# Recurrent neural network models reveal unified mechanisms generating event-related potentials from MMN to P300

*Anthony Strock[1,*], Trang-Anh E. Nghiem[1,*], Nathan Trouvain[4,5,6], Percy K. Mistry[1], Xavier Hinaut[4,5,6], Vinod Menon[1,2,3]*

Department of Psychiatry & Behavioral Sciences[1]
Department of Neurology & Neurological Sciences[2]
Wu Tsai Neurosciences Institute[3]
Stanford University School of Medicine
Stanford, CA 94305

INRIA Centre of Bordeaux University[4]
LaBRI, CNRS UMR 5800[5]
IMN, CNRS UMR 5293[6]
Bordeaux, France

* Equal Contribution

**Address for correspondence:**

Anthony Strock (astrock@stanford.edu), Vinod Menon (menon@stanford.edu)

## Abstract

The brain's ability to detect behaviorally relevant stimuli from sensory inputs is fundamental to cognition, yet the neural mechanisms linking synaptic processes to event-related potential (ERP) signatures remain unclear. Here, we develop a recurrent neural network (RNN) models of ERP responses demonstrating that short-term synaptic depression – a ubiquitous plasticity mechanism – provides a unified computational framework for mismatch negativity (MMN) and P300 responses across passive and active oddball paradigms. Our models reveal that neural populations spontaneously organize stimulus representations into probability-dependent geometric manifolds, where rare events occupy expanded state space. Hierarchical connectivity creates 9-fold signal amplification with enhanced noise robustness, explaining P300's functional advantages over sensory responses. Critically, the same synaptic mechanism accounts for attentional modulation of behaviorally relevant stimuli, providing the first unified explanation bridging automatic and controlled attention. This framework offers for how synaptic and connectivity disruptions manifest as altered MMN and P300 characteristics in neuropsychiatric disorders.

## Introduction

The human brain is bombarded by an immense stream of sensory data from which it selectively prioritizes stimuli that are rare and behaviorally relevant. The capacity to detect and respond to novel, unexpected, or behaviorally significant stimuli represents one of the most fundamental operations of neural systems—enabling rapid adaptation to changing environments while filtering out irrelevant background information and allocating resources to salient stimuli[1-5]. William James famously described attention as "taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought"[6]. The failure of this attentional system has profound consequences, contributing to core deficits observed in neuropsychiatric disorders such as schizophrenia, where individuals show impaired ability to distinguish behaviorally-relevant signals from neural and environmental noise[7]. Understanding the neural mechanisms underlying this process is therefore crucial for both basic neuroscience and clinical research.

Event-related potentials (ERPs) have proven invaluable for probing the temporal dynamics of orienting attention in the human brain [8-10]. Among the brain's electrophysiological signatures, few phenomena are as robust and well-characterized as the neural responses to unexpected stimuli. The discovery of ERPs in response to salient cues represents one of the oldest and most consistently replicated findings in the history of human neuroscience[11-14], with these responses observed across sensory modalities and experimental paradigms.

The oddball paradigm—where sequences of frequent standard stimuli are occasionally interrupted by rare deviant stimuli—has proven particularly powerful for isolating neural mechanisms of salience detection[15,16]. This simple but powerful experimental design effectively isolates brain responses to unexpected events and has been instrumental in characterizing two key ERP components that represent distinct stages of uncertainty and attention processing[17].

The mismatch negativity (MMN), typically occurring 150-250 milliseconds post-stimulus, is thought to reflect the brain's automatic, pre-attentive detection of deviations from established patterns[15]. Originating primarily in sensory cortical areas, the MMN represents a fundamental change-detection mechanism that operates below the threshold of conscious awareness [18]. Critically, MMN amplitude scales inversely with stimulus probability, providing a direct neural readout of the brain's internal statistical model of the environment[19].

The P300 response, peaking approximately 300 milliseconds post-stimulus onset, represents a later, more complex stage of attention processing[20]. Unlike the MMN, the P300 is maximal in frontal and parietal regions and is strongly modulated by attention and behavioral relevance[21]. The P300 comprises two functionally distinct subcomponents: an early component elicited by novel unexpected stimuli regardless of their behavioral significance, and a later component, which emerges specifically when stimuli require overt attention and behavioral responses[20]. This temporal and anatomical progression from MMN to P300 is thought to index a hierarchical processing cascade from automatic detection to conscious evaluation[12].

While MMN and P300 have been extensively studied as biomarkers of cognitive function and are consistently altered in neuropsychiatric conditions[22,23], the underlying neural mechanisms

remain poorly understood. This knowledge gap limits our ability to understand how cognitive functions emerge from neural circuits and prevents us from developing principled interventions for disorders characterized by attention processing deficits. Recent advances in computational neuroscience suggest that short-term synaptic depression (STD)—a ubiquitous form of activity-dependent plasticity—may provide a crucial piece of this puzzle[24]. STD causes synapses to weaken dynamically in response to repeated activation, creating a natural mechanism for distinguishing rare from common events. When stimuli are presented frequently, their corresponding synapses become depleted, leading to diminished neural responses. Conversely, synapses transmitting rare stimuli remain active, producing stronger neural responses. This mechanism has been proposed to explain MMN in sensory regions[19], but its potential role in generating P300 responses and enabling hierarchical attention processing across brain regions remains largely explored.

To bridge the gap between synaptic mechanisms and system-level cognition, we developed a computational framework using recurrent neural networks (RNNs) with biologically-constrained dynamics. RNNs have emerged as powerful tools for bridging explanatory levels in neuroscience, providing a flexible framework for modeling how neural dynamics give rise to cognition and behavior[25-27]. The RNN formalism is particularly powerful because enables systematic exploration of how basic synaptic mechanisms can generate complex neural signatures. This computational approach has proven valuable for discovering the minimal mechanisms sufficient to explain cognitive phenomena[25-27], making RNNs ideally suited for investigating the synaptic origins of attention to deviant oddball stimuli. While recent work has begun extending RNNs beyond individual brain regions[28,29], there remains significant potential in RNN development to explore neural mechanisms underlying ERPs and multi-regional representations of salience. Our work advances this frontier by developing hierarchical RNN architectures that model coordinated dynamics across sensory and association cortices, incorporating biologically-realistic synaptic plasticity mechanisms, and systematically characterizing the representational geometry that emerges from these dynamics.

Several fundamental questions about attention processing and P300 generation remain unresolved. A key question is how inter-regional connectivity influences the emergence of P300 signals. While it is well established that P300 responses involve frontal association cortex, the specific role of inter-regional connectivity in shaping these responses is not understood[21]. Another unresolved issue is whether MMN and P300 can arise from a shared neural mechanism. While STD has been shown to explain MMN in sensory regions, it remains unclear whether the same mechanism can account for P300 in association cortex. The role of feedforward and attentional modulation on P300 dynamics also remains an open question. Although P300 amplitude increases when stimuli are behaviorally relevant, the underlying neural processes responsible for this enhancement are also not known[20]. Feedback mechanisms from decision-making and attention systems may rely on the same synaptic mechanisms that generate passive responses to oddball stimuli, but this hypothesis has not been computationally tested[30,31]. Evidence accumulation processes may further contribute by sustaining neural representations over time, potentially modulating both the duration and robustness of salient oddball responses[32].

We had three fundamental goals which address these longstanding unresolved questions. Our first goal was to validate STD as a sufficient mechanism for sensory-level deviance detection.

We systematically tested whether short-term synaptic depression can account for MMN properties observed in sensory regions. While STD has been implicated in sensory adaptation, no previous work has demonstrated its sufficiency for explaining MMN's characteristic timing (150-250ms latency), probability scaling (inverse relationship with deviant frequency), and temporal sensitivity (dependence on interstimulus intervals). By implementing STD in a minimal sensory network and comparing model predictions to empirical MMN properties, we establish whether this ubiquitous synaptic mechanism provides a complete account of automatic change detection.

Our second goal was to elucidate how hierarchical connectivity transforms local change detection into global attention signals across sensory and association cortex. We investigated whether the propagation of sensory signals to frontal association regions through feedforward connectivity can explain the transformation from brief, localized MMN responses to the amplified, sustained P300 responses observed in association cortex. Crucially, we determined the quantitative relationship between structural connectivity and functional amplification, and whether hierarchical processing confers computational advantages such as enhanced noise robustness.

Our third goal was to determine whether passive and active attention reflect unified mechanisms. We tested the hypothesis that the same STD mechanisms explaining passive MMN and P300 responses can account for attentional enhancement of behaviorally relevant stimuli in active oddball paradigms. This goal directly addresses the longstanding question of whether automatic and controlled P300 components reflect fundamentally different neural systems or represent different expressions of shared computational principles.

Importantly, across each of these goals, we sought to characterize the representational geometry of hierarchical attention processing. We investigated how stimulus representations organize in high-dimensional neural state space, with particular focus on: (a) how representational manifolds scale with stimulus probability and frequency, (b) how hierarchical processing transforms the geometry of these manifolds, and (c) whether geometric changes enhance the robustness of stimulus discrimination under noise. This analysis addresses the crucial but understudied question of how information is spatially organized in neural networks and how this organization supports cognitive function, a question that traditional amplitude-based ERP analyses cannot address.

Our findings demonstrate that a parsimonious RNN model incorporating STD with feedforward and feedback connectivity can explain the full spectrum of neural responses to salient stimuli observed in both passive and active oddball paradigms. Our findings reveal that remarkably simple principles—STD combined with structured connectivity—can explain the emergence of complex neural signatures that have puzzled neuroscientists for decades. More broadly, this work demonstrates how computational approaches can illuminate the fundamental principles that govern the transformation of neural activity into cognitive function, providing a template for understanding other complex brain functions.

## Results

### A. RNN modelling of deviancy detection

***Short-term synaptic depression explains enhanced responses to deviant stimuli***

To establish a foundational mechanism for deviancy detection, we first investigated whether STD could account for the enhanced neural responses to rare stimuli observed in sensory regions during oddball paradigms. We modeled a sensory cortical network receiving inputs from two stimulus-specific input units (**Figure 1A**). The model incorporated STD at synapses between input units and the RNN, consisting of 80 excitatory and 20 inhibitory interconnected integrate-and-fire neurons. One of two stimuli ($X$ and $Y$) was presented at regular 2-second intervals, with a controlled probability of occurrence ranging from 0.1 to 0.9. Standard stimuli occurred with higher probability than deviant ones, typically 0.8 and 0.2 respectively.

We observed that STD at input synapses systematically modulated neural responses in a probability-dependent manner. Standard stimuli, due to their frequent presentation, induced progressive neurotransmitter depletion at their corresponding synapses, resulting in diminished response amplitudes over time (**Figure 1B**). In contrast, synapses transmitting deviant stimuli had more time to recover between presentations, better preserving their efficacy, and producing stronger neural responses.

The difference in the firing rate of sensory neurons between deviant and standard stimuli was statistically significant ($t = 18.46,\ p < 10^{-16}$, **Figure 1C**), but was no longer significant when STD was removed ($t = -0.50,\ p = 0.96$, **Figure 1C**). The peak differentiation consistently occurred at 190 ms post-stimulus onset (**Figure 1D-E**), coinciding roughly with stimulus offset. Moreover, this difference scaled inversely with deviant probability difference ($r = -0.95,\ p < 10^{-135}$, **Figure 1F**) and interstimulus interval ($r = -0.39,\ p < 10^{-10}$, **Figure 1I**), matching known MMN characteristics.

Further analysis confirmed significant negative correlations between firing rate and stimulus probability for both stimuli (for stimuli $X$, $r = -0.65,\ p < 10^{323}$, **Figure 1G**, for stimuli $Y$, $r = -0.70,\ p < 10^{-323}$, **Figures 1H**), demonstrating that single-neuron response inherently encode stimulus probability. This explains why the difference between average responses to stimuli B and A is negatively correlated to $p_Y - p_X = 2p_Y - 1$, and therefore highly negatively correlated to $p_Y$ (**Figure 1E**).

In contrast, firing rate showed a positive correlation with interstimulus interval for both stimulus types (for stimuli $X$, $r = 0.73,\ p < 10^{-323}$, **Figure 1G**, for stimuli $Y$, $r = 0.48,\ p < 10^{-276}$, **Figures 1H**), but faster increase for standard stimuli (**Figure 1G**) than for deviant stimuli (**Figures 1H**). Since, the average response for both stimulus types tends toward the overall firing rate when STD is removed, and the response to standard stimuli is lower than that to deviant stimuli at a 2-second inter-stimulus interval, this explains why the difference between the average responses to stimuli $Y$ and $X$ decreases as the inter-stimulus interval increases.

These results establish that STD at input synapses is sufficient to explain the characteristic properties of MMN responses in sensory cortex, including appropriate latency (150-250ms),

inverse scaling with deviant probability, and sensitivity to interstimulus intervals. Moreover, our findings reveal that probability encoding can naturally emerge from this simple synaptic mechanism.

***Frequency-dependent neural representation geometry emerges from STD***

To investigate how STD shapes the distributed representation of stimuli at the population level, we analyzed the high-dimensional patterns of neural activity in the sensory node. Neural activity from all 80 excitatory neurons in the sensory network was recorded during 100 trials of the oddball paradigm, varying either the probability of occurrence of stimuli or the interstimulus interval, both of which affect the frequency of occurrence of stimuli.

We applied two distinct principal component analyses (PCA) to simulated population activity matrices. Activity matrices were of dimension $N \times T$ with $N = 80$ (number of neurons) and $T = 1800s$ (number of timesteps following first stimulus onset, with 10ms resolution, when varying probability of occurrence) or $T = 2700s$ (number of timesteps following first stimulus when varying interstimulus interval). PCA of neural population activity revealed distinct representational patterns for stimuli in the oddball paradigm. The temporal trajectories of neural responses formed two non-overlapping semi-filled elliptical patterns in the reduced dimensional space, with one ellipse corresponding to stimulus $X$ and the other to stimulus $Y$ (**Figure 2A**). This separation indicates that at the population level, the network maintains distinct neural codes for stimuli based on their identity.

When we systematically varied the probability of the stimuli, we observed that the size of the elliptical trajectories was inversely related to stimulus occurrence probability (**Figure 2B**). Specifically, by fitting ellipses to the first second of the last 50 trials (**Figure 2C**), we observe that while the orientation of the ellipses does not vary significantly with probability of stimuli ($r = -0.006,\ p = 0.92$, **Figure 2D**), the semi-major axes of the ellipse decrease with the probability of occurrence for both stimuli (for stimulus $X$, $r = -0.90,\ p < 10^{-96}$, **Figure 2E**, for stimulus $Y$, $r = -0.92,\ p < 10^{-110}$, **Figures 2F**).

Similarly, when we systematically varied the interstimulus interval, we observed that the size of the elliptical trajectories was related to the interstimulus interval (**Figure 2H**). Specifically, we observed that while the orientation of the ellipses does not seem to vary with interstimulus interval ($r = -0.03,\ p = 0.58$, **Figure 2J**), the semi-major axes of the ellipse increased with the interstimulus interval for both stimulus type (for standard stimuli $X$, $r = 0.89,\ p < 10^{-94}$, **Figure 2K**, for stimuli $Y$, $r = 0.55,\ p < 10^{-21}$, **Figures 2L**).

These results demonstrate that STD can naturally give rise to emergent neural representations of stimulus frequency, where rarer events occupy larger representational areas in neural state space. This geometric property of neural representations provides a potential mechanism for the amplified processing of surprising or unexpected events at the neural population level.

**B. RNN modeling of passive oddball task**

***Inter-regional connectivity produces amplified P300-like responses in frontal regions***

Having established the mechanism for sensory-level deviance detection, we next investigated how these signals propagate and are transformed to generate the P300 component, which originates primarily from (non-sensory) association cortex and exhibits distinct temporal characteristics from the MMN. To accomplish this, we extended our model by adding a second RNN representing "frontal" cortical areas (**Figure 3A**). This frontal network received no direct sensory input but was driven by excitatory projections from the sensory network. We systematically varied the strength and density of these inter-regional connections to examine their impact on frontal response dynamics.

With appropriate connectivity parameters (e.g. full density, connection strength G=0.1), neural responses in the frontal network exhibited the hallmark characteristics of P300 signals (**Figure 3B**): they were delayed, amplified, and prolonged compared to their sensory counterparts (**Figure 3C**). Specifically, the peak difference between deviant and standard responses occurred at 310ms in frontal regions compared to 190ms in sensory regions – a 120ms delay (**Figure 3D**). Furthermore, this difference was significantly amplified in frontal areas by a factor of 9.34 ± 0.09 ($t = 16.59,\ p < 10^{-15}$, **Figure 3E**).

As with sensory responses, frontal response amplitudes decreased with increasing deviant stimulus probability ($r = -0.84,\ p < 10^{-40}$, **Figure 3F**) and interstimulus interval ($r = -0.82,\ p < 10^{-66}$, **Figure 3G**), consistent with empirical P300 characteristics. Importantly, we found that the amplification ratio between sensory and frontal responses was directly proportional to the product of inter-regional connection strength and probability ($r = 1.0,\ p < 10^{-323}$, **Figure 3H**), revealing a simple relationship between structural connectivity and functional amplification.

These results demonstrate that the characteristic properties of P300 response – their delayed timing, enhanced amplitude – can emerge naturally from the propagation of sensory signals through hierarchical brain networks with appropriate connectivity strengths. This mechanism requires no additional computational components beyond STD suggesting a fundamental principle for distributed attentional processing in the brain.

***Amplified response in frontal region improves robustness to noise***

Having established that the mechanism underlying P300 component would also cause amplification in frontal regions, we next examined its functional advantage. Specifically, we investigated the downstream processing of salient information by evaluating the conditions of external noise under which the two stimuli could still be discriminated. To this end, we added zero-mean Gaussian noise to the neuronal responses – without altering the underlying neural dynamics – and trained a linear classifier to decode stimulus identity within a 200ms sliding time window.

In the absence of noise, perfect classification accuracy (i.e. 100%) was achieved in both sensory and frontal regions. As the noise standard deviation was systematically increased, classification accuracy degraded rapidly in both the sensory (**Figure 4A&D**) and frontal (**Figure 4B&E**)

regions, eventually approaching the performance of a classifier that always predicts the standard stimulus (i.e. 80% given $p_X = 0.8$). However, we found that as long as the noise standard deviation $\sigma$ remained within reasonable bounds ($10^{-6} \leq \sigma \leq 1$), there existed a time point at which classification accuracy in the frontal network surpassed from 5 to 12% that of the sensory network ($t \geq 6.92,\ p < 10^{-7}$, **Figure 4C**). When considering the entire trial, classification accuracy remained significantly greater in the frontal network than in the sensory network for an even bigger noise standard deviation ($10^{-6} \leq \sigma \leq 10$, $t \geq 13.05,\ p < 10^{-35}$, **Figure 4F**).

These results demonstrate that the amplification of neural representations in frontal areas can provide a functional advantage by increasing robustness to noise, facilitating more reliable downstream processing of salient information.

***Frequency-dependent neural representation geometry propagates to frontal regions***

To investigate how the frequency-dependent neural representation from the sensory network propagates in the frontal network, we analyzed the high-dimensional patterns of neural activity in the frontal node. Neural activity from all 80 excitatory neurons in the "frontal" node was recorded during 100 trials of the oddball paradigm, while varying either the probability of occurrence of stimuli or the interstimulus interval, both of which affect the frequency of occurrence of stimuli.

We applied two distinct principal component analyses (PCA) to the population activity matrices of dimension $N \times T$. PCA of neural population activity again revealed distinct representational patterns for stimuli in the oddball paradigm. We observed similar temporal trajectories of neural responses forming two non-overlapping semi-filled elliptical patterns in the reduced dimensional space, with one ellipse corresponding to stimulus $X$ and the other to stimulus $Y$ (**Figure 5A**). However, in frontal areas, the first component explained even more variance than in sensory areas (~95% in sensory vs. 99.9% in frontal), and only the third component allowed for stimulus distinction, accounting for just 0.01% of the variance (compared to ~4% explained by the second component in sensory areas). These results suggest a progressive loss of stimulus-specific information along the processing hierarchy at the population level.

As in the sensory network, when we systematically varied stimulus probability, we observed that the size of the elliptical trajectories in the frontal network was inversely related to stimulus occurrence probability (**Figure 5B**). Specifically, by fitting ellipses to the first second of the last 50 trials (**Figure 5C**), we observe that while the orientation of the ellipses does not significantly vary with probability of stimuli ($r = -0.002,\ p = 0.97$, **Figure 5D**), the ellipses' semi-major axis length decrease with the probability of occurrence for both stimuli (for stimulus $X$, $r = -0.97,\ p < 10^{-82}$, **Figure 5E**, for stimulus $Y$, $r = -0.89,\ p < 10^{-93}$, **Figures 5F**).

When systematically varied the interstimulus interval, we observed that the size of the elliptical trajectories in the frontal network was related to the interstimulus interval (**Figure 5H**). Specifically, by fitting ellipses to the first second of the last 50 trials (**Figure 5I**), we observed that while the orientation of the ellipses does not seem to vary with interstimulus interval ($r = -0.09,\ p = 0.13$, **Figure 5J**), the semi-major axes of the ellipse increased with the interstimulus

interval for both stimulus type (for standard stimuli $X$, $r = 0.86$, $p < 10^{-80}$, **Figure 5K**, for stimuli $Y$, $r = 0.51$, $p < 10^{-18}$, **Figures 5L**).

These results demonstrate that the frequency-dependent neural representation that emerges in sensory network can naturally be transmitted to the frontal network and explains the observed frequency dependency of P300 amplitude.

## C. RNN modelling of active oddball task

### *Attentional modulation drives enhanced neural responses in active oddball tasks*

Having established that STD and inter-regional connectivity can account for the passive oddball P300 response, we next investigated whether the same mechanisms could be extended to explain the enhancement of neural responses in the active task paradigm. Empirical studies consistently show that P300 responses are further enhanced when participants actively respond to target stimuli compared to when they passively observe the same stimuli. This phenomenon, reflected in the P3b ERP component – has been linked to attentional modulation and behavioral relevance. We hypothesized that attentional modulation could account for this enhancement while maintaining the same core STD mechanism that explained passive oddball responses.

To model the active oddball paradigm, we extended our network architecture by adding two readout units connected to the frontal network (**Figure 6A**), which were trained to discriminate between standard and deviant stimuli throughout the whole duration of a trial. These units represented the two possible behavioral responses: producing a motor output (in response to target deviant stimuli) or withholding response (for standard stimuli). The feedforward connections from the frontal network to these readout units were trained using FORCE learning[33] to ensure that the appropriate unit activated following each stimulus type. Critically, we implemented feedback connections from these readout units back to the frontal network, with these connections also subject to STD. This created an attentional modulation where behavior-related signals could modulate frontal activity patterns (**Figure 6B**).

After the training phase, the RNN model successfully reproduced the key characteristics of neural responses in active oddball paradigms. As shown in the confusion matrix, classification accuracy was high (**Figure 6C**), demonstrating that the model reliably discriminated between stimulus categories despite the additional processing stage. The readout units successfully discriminated between stimulus types. The non-target unit consistently activated in response to standard stimuli, while the target unit preferentially responded to deviant stimuli, as the units were trained to. Specifically, stimulus $X$ (standard) was correctly classified in 99% of time steps, while stimulus $Y$ (deviant) was correctly identified 92% of the time, with 8% of instances misclassified as stimulus A (**Figure 6C**).

We found that the temporal activation profiles of the readout units (**Figure 6D**) showed clear differentiation between stimulus types. The non-target unit responded strongly to both standard and deviant stimuli at stimulus onset but maintained its response only when standard stimulus was presented and decayed rapidly when deviant stimulus was presented. The target unit responded strongly only to deviant stimuli and decayed slowly throughout the trial. This

differential activation pattern led to distinct feedback effects on the frontal network. When deviant target stimuli activated the target readout unit, the feedback connections enhanced firing rates within the frontal network, producing an amplified response profile (**Figure 6E**). Conversely, when standard stimuli activated the non-target readout unit, the feedback effect was attenuated due to synaptic depression at the more frequently used feedback connections. This resulted in a significantly larger deviant-standard difference in firing rates compared to the passive condition without feedback. In frontal networks, we observe similar profile in the firing rate response to both standard and deviant stimulus, with amplified response in the active condition (**Figure 6F**. The peak difference between responses to deviant and standard stimuli occurred at approximately 300ms (**Figure 6G**), aligning remarkably well with the typical latency of the P3b component observed in human EEG studies. This suggests that the additional processing time required for evidence accumulation naturally accounts for the temporal characteristics of the P3b response.

Importantly, although the difference in firing rate between deviant and standard stimuli did not differ significantly between passive and active condition in the sensory network ($t = 0.47,\ p = 0.64$), in frontal network, the difference was higher in the active condition than in the passive condition ($t = 8.95,\ p < 10^{-11}$, **Figure 6H**).

This demonstrates that feedback from attentional mechanisms can enhance amplification of responses to salient stimuli, providing a possible mechanistic explanation for the increased amplitude of P3b responses to behaviorally relevant stimuli.

***Attentional modulation preserves robustness to noise in frontal region***

Next, we investigated how attentional modulation affects the resilience to noise on downstream processing of salient information in the frontal network. As before, we evaluated the conditions of external noise under which the two stimuli could still be discriminated by adding a zero-mean Gaussian noise to the neuronal responses – without altering the underlying neural dynamics – and trained a linear classifier to decode stimulus identity within a 200ms sliding time window.

As the noise standard deviation increased, classification accuracy declined rapidly in both the passive (**Figure 7A&D**) and active (**Figure 7B&E**) conditions, eventually converging toward the performance of a classifier that consistently predicts the standard stimulus (i.e. 80% given $p_A = 0.8$). We observed that when the noise amplitude remained within reasonable bounds ($10^{-5} \leq \sigma \leq 1$), the classification accuracy remained high throughout the whole trial in the active condition showing a clear advantage of active over passive condition ($t \geq 7.88,\ p < 10^{-8}$). This pattern also held when looking across the entire trial: classification from the frontal network in the active condition continued to outperform the passive condition ($10^{-6} \leq \sigma \leq 10$, $t \geq 13.72,\ p < 10^{-38}$, **Figure 7F**).

These results demonstrate that attentional modulation can preserve the advantage of the simple amplification observed from sensory to frontal regions in the passive oddball. Consistent with previous results, the amplification of neural representations in frontal areas in active oddball

provides a functional benefit by enhancing robustness to noise, thereby supporting more reliable downstream processing of salient information.

***Frequency-dependent neural representation geometry are preserved by attentional modulation***

To investigate how the frequency-dependent neural representation from the sensory network propagates in the frontal network, we analyzed the high-dimensional patterns of neural activity in the frontal node. Neural activity from all 80 excitatory neurons in the frontal network was recorded during 100 trials of the oddball paradigm, varying either the probability of occurrence of stimuli or the interstimulus interval, both of which affect the frequency of occurrence of stimuli.

As before, we applied PCA to the population activity, which again revealed distinct representational patterns for stimuli in the oddball paradigm. We found that the feedback distorted the geometrical elliptical shape (**Figure 8A**). The first component explained less variance in the active than in the passive condition (~99.9% in passive vs ~99% in active), and the second component, accounting for ~1% of the variance, allowed again to discriminate the stimulus identity in the active condition. These results suggest that the feedback from the active condition prevents the loss of stimulus-specific information along the processing hierarchy at the population level.

Similar to the active condition, when we systematically varied stimulus probability, we observed that the size of the elliptical contour of the representation in the frontal network was inversely related to stimulus occurrence probability (**Figure 8B**). Specifically, by fitting ellipses to the convex hull of the last 50 trials (**Figure 8C**), we observe that while the orientation of the ellipses does not seem to vary with probability of stimuli ($r = -0.12,\ p = 0.058$, **Figure 8D**), the semi-major axis length decrease with the probability of occurrence for both stimulus type (for stimuli $X$, $r = -0.76,\ p < 10^{-52}$, **Figure 8E**, for stimuli $Y$, $r = -0.86,\ p < 10^{-81}$, **Figures 8F**).

When we systematically varied the interstimulus interval, we observed that the size of the elliptical contour of the representation in the frontal network was related to the interstimulus interval (**Figure 8H**). Specifically, by fitting ellipses to the convex hull of the last 50 trials (**Figure 8I**), we observed that the orientation of the ellipses slightly vary with interstimulus interval ($r = -0.23,\ p < 10^{-3}$, **Figure 8J**), the semi-major axes of the ellipse increased with the interstimulus interval for both stimulus type (for standard stimuli $X$, $r = 0.56,\ p < 10^{-23}$, **Figure 8K**, for stimuli $Y$, $r = 0.49,\ p < 10^{-17}$, **Figures 8L**).

These results demonstrate that the frequency-dependent neural representation that emerges in sensory network and is transmitted to the frontal network, can be maintained in the active condition and explain the observed frequency dependency of P300 amplitude.

## Discussion

We developed recurrent neural network (RNN) models to bridge the fundamental explanatory gap between cellular mechanisms and system-level cognitive functions. Our biologically-

constrained modeling approach systematically tested how basic synaptic plasticity and network connectivity can give rise to the neural signatures observed in event-related potential (ERP) studies of oddball paradigms. Through hierarchical RNN architectures that incorporate sensory and frontal regions with structured feedforward and feedback connections, we reproduced the complete spectrum of empirical phenomena—from automatic change detection in sensory regions to amplified, sustained responses in association cortex, and their modulation by attention and behavioral relevance.

Our results demonstrate that short-term synaptic depression (STD) combined with structured inter-regional connectivity and feedback loops can account for a wide spectrum of ERP responses across passive and active oddball paradigms, including their modulation by stimulus probability and behavioral relevance on neural response amplitude and duration. Representational and manifold analysis revealed that RNN populations spontaneously organize into probability-dependent geometric manifolds that enhance noise robustness through hierarchical amplification. These mechanisms generate the rich repertoire of neural responses that emerge during deviant "oddball" processing, from the automatic change detection reflected in mismatch negativity to the complex attentional modulations observed in P300 responses. This represents a fundamental shift from models that invoke separate mechanisms for different cognitive functions toward a unified framework based on shared computational principles operating across multiple regions and scales of brain organization.

***Short-term synaptic depression as a mechanism for sensory-level detection of salient stimuli***

Our first goal was to use RNNs to validate STD as a sufficient mechanism for explaining the complete spectrum of mismatch negativity properties in sensory regions. We show that STD can account for the emergence of enhanced responses to deviant stimuli during passive oddball paradigms with remarkable quantitative precision. The model reproduced MMN's characteristic timing (190ms peak), probability scaling ($r = -0.95$), and temporal sensitivity ($r = -0.39$), establishing STD as a viable mechanism for explaining pre-attentive change detection.

A note on polarity: It should be noted that the MMN deflections observed in our RNN model are positive, similar to other previous computational modeling studies[34]. In contrast, empirical MMN manifests as negative deflections in electric field recordings[12-14,31,35], reflecting the complex geometry of cortical dipole sources and volume conduction[36]. While our model focuses on the computational logic rather than the biophysical details of electric field generation, it provides a foundation for future work incorporating cortical laminar organization and realistic head models to explain the negative polarity. Crucially, the functional significance—enhanced processing of unexpected events—is fully captured by our STD mechanism.

Our findings represent a significant theoretical advance beyond previous modeling work. While earlier studies suggested STD's involvement in sensory adaptation[34,37-40], none demonstrated its sufficiency for explaining a complete range of empirically-observed MMN phenomenon. Our minimal network architecture—comprising only two input units, STD exclusively on input synapses, and RNN excitatory-inhibitory dynamics—proves that complex cognitive signatures can emerge from elementary synaptic mechanisms. This parsimony challenges assumptions about the need for complex mechanisms. While alternative mechanisms have been proposed,

including intracellular potassium depletion[41-43] and lateral inhibition in retinal circuits[44], but our results demonstrate that STD as a more parsimonious model[16]. The robustness of our findings across parameter variations and the precise quantitative match to empirical data support STD as the primary mechanism underlying automatic deviance detection. This has important implications for understanding evolutionary conservation—STD is ubiquitous across neural systems, suggesting that deviance detection may represent a fundamental computational primitive present in all nervous systems.

**Hierarchical connectivity creates robust salience signals in association cortex**

Our second goal addressed how local sensory change-detection signals transform into the amplified, sustained P300 responses characteristic of attentive processing[20,45]. We systematically varied the strength and density of feedforward connections between sensory and "frontal" networks and quantified the relationship between structural connectivity and functional amplification.

Our results revealed that hierarchical connectivity produces a striking transformation of neural dynamics. When inter-regional connections were sufficiently strong and dense (full connectivity with synaptic strength G=0.1), frontal networks exhibited all hallmark characteristics of empirical P300 responses: delayed timing (310ms peak versus 190ms in sensory regions), dramatic amplification (9.34-fold enhancement of the deviant-standard difference), and extended duration compared to brief sensory responses. Critically, we discovered a perfect linear relationship between the product of connection strength and connection probability and the magnitude of functional amplification ($r = 1.0$). This relationship held across the full range of connectivity parameters we tested, from weak sparse connections that produced attenuation to strong dense connections yielding substantial amplification.

We also observed that hierarchical processing creates systematic temporal delays that emerge from network dynamics rather than built-in delays. The 120ms delay between sensory and frontal peak responses arose naturally from the filtering properties of recurrent networks, with each processing stage acting as both an amplifier and low-pass filter. This emergent temporal structure matched empirical observations and created increasingly slower dynamics at higher levels of the cortical hierarchy, consistent with longer intrinsic timescales recorded in frontal regions[46,47].

These emergent properties of temporal hierarchy provide new insights into how cortical organization supports cognitive function. Rather than requiring specialized cellular mechanisms to create different timescales across brain regions, our results suggest that hierarchical network structure alone can generate the temporal dynamics observed empirically. Each recurrent network in the processing hierarchy naturally implements low-pass filtering that progressively slows responses while amplifying relevant signals. This principle may generalize beyond salience processing to explain temporal hierarchy throughout cortex, suggesting that the characteristic timescales of different brain regions reflect their position in processing hierarchies rather than solely intrinsic cellular properties[48-50].

The linear relationship between connectivity parameters and functional amplification suggests that P300 amplitude suggests a link with white matter integrity. This has implications for

understanding individual differences—variations in P300 amplitude may directly reflect differences in connectivity strength between sensory and association cortices. Similarly, clinical conditions affecting white matter (such as multiple sclerosis, schizophrenia, or healthy aging) should produce predictable changes in ERP characteristics that can be predicted using our framework.

To assess the functional significance of hierarchical amplification, we systematically tested noise robustness. We found that frontal networks maintained superior classification performance compared to sensory networks across a wide range of noise conditions (noise standard deviations from $10^{-2}$ to 10), with the advantage being particularly pronounced for sustained processing across entire trials. This enhanced robustness persisted even as overall performance degraded with increasing noise levels. The noise robustness suggests a particular advantage of hierarchical brain architectures. By progressively enhancing signal-to-noise ratios through amplification, hierarchical processing creates representations that remain reliably decodable despite the ubiquitous neural variability that characterizes biological systems.

**Unified mechanisms for automatic and controlled attention processing**

Our third goal tested whether the same STD mechanisms explaining passive responses could account for attentional enhancement in active behavioral paradigms. By extending STD to feedback connections from behavioral readout units, we successfully modeled the transition from automatic to controlled processing within a unified framework. The longstanding distinction between P300 elicited by these processes has been interpreted as evidence for fundamentally separate neural systems—automatic versus controlled processing mechanisms[20,51]. Our results demonstrate instead that these phenomena reflect the same underlying STD mechanism operating in different anatomical contexts. The key insight is that behavioral feedback creates asymmetric STD effects: frequent non-target feedback becomes depressed while rare target feedback remains strong, producing selective enhancement for behaviorally relevant stimuli.

Remarkably, the temporal precision of our P300 model (300ms peak latency) emerged naturally from feedback dynamics without parameter tuning. This temporal match to empirical data validates our mechanistic account and suggests that evidence accumulation and feedback naturally create the delays characteristic of controlled processing. The sustained activation of target units, made biologically plausible by longer intrinsic timescales in cortical regions[46], provides the temporal extension that distinguishes later P300 from earlier components.

**Neural representational geometry of attention processing**

Our fourth goal was to characterize neural representational geometry underlying attention processing. We discovered that neural populations spontaneously organize stimulus representations into probability-dependent geometric manifolds, where rare events occupy expanded regions in high-dimensional neural state space. This geometric organization provides a principled mechanism for pre-attentive and attentive processing that scales with statistical surprise.

The inverse relationship between stimulus probability and representational volume suggests that neural networks naturally implement optimal coding strategies that maximize information transmission about unexpected events. This geometric principle operates automatically through STD dynamics, requiring no explicit probability calculations. Our discovery of probability-dependent manifold scaling may provide a novel principle of neural coding.

Our analysis also revealed that hierarchical processing refines these geometric representations. While preserving probability-dependent scaling, frontal networks show increased concentration of variance along primary dimensions (99.9% vs 95% in sensory regions) with stimulus-specific information becoming more refined. This suggests that cortical hierarchy serves not only to amplify signals but to extract and concentrate the most behaviorally relevant features while filtering irrelevant variation.

The preservation of noise robustness in active conditions demonstrates that attentional enhancement does not compromise representational quality. Unlike simple amplification schemes that might increase noise susceptibility, our STD-based feedback mechanism enhances robustness by creating sustained, stable representations. This provides a principled explanation for how feedback-enhanced attention improves both signal strength and reliability.

These geometric principles may represent fundamental constraints on neural representation that extend beyond attention processing. The organization of neural activity into probability-dependent manifolds could be a general feature of neural networks that must represent statistical structure, with implications for understanding learning, memory, and decision-making. We suggest that by creating representations that remain discriminable under noise while encoding statistical information, these geometric principles enable robust behavior in uncertain environments.

## Conclusions

We demonstrate that recurrent neural networks with short-term synaptic depression, a ubiquitous synaptic mechanism, provides a unified computational framework for automatic change detection and controlled attention. Combined with hierarchical connectivity, short-term synaptic depression generates the complete spectrum of neural responses observed in oddball paradigms, including its canonical MMN and P300 components of event-related potentials. This shows that a parsimonious principle operating across multiple brain scales can explain complex neural signatures of salience detection. Our discovery that neural populations spontaneously organize stimulus representations into probability-dependent geometric manifolds reveals a novel coding strategy where rare events automatically occupy expanded representational space. This geometric organization emerges naturally from RNN dynamics and provides an elegant solution to the fundamental challenge of extracting behaviorally crucial information from complex sensory environments. Our unified framework linking synaptic mechanisms to event-related potential characteristics provides a quantitative computational basis for understanding psychiatric disorders characterized by attention processing deficits. In conditions such as schizophrenia and autism, where P300 abnormalities are well-documented biomarkers[22,52-55].

## Methods

### Stimuli and task — Oddball paradigm

To generate stimuli as in the oddball paradigm, we randomly draw one of two stimuli $X$ and $Y$ with probabilities $p_X$ and $p_Y = 1 - p_X$. When using a fixed stimulus probability, we set $p_X = 0.8$ and $p_Y = 0.2$. For conditions involving varying stimulus probabilities, we systematically varied $p_X$ from 0.1 to 0.9 in increments of 0.1. We refer to the stimulus with the lower probability as the deviant and the one with the higher probability as the standard.

Each stimulus was presented 100 times, with each presentation lasting 200ms. In the fixed interstimulus interval condition, stimuli were presented every 2s, resulting in a total session duration of 200s. For conditions involving variable interstimulus intervals, the interval was systematically varied from 1.0 to 5.0s in increments of 0.5s.

In the passive version of the task, models received stimuli and no outputs were decoded from simulated dynamics. In the active version of the task, model received stimuli and a decoder was trained to reproduce the input from simulated dynamics for the whole duration of the trial, thus mimicking a task detecting the deviant stimuli.

### Model

#### *A. Recurrent network model of sensory responses to stimuli in passive oddball paradigm*

To model thalamic inputs sending stimulus-related signals to the sensory cortex, we simulate two input units corresponding to $X$ and $Y$ stimuli, whose activity levels are described by a one-hot binary 2-dimensional vector $u$ i.e. $u = (1,0)$ when stimulus $X$ is presented and $u = (0,1)$ when $Y$ is presented.

We model the sensory cortex as a recurrent network of 80 excitatory and 20 inhibitory leaky integrate-and-fire neurons subject to short-term depression of input synaptic weights, the dynamics of which can be described by:

$$\tau_I \dot{r}_I = -r_I + f(W_{E,I} r_E - W_{I,I} r_I + W_{in,I} x u)$$
$$\tau_E \dot{r}_E = -r_E + f\left(W_{E,E} r_E - W_{I,E} r_I + W_{in,E} x u\right)$$
$$\dot{x} = \frac{1 - x}{\tau_x} - x u$$

where $r_I$ and $r_E$ represent the firing rate activity of inhibitory and excitatory neurons respectively, $u$ represents the 2 excitatory input neurons, one each for stimuli $X$ and $Y$, and $x$ represents the fraction of neurotransmitters within each input neuron modulating the strength of signal transmission from input units to sensory neurons.

$W_{in,I}$ and $W_{in,E}$ represent the synaptic input weights and are uniformly randomly sampled between 0 and 1. $W_{E,I}$, $W_{I,I}$, $W_{E,E}$ and $W_{I,E}$ represent the recurrent synaptic weights and are first uniformly randomly sampled between 0 and 1, then globally multiplied by a scaling factor such

that the spectral radius, that is, the eigenvalue product of $\begin{pmatrix} -W_{I,I} & W_{E,I} \\ -W_{I,E} & W_{E,E} \end{pmatrix}$ is fixed to 1, and finally scaled individually so that $\frac{sum(W_{E,I})}{sum(W_{I,I})} = \frac{sum(W_{E,E})}{sum(W_{I,E})} = 1$.

$\tau_I = \tau_E = 100ms$ represent the time constant of inhibitory and excitatory neurons respectively, $\tau_x = 20s$ represents the adaptation time constant of the input, and finally $f$ represents the non-linear neural response function, and is a rectified linear function (i.e. $f(x) = \max(0, x)$).

To simplify the written representation of the model, we denote the firing rate dynamics as follows:

$$\tau\dot{r} = -r + f(Wr + W_{in}xu)$$

where $r = (r_I \quad r_E)$, $W = \begin{pmatrix} -W_{I,I} & W_{E,I} \\ -W_{I,E} & W_{E,E} \end{pmatrix}$, $W_{in} = (W_{in,I} \quad W_{in,E})$, and $\tau = \tau_I = \tau_E$.

***B. Two-network model of sensory and frontal responses to stimuli in passive oddball paradigm***

To model signal transmission from sensory to frontal areas, we simulate the sensory cortex and frontal cortex as two recurrent networks of 80 excitatory and 20 inhibitory leaky integrate-and-fire neurons each, connected in a feedforward way from the sensory cortex to the frontal cortex, subject to short-term depression only for the input synaptic weights of the sensory cortex. The dynamics can be described by the following:

$$\tau\dot{r_1} = -r_1 + f(W_1 r_1 + W_{in}xu)$$
$$\tau\dot{r_2} = -r_2 + f(W_2 r_2 + W_{1,2} r_1)$$
$$\dot{x} = \frac{1-x}{\tau_x} - xu$$

where $r_1$ and $r_2$ represent the firing rate of the sensory and frontal cortex respectively, $u$ represents input unit activity delivered only to the sensory network, and $x$ represents the fraction of neurotransmitters within each input neuron modulating the intensity of the input. $W_1$ and $W_2$ represent the recurrent synaptic weights of sensory and frontal cortex respectively, and are sampled in the same was $W$ in the previous model. $W_{1,2}$ represents the forward synaptic weights from sensory to frontal cortex. Only excitatory to excitatory connections were allowed between cortices. All non-zero weights were randomly drawn from a uniform distribution between 0 and $g$ where $g$ is a positive gain parameter.

***C. Two-network model of sensory, frontal, and behavioral responses in active oddball paradigm***

To investigate the impact of the behavioral relevance of stimuli on frontal responses, we model the behavioral target as an output from the frontal cortex, which is also fed back to the frontal cortex. This feedback is also subject to short-term depression.

To model frontal output detecting the target, we use two output units corresponding to $X$ and $Y$ stimuli, whose activity levels are described by a one-hot binary 2-dimensional vector $y$ i.e. $y = (1,0)$ when stimulus $X$ is presented and $y = (0,1)$ when $Y$ is presented.

The overall dynamics of the model can be described by:

$$\tau \dot{r_1} = -r_1 + f(W_1 r_1 + W_{in} x_u u)$$
$$\tau \dot{r_2} = -r_2 + f(W_2 r_2 + W_{1,2} r_1 + W_{fb} x_y y)$$
$$y = g(W_{out} r_2)$$
$$\dot{x_\alpha} = \frac{1 - x_\alpha}{\tau_{x_\alpha}} - x_\alpha \alpha$$

where $r_1$ and $r_2$ represent the firing rate of the sensory and frontal cortex respectively, $u$ represents the 2 excitatory input neurons, one for each stimuli $X$ and $Y$, and y represents 2 excitatory neurons detecting whether the target is present. $x_u$ and $x_y$ represent the fraction of neurotransmitters modulating the intensity of the inputs (within the input neuron) and the target prediction feedback (within the readout neuron) respectively. $W_1$ and $W_2$ represent the recurrent synaptic weights of the sensory and frontal cortex respectively, and are sampled similar to $W$ in the previous model. $W_{1,2}$ represents the forward synaptic weights from sensory to frontal cortex and is defined as in the previous model. $W_{out}$ represents the synaptic weights allowing the detection of the target from frontal response. $W_{fb}$ represents the feedback synaptic weights from target detection units to frontal cortex, and are sampled similarly to $W_{1,2}$; only target to excitatory neurons connections are allowed and corresponding synaptic weights are randomly uniformly sampled between 0 and $g_{fb}$, a posititve feedback gain.

$W_{out}$ is learnt with FORCE learning[33], a Hebbian learning rule which implements the Recursive Least Squares algorithm[56]. While learning occurs, the feedback that $r_2$ receives is replaced by the target value for $y$ referred as $\tilde{y}$. The overall learning dynamics can be described as:

$$\Delta W_{out} = W_{out}(t) - W_{out}(t - dt) = -e(t)P(t)x(t)$$
$$\Delta P = P(t) - P(t - dt) = -\frac{P(t - dt)x(t)x(t)^T P(t - dt)}{1 + x^T(t)P(t - dt)x(t)}$$
$$\mathrm{e}(t) = W_{out}(t)x(t) - \tilde{y}$$
$$W_{out}(0) = 0$$
$$P(0) = \frac{I}{\alpha}$$

Where $\alpha$ is a regularization parameter.

### *D. Models hyperparameters*

We ran 30 instances of every variant of model described above by initializing with different seeds but using the following common hyper-parameters (**Table 1**).

| Parameter | Symbol(s) | Value |
|---|---|---|
| Sampling step | dt | 10 ms |
| Neuron time constant | $\tau_I$, $\tau_E$, $\tau$ | 100 ms |
| Short-term depression time constant | $\tau_x$, $\tau_{x_X}$ | 20 s |

**Table 1. Parameters of models and task.**

## Analysis

### *Dimensionality reduction*

To study the underlying low-dimensional representation we used principal component analyses (PCA)[57]. Specifically, we applied PCA separately to the sensory and frontal networks, while either varying only the probability of occurrence of stimuli, or the interstimulus interval.

### *Convex hull ellipse fitting*

As neural representations of simulated neural responses in principal components resemble ellipses, we quantitatively measure the distortion caused by frequency of stimuli (probability or interstimulus interval) by fitting ellipses to the convex hull of two principal components of the firing rate response of excitatory neurons. We use only the last 50 trials to allow for time for the short-term depression to produce stable representations. To fit the ellipses we first perform a linear fit:

$$Ax^2 + Bxy + Cy^2 + Dx + Ey = 1$$

Where $x$ and $y$ are the coordinates of the convex hull of two principal components, $A$, $B$, $C$, $D$ and $E$ are the linear weights obtained through least square linear regression.

The we compute the parameters of the ellipses as follow:

$$\Delta = B^2 - 4AC$$

$$x_0 = \frac{2CD - BE}{\Delta}$$

$$y_0 = \frac{2AE - BD}{\Delta}$$

$$\theta = atan2(-B, C - A)$$

$$a = \frac{-\sqrt{2(AE^2 + CD^2 - BDE - \Delta)\left(A + C + \sqrt{(A + C)^2 + \Delta}\right)}}{\Delta}$$

$$b = \frac{-\sqrt{2(AE^2 + CD^2 - BDE - \Delta)\left(A + C - \sqrt{(A + C)^2 + \Delta}\right)}}{\Delta}$$

Where $x_0$ and $y_0$ are the coordinates of the center of the ellipse, $\theta$ is the rotation of the ellipse, $a$ and $b$ are the semi major and minor axes of the ellipse.

When the convex hull is a triangle, the initial ellipse fit occasionally fails — a rare occurrence observed only in the active condition within the frontal region. To address this issue, we introduce isotropic Gaussian noise to the convex hull coordinates only when the initial fit fails. The noise has zero mean and a standard deviation set to 0.01 times the average range of the first and second coordinates.

**Acknowledgement**

This work was supported by the National Institutes of Health (NS086085, MH137325), and by the France-Stanford Center for Interdisciplinary Studies.

## References


1 Lien, M.-C., Ruthruff, E. & Johnston, J. C. Attentional capture with rapidly changing attentional control settings. *Journal of Experimental Psychology: Human Perception and Performance* **36**, 1 (2010).

2 Nobre, A. C. K. & Mesulam, M. Large-scale networks for attentional biases. (2014).

3 Pashler, H. The psychology of attention. *JOP* **14** (2000).

4 Pashler, H., Johnston, J. C. & Ruthruff, E. Attention and performance. *Annual review of psychology* **52**, 629-651 (2001).

5 Yantis, S. in *Visual attention mechanisms* 125-134 (Springer, 2002).

6 James, W. The principles of psychology. *Henry Holt* (1890).

7 Hoonakker, M., Doignon-Camus, N. & Bonnefond, A. Sustaining attention to simple visual tasks: a central deficit in schizophrenia? A systematic review. *Annals of the New York Academy of Sciences* **1408**, 32-45 (2017).

8 Hopfingeris, J. B. & Slotnick, S. D. The Cognitive Neuroscience of Attention.

9 Mangun, G. R. *Cognitive electrophysiology of attention: signals of the mind*. (Academic Press, 2013).

10 Nobre, K. & Kastner, S. *The Oxford handbook of attention*. (Oxford University Press, USA, 2014).

11 Fitzgerald, K. & Todd, J. Making sense of mismatch negativity. *Frontiers in psychiatry* **11**, 468 (2020).

12 Picton, T. W. The P300 wave of the human event-related potential. *Journal of clinical neurophysiology* **9**, 456-456 (1992).

13 Polich, J. Neuropsychology of P300. *The Oxford handbook of event-related potential components* **641**, 159-188 (2012).

14 Pritchard, W. S. Psychophysiology of P300. *Psychological bulletin* **89**, 506 (1981).

15 Näätänen, R., Paavilainen, P., Rinne, T. & Alho, K. The mismatch negativity (MMN) in basic research of central auditory processing: a review. *Clinical neurophysiology* **118**, 2544-2590 (2007).

16 Nelken, I. & Ulanovsky, N. Mismatch negativity and stimulus-specific adaptation in animal models. *Journal of Psychophysiology* **21**, 214-223 (2007).

17 Sutton, S., Braren, M., Zubin, J. & John, E. Evoked-potential correlates of stimulus uncertainty. *Science* **150**, 1187-1188 (1965).

18 Paavilainen, P. The mismatch-negativity (MMN) component of the auditory event-related potential to violations of abstract regularities: a review. *International journal of psychophysiology* **88**, 109-123 (2013).

19 Garrido, M. I., Kilner, J. M., Stephan, K. E. & Friston, K. J. The mismatch negativity: a review of underlying mechanisms. *Clinical neurophysiology* **120**, 453-463 (2009).

20 Polich, J. Updating P300: an integrative theory of P3a and P3b. *Clin Neurophysiol* **118**, 2128-2148 (2007). https://doi.org:10.1016/j.clinph.2007.04.019

21 Linden, D. E. The P300: where in the brain is it produced and what does it tell us? *The Neuroscientist* **11**, 563-576 (2005).

22 Cui, T., Wang, P. P., Liu, S. & Zhang, X. P300 amplitude and latency in autism spectrum disorder: a meta-analysis. *European child & adolescent psychiatry* **26**, 177-190 (2017).

23 Tada, M. *et al.* Mismatch negativity (MMN) as a tool for translational investigations into early psychosis: A review. *International Journal of Psychophysiology* **145**, 5-14 (2019).

24 Abbott, L. F. & Regehr, W. G. Synaptic computation. *Nature* **431**, 796-803 (2004).
25 Ji-An, L., Benna, M. K. & Mattar, M. G. Discovering cognitive strategies with tiny recurrent neural networks. *Nature*, 1-9 (2025).
26 Richards, B. A. *et al.* A deep learning framework for neuroscience. *Nature neuroscience* **22**, 1761-1770 (2019).
27 Yang, G. R. & Wang, X.-J. Artificial neural networks for neuroscientists: a primer. *Neuron* **107**, 1048-1070 (2020).
28 Perich, M. G. *et al.* Inferring brain-wide interactions using data-constrained recurrent neural network models. *BioRxiv*, 2020.2012. 2018.423348 (2020).
29 Soo, W., Battista, A., Radmard, P. & Wang, X.-J. Recurrent neural network dynamical systems for biological vision. *Advances in Neural Information Processing Systems* **37**, 135966-135982 (2024).
30 Mertens, R. & Polich, J. P300 from a single-stimulus paradigm: passive versus active tasks and stimulus modality. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section* **104**, 488-497 (1997).
31 Tiitinen, H., May, P., Reinikainen, K. & Näätänen, R. Attentive novelty detection in humans is governed by pre-attentive sensory memory. *Nature* **372**, 90-92 (1994).
32 Donchin, E. & Coles, M. G. Is the P300 component a manifestation of context updating? *Behavioral and brain sciences* **11**, 357-374 (1988).
33 Sussillo, D. & Abbott, L. F. Generating coherent patterns of activity from chaotic neural networks. *Neuron* **63**, 544-557 (2009).
34 Yarden, T. S. & Nelken, I. Stimulus-specific adaptation in a recurrent network model of primary auditory cortex. *PLoS computational biology* **13**, e1005437 (2017).
35 Hamm, J. P. & Yuste, R. Somatostatin interneurons control a key component of mismatch negativity in mouse visual cortex. *Cell reports* **16**, 597-604 (2016).
36 Nunez, P. L. & Srinivasan, R. *Electric fields of the brain: the neurophysics of EEG*. (Oxford university press, 2006).
37 Mill, R., Coath, M., Wennekers, T. & Denham, S. L. A neurocomputational model of stimulus-specific adaptation to oddball and Markov sequences. *PLoS computational biology* **7**, e1002117 (2011).
38 Seay, M. J., Natan, R. G., Geffen, M. N. & Buonomano, D. V. Differential short-term plasticity of PV and SST neurons accounts for adaptation and facilitation of cortical neurons to auditory tones. *Journal of Neuroscience* **40**, 9224-9235 (2020).
39 Shymkiv, Y., Hamm, J. P., Escola, S. & Yuste, R. Slow cortical dynamics generate context processing and novelty detection. *Neuron* **113**, 847-857. e848 (2025).
40 Yarden, T. S., Mizrahi, A. & Nelken, I. Context-dependent inhibitory control of stimulus-specific adaptation. *Journal of Neuroscience* **42**, 4629-4651 (2022).
41 Brette, R. & Gerstner, W. Adaptive exponential integrate-and-fire model as an effective description of neuronal activity. *Journal of neurophysiology* **94**, 3637-3642 (2005).
42 Jercog, D. *et al.* UP-DOWN cortical dynamics reflect state transitions in a bistable network. *elife* **6**, e22425 (2017).
43 Nghiem, T.-A. E. *et al.* Cholinergic switch between two types of slow waves in cerebral cortex. *Cerebral Cortex* **30**, 3451-3466 (2020).
44 Ebert, S., Buffet, T., Sermet, B. S., Marre, O. & Cessac, B. Temporal pattern recognition in retinal ganglion cells is mediated by dynamical inhibitory synapses. *Nature communications* **15**, 6118 (2024).

45 Polich, J. 50+ years of P300: Where are we now? *Psychophysiology* **57** (2020).
46 Golesorkhi, M. *et al.* The brain and its time: intrinsic neural timescales are key for input processing. *Communications biology* **4**, 970 (2021).
47 Kiebel, S. J., Daunizeau, J. & Friston, K. J. A hierarchy of time-scales and the brain. *PLoS computational biology* **4**, e1000209 (2008).
48 Chaudhuri, R., Knoblauch, K., Gariel, M.-A., Kennedy, H. & Wang, X.-J. A large-scale circuit mechanism for hierarchical dynamical processing in the primate cortex. *Neuron* **88**, 419-431 (2015).
49 Murray, J. D. *et al.* A hierarchy of intrinsic timescales across primate cortex. *Nature neuroscience* **17**, 1661-1663 (2014).
50 Wang, X.-J. Probabilistic decision making by slow reverberation in cortical circuits. *Neuron* **36**, 955-968 (2002).
51 Polich, J. & Criado, J. R. Neuropsychology and neuropharmacology of P3a and P3b. *International Journal of Psychophysiology* **60**, 172-185 (2006).
52 Bramon, E., Rabe-Hesketh, S., Sham, P., Murray, R. M. & Frangou, S. Meta-analysis of the P300 and P50 waveforms in schizophrenia. *Schizophrenia research* **70**, 315-329 (2004).
53 Foss-Feig, J. H. *et al.* Abnormally large baseline P300 amplitude is associated with conversion to psychosis in clinical high risk individuals with a history of autism: a pilot study. *Frontiers in psychiatry* **12**, 591127 (2021).
54 Hamilton, H. K., Mathalon, D. H. & Ford, J. M. P300 in schizophrenia: Then and now. *Biological psychology* **187**, 108757 (2024).
55 Jeon, Y. W. & Polich, J. Meta-analysis of P300 and schizophrenia: Patients, paradigms, and practical implications. *Psychophysiology* **40**, 684-701 (2003).
56 Hayes, M. H. in *Statistical digital signal processing and modeling* Ch. 9.4: Recursive Least Squares, (John Wiley & Sons, 1996).
57 Jolliffe, I. T. & Cadima, J. Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences* **374**, 20150202 (2016).

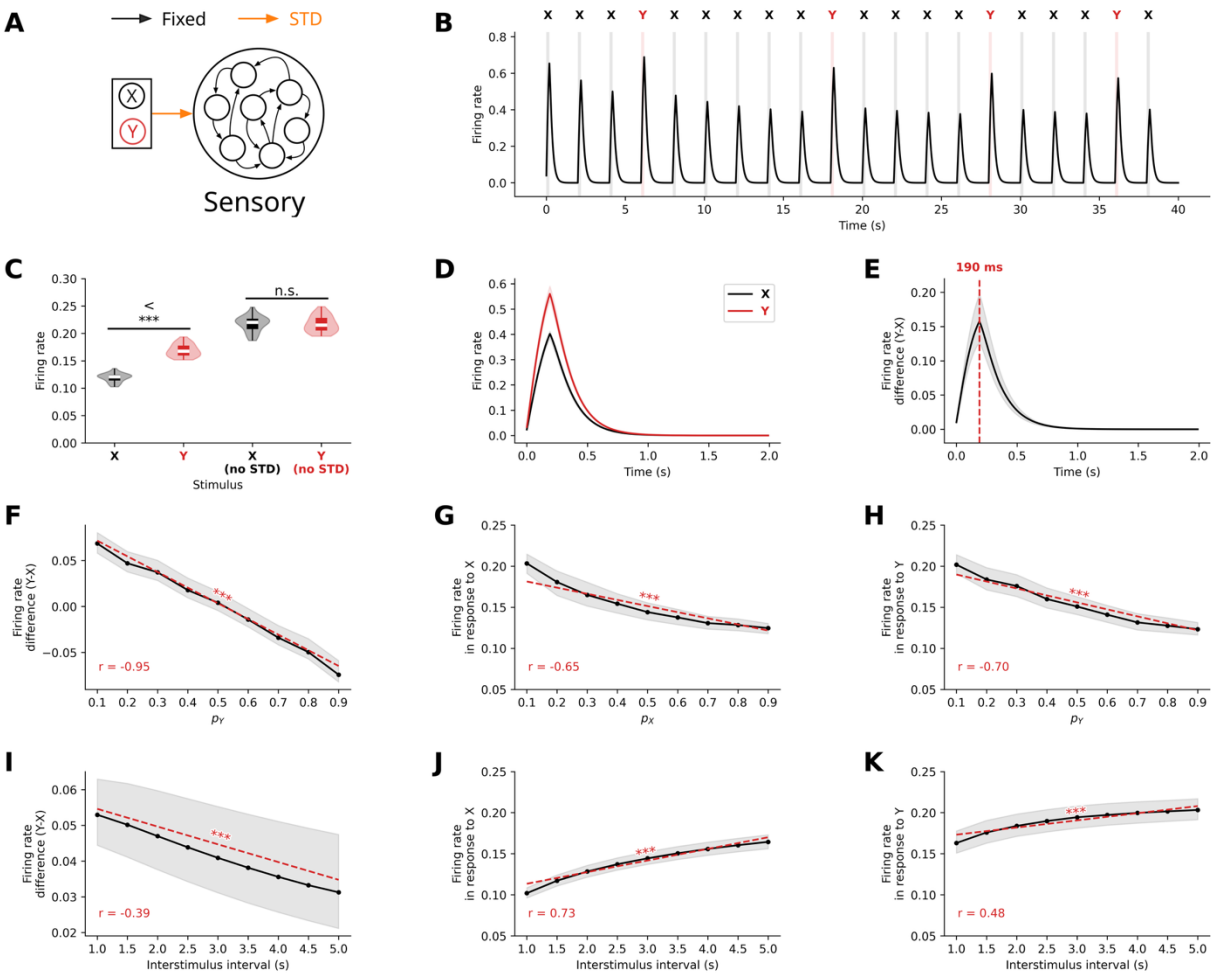


**Figure 1. Short-term synaptic depression (STD) accounts for enhanced sensory responses to deviant stimuli in a passive oddball paradigm. (A)** Schematic of the recurrent neural network model with STD modulating input connection strength. **(B)** Representative time series showing mean firing rate within the sensory network during oddball stimulus presentations. Gray bars indicate stimulus presentations with stimulus identity labeled above (X = standard, Y = deviant). **(C)** Comparison of mean firing rates for standard (X) and deviant (Y) stimuli with and without STD implementation. STD is essential for generating differential responses ($p < 0.001$). **(D)** Population-averaged firing rate responses to standard and deviant stimuli over time, showing enhanced responses to deviant stimuli. **(E)** Time course of the deviant-standard response difference, with peak occurring at 190ms post-stimulus onset (median, red dashed line). **(F)** Mean deviant-standard response difference as a function of deviant stimulus probability, showing inverse relationship. **(G-H)** Individual stimulus responses as functions of their respective probabilities, demonstrating that both stimulus types show probability-dependent response amplitudes. **(I)** Mean deviant-standard response difference as a function of interstimulus interval, showing decay with longer intervals **(J-K)** Individual stimulus responses as functions of interstimulus interval, showing differential increases to standard and deviant stimuli. Data represent median ± interquartile range.

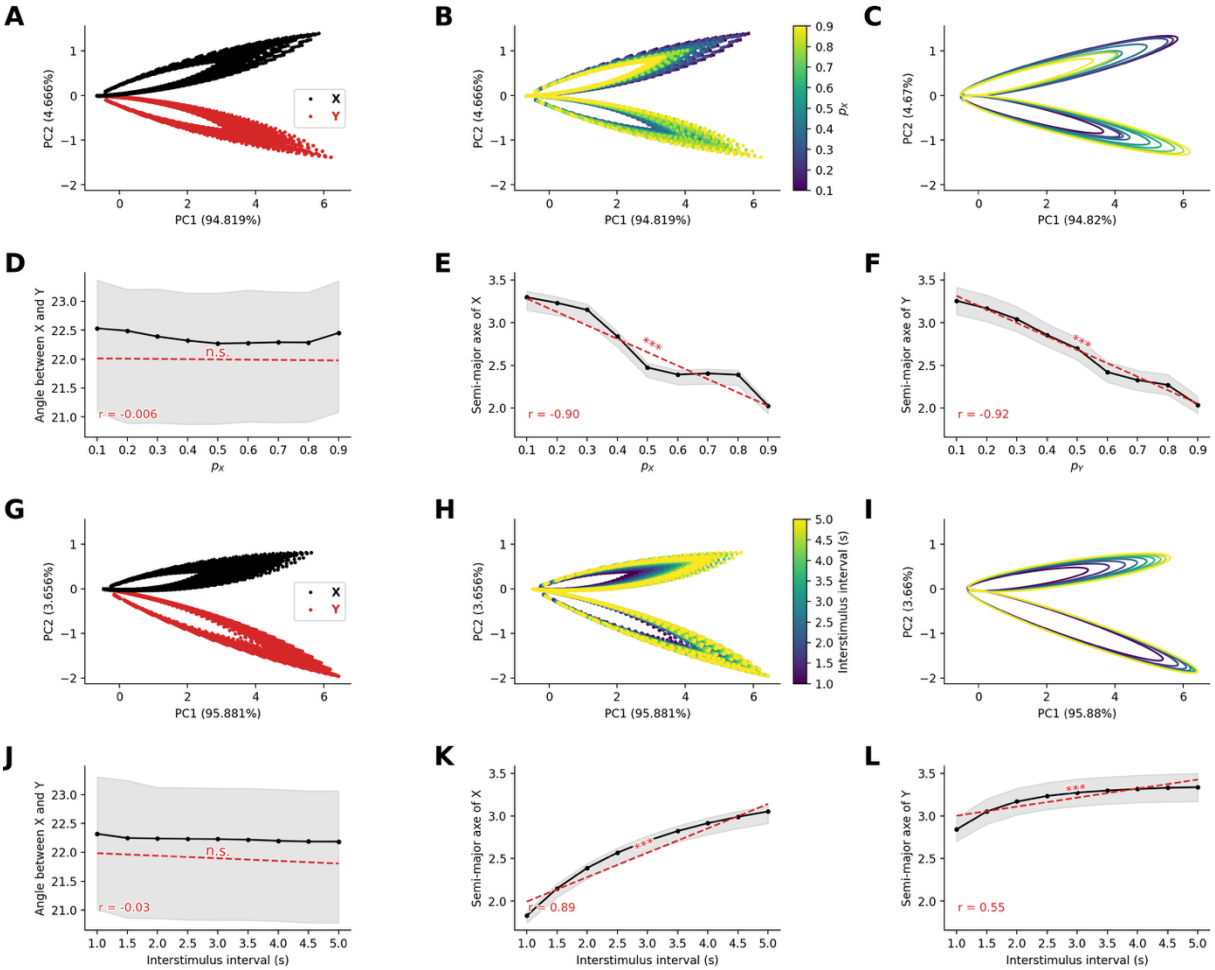


**Figure 2. Stable neural representation geometry emerges from short-term synaptic depression in sensory networks**. **(A)** Principal component analysis (PCA) of sensory network population responses to standard and deviant stimuli. Neural trajectories form distinct elliptical patterns in PC space, colored by stimulus identity (red = standards A, black = deviants B). **(B)** Principal component analysis of sensory network population responses to standard stimuli, colored by stimulus probability (color scale). **(C)** Fitted ellipses to stimulus representations show probability-dependent geometric properties. **(D)** Angular separation between stimulus ellipses remains constant across probabilities. **(E-F)** Semi-major axes of fitted ellipses decrease with increasing stimulus probability for standard and deviant stimuli, encoding frequency information in representational geometry. **(G-H)** PCA with varying interstimulus intervals at stimulus probabilities of $p_X = 0.8$ and $p_Y = 0.2$, showing similar elliptical trajectories colored by (G) stimulus identity and (H) interstimulus interval duration. **(I)** Fitted ellipses for different interstimulus intervals. **(J)** Angular relationships between standard and deviant remain stable across interstimulus intervals. **(K-L)** Semi-major axes increase with longer interstimulus intervals, reflecting enhanced neural differentiation with more recovery time between stimuli. Results demonstrate emergent encoding of stimulus frequency in population-level representational geometry.

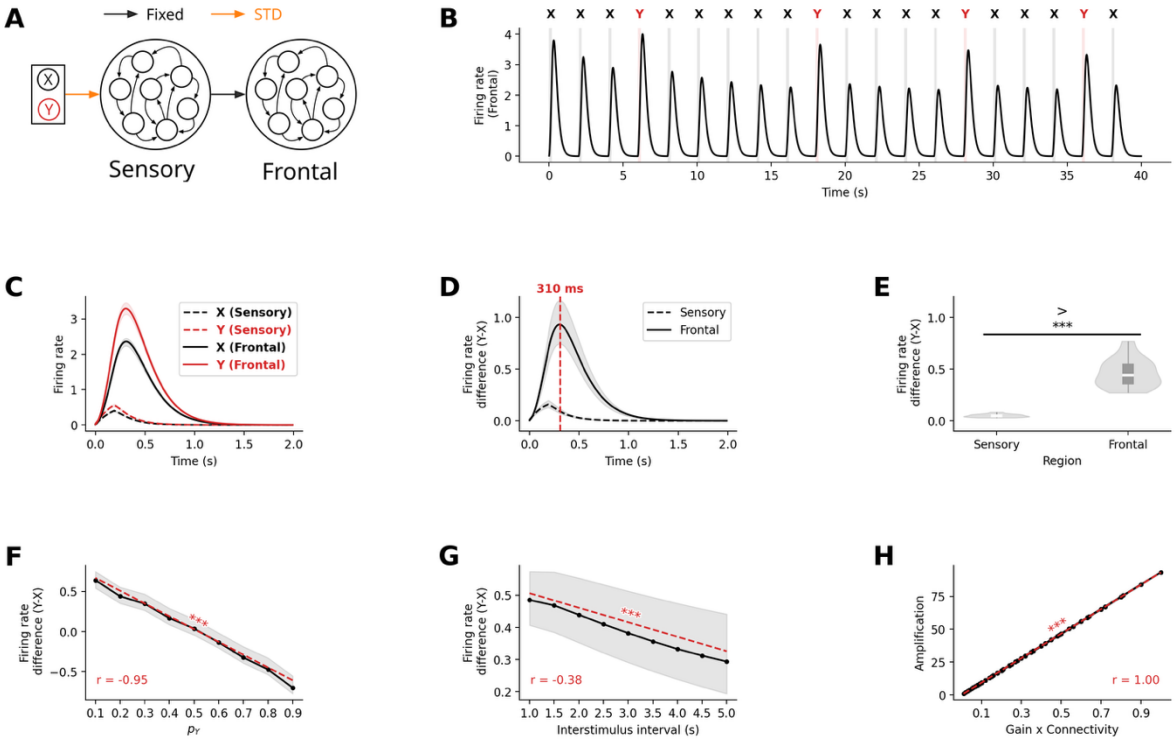


**Figure 3. Amplification of deviant-standard response differences from sensory to frontal cortex. (A)** Extended model architecture with separate sensory and frontal recurrent networks. Frontal network receives excitatory projections from sensory network but no direct stimulus input. **(B)** Mean firing rate within the frontal network during oddball stimulus presentation. **(C)** Direct comparison of firing rates between networks for standard and deviant stimuli, showing delayed, amplified, and prolonged responses compared to sensory responses. **(D)** Temporal evolution of deviant minus standard differences in both networks, with frontal peak delayed by ~120ms (310ms vs 190ms) and amplified ~9-fold. **(E)** Quantification of significant response amplification from sensory to frontal regions ($p << 10^{-3}$). **(F)** Frontal deviant-standard difference maintains inverse relationship with deviant probability ($p << 10^{-3}$). **(G)** Similar inverse relationship with interstimulus interval ($p << 10^{-3}$). **(H)** Sensory to frontal amplification ratio scales linearly with the product of inter-regional connection strength and probability ($p << 10^{-3}$), revealing the structural basis of functional amplification. This demonstrates that P300-like responses can emerge naturally from hierarchical connectivity without requiring additional mechanisms beyond STD.

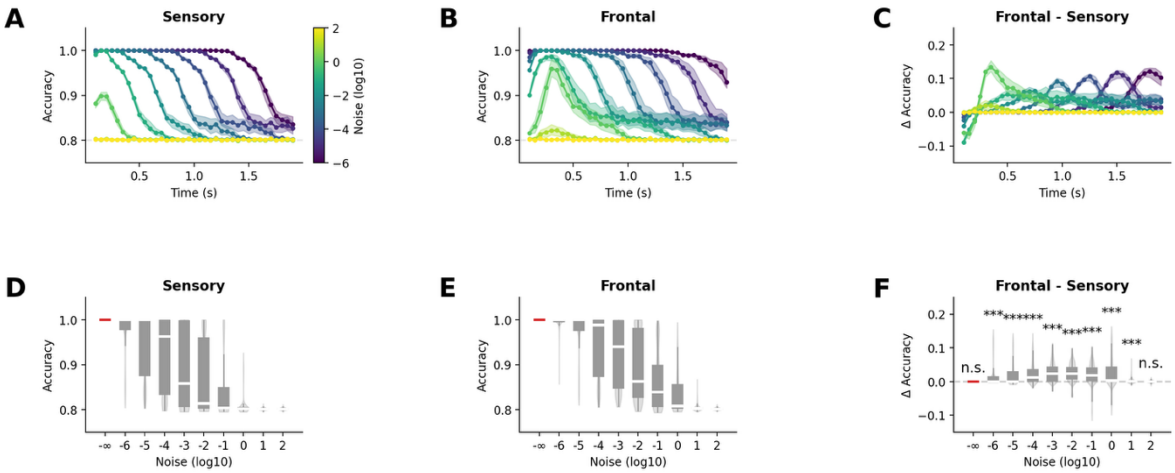


**Figure 4. Frontal representations show enhanced robustness to noise compared to sensory regions. (A-B)** Stimulus decoding accuracy within 200ms sliding windows as a function of post-simulation Gaussian noise amplitude for sensory (A) and frontal (B) networks. Color scale indicates noise standard deviation. **(C)** Difference in decoding accuracy between frontal and sensory networks. **(D-E)** Decoding accuracy across all 200ms windows in 2s trials for varying noise levels in sensory (D) and frontal (E) networks. **(F)** Sustained advantage of frontal over sensory decoding across trial duration, demonstrating functional benefits of response amplification. Red dashed lines indicate constant accuracy levels. Frontal amplification provides significant decoding advantages, supporting the functional role of P300 responses in robust information processing under noisy conditions.

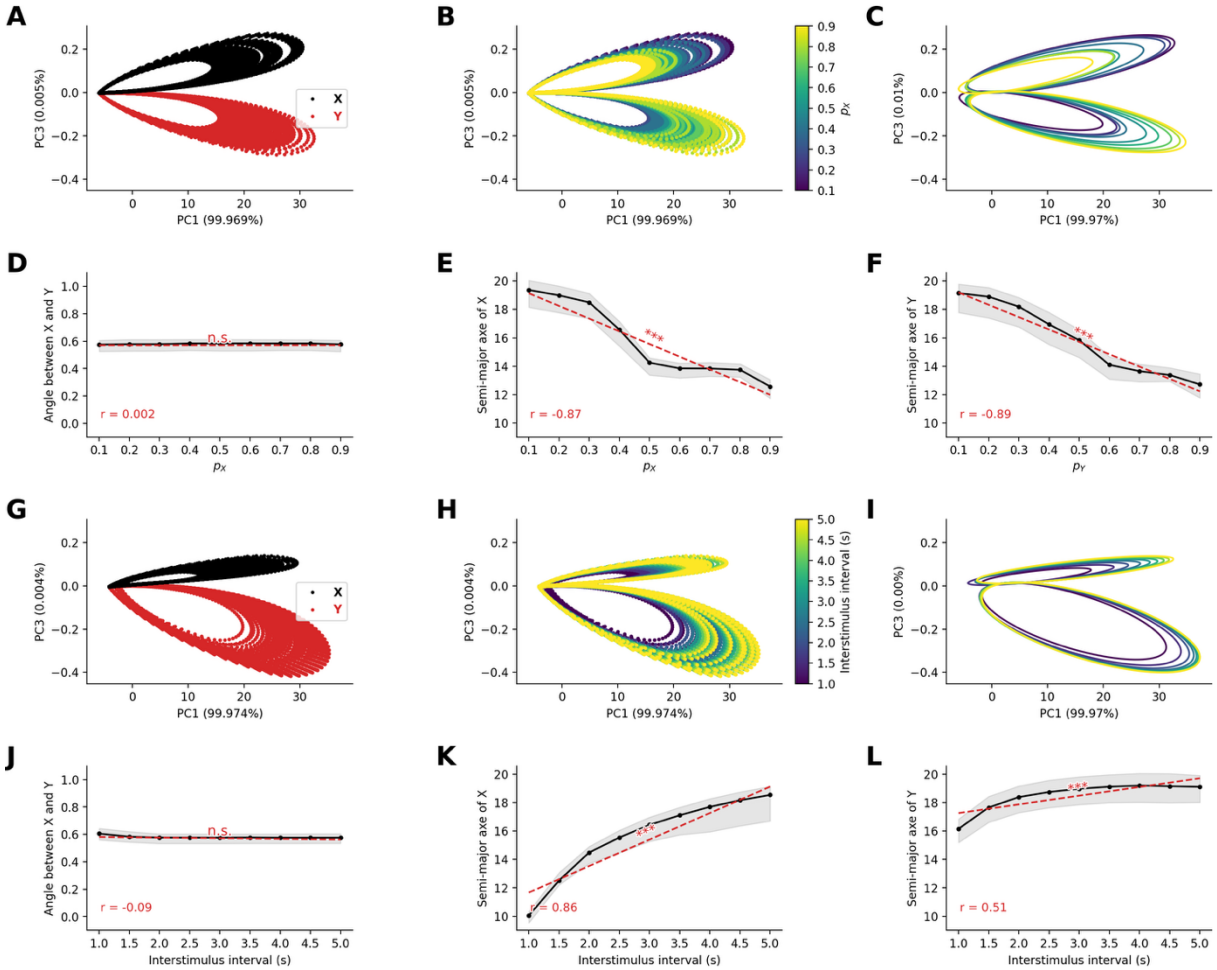


**Figure 5. Stable neural representation geometry in frontal networks. (A-B)** Principal component analysis of frontal network responses with varying stimulus probabilities, showing stability of elliptical representational structure from sensory regions across stimulus intensities (A) and stimulus probabilities (B). **(C)** Ellipse fits to frontal representations preserve probability-dependent geometry. **(D)** Angular relationships remain stable across varying stimulus probabilities. **(E-F)** Semi-major axes of standard (E) and deviant (F) stimuli maintain inverse relationship with probability, demonstrating preserved stimulus frequency encoding. **(G-H)** Analysis with varying interstimulus intervals shows similar preservation of representational structure. **(I)** Ellipse fits across interstimulus intervals. **(J)** Stable angular relationships across interstimulus intervals. **(K-L)** Preserved stimulus interval-dependent scaling of representational size. Results demonstrate that frequency-dependent geometric properties established in sensory regions through STD are maintained and amplified in frontal processing, providing the representational basis for probability-sensitive P300 responses.

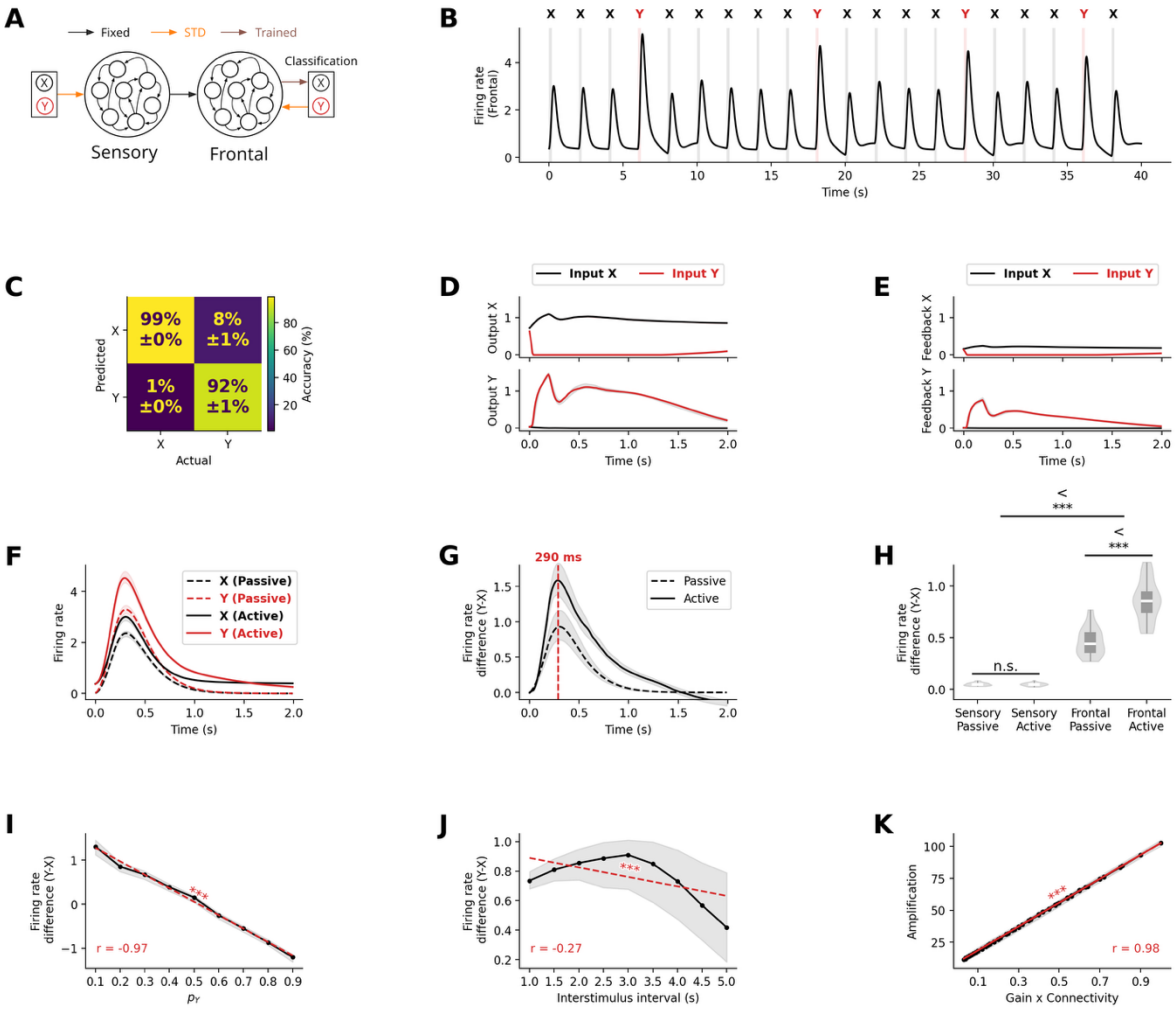


**Figure 6. Attentional modulation via feedback STD from decoder enhances frontal network neural response in active oddball paradigms. (A)** Extended model architecture for active oddball task, incorporating trained readout units with feedback connections subject to STD. **(B)** Frontal network activity during active paradigm shows similar temporal dynamics with additional feedback modulation. **(C)** Confusion matrix demonstrating high classification accuracy (99% for standard stimuli X, 92% for deviant stimuli Y). **(D)** Frontal population responses show differential activations to deviant and standard stimuli. **(E)** Effective feedback strength after STD, showing differential modulation of deviant and standard stimuli. **(F)** Comparison of neural responses in frontal network in active vs passive oddball conditions, demonstrating hierarchical amplification with additional feedback enhancement. **(G)** Deviant minus standard neural responses peak at ~300ms in frontal network, consistent with P3b timing. **(H)** Significant enhancement in frontal ($p << 10^{-3}$) but not sensory ($p = 0.64$) network during active compared to passive oddball conditions. **(I)** Deviant minus standard stimulus neural response with varying probability of deviant stimulus presentation. **(J)** Deviant minus standard stimulus neural response across different inter-stimulus intensities. **(K)** Firing rate response amplification ratio between the frontal and sensory network as a function of the product between

synaptic gain and connection probability between networks, showing that linear relationship between connectivity parameters and amplification is maintained during the active oddball task. Results demonstrate that the same STD mechanism explains both passive MMN/P300 responses and active P3b enhancement through differential feedback modulation.

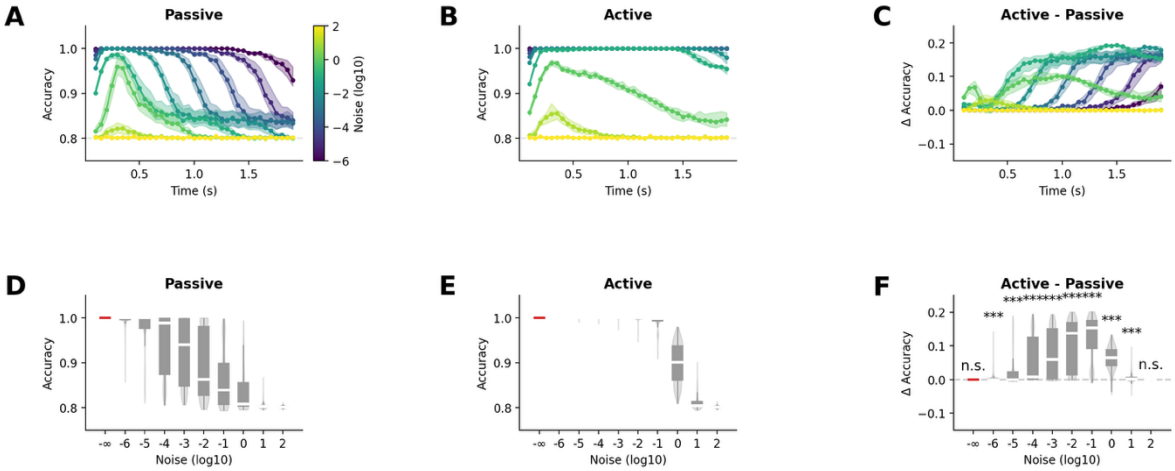


**Figure 7. Active oddball paradigm maintains enhanced noise robustness in frontal network. (A-B)** Decoding accuracy analysis comparing (A) passive and (B) active oddball conditions under varying noise levels. Active condition shows sustained high performance. **(C)** Temporal advantage of active over passive processing, demonstrating sustained functional benefits of amplification with feedback. **(D-E)** Full-trial decoding accuracy for (D) passive and (E) active conditions across noise levels. **(F)** Difference in decoding accuracy between active and passive deviant processing, showing significant improvement across wide range of noise conditions ($p << 10^{-3}$). Results confirm that attentional modulation through higher-order feedback connections preserves and enhances the functional advantages of frontal response amplification, supporting robust stimulus discrimination under noisy conditions essential for behavioral responses.

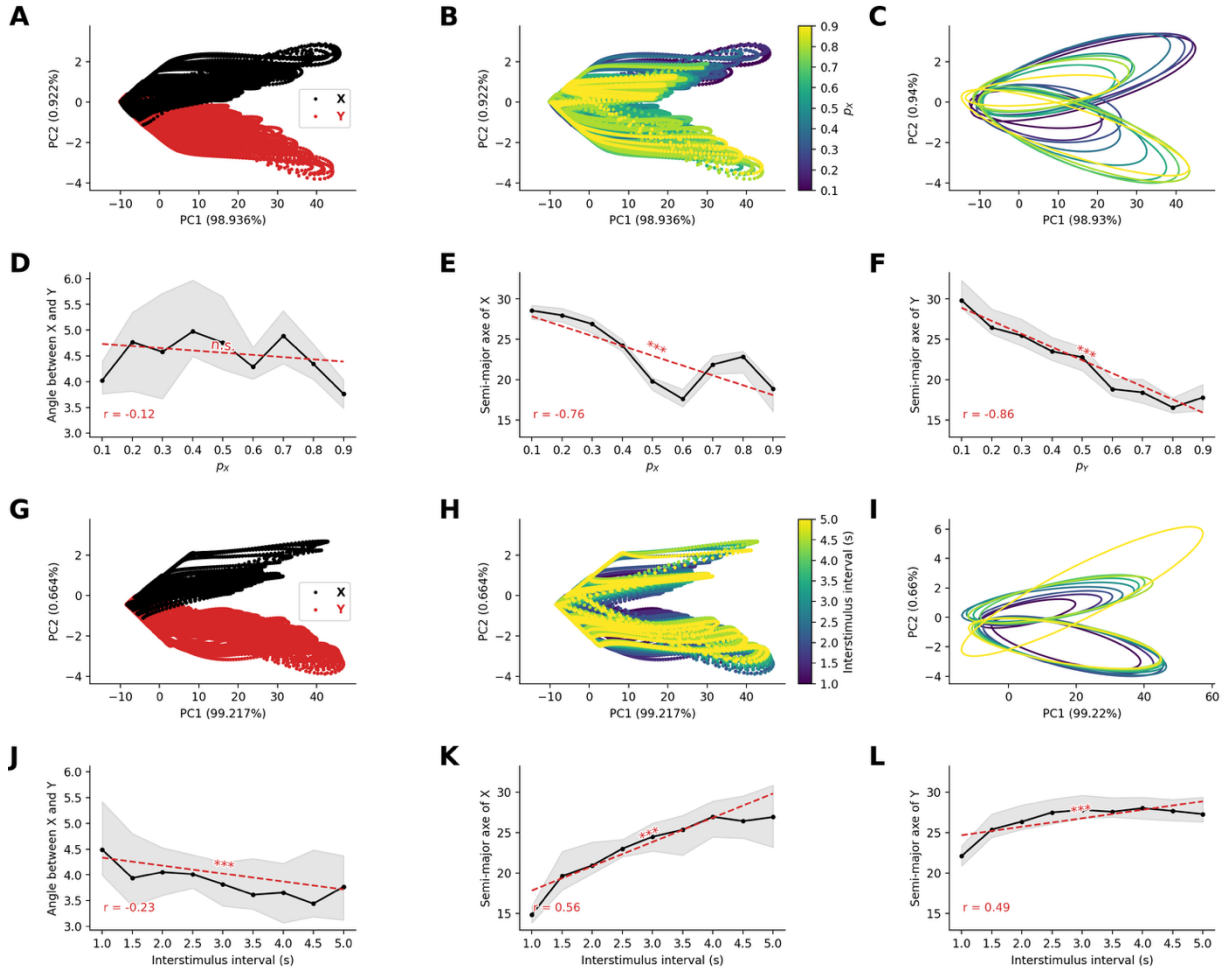


**Figure 8. Neural representation geometry is preserved under attentional modulation in active oddball paradigms. (A-B)** Principal component analysis of frontal responses in active oddball condition with varying stimulus probabilities. Feedback modulation distorts but preserves essential geometric structure, with (A) stimulus identity and (B) probability information maintained. **(C)** Ellipse fits to active condition representations show maintained probability dependencies. **(D)** Angular relationships show slight variation but remain relatively stable. **(E-F)** Semi-major axes preserve inverse relationship with probability. **(G-H)** Analysis with varying interstimulus intervals demonstrates maintained representational structure under attentional modulation. **(I)** Ellipse fits across intervals in active condition. **(J)** Angular relationships show increased sensitivity to interval duration ($p < 10^{-3}$). **(K-L)** Preserved scaling relationships between interstimulus intervals and representational geometry. Results demonstrate that attentional feedback preserves the fundamental frequency-encoding properties of neural representations and enhance stimulus-specific information processing, providing a mechanistic basis for the enhanced responses to behaviorally relevant stimuli.