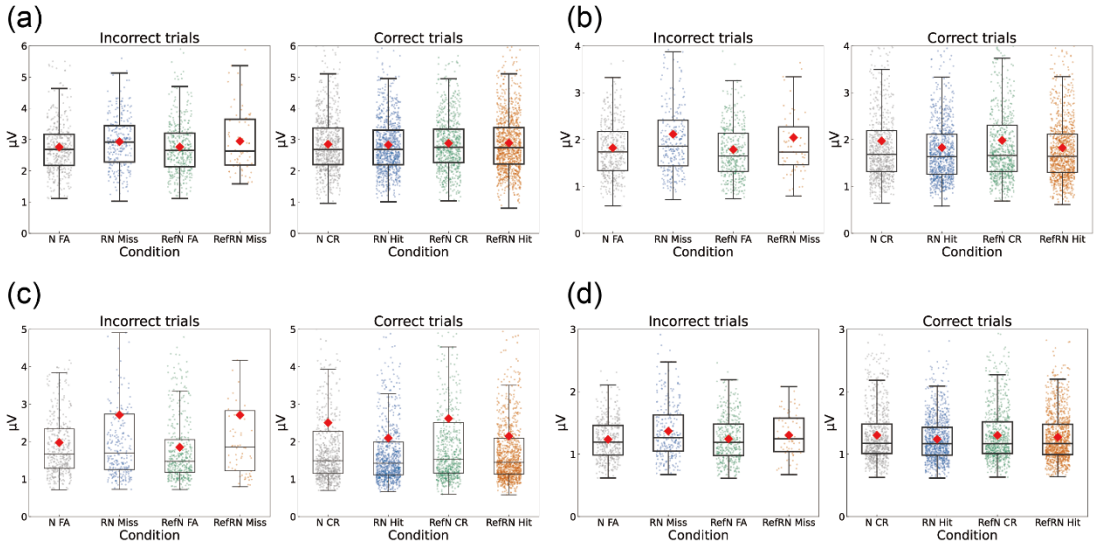

Figure S6. L1 frontal ROI.

The well-learners' temporal ROI L1 distances of each condition are presented separately according to the correctness of their response. The left panel shows incorrect trials and the right panel shows correct trials, plotted separately for each of the four conditions. Thus, false alarms in N and RefN, as well as hits in RN and RefRN, correspond to *Perceive* trials in which participants reported hearing repetition. Dots represent all pooled trials from all participants. The black boxplots for each condition show Q1–Q3, with the horizontal line indicating the median; whiskers extend to 1.5×IQR. The red diamonds represent the mean values. (a) Delta FOI, (b) Theta FOI, (c) Alpha FOI, (d) Beta FOI.

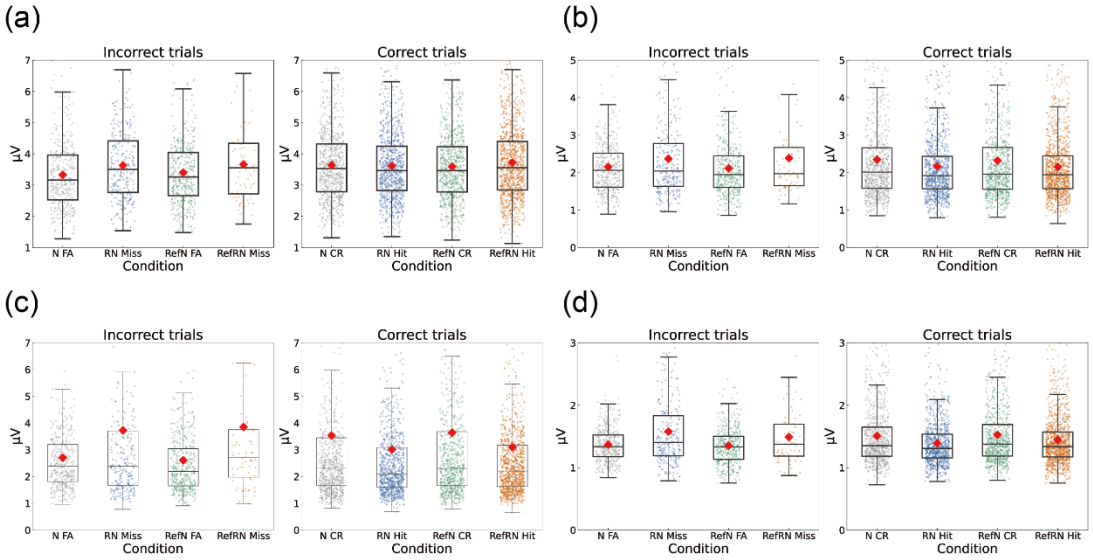

Figure S7. L1 parietal ROI.

The well-learners' temporal ROI L1 distances of each condition are presented separately according to the correctness of their response. The left panel shows incorrect trials and the right panel shows correct trials, plotted separately for each of the four conditions. Thus, false alarms in N and RefN, as well as hits in RN and RefRN, correspond to *Perceive* trials in which participants reported hearing repetition. Dots represent all pooled trials from all participants. The black boxplots for each condition show Q1–Q3, with the horizontal line indicating the median; whiskers extend to 1.5×IQR. The red diamonds represent the mean values. (a) Delta FOI, (b) Theta FOI, (c) Alpha FOI, (d) Beta FOI.

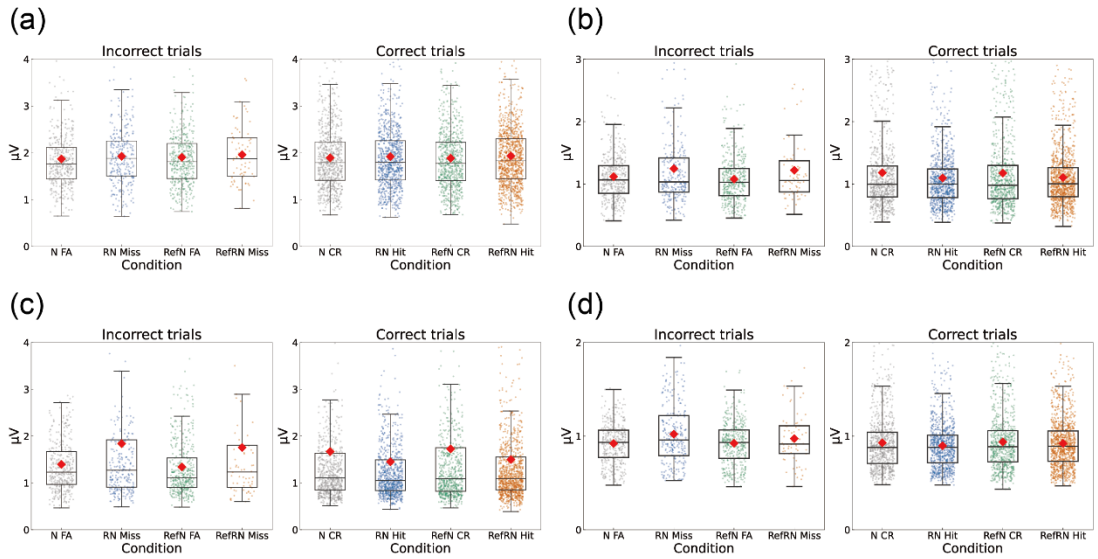

Figure S8. L1 temporal ROI.

The well-learners' temporal ROI L1 distances of each condition are presented separately according to the correctness of their response. The left panel shows incorrect trials and the right panel shows correct trials, plotted separately for each of the four conditions. Thus, false alarms in N and RefN, as well as hits in RN and RefRN, correspond to *Perceive* trials in which participants reported hearing repetition. Dots represent all pooled trials from all participants. The black boxplots for each condition show Q1–Q3, with the horizontal line indicating the median; whiskers extend to 1.5×IQR. The red diamonds represent the mean values. (a) Delta FOI, (b) Theta FOI, (c) Alpha FOI, (d) Beta FOI.

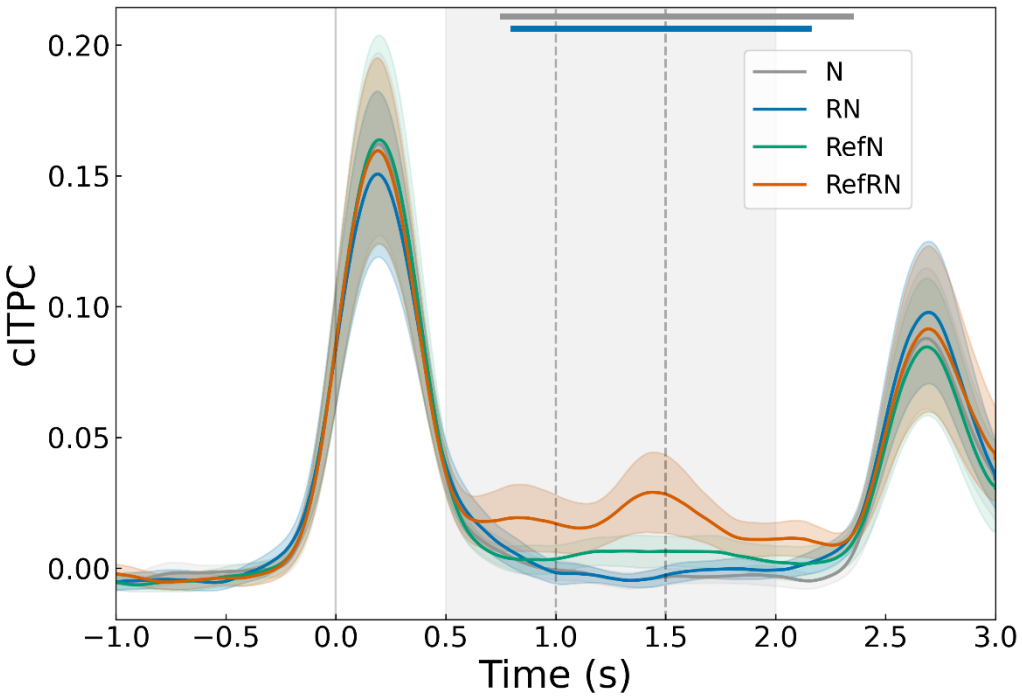

Figure S9. The delta-band cITPC from pooled-ROIs for each stimulus type.

Each coloured line shows cITPC of each stimulus type (grey: N, blue: RN, green: RefN, orange: RefRN) calculated using all trials regardless of the correctness. There were significant differences between RefRN and N (0.7–2.4-s, $p_{cluster} = 0.001$), RefRN and RN (0.8–2.2-s, $p_{cluster} = 0.001$), and tendency between RefRN and RefN (0.7–1.0-s, $p_{cluster} = 0.061$; 1.2–1.6-s, $p_{cluster} = 0.054$).

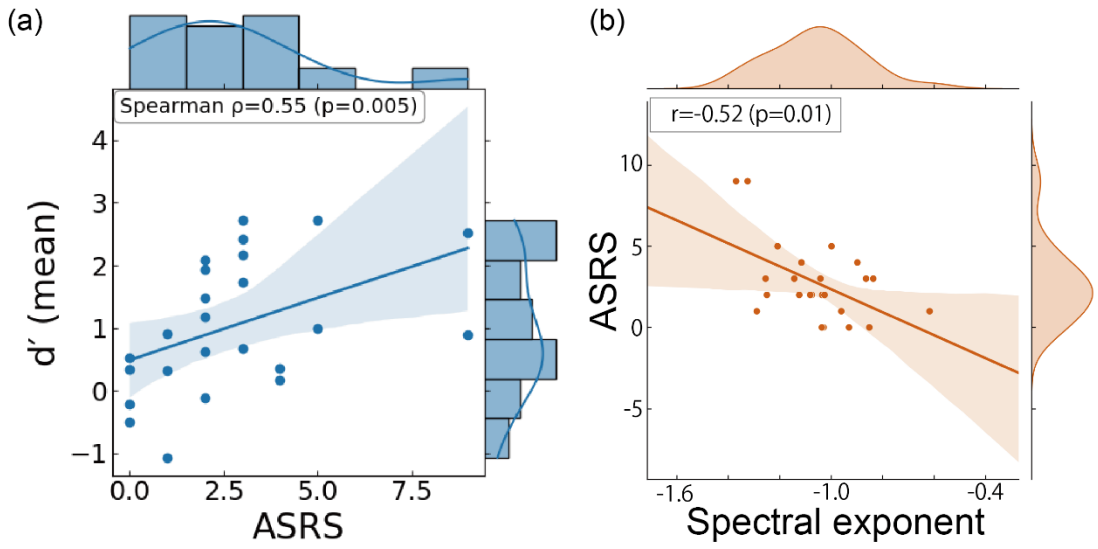

Figure S10. ASRS score and d', spectral radius

ASRS scores showed a positive correlation with overall task sensitivity (mean d′ ; $r = 0.55$, $p = 0.005$). Also, it was negatively correlated with SE ($r = -0.52$, $p = 0.01$).

6.4. Glossary of Key Terms

Dynamical System

A dynamical system specifies a set of variables $\boldsymbol{x}(t)$ that represent the system's current state and a rule of evolution that determines how that state changes over time. In neuroscience, the state variables might be membrane potential and gating variables, firing rates, synaptic currents, EEG, BOLD, etc., with the key idea that the future state is determined by the present state (and any input).

The evolution rule is typically written as an ODE in continuous time, $\frac{d\boldsymbol{x}}{dt} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}, t)$, or as an update map in discrete time, $\boldsymbol{x}_{t+1} = \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{u}_t)$.

Chaos

Chaos is commonly defined as a property of a deterministic dynamical system that nevertheless shows aperiodic, irregular-looking long-term behaviour together with sensitive dependence on initial conditions, meaning that arbitrarily small differences in the initial state can grow rapidly over time. This sensitive dependence is often illustrated by the "butterfly effect" and is a key reason why chaotic dynamics can appear random.

A standard quantitative characterisation is the exponential separation of nearby trajectories: if $\delta(t)$ denotes the distance between two initially close but different trajectories, we can write as:

$$\delta(t) \approx \delta(0)\mathrm{e}^{\lambda_{\max} t}, \qquad \text{(eq. S1)}$$

where the largest Lyapunov exponent $\lambda_{\max}$ is used as an indicator (with $\lambda_{\max} > 0$ commonly taken as evidence on chaos).

Lyapunov Exponent

A Lyapunov exponent in a dynamical system quantifies the average exponential rate at which an infinitesimal difference in initial conditions $\delta(0)$ grows or decays along a trajectory. For a continuous-time system $\frac{dx}{dt} = \boldsymbol{F}(\boldsymbol{x})$, the perturbation obeys the variational equation $\frac{dx}{dt} = \mathbf{J}\big(\boldsymbol{x}(t)\big)\delta t$ to first order, where $\mathbf{J}$ is the Jacobian of function **F**. A common definition of the maximal Lyapunov exponent is

$$\lambda_{\max} = \lim_{t\to\infty} \frac{1}{t} \log \frac{\|\delta x(t)\|}{\|\delta x(0)\|} \qquad \text{(eq. S2)}$$

so that Eq. S1 (Strogatz, 2024). In this sense, $\lambda > 0$ indicates exponential separation of nearby trajectories(=chaos), whereas $\lambda < 0$ indicates average contraction.

Jacobian

The Jacobian (Jacobian matrix) is the matrix form of the first derivative of a multivariate vector-valued function $\mathrm{f}: \mathbb{R}^n \to \mathbb{R}^m$, defined component-wise as

$$\mathrm{J}_f(\boldsymbol{x}) = \left[\frac{\partial \mathrm{f_i}}{\partial x_j}\right]_{i=1\ldots m, j=1\ldots n}. \qquad \text{(eq. S3)}$$

It provides the standard local linear approximation of f near a point: for a small perturbation δx, one approximates $\delta f \approx J_f(x)\delta x$ (Strogatz, 2024).

Edge of Chaos

The edge of chaos refers to the transition region near the boundary between ordered dynamics (perturbations rapidly decay; trajectories tend to settle to fixed points or periodic orbits) and chaotic dynamics (perturbations grow exponentially; trajectories rapidly diverge).

In neural network models and reservoir computing, the edge of chaos is often treated operationally as the border between stable (convergent) and unstable (divergent) regimes; in Lyapunov terms, one common description is that it lies near $\lambda_{\max} \approx 0$. This region is usually known as the maximum computational processing capacity (Bertschinger & Natschläger, 2004; Boedecker et al., 2012; Chua et al., 2012).

Criticality

Criticality refers to a system being at, or near, the critical point of a continuous (second-order) phase transition near criticality, the correlation length ξ diverges and one expects scale invariance (no characteristic scale), power-law behaviour, and universality in a standard statistical-physics sense. As similar as edge of chaos, usually a system operating near criticality is known to maximise its computational performance (Del Papa et al., 2017; Hesse & Gross, 2014; Shew & Plenz, 2013).

In neuroscience, criticality is often used within the critical brain / criticality hypothesis, which frames cortical dynamics as operating near a phase-transition boundary—informally, between overly ordered and overly unstable/chaotic regimes (Beggs, 2019; Beggs & Plenz, 2003; Cocchi et al., 2017; de Arcangelis & Herrmann, 2010; Friedman et al., 2012).

Hebbian Plasticity

Hebbian plasticity refers to a family of activity-dependent synaptic plasticity rules in which changes in synaptic efficacy depend on the relationship between presynaptic and postsynaptic activity (often framed as co-activation/correlation, and in a stricter reading, a causal contribution of the presynaptic neuron to postsynaptic firing). The idea traces back to Hebb (1949) and is frequently summarised as "*fire together, wire together"* (Sejnowski, 1999).

In computational treatments, a minimal rate-based form updates the synaptic weight W_{ij} (from presynaptic neuron j to postsynaptic neuron i) in proportion to the product of presynaptic activity x_{j} and postsynaptic activity x_{i} (or, more generally, their correlation/covariance):

$$\Delta W_{\mathrm{ij}} = \eta x_j x_i \qquad \text{(eq. S4)}$$

with learning rate η. This captures a core point: synapses can change without an explicit teacher signal, driven by activity statistics, making Hebbian plasticity a standard building block for associative learning and representation formation in theoretical neuroscience.

Oja's Hebbian Rule

Oja's Hebbian rule (Oja's rule) is a Hebbian learning rule (eq. S4) that augments the basic update $\Delta W_{\mathrm{ij}} = \eta x_j x_i$ with an additional normalising term so that the weight vector $\boldsymbol{W}$ does not grow without bound (Oja, 1982). For a linear neuron $\boldsymbol{x} = \boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}$, a standard discrete form is

$$\Delta W_{ij} = \alpha x_i\left(x_j - x_i W_{ij}\right). \qquad \text{(eq. S5)}$$

Conceptually, the subtractive term $-x_i^2 W_{ij}$ counteracts the pure Hebbian growth term, stabilising the norm of $\boldsymbol{W}$ during learning. In this thesis, this normalising term may be the reason for the selective convergence of the Jacobian during learning. See 4.4.1.

Self-organised Criticality

Self-organised criticality (SOC) is the idea that certain spatially extended, driven–dissipative systems can self-tune to a state corresponding to a critical point of a second-order phase transition without fine-tuning of control parameters. In the classic formulation, the system naturally evolves towards a barely stable critical state, exhibiting $1/f$-like temporal fluctuations and scale-invariant spatial structure (Bak et al., 1987). Operationally, SOC is often characterised by cascade-like avalanches whose size and duration lack a characteristic scale and are frequently approximated by power laws.

In n neuroscience, SOC has been discussed in connection with neuronal avalanches and broader critical brain hypotheses, where avalanche statistics and related signatures are interpreted in terms of proximity to critical dynamics

A-weighting Filter

An A-weighting filter is a standard frequency-weighting used to compress a sound's spectrum into a single level value in a way that roughly reflects human perceived "loudness" as a function of frequency, rather than treating all frequencies equally (Fletcher & Munson, 1933; Houser et al., 2017). Because human hearing is relatively less sensitive at low and very high frequencies and more sensitive in the mid range, the A-weighting response strongly attenuates low and high frequencies (Fig. S11).

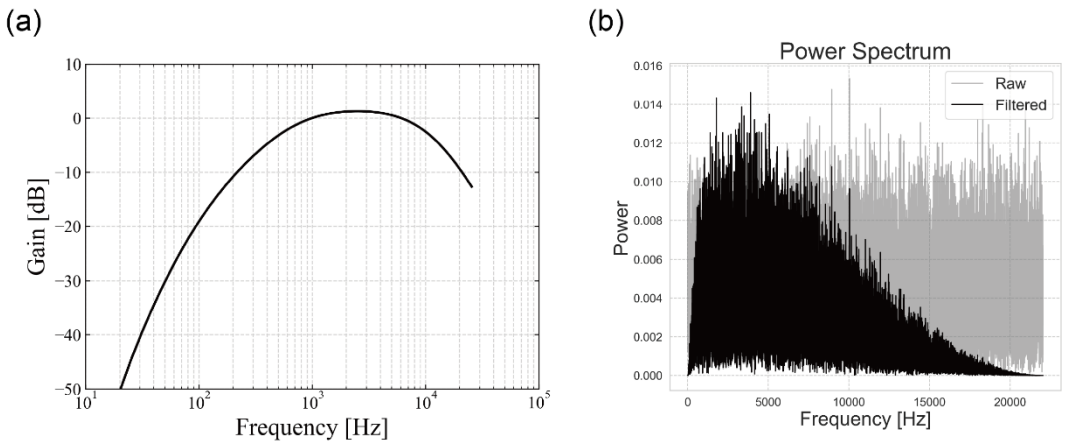

Figure S11. A-weighting filter

(a) Frequency gain of the A-weighting filter. (b) Power spectrum of the raw white-noise signal (grey) and its A-weighting-filtered time series (black).

Reservoir Computing

Reservoir computing is a framework for processing time-dependent inputs $\boldsymbol{u}(t)$ in which the recurrent part of an RNN (the reservoir) is kept largely fixed, while the input-driven high-dimensional state $\boldsymbol{x}(t)$ is used as a feature representation and only a readout layer is trained to produce the target output. It is commonly associated with Echo State Networks (ESNs) and Liquid State Machines (LSMs), and is characterised by replacing full RNN training (e.g., through backpropagation through time) with efficient readout learning such as linear regression (Jaeger & Haas, 2004; Lukoševičius, 2012; Lukoševičius & Jaeger, 2009).

In the ESN line of work, the echo state property—that the reservoir state is determined by the input history rather than initial conditions—is often highlighted as a key condition for stable use and training.

Echo State Network

An echo state network (ESN) is a reservoir computing architecture and training principle in which an input-driven high-dimensional internal state $\boldsymbol{x}(t)$ is used as a feature representation, while learning is largely confined to a readout layer. In typical ESNs, the recurrent reservoir weights $\boldsymbol{W}$ and input weights $\boldsymbol{W_{in}}$ are randomly initialised and kept fixed, and supervised learning is performed by estimating $\boldsymbol{W_{out}}$ using linear regression methods. Therefore, it is ideal model when we separate "representation" and "decision/classification/computation". In this study, I modified it by introducing

Hebbian plasticity in $\boldsymbol{W_{in}}$, but still with no explicit optimisation.

Several studies have demonstrated that the introduction of a variable called the "leaking rate" can enhance the computational performance of ESN (Jaeger, 2001). However, my research aims to examine changes in the system's behaviour induced by plasticity, rather than its computational performance; therefore, to simplify the problem, I did not introduce the leaking rate.

Spectral Radius

The spectral radius is an important hyperparameter that controls the connection strength in the reservoir. Specifically, it refers to the maximum absolute value of the eigenvalues of $\boldsymbol{W}$. In general, if the activation function f of the reservoir is $\tanh$,

$$\rho(\boldsymbol{W}) = \max_i(|\lambda_i|) < 1 \qquad \text{(eq. S6)}$$

is a necessary condition for a reservoir to have ESP. Therefore, by adjusting the value of the spectral radius, I prepared various reservoirs near the edge of chaos (Legenstein & Maass, 2007) (the edge of having and lacking ESP).

Minibatch-based Gradient Descent

The error function that the gradient descent algorithm minimises is written as follows:

$$E(\boldsymbol{W}^{out}) = \frac{1}{2}\sum_{t=1}^{\mathrm{T}} \| d(t) - y(\boldsymbol{x}(t); \boldsymbol{W}^{out}) \|^2, \qquad \text{(eq. S7)}$$

where $d(t)$ is the training data at time point t, and T is the total number of time points. The goal of the learning is to find the following:

$$\boldsymbol{W}^{out} = \mathrm{argmin}_{\boldsymbol{W}} E(\boldsymbol{W}). \qquad \text{(eq. S8)}$$

To solve the above problem, the gradient $\nabla \boldsymbol{E}$ is acquired. Then, the algorithm changes the variables $\boldsymbol{W}$ for the negative direction of the gradient $-\nabla \boldsymbol{E}$ proportionally to the learning rate η:

$$\Delta \boldsymbol{W}^{out} = -\eta \nabla \boldsymbol{E}. \qquad \text{(eq. S9)}$$

Hyperparameter η controls how rapidly the descent progresses, which is generally set at 0.01–0.00001. For my model, 20 statistically independent noisy inputs were given to reproduce the perceptual learning of humans, and I repeatedly applied the gradient descent method for each trial. This kind of pseudo-online gradient descent method is called "the minibatch gradient descent method" because the training data are divided into "minibatches". The error of the minibatch number n, E_n, is calculated as follows:

$$E_n(\boldsymbol{W}^{out}) = \frac{1}{\mathrm{T_n}} \sum_{t \in D_n} E_k(\boldsymbol{W}^{out}), \qquad \text{(eq. S10)}$$

where E_n reproduces the error of the minibatch number n, and $\mathrm{T_n}$ is the sample size of the minibatch data D_n. For normalisation, the summation of error at each time point is divided by the sample size. The learning rate η was set to 0.01.

Inter-trial Phase Coherence (ITPC), Phase-Locking Value (PLV)

Let the instantaneous phase at time t, frequency f, and channel ch on trial $k \in \{1, \ldots, \mathrm{K}\}$ be $\theta_k(t, f, c)$, and define the unit phasor

$$\mathrm{z}_k(t, f, ch) = e^{i\theta_k(t,f,ch)}. \qquad \text{(eq. S11)}$$

The basic inter-trial phase coherence (ITPC; also called the mean resultant length) is

$$\mathrm{ITPC}(t, f, ch) = \left| \frac{1}{K} \sum_{k=1}^{K} z_k(t, f, ch) \right|. \qquad \text{(eq. S12)}$$

It ranges from 0 to 1: values near 0 indicate high phase variability across trials, whereas values near 1 indicate strong phase alignment. However, ITPC/PLV can show positive bias for finite trial counts (M. X. Cohen, 2014; van Diepen & Mazaheri, 2018).

cITPC: Corrected ITPC Value in This Study (Fig. S4)

Even if phases are not locked across trials—i.e. θ_n are i.i.d. uniform on [0,2π)—the sample mean of random unit vectors has a non-zero expected magnitude for finite K. This is easiest to see by considering ITPC^2:

$$\mathrm{ITPC}^2 = \left(\frac{1}{K} \sum_{k=1}^{K} z_k \right) \left(\frac{1}{K} \sum_{l=1}^{K} z_l \right)^* = \frac{1}{K^2} \sum_{k=1}^{K} \sum_{k=1}^{K} z_k z_l^*. \qquad \text{(eq. S13)}$$

Taking the expectation under i.i.d. uniform phases: for $k = l$: $z_k z_l^* = |z_k|^2 = 1$, for $k \neq l$: $\mathrm{E}[z_k z_l^*] = \mathrm{E}[e^{i\theta_k}]\mathrm{E}[e^{-i\theta_l}] = 0$, because $\mathrm{E}[e^{i\theta_k}] = 0$ for $\theta \sim \mathrm{Unif}(0, 2\pi)$. Therefore,

$$\mathrm{E}[\mathrm{ITPC}^2] = \frac{1}{K^2} \left(\sum_{k=1}^{K} 1 + \sum_{k \neq l} 0 \right) = \frac{1}{K}. \qquad \text{(eq. S14)}$$

This equation shows an exact finite-sample bias in ITPC^2: even with perfectly random phases, the

expected squared ITPC is $1/K$ rather than 0. Consequently, ITPC itself is also positively biased and decreases approximately as $K^{-1/2}$ under the null.

Thus, the desired corrected phase-consistency measure cITPC is an affine transform of ITPC^2 and satisfies (i) $\mathrm{cITPC} = 0$ when $\mathrm{ITPC}^2 = \frac{1}{K}$, (ii) $\mathrm{cITPC} = 1$ when $\mathrm{ITPC}^2 = 1$. And as the unique affine mapping that sends $x = \frac{1}{K} \mapsto 0$ and $x = 1 \mapsto 1$, we get

$$\mathrm{cITPC}(t,f,ch) = \frac{K}{K-1}\left(\mathrm{ITPC}^2(t,f,ch) - \frac{1}{K}\right). \qquad \text{(eq. S15)}$$

This conversion was proposed earlier (Aydore et al., 2013) and is equivalent to PPC (Vinck et al., 2010).

Pairwise Phase Consistency (PPC)

PPC quantifies how consistent a set of trial-wise phases $\{\theta_k(t,f,c)\}^K{}_{k=1}$ is by averaging, across all trial pairs, how aligned their phase differences are. Using eq. Sa, PPC can be defined as the average cosine of pairwise phase differences:

$$\mathrm{PPC} = \frac{2}{K(K-1)} \sum_{1 \le k < l \le K} \cos(\theta_k - \theta_l)\,. \qquad \text{(eq. S16)}$$

Under random (independent, uniform) phases its expectation is 0, and PPC was introduced as an estimator that avoids the positive finite-sample bias that can affect ITPC/PLV-type measures based on vector averaging (Vinck et al., 2010).

Using the identity

$$\left|\sum_{k=1}^{K} z_k\right|^2 = \sum_{k=1}^{K}\sum_{l=1}^{K} z_k z_l^* = K + 2\sum_{k<l} \cos(\theta_k - \theta_l)\,, \qquad \text{(eq. S17)}$$

we obtain

$$\mathrm{PPC} = \frac{\left|\sum_{k=1}^{K} z_k\right|^2 - K}{K(K-1)}. \qquad \text{(eq. S18)}$$

Since $\left|\sum_{k=1}^{K} z_k\right|^2 = K^2\mathrm{ITPC}^2$ from Eq.S12, substituting into the equation above yields

$$\mathrm{PPC} = \frac{K^2\mathrm{ITPC}^2 - K}{K(K-1)} = \frac{K\mathrm{ITPC}^2 - 1}{K-1}, \qquad \text{(eq. S19)}$$

which is identical to the constraint-derived ITPC correction (cITPC) of Eq.S15.

Spectral Exponent

In a dynamical system poised near a criticality, long-range correlations in time and space can become scale-invariant, yielding power-law statistics without a characteristic scale. When temporal correlations follow a power-law form, the corresponding power spectrum can adopt a broadband $(f) \propto f^{-\alpha}$ scaling, making the spectral exponent α compact summary of scale-free temporal structure. Spectral exponent quantifies how quickly the (aperiodic) background of a neural power spectrum decreases with frequency. Within the self-organised criticality framework, 1/f-like spectra have been proposed as a characteristic temporal fingerprint of critical-state dynamics (Bak et al., 1987).

Many electrophysiological signals show an approximate scale-free form

$$P(f) \approx Af^{\beta} \ (\beta < 0), \qquad \text{(eq. S20)}$$

or equivalently

$$P(f) \approx \frac{A}{f^{\alpha}} \ (\alpha > 0, \alpha = -\beta). \qquad \text{(eq. S21)}$$

On log-log axes, the exponent is the slope:

$$\log P(f) = \beta \log f + \log A\,. \qquad \text{(eq. S22)}$$

A more negative β (larger α) corresponds to a steeper spectrum (relatively more low-frequency power); a less negative β (smaller α) corresponds to a flatter spectrum (relatively more high-frequency power). This aperiodic component is commonly interpreted as reflecting scale-free or arrhythmic activity rather than narrowband oscillations.

In non-invasive EEG/MEG, apart from oscillatory peaks, a broadband aperiodic (1/f-like) component is consistently observed as background signal, which can be summarised as the spectral exponent α (He, 2014b). Therefore, the spectral exponent in EEG is not so much an indicator directly identifying "criticality itself", but rather serves as a practical proxy, enabling comparison via a single parameter of whether the system exhibits more scale-invariant dynamics (i.e., consistent with a state "close" to criticality).

However, crucially, a 1/f-like spectrum is not a definitive test of criticality. 1/f noise can arise from multiple non-critical mechanisms (e.g., mixtures of processes with widely distributed time constants), as established in classic reviews of 1/f phenomena. In the neuroscience context, it has also been argued—using simultaneous global and neuronal measurements—that 1/f scaling in global signals (including EEG) does not, by itself, imply that the underlying neuronal dynamics are in a critical state (Bédard et al., 2006). More generally, the critical brain hypothesis remains an active area with

established results and ongoing conceptual/empirical controversies (Wilting & Priesemann, 2019). Accordingly, in this thesis, the EEG spectral exponent is used as a pragmatic proxy for scale-free dynamics that may covary with proximity to critical regimes, rather than as standalone evidence of criticality.

Intraclass Correlation Coefficients

The intraclass correlation coefficient (ICC) summarises how reliably multiple measurements obtained on the same targets (e.g., repeated sessions,) reflect stable between-target differences. Conceptually, ICC is constructed by decomposing the total variance into a between-target component $\sigma^2_{between}$ and a within-target (error) component σ^2_{within}, and expressing ICC as the proportion of total variance attributable to between-target differences. So a basic variance components form is

$$\mathrm{ICC} = \frac{\sigma^2_{between}}{\sigma^2_{between} + \sigma^2_{within}}. \qquad (\mathrm{eq.\,S23})$$

There are multiple ICC "forms" matched to the study design. Shrout & Fleiss (1979) formalised choices based on the underlying ANOVA/mixed model (e.g., one-way vs two-way; raters treated as fixed vs random) and on whether one wants reliability of a single measurement or of the mean of k measurements (often denoted ICC(1/2/3, 1/k)). Therefore, ICC(3,k) in this study means the performance reliability across three sessions.

Cluster-based Monte-Carlo Permutation Test

Time–frequency cITPC effects were assessed using nonparametric cluster-based Monte Carlo permutation tests to control the family-wise error rate across the time–frequency dimensions. In brief, condition contrasts were computed at each time–frequency bin, bins exceeding a cluster-forming threshold were grouped into contiguous clusters, and cluster mass (sum of test statistics) was used as the cluster-level statistic. A null distribution of maximum cluster mass was generated by repeatedly permuting condition labels within participants (and within session when applicable) and recomputing the maximum cluster statistic for each permutation. Observed clusters were declared significant if their cluster mass exceeded the (97.5) quantile of the permutation null distribution (Maris & Oostenveld, 2007).

Lind–Mehlum U-test

The Lind–Mehlum U-test is designed to avoid the common mistake of inferring a U- (or inverted-

U-, like in this study) shape from a significant quadratic term alone. Instead, it tests whether the slope changes sign over a specified interval $[x_L, x_R]$.

For a quadratic $y = \alpha + \beta x + \gamma x^2 + \cdots$, the slope is $\frac{dy}{dx} = \beta + 2\gamma x$. Therefore, an inversed U-shape requires

$$\beta + 2\gamma x_L > 0, \qquad \beta + 2\gamma x_R < 0. \tag{eq. S24}$$

This is implemented as an intersection–union test: both one-sided endpoint-slope tests must be significant (Lind & Mehlum, 2010).

Predictive Coding and Free Energy Principle

Among contemporary computational theories of perception and perceptual learning, Bayesian inference–based frameworks are currently the most widely accepted. The most comprehensive and influential of these is Friston's Free Energy Principle (FEP) (Friston, 2005; Friston et al., 2006). In the present thesis, while remaining mindful of this framework, I do not restrict the discussion to it alone. Instead, I situate the argument within the broader family of Bayesian theories—including the Bayesian Brain hypothesis (Knill & Pouget, 2004) and predictive coding (Friston, 2005; Rao & Ballard, 1999)—and outline how perception and perceptual learning are conceptualised under these approaches.

Since Helmholtz's seminal observations (von Helmholtz, 1924), perception has been understood as the process by which the brain infers the state of the external world from sensory signals. Because sensory organs have inherently limited resolution and reliability, it is impossible to access the physical state of the environment directly. As a result, perception constitutes a fundamentally ill-posed inverse problem, one that necessarily requires inference. In this sense, the "environment" we experience is not the physical world itself, but rather the outcome of neural inference performed by the brain.

Today, it is widely held that this inference process is Bayesian. I therefore begin by providing an overview of this perspective. Let s denote the physical state of the environment (and its stimuli), which is not directly observable, and let x denote the sensory neural activity generated by that environment. The quantity to be inferred is therefore the state of the environment that caused the observed sensory input, formalised as the posterior distribution $p(s|x)$. s, x are given as

$$\frac{ds(t)}{dt} = f\big(s(t), v(t)\big) + \varepsilon_s(t), \tag{eq. S25}$$

$$x(t) = g\big(s(t), v(t)\big) + \varepsilon_u(t). \tag{eq. S26}$$

Here, v denotes a hidden cause of the environment—an unobserved variable that either remains

constant over time or drives transitions between environmental states. For simplicity, v will be omitted from the formulation below; it will be needed for hierarchical models. Equation n is referred to as the state equation, and equation m as the observation equation. The state equation describes how the environment evolves, whereas the observation equation describes how sensory signals are generated from the environmental state. The terms ε represent uncertainty in environmental dynamics and in neural observation, respectively, and are assumed to be independent of other variables and to follow Gaussian distributions.

The goal of the brain is to infer the state of the external world from the available sensory signals, that is, to compute

$$\mathrm{p}(s|x) = \frac{\mathrm{p}(x,s)}{\mathrm{p}(x)} = \frac{\mathrm{p}(x|s)\mathrm{p}(s)}{\mathrm{p}(x)} \qquad \text{(eq. S27)}$$

Because $\mathrm{p}(x,s)$ specifies the correspondence between environmental states and sensory signals, knowing this joint probability distribution would imply complete knowledge of the environment. In this sense, the joint distribution $\mathrm{p}(x,s)$ is referred to as the generative model of the environment.

By the rules of probability, the generative model can also be factorised, as shown on the middle and right-hand sides. Here, $\mathrm{p}(s|x)$ can be regarded as corresponding to the observation equation (Eq. S25), and $\mathrm{p}(s)$ to the state equation (Eq. S26). Because the brain cannot represent the world perfectly, perception is not identified with the true posterior distribution $\mathrm{p}(s|x)$, but rather with an approximate posterior $q(s)$. Perceptual learning can therefore be characterised as the process by which this approximate posterior $q(s)$ is brought closer to the true posterior $\mathrm{p}(s|x)$.

Introducing the Kullback–Leibler divergence, which quantifies the distance between two probability distributions, the discrepancy between these distributions can be expressed as follows.

$$\mathrm{D}_{\mathrm{KL}}(q(s)||p(s|x)) = \int_{-\infty}^{\infty} q(s)\log\left(\frac{q(s)}{p(s|x)}\right)ds$$

$$= \int_{-\infty}^{\infty} q(s)\log\left(\frac{q(s)}{p(x,s)}\right)ds - (-\log p(x)). \qquad \text{(eq. S28)}$$

In this formulation, the first and second terms are referred to in the FEP as free energy and surprise, respectively (Friston et al., 2006). Surprise corresponds to Shannon's self-information and is therefore also called Shannon surprise (Shannon, 1948).

Because surprise does not depend on the inferred environmental state s, obtaining a better approximation $q(s)$ that more faithfully reflects the environment can be achieved by minimising free energy. This is the origin of the term *Free Energy Principle* (Friston et al., 2006). In contrast, free energy is determined by the recognition distribution $q(s)$ and the generative model $\mathrm{p}(x,s)$.

$$F(q,p;x) = \int_{-\infty}^{\infty} q(s)\log\left(\frac{q(s)}{p(x,s)}\right)ds \qquad \text{(eq. S29)}$$

Introducing the internal energy $U(s;\ x)$, defined as:

$$U(s;x) = -\log p(x,s) = -\log p(x|s)\ -\log p(s), \qquad \text{(eq. S30)}$$

the free energy F can be rewritten as follows.

$$F(q,p;x) = \int_{-\infty}^{\infty} q(s)U(s;x)ds + \int_{-\infty}^{\infty} q(s)\log q(s)\,ds \qquad \text{(eq. S31)}$$

Within the FEP, the Laplace approximation is typically adopted, such that the recognition distribution $q(s)$ is assumed to be Gaussian. Accordingly, $q(s)$ is fully specified by its mean μ and covariance Σ, that is, $q(s) \backsim N(\mu,\Sigma)$. Taking this assumption into account, the free energy can be further simplified, yielding the following expression.

$$F(\mu,\Sigma) = U(\mu) + \frac{1}{2}\mathrm{tr}\big(\Sigma\nabla^2 U(\mu)\big) - \frac{1}{2}\log|\Sigma| - \frac{n}{2}\log(2\pi e) \qquad \text{(eq. S32)}$$

Here, the explicit dependence on the generative model p and the sensory input u has been omitted for notational simplicity. The values of μ and Σ that minimise this expression are obtained by gradient descent. Noting that, at the minimum of F, the gradient with respect to the parameters vanishes, the optimal solution for the covariance can be expressed as a function of μ, as

$$\Sigma_{\mathrm{opt}} = \big(\nabla^2 U(\mu)\big)^{-1}. \qquad \text{(eq. S33)}$$

Consequently, the free-energy expression can be reduced to the form given as

$$F(\mu) = U(\mu) - \frac{1}{2}\log|\Sigma| - \frac{n}{2}\log(2\pi)\,, \qquad \text{(eq. S34)}$$

which depends only on the mean parameter μ. Within the FEP, this mean-parameter μ is interpreted as being encoded in the activity of neural populations.

Taken together, under a Bayesian interpretation, perceptual learning can be described as the process of estimating the optimal parameter μ_{opt} that minimises the free energy F under the approximate posterior $q(s)$, typically via gradient descent. Recalling that the sensory input is generated as

$$u(\mathrm{t}) = g\big(s(t), v(t)\big) + \varepsilon_x(t), \varepsilon(\mathrm{t}) \backsim \mathrm{N}(0,\sigma_x^2), \qquad \text{(eq. S35)}$$

this leads to the update rule given below.

$$\frac{d\mu}{dt} = -\frac{\partial F(\mu)}{\partial \mu} = \frac{x_1 - g(\mu)}{\sigma_x^2} g'(\mu) + \frac{\mu_0 - \mu}{\sigma_s^2} \quad \text{(eq. S36)}$$

In particular, by introducing precision Π (the inverse of variance) and the prediction error ϵ,

$$\frac{x_1 - g(\mu)}{\sigma_x^2} = \Pi_\varepsilon \epsilon_x, \quad \text{(eq. S37)}$$

it becomes clear that the prediction error and the precision weight the magnitude of model updates.

Funding

This study was supported in part by JSPS KAKENHI, JP24KJ1151, JST SPRING, JPMJSP210, JST Moonshot R&D, JPMJMS2292, and The Encouraging Grants for Young Researchers at NIPS. I am sincerely grateful for this financial support.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Prof. Keiichi Kitajo, for his continuous, encouraging, and patient guidance throughout my 5-year PhD course. Without his and our lab members' daily, insightful, high standards, and generous support—I believe I could never have found anywhere else—I would not have been able to complete this work and be an independent researcher. Not only researchers, but also technical staff and secretaries.

I am also grateful to the members of my thesis committee, Prof. Hiromasa Takemura, Prof. Ryo Sasaki, and Prof. Okito Yamashita, for their constructive comments and valuable suggestions, which have greatly improved the quality of this work. I would further like to thank all researchers and students at the National Institute for Physiological Sciences for providing important advice, critical questions, and comments, which have also yielded many important insights.

I want to thank my colleague, Dr Makoto Hagihara, who collaborated on Project 2, for his tremendous technical support. I would also like to express my gratitude to the many researchers and students who, though not co-authors, offered valuable opinions — sometimes led to significant direction shifts — through discussions in corridors, at parties, and during my presentation at

conferences.

Finally, I would like to express my most profound appreciation to my family and friends. Despite the many difficulties they were facing, my family accepted my decision to pursue the uncertain and precarious path of graduate study far from home, warmly sent me off, and have continued to support me emotionally throughout this journey. Being born into such a family is the most excellent fortune of my life. I am also profoundly grateful to my friends all across Japan who have understood my choices and encouraged me, sharing both worries and hopes along the way. Without their patience and belief in me, this dissertation would not have been possible.

This thesis is dedicated to the memory of my late father, Hideki Goto.