Received: 29 April 2022 | Revised: 2 June 2022 | Accepted: 7 June 2022

DOI: 10.1111/ejn.15736

RESEARCH REPORT

EJN European Journal of Neuroscience FENS WILEY

Decoding violated sensory expectations from the auditory cortex of anaesthetised mice: Hierarchical recurrent neural network depicts separate 'danger' and 'safety' units

Jamie A. O'Reilly[1] | **Thanate Angsuwatanakul**[1] | **Jordan Wehrman**[2]

[1]College of Biomedical Engineering, Rangsit University, Lak Hok, Thailand

[2]Brain and Mind Centre, University of Sydney, Camperdown, New South Wales, Australia

Correspondence
Jordan Wehrman, Brain and Mind Centre, University of Sydney, Camperdown, New South Wales, Australia.
Email: jordan.wehrman@sydney.edu.au

Funding information
Research Institute of Rangsit University, Grant/Award Number: 90/2561; Engineering and Physical Sciences Research Council, Grant/Award Number: EP/F50036X/1

Edited by: John Foxe

Abstract

The ability to respond appropriately to sensory information received from the external environment is among the most fundamental capabilities of central nervous systems. In the auditory domain, processes underlying this behaviour are studied by measuring auditory-evoked electrophysiology during sequences of sounds with predetermined regularities. Identifying neural correlates of ensuing auditory novelty responses is supported by research in experimental animals. In the present study, we reanalysed epidural field potential recordings from the auditory cortex of anaesthetised mice during frequency and intensity oddball stimulation. Multivariate pattern analysis (MVPA) and hierarchical recurrent neural network (RNN) modelling were adopted to explore these data with greater resolution than previously considered using conventional methods. Time-wise and generalised temporal decoding MVPA approaches revealed previously underestimated asymmetry between responses to sound-level transitions in the intensity oddball paradigm, in contrast with tone frequency changes. After training, the cross-validated RNN model architecture with four hidden layers produced output waveforms in response to simulated auditory inputs that were strongly correlated with grand-average auditory-evoked potential waveforms ($r^2 > .9$). Units in hidden layers were classified based on their temporal response properties and characterised using principal component analysis and sample entropy. These demonstrated spontaneous alpha rhythms, sound onset and offset responses and putative 'safety' and 'danger' units activated by relatively inconspicuous and salient changes in auditory inputs, respectively.

Abbreviations: Adam, adaptive moment estimation; D, deviant stimuli; D1, increasing deviant stimulus; D2, decreasing deviant stimulus; EEG, electroencephalography; ERP, event-related potential; fD1, ascending frequency deviant stimulus condition; fD1D2, frequency oddball paradigm D1 vs. D2 decoding; fD2, descending frequency deviant stimulus condition; fSD, frequency oddball paradigm S vs. D decoding; fSD1, frequency oddball paradigm S vs. D1 decoding; fSD2, frequency oddball paradigm S vs. D2 decoding; iD1, rising intensity deviant stimulus condition; iD1D2, intensity oddball paradigm D1 vs. D2 decoding; iD2, falling intensity deviant stimulus condition; iSD, intensity oddball paradigm S vs. D decoding; iSD1, intensity oddball paradigm S vs. D1 decoding; iSD2, intensity oddball paradigm S vs. D2 decoding; MMN, mismatch negativity; MSE, mean squared error; MVPA, multivariate pattern analysis; NMDA, n-methyl-d-aspartate; P3a, involuntary P300 component; PCA, principal component analysis; RNN, recurrent neural network; RON, reorienting negativity; S, standard; SPL, sound pressure level; STFT, short time Fourier transform; VM, support vector machine; σ, standard deviation.

The hypothesised existence of corresponding biological neural sources is naturally derived from this model. If proven, this could have significant implications for prevailing theories of auditory processing.

KEYWORDS

auditory novelty, emergent computational physiology, mismatch negativity, multivariate pattern analysis, recurrent neural network, sensory processing

1 | INTRODUCTION

The survival of humans and animals relies on their ability to detect and respond appropriately to the environment. This is a fundamental behaviour shared by many species. When residing in a dark forest, for example, the mammalian central auditory system automatically processes salient changes in the acoustic environment and redirects the organism's attention towards any potential sources of danger or reward. These processes can be studied by recording electrophysiological responses to expected and unexpected sounds in a passive auditory oddball sequence. In humans, this is typically performed while recording electroencephalography (EEG), after which averaging responses measured over multiple presentations of the same stimulus condition produces a stereotypical pattern of components in the event-related potential (ERP). These auditory novelty responses are interpreted to reflect surprise or prediction-error signalling and include mismatch negativity (MMN), P3a and reorienting negativity (RON), which are distinct from obligatory components of the auditory-evoked response that are elicited by all perceptible stimuli, regardless of their context (e.g., P50 or N1).

One of the most widely studied ERP components related to surprise is the MMN (Näätänen et al., 1978). This component occurs between 100 and 200 ms after the onset of a surprising stimulus at fronto-central electrodes (Näätänen et al., 2007) and is calculated from the difference between the responses to expected 'standard' and unexpected 'deviant' stimuli. The larger the MMN, the greater the level of surprise (De-Wit et al., 2010; Friston & Kiebel, 2009; Garrido et al., 2009; Heilbron & Chait, 2018; Mathys et al., 2011). Supporting this, MMN has been found to be relatively diminished under several conditions in which sensory predictions are thought to be compromised, for example, in schizophrenia (Umbricht & Krljes, 2005), Parkinson's disease (Brønnick et al., 2010) and when subjects are administered ketamine (Rosburg & Kreitschmann-Andermahr, 2016). It has also been observed in unresponsive patients, and there is evidence supporting the effectiveness of both MMN and P3a in evaluating conserved physiological function in unconscious humans (Morlet & Fischer, 2014).

The P3a is a positive amplitude response to unexpected, environmentally salient stimuli that peaks approximately 250 to 300 ms after stimulus presentation (Friedman et al., 2001; Polich, 2007). While MMN is an early index of surprise, P3a is thought to reflect the attention directed towards the surprising stimulus (Friedman et al., 2001). A loud growling noise in the forest, for example, would likely draw our attention, resulting in a larger P3a. However, if the growling noise is quickly determined to be that of a smaller, non-threatening source, our attention would return to whichever task was at hand prior to the noise disturbance. This reorientation to prior task elicits a much later negative ERP component, the RON. As per the P3a and MMN, this response occurs at fronto-central electrodes, though much later, around 400- to 600-ms post-stimulus (Otten et al., 2000; Schröger & Wolff, 1998). Similar to the MMN, both the P3a and RON are affected by neurophysiological deficits, for example, in schizophrenia patients (Higuchi et al., 2014). Further, it is worth noting that, while related, each of these three components may be separated experimentally (Horváth et al., 2008).

Neural activity associated with these processes is examined in animal models on the basis that their auditory systems are homologous to those in humans. It remains challenging, however, to directly associate signals measured from animals with equivalent components observed in human ERP waveforms, due to anatomical and physiological differences. Ambiguity thereby arises concerning the latency of obligatory and context-dependent components of the auditory response in animals, which may be problematic to dissociate (Parras et al., 2017; Taaseh et al., 2011). Some data from anaesthetised mice suggest that earlier obligatory components are predominantly influenced by adaptation and the physical properties of stimuli (O'Reilly & Conway, 2021), whereas later, context-dependent components are more closely associated with violation of a sensory regularity (Casado-Román et al., 2020; Chen et al., 2015; Kurkela et al., 2018; O'Reilly, 2019a; O'Reilly & Angsuwatanakul, 2021). It has also been shown that this

late component in anaesthetised mice relies on intact NMDA receptor function (Chen et al., 2015), which is also the case for human MMN (Avissar & Javitt, 2018; Rosburg & Kreitschmann-Andermahr, 2016). Curiously, obligatory components of the auditory response in mice generally occur earlier than those recorded in humans, consistent with the theory that sensory response latencies are determined by anatomical size and complexity (Itoh et al., 2022; Javitt et al., 1992; Komatsu et al., 2015), whereas components associated with auditory novelty occur at comparatively similar latencies, apparently contradicting this theory (Chen et al., 2015; O'Reilly, 2019a; O'Reilly & Angsuwatanakul, 2021). As such, the relationships between physiology and cortical auditory-evoked potential components in different species remain an unsolved puzzle.

Auditory novelty responses have been studied extensively in basic, clinical and preclinical animal investigations, although despite this there remains substantial debate concerning their underlying neurophysiology (May, 2021; May & Tiitinen, 2010; Näätänen et al., 2005; O'Reilly & O'Reilly, 2021). This stems from confounds or alternative explanations for differences between responses to sequences of physically different stimuli that have been challenging to completely dissociate, such as adaptation (May, 2021; May & Tiitinen, 2010) or the inherent modulation of overlapping ERP components by the physical properties of sounds (O'Reilly & Conway, 2021; O'Reilly & O'Reilly, 2021; Takegata et al., 2008; Todd et al., 2008). In the present study, we revisit data recorded from anaesthetised mice to ask how cortical auditory reflexes in this preparation encode violations of sensory expectations during passive frequency and intensity oddball paradigms. To support this, multivariate pattern analysis (MVPA) and hierarchical recurrent neural network (RNN) modelling are adopted to characterise cortical signal dynamics associated with auditory novelty responses. The MVPA approach was selected to explore subtle effects of stimulus conditions that may be missed by conventional methods (Bae et al., 2020), whereas modelling with a hierarchical RNN provides a tool for studying potential computational principles that underlie the generation of cortical auditory-evoked responses (Barak, 2017; Barrett et al., 2019; Yang & Molano-Mazón, 2021).

2 | MATERIALS AND METHODS

2.1 | Data

The data used in this study have been described elsewhere (O'Reilly, 2019a). These experiments were approved by the Animal Welfare and Ethical Review Body, University of Strathclyde, and performed in accordance with the UK Animals (Scientific Procedures) Act 1986.

To summarise, 14 urethane-anaesthetised mice were presented with frequency and intensity oddball sequences while epidural field potentials were recorded bilaterally from electrodes positioned above their primary auditory cortices (as shown in fig. 1 of O'Reilly, 2019b). In both of the oddball paradigms, standard stimuli were 100 ms, 10 kHz, 80 dB, monophonic pure-tones. Decibel units throughout this article refer to sound pressure level (SPL). In the frequency oddball paradigm, deviant stimuli deviated by ±2.5 kHz, and in the intensity oddball paradigm they deviated by ±10 dB, above and below the standard. These were played with a constant offset to onset inter-stimulus interval of 450 ms. Each sequence included 800 standards (S), 100 increasing deviants (D1) and 100 decreasing deviants (D2). To balance the numbers of trials, standards preceding deviants were extracted, reducing the number of standard trials to 200 for each animal.

Signals were recorded at a frequency of 1000 Hz and subsequently band-pass filtered between .1 and 30 Hz. Trials containing two consecutive stimulus responses were extracted, that is, either responses to two repeated standard stimuli or responses to an unexpected deviant followed by a standard stimulus, as described previously (O'Reilly, 2019a; O'Reilly & Angsuwatanakul, 2021). These segments spanned from .1 s before to 1 s after the first stimulus, capturing two consecutive auditory-evoked responses. Baseline correction was applied by subtracting the average amplitude measured within the .1-s pre-first-stimulus period from the whole trial. Signals were then resampled to 100 Hz for MVPA and modelling.

2.2 | ERP decoding

2.2.1 | Time-wise

To perform MVPA on ERPs, also referred to as ERP decoding (Grootswagers et al., 2017), data from each animal were analysed separately before averaging the results. A linear support vector machine (SVM) was trained to perform binary classifications between ERPs at each time-point, down-sampled to 100 Hz. Four different binary classifications were assessed: standards or deviants (S vs. D; fSD or iSD), standards or increasing-deviants (S vs. D1; fSD1 or iSD1), standards or decreasing-deviants (S vs. D2; fSD2 or iSD2) and increasing-deviants or decreasing-deviants (D1 vs. D2; fD1D2 or iD1D2).

The number of trials presented were equalised (e.g., in the case of S vs. D2, S was under-sampled randomly in order that both conditions had equal trial numbers), and 10-fold cross-validation was performed. This cross-validation was then repeated five times, each time the under-sampling and fold allocation were performed again. The results across all these were then averaged to get an accuracy measure of the decoding at each time point. The MVPA was performed using MVPA-Light (Treder, 2020) and Fieldtrip (Oostenveld et al., 2011). After calculating the time-wise decoding accuracy for each animal, these data were evaluated using *t* test statistics, with control for multiple comparisons using cluster-based permutation tests, with the threshold for statistical significance conventionally set to $\alpha = .05$.

2.2.2 | Temporal generalisation method

To evaluate potential correlations and interactions between neural responses occurring at different latencies, time-generalised decoding was performed (King & Dehaene, 2014). The same eight binary classifications (fSD, fSD1, fSD2, fD1D2, iSD, iSD1, iSD2 and iD1D2) were performed separately using a linear SVM, although in the temporal generalisation method the model was trained on data from one time point and then tested across all time points. This produced two-dimensional matrices of decoding accuracy by training and test time for each animal. Otherwise, analysis was as per above.

2.3 | Modelling

2.3.1 | Inputs

Sound waveforms comparable with those presented during the in vivo experiment were produced with a sampling frequency of 100 kHz. These included standard (S), ascending frequency deviant (fD1), descending frequency deviant (fD2), louder intensity deviant (iD1) and quieter intensity deviant (iD2) stimulus conditions. To simulate sound intensity, amplitudes were normalised to 80 dB (i.e., 80-dB sounds had a peak amplitude of 1). These were converted into the time-frequency domain using the short-time Fourier transform (STFT), producing a representation of cortical input from the ascending auditory pathway (Rahman et al., 2020). The STFT was performed on Hann-windowed segments of 200 samples with an overlap of 100 samples. This evaluated frequencies from 0 to 50 kHz, approximating the spectral hearing range of mice, with linear spacing of .5 kHz. Magnitudes of the complex STFT output for each trial type were down-sampled to 100 Hz and provided to the model as input features.

Target outputs for the model were obtained by averaging the trials from all of the animals to produce a single 'idealised experiment', depicting the most salient electrophysiological features with reduced noise and variability. Left and right auditory cortex channels were also averaged together. This process generated 200 standard, 100 increasing-deviant, and 100 decreasing-deviant trials from frequency and intensity oddball paradigms, equalling 800 trials in total. Model inputs and target outputs were both formatted as sequences of 111 time-samples for training the model in a supervised learning paradigm.

2.3.2 | Model architecture and training

A hierarchical RNN was selected to model auditory-evoked potential data. This may be interpreted as a firing rate model that converges towards one of the possible solutions to the inverse source problem. It had 101 input units, each associated with a frequency component of the STFT output. These input units connected to the first of four hidden layers of 64 recurrent units, whose outputs are determined by inputs from the current time-step and feedback from their outputs in the previous time-step. Recurrent connections allow RNNs to learn from sequences of inputs, and are loosely analogous to feedback connections in biological neural networks. The output layer consisted of a single recurrent unit that produced signals in response to simulated auditory inputs.

Input features were fed through the model and its parameters (weights and biases) were optimised to minimise mean-squared-error (MSE) loss between model outputs and target evoked responses. Adaptive moment estimation (Adam) optimization was used with a learning rate of .001, beta-1 equal to .9 and beta-2 equal to .99. Connection weights between layers were initialised from a Glorot uniform distribution (Glorot & Bengio, 2010) and recurrent weights were initialised as an orthogonal matrix from a normal distribution (Saxe et al., 2013). Generalisation of this model architecture was verified by cross-validating over 10-folds, each with 10% of trials of each stimulus type held back for evaluation. This resulted in average terminal MSE for the training set of 114.2 (1.48 SD) and for the validation set of 109.1 (9.71 SD), demonstrating adequate ability of the model architecture to generalise to withheld data. Five identical models were then trained for 500 epochs and evaluated in terms of MSE and Pearson's correlation coefficient (r^2) between model outputs in response to each stimulus condition and their associated grand-average ERP waveforms. The

best performing model in terms of minimising MSE and maximising r^2 was subsequently analysed and used in simulated experiments.

2.3.3 | Hidden unit categorisation

After establishing the best performing model, its units were categorised based on their time-domain response properties. Units with peak activation latency before stimulus onset, or with activations exceeding three standard deviations above the mean during the pre-stimulus period and having a post-stimulus peak activity less than half of the pre-stimulus peak activity, were categorised as 'alpha' units, as these were found to display periodic activity within the alpha frequency range. Units with peak activity during stimulus-on times were categorised as 'onset' units, and those with peak activity up to 50 ms after stimulus-off times were categorised as 'offset' units. Units with activity peaking between 50 and 150 ms after stimulus-off times were tentatively categorised as 'safety' units, while those that peaked between 150 and 450 ms after the first stimulus were categorised as 'danger' units. This nomenclature was derived from an interpretation of the behavioural significance of environmentally inconspicuous and salient stimuli, respectively signalling safety or potential danger, which evoked stereotyped patterns of responses from specific hidden units during these latency ranges. Units that consistently had zero activation across all stimulus conditions were classed as 'zero' units. Any remaining units that did not fall within these categories were labelled as 'other' units, of which there was only one. This time-domain categorisation procedure was applied to all of the hidden units in response to five stimulus conditions (S, fD1, fD2, iD1, iD2) and the most frequent categorisation across stimulus conditions was selected for each unit.

2.3.4 | Principal component analysis

Principal component analysis (PCA) provides a dimensionality reduction technique that enables analysis and visualisation of a multidimensional dataset in two dimensions while preserving the largest portions of variance from the original dataset. Hidden unit activations were transformed into principal component space to examine their principal modes of variance, providing a means of the analysing model, layer and unit responses. Three different aspects of the data were explored using PCA. (1) Model responses to five stimulus conditions, a 5-by-28,416 (4 layers, 64 units, 111 time-samples) matrix, were transformed into two principal components that explained 84% of the variance in the data. (2) Layer responses, a $4 \times 35{,}520$ (5 stimulus conditions, 64 units, 111 time-samples) matrix, were transformed into two components that accounted for 97% of the variance in the data. (3) Units in each layer, four 64-by-555 (5 stimulus conditions, 111 time-samples) matrices, were each transformed into two components that accounted for 68%, 41%, 55% and 39% of the variance in layer 1, 2, 3 and 4, respectively. Only two principal components were selected for effective visualisation.

2.3.5 | Entropy analysis

Sample entropy was calculated from model hidden unit activation signals in response to different stimuli using the method proposed by (Richman & Moorman, 2000), with subsequence length, $m = 2$, and tolerance, $r = .15$. Higher sample entropy is interpreted to reflect greater levels of information production in a dynamic system, while lower sample entropy reflects the opposite. This is another perspective on the conventional view of entropy as being proportional to disorganisation, given that disorganised (changing) signals potentially contain more information than completely organised (unchanging) signals. Entropy measures have been used successfully to describe biological neural signals (Angsuwatanakul et al., 2020; Phukhachee et al., 2019); thus, this technique is appropriate for characterising the behaviour of artificial neural signals generated by the model.

2.3.6 | Simulated experiments

Audio waveforms of stimuli varying in duration, frequency, and intensity were generated for use in simulated experiments. Different duration stimuli (100, 200, 300, 400 and 500 ms) were all 10 kHz, 80-dB pure-tones; frequency stimuli (5, 7.5, 10, 12.5 and 15 kHz) were 100-ms duration and 80-dB intensity; and intensity stimuli (60, 70, 80, 90 and 100 dB) were 100 ms, 10-kHz pure-tones. Each simulated tone was converted into a time-frequency domain representation, as described above, and applied as model input to examine whether the resulting output would meet logical expectations based on neurophysiological findings reported in the literature.

2.3.7 | Software

Python 3, MNE 0.23.4, Scikit-learn 0.22.2 and Tensorflow 2.4.1 were used for data processing, MVPA and RNN

modelling. The Matlab toolboxes Fieldtrip and MVPA-Light were also used to perform statistical analyses.

3 | RESULTS

3.1 | Tone frequency changes and rising sound-level transitions trigger positive long-latency response

Decoding of stimulus conditions from ERP waveforms evoked during frequency and intensity oddball paradigms is shown in Figure 1. Grand-average ERP waveforms evoked by frequency oddball paradigm stimuli are plotted in Figure 1a, and those from the intensity oddball sequence are plotted in Figure 1b. The obligatory auditory response is characterised by a negative stimulus-onset peak and a positive stimulus-offset peak, both of which are influenced by the physical properties of eliciting stimuli and are evoked irrespective of stimulus context (e.g., whether expected or unexpected). Context-dependent, long-latency components of the ERP waveform are also apparent. Both frequency and increasing intensity deviant stimuli induce a positive amplitude feature that peaks at approximately .3 to .5 s. This is followed by a negative-going feature that coincides with presentation of the following standard stimuli at .55 s through to approximately .8 s. Responses to standard stimuli that follow quieter deviant stimuli, shown in Figure 1b, also appear to exhibit a positive-going

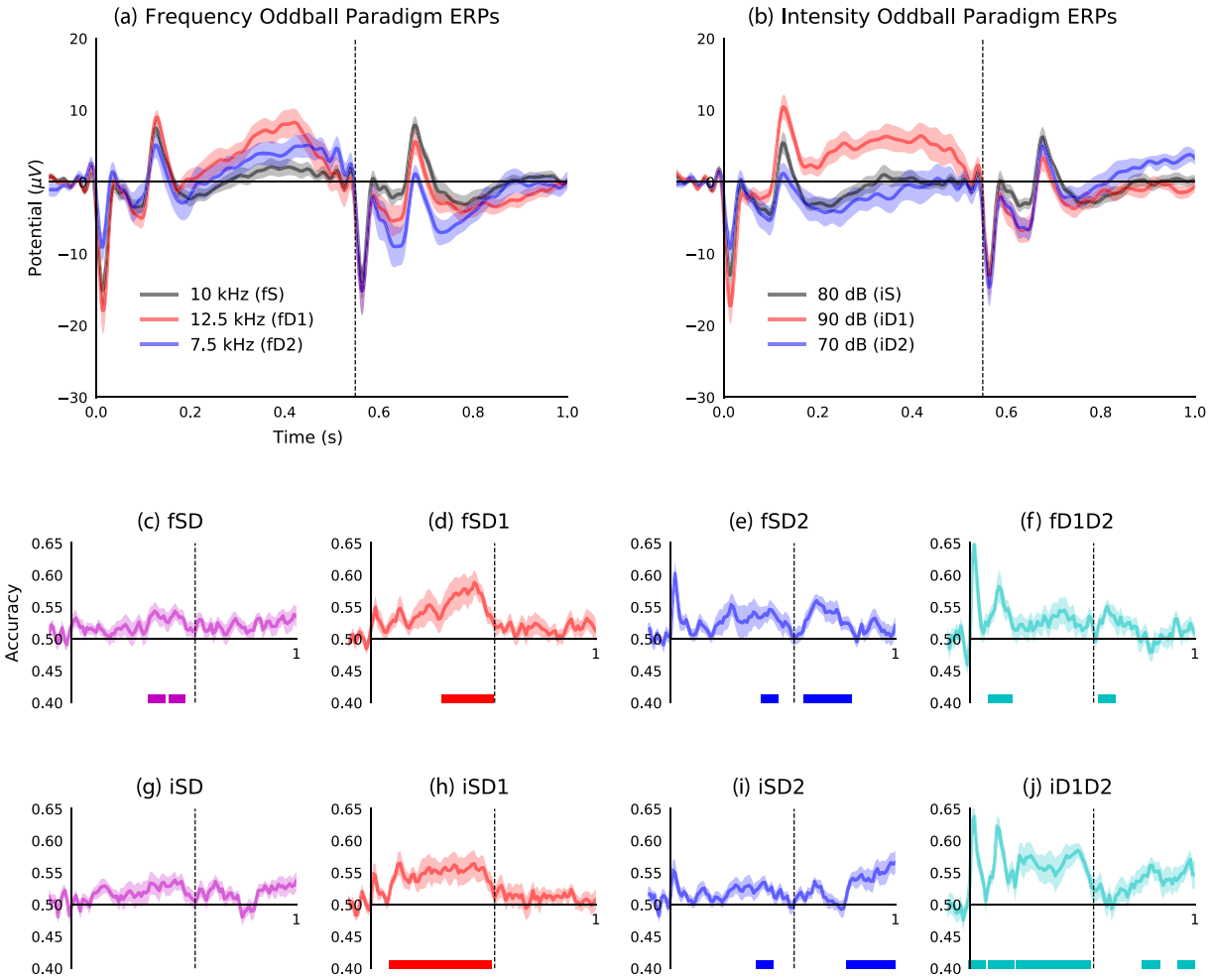

FIGURE 1 Tone frequency changes and rising sound-level transitions trigger positive long-latency response. (a) Grand-average event-related potential (ERP) waveforms from frequency oddball paradigm stimuli. (b) Grand-average ERP waveforms from intensity oddball paradigm stimuli. (c–f) Decoding accuracy for frequency oddball paradigm stimuli: Standard versus deviant (fSD), versus ascending deviant (fSD1), versus descending deviant (fSD2) and ascending versus descending deviant (fD1D2). (g–j) Decoding accuracy for intensity oddball paradigm stimuli: Standard versus deviant (iSD), versus louder deviant (iSD1), versus quieter deviant (iSD2) and louder versus quieter deviant (iD1D2). Dashed vertical lines at .55 s indicate onset of the second (standard) stimuli. Coloured bars at the bottom of each plot represent statistically significant decoding accuracy ($p < \alpha$; corrected for multiple comparisons).

trajectory, resembling the long-latency response evoked by frequency or louder deviant stimuli, which was not recognised in previous analysis of these data. Statistics related to this qualitative description can be found in (O'Reilly, 2019a).

Accuracy of decoding stimulus conditions from responses to frequency oddball paradigm stimuli reinforces observations from grand-average ERP waveform morphologies. Responses to standard stimuli compared with those to both frequency deviants (fSD; Figure 1c) demonstrate significant decoding accuracy over .35 to .41 s ($p = .031$) and .44 to .50 s ($p = .026$), aligning with the positive long-latency component. This corresponds to overlapping latencies of significant decoding accuracy in response to ascending (fSD1: .32 to .54 s, $p = .001$; Figure 1d) and descending frequency deviants (fSD2: .41 to .47 s, $p = .037$; Figure 1e); although the subsequent negative amplitude feature in descending frequency deviant trials also produced significant decoding accuracy from .60 to .80 s ($p = .002$). Despite their somewhat similarity trajectories, ascending and descending frequency deviant responses are decoded with significant accuracy (fD1D2; Figure 1f) extending over the stimulus-offset response peak latency, from .09 to .18 s ($p = .003$) and again from .58 to .64 s ($p = .014$).

Comparable MVPA applied to stimulus conditions in the intensity oddball sequence reveals a more complex pattern of responses. Starting with the two deviants (iD1D2; Figure 1j), regions of significant decoding accuracy occur during stimulus-onset (0 to .06 s, $p = .023$) and stimulus-offset (.09 to .19 s, $p = .005$) response peak latencies, the positive portion of the long-latency response evoked by louder deviants (.21 to .37 s, $p = .004$; .39 to .53 s; $p = .001$) and the late positive amplitude response evoked by standards following quieter deviants (.77 to .84 s; $p = .022$; .93 to 1.0 s, $p = .025$). Relative to the standard, louder deviant stimuli produced significant decoding accuracy between stimulus-offset through to presentation of the second standard stimulus (iSD1: .09 to .53 s, $p < .001$; Figure 1h). In contrast, classifying standard versus quieter deviant stimuli (iSD2; Figure 1i) produced a region of significant decoding accuracy from .39 to .45 s ($p = .046$) and later after the return to the standard tone from .79 to 1.0 s ($p < .001$). It is noteworthy that grand-average ERP waveforms (Figure 1b) indicate that this late response is caused by the 80-dB standard tone that follows the 70-dB deviant. These distinctions can help to explain the absence of significant decoding accuracy between standard and combined deviant conditions in the intensity oddball paradigm (iSD; Figure 1g).

In comparisons between standard and deviant (Figures 1d, 1e, 1h and 1i) and both frequency deviant conditions (Figure 1f), discrimination accuracies during initial onset peaks were not found to be statistically significant. This does not suggest that there were no differences between corresponding ERP amplitudes during these latencies, as onset and offset responses both appear to be modulated by tone frequency and intensity. Rather, these only register as statistically significant when comparing the responses to louder and quieter deviant tones (Figure 1j), for which these differences are most pronounced. The reason why this decoding accuracy does not compute as significantly above chance is likely due to its transience; cluster-based statistics add together sequential time points to reach a cluster statistic, and because the time period of the uptick in decoding accuracy is so brief, the cluster statistic does not reach the significance threshold.

3.2 | Intensity oddball stimuli influence the auditory response asymmetrically

Generalised decoding analysis is presented in Figure 2. Performance of classifying stimulus conditions based on responses evoked by frequency oddball paradigm stimuli (Figure 2a–d) largely reinforce findings from ERP decoding analysis presented in Figure 1c–h, with regions of significant decoding accuracy correlating with long-latency features of the ERP. The seemingly biphasic nature of these long-latency components is highlighted by opposing changes in decoding accuracy when trained and tested over time-points within the ranges of .3 to .5 s and .6 to .8 s. For example, when samples from the .3- to .5-s range are used for training, this increases decoding accuracy above chance for testing over the same interval, while simultaneously driving decoding accuracy below chance over the .6- to .8-s window, and vice versa. This reflects the positive amplitude shift evoked by frequency-deviant stimuli and the negative amplitude shift that occurs following the subsequent standard stimulus. Generally, significant decoding comparing standard to all deviant stimuli occurred from .37 to .53 s in training times, and .31 to 57 s in the test times ($p = .024$), similar to in the standard versus higher frequency deviance occurring from .35 to .53 s for the training times and .31 to .60 s for the test times ($p = .022$). Decoding differentiating the standards to the lower frequency tones was somewhat later, with significant training times occurring from .64 to .80 s and testing times from .60 to .82 s ($p = .012$).

Generalised decoding of stimulus condition from responses to intensity oddball paradigm stimuli (Figure 2e–h) is complicated by asymmetries between

louder and quieter deviant sounds, and their influence on the sound-level transition between the first and second stimuli in each tone-pair. These complications are emphasised by statistically significant off-diagonal effects when decoding between standard and deviant stimuli (iSD: training = .82 to 1.0 s; .80 to .21 s; .25 to .59 s, testing = .27 to .57 s; .28 to .54 s; .82 to 1.0 s; p = .033, .046, .048, respectively; Figure 2e). Note that these off-centre effects are particularly notable as there was no significant decoding in the standard time decoding of these stimuli. Regions of significant decoding accuracy between standard and louder intensity deviants (iSD1; training = .40 to .54 s, testing = .06 to .57 s, p = .002, Figure 2f), and standard and quieter deviant (iSD2; training = .05 to .54 s, .78 to 1.0 s, testing = .23 to .71 s, .75 to 1.0 s; p = .012, .038, respectively; Figure 2g) in time-points covering the range of approximately .05 to .5 s are thought to result from deviant stimuli evoking amplitude shifts in

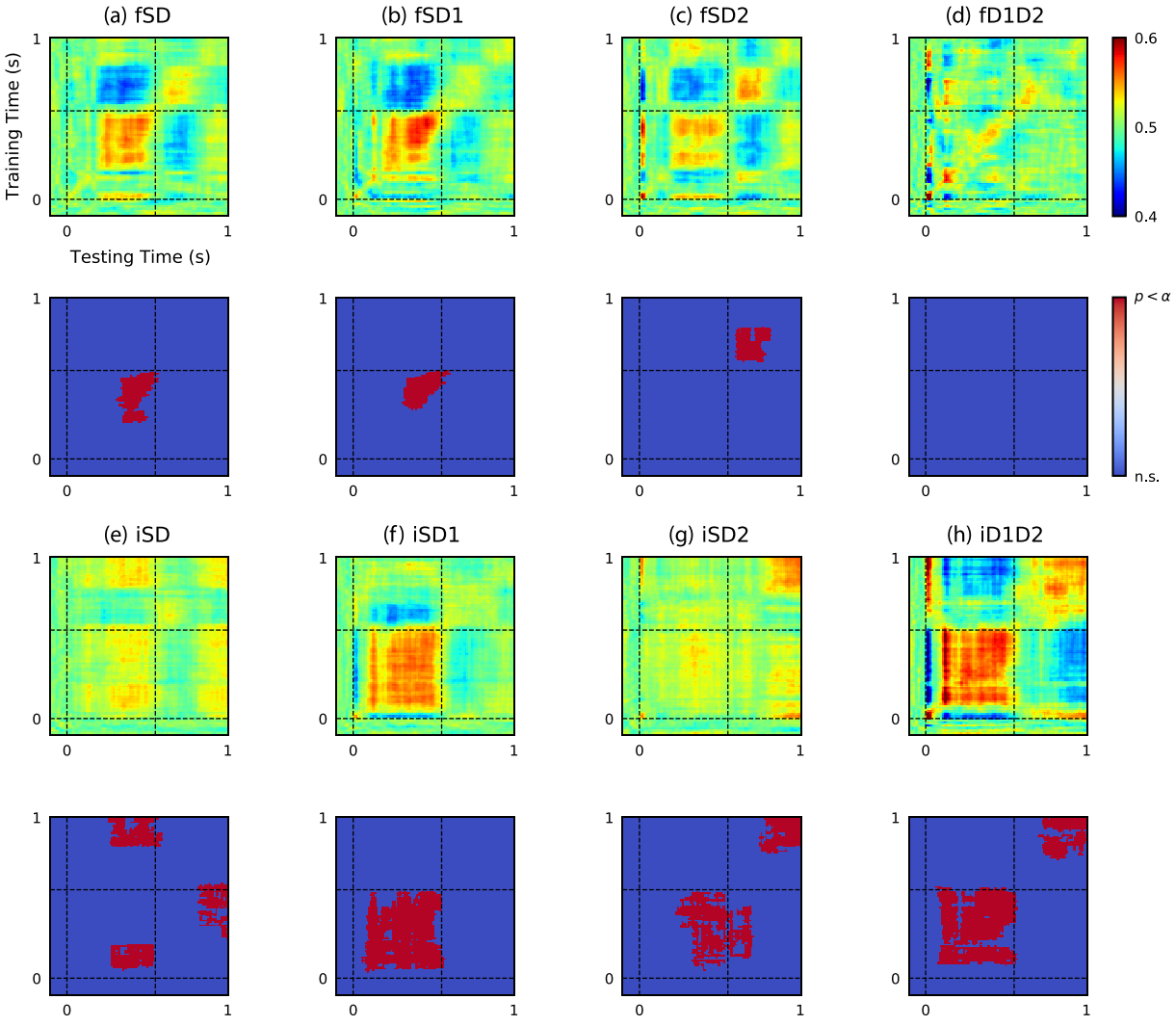

FIGURE 2 Intensity oddball stimuli influence the auditory response asymmetrically. The top two rows (a–d) display results of pairwise decoding between stimulus conditions in the frequency oddball paradigm. Both frequency deviant stimuli produced significant decoding accuracy during the long-latency window when contrasted with the standard, and were comparatively similar. Decoding accuracy of the descending frequency deviant (fD2) is more pronounced and crosses-over into the second stimulus response. The bottom two rows (e–h) show the results of applying this analysis to intensity oddball paradigm data. The two intensity-deviant stimuli also produced significant decoding accuracies within the window of the first stimulus response, although by inspecting the event-related potentials (ERPs) (Figure 1b), it can be seen that this was due to amplitude shifts in opposite directions, which nullified decoding accuracy between standards and deviants (iSD) over the first stimulus response window. There is also a late portion of significant decoding accuracy cause by the iD2 condition, which is presumably caused by the sound-level transition between a 70-dB deviant stimulus and an 80-dB standard stimulus. Stimulus onset times at 0 and .55 s are denoted with dashed lines. In statistical plots below each decoding matrix, red represents statistically significant decoding accuracy following an adjustment for multiple comparisons using cluster-based corrections.

opposite directions, thereby cancelling each other out when pooled together (Figure 2e). Furthermore, trials where quieter deviants are immediately followed by an increasing sound-level transition (i.e., the following standard, which was 80 dB), produced statistically significant, off-diagonal activity when trained on early time-samples coinciding with the long latency response and tested on onset response activity from the second tone, as seen in Figure 2g. Note this is part of the early cluster in iSD2. There is also significant later decoding in the later time point in the iSD2 condition. This is considered to result from relative differences in amplitude between standard and quieter deviant trials during these distant time-windows that are predictive for stimulus condition. Finally, comparing the two deviant trials (iD1D2) resulted in significant early (training = .08 to .57 s, testing: .06 to .57 s, $p < .001$) and later (training = .74 to 1.0 s, testing = .70 to 1.0 s, $p = .038$) time points.

3.3 | Hierarchical recurrent neural network fitted to idealised mouse model cortical auditory evoked response

Hierarchical RNNs were trained to generate output sequences matching cortical evoked potentials in response to simulated auditory inputs computed from a STFT. This process is illustrated in Figure 3a. By

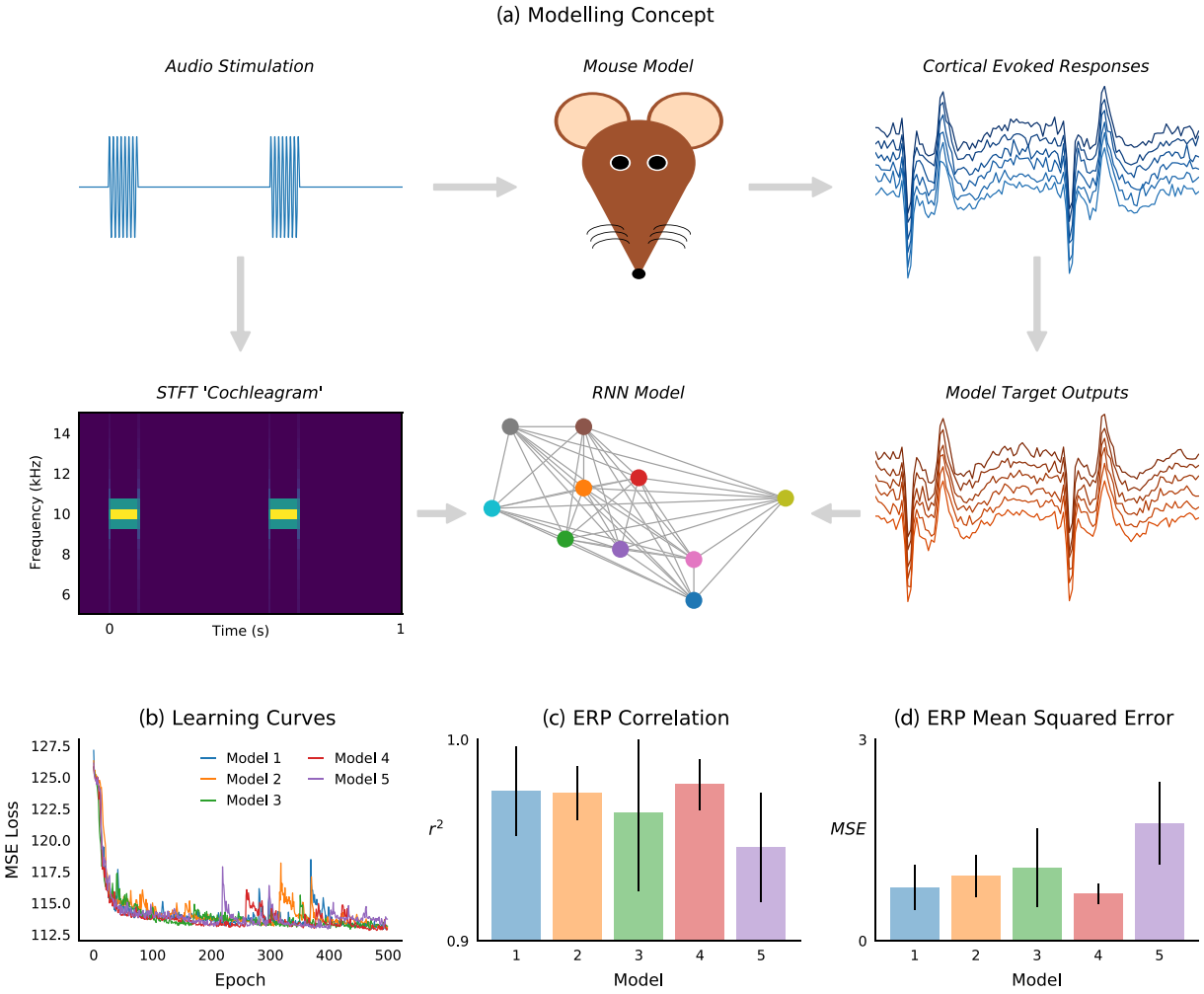

FIGURE 3 Hierarchical recurrent neural network fitted to idealised mouse model cortical auditory evoked response. (a) Data came from an in vivo experiment in which audio stimulation was applied to the mouse model while recording cortical evoked responses. Sound waveforms were transformed into time-frequency domain 'cochleagrams' using the short time Fourier transform (STFT) and used to train an recurrent neural network (RNN) model to generate signals equivalent to target outputs from the mouse cortical evoked responses. The RNN model graphic is illustrative, as it actually consisted of four hidden layers, each with 64 units, and a single output unit. (b) Learning curves from five models trained over 500 epochs. (c) Correlation, and (d) MSE, between model outputs and grand-average event-related potentials (ERP) waveforms from five stimulus conditions (S, fD1, fD2, iD1 and iD2); from this analysis, model 4 was identified as the best model. Bar charts display mean with standard deviation.

minimising MSE loss, model outputs became strongly correlated with grand-average ERP waveforms. Learning curves from training five models, shown in Figure 3b, converge towards a common point, comparable with model performance in cross-validation. After training, the best model was selected based on evaluation between its outputs in response to the five stimulus conditions (S, fD1, fD2, iD1 and iD2) and grand-average ERP waveforms. Model 4 outputs had the highest correlation ($r^2 = .977$, $\sigma = .013$; Figure 3c) and lowest error (MSE = .703, $\sigma = .153$; Figure 3d) measured across stimulus conditions.

Best model hidden unit activations and outputs in response to input stimuli are displayed in Figure 4. Layers 1 and 4 display several units with phasic activations, although in layer 1 these are non-specific, whereas in layer 4 these appear to be context-dependent, occurring during the positive component of the long-latency response evoked by ascending frequency (fourth row of Figure 4b), descending frequency (fourth row of Figure 4c) and louder intensity (fourth row of Figure 4d) stimuli. Across all layers and stimulus conditions there are pronounced unit activations time-locked to auditory input, analogous to in vivo recordings of biological

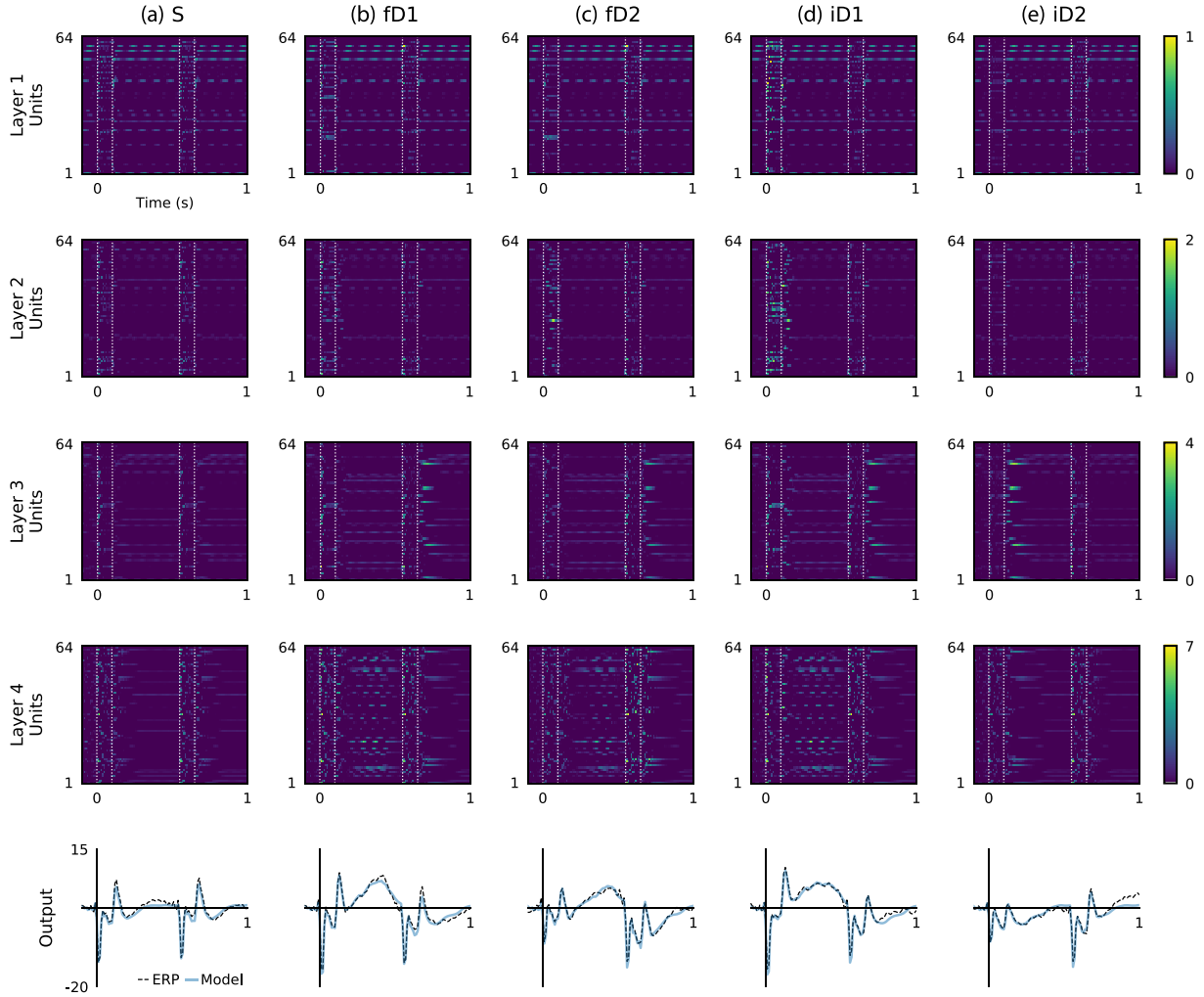

FIGURE 4 Model hidden unit activations interact to generate outputs comparable with event-related potentials (ERP) waveforms. The upper four rows represent hidden unit activations, while the lower row displays model outputs alongside grand-average ERP waveforms, representing the ground truth. (a) Responses to standard stimuli (S). (b) Responses to ascending frequency stimuli (fD1). (c) Responses to descending frequency stimuli (fD2). (d) Responses to higher intensity stimuli (iD1). (e) Responses to lower intensity stimuli (iD2). Light colouring in the hidden layer plots reflect higher unit activity. Stimulus onset and offset times are annotated with vertical white dotted lines on the unit activation plots; the second stimulus in each condition was a standard. The model outputs generally match the features of grand-average ERPs; however, they do not capture the positive amplitude response towards the end of the iD2 condition caused by an increasing sound level transition introduced by the second stimulus. Subtle amplitude differences between hidden unit activations may be difficult to discern from these raster images, and may be examined more closely in the equivalent time-domain signals plotted in Figure 5

neurons of the auditory cortex in response to sound stimulation.

More detail is apparent from the time-domain dynamics of hidden unit activations plotted in Figure 5. Traces are coloured to reflect their categorisation based on temporal response fields, as described in the methods section and outlined in Table 1. This highlights periodic activity of units repeating at approximately 10 Hz, labelled as 'alpha' units, as this lies within the alpha range of EEG signal frequencies. Phasic activations are pronounced in layer 1 and are present to a lesser degree in layers 2 and 3. The majority of units are categorised as 'onset' units, because their activity peaked during the envelope of auditory stimulation. Together these reliably capture the influence of stimulus frequency and intensity on stimulus onset responses observed in the training data. Comparatively fewer units are categorised as 'offset' units, which are responsible for reproducing the offset response feature of the ERP waveform, also present in all of the layers. In contrast, units with peak activity inside the time-window of the positive amplitude long-latency response, categorised as 'danger' units, are only present in layer 4. Similarly, although seemingly encoding the opposite phenomena, units with peak activity within the latency of the negative amplitude feature following quieter deviants or transitions from deviants to standards are categorised as 'safety' units, which are present in layers 3 and 4.

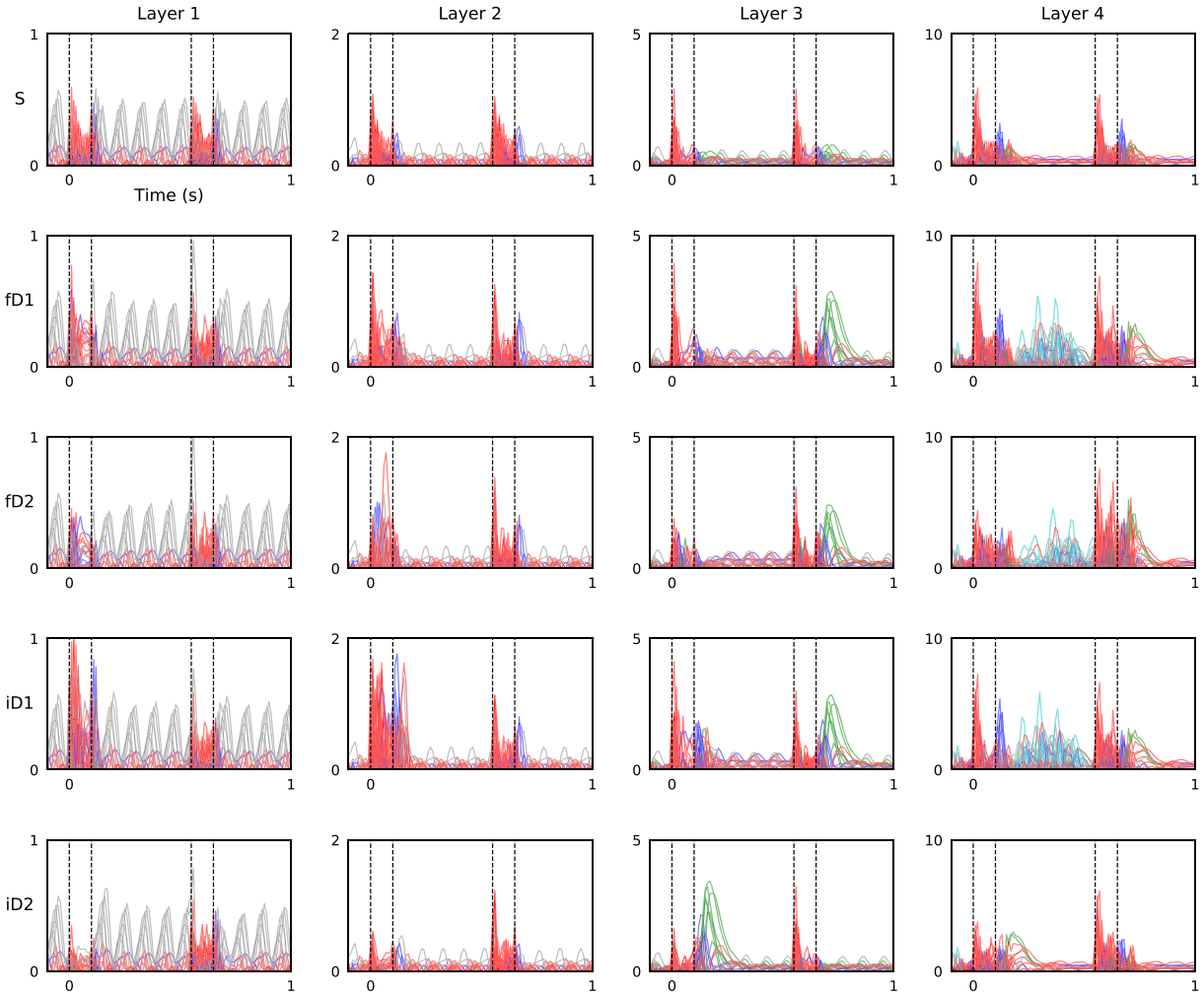

FIGURE 5 Hidden units classified by their temporal response properties. Five categories of units were defined based on their latency of maximum responsiveness. Vertical axes represent hidden unit activation amplitude. Units that were highly active during the pre-stimulus baseline period are coloured grey, those whose response peaked during the stimulus-on period are coloured red, those that peaked in the first 50 ms after the stimulus-on period are coloured blue, those that peaked from 50 to 150 ms after the stimulus-on period are coloured green and those that peaked between 150 and 450 ms after the stimulus-on period are coloured cyan. These five groups are categorised as alpha, onset, offset, safety, and danger, respectively. Remaining units that were unchanging across all stimulus conditions or otherwise did not fall into these categories are coloured black.

TABLE 1 Hidden unit categorisations based on activation peak latency

Type[a]	Limits	Layer 1	Layer 2	Layer 3	Layer 4	Total
Zero	Unchanging	11	7	2	3	23
Alpha	<0 s	11	8	7	5	31
Onset	0–.1 s .55–.65 s	32	40	30	34	136
Offset	.1–.15 s .65–.7 s	10	9	15	10	44
Safety	.15–.25 s .7–.8 s	0	0	9	4	13
Danger	.25–.55 s	0	0	0	8	8
Other	Otherwise	0	0	1	0	1
Total	—	64	64	64	64	256

[a]Most frequent categorisation across all stimulus conditions.

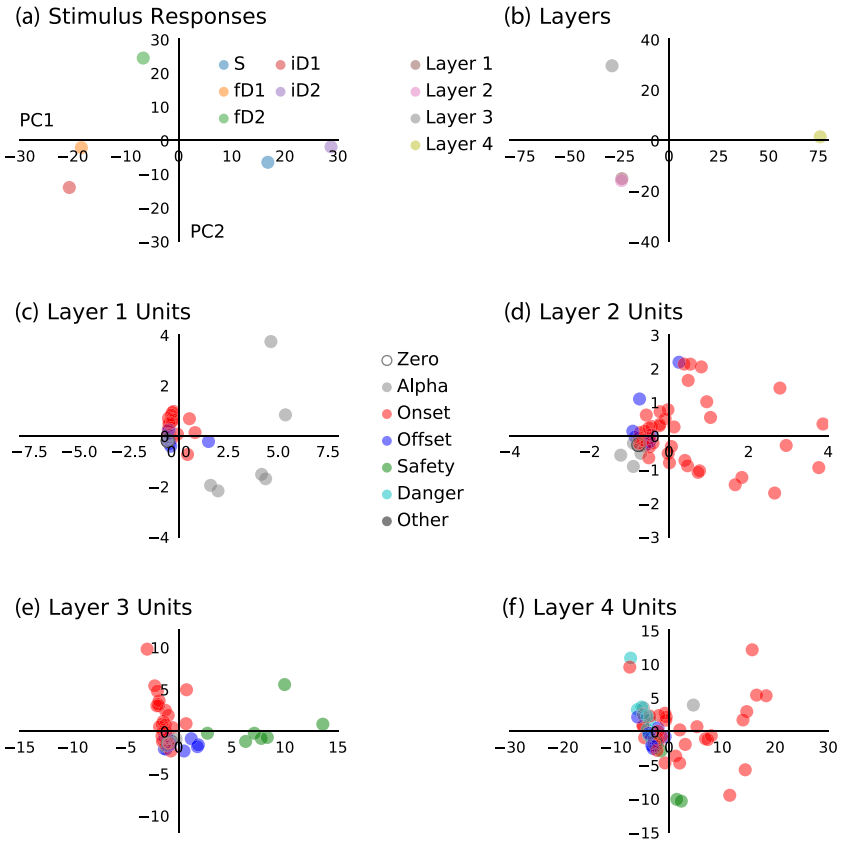

FIGURE 6 Principal component analysis groups model stimulus responses, layers and hidden units by temporal response classification. (a) Whole-model responses to five stimulus conditions. This shows clear separation between responses to stimuli that did (fD1, fD2 and iD1) and did not (S and iD2) produce a positive amplitude long-latency response over .3- to .5-s post-stimulus. (b) Responses of model layers across all stimulus conditions. Layers 1 and 2 appear to be very similar, whereas layers 3 and 4 account for most of the variance in the first two principal components. Responses of layer 1–4 hidden units are plotted in (c)–(f), respectively. In (c)–(f), units are coloured according to their categorisation based on activation peak latency. Time-domain categorisations reflect partial clustering of units in two-dimensional principal component space.

3.4 | Principal component analysis groups model stimulus responses, layers, and hidden units by temporal classification

To further examine model behaviour, hidden unit activations were transformed into principal components, presented in Figure 6. Activations of all units in response to the five stimulus conditions represented in principal component space (Figure 6a) exhibit separation between stimuli that evoked a positive amplitude response from .3 to .5 s (i.e., fD1, fD2 and iD1) and those that did not (i.e., S and iD2). Principal components of activations grouped by layer (Figure 6b) demonstrate the relative similarity between more superficial layers 1 and 2, and dissimilarity between those and deeper layers 3 and 4. Moreover, individual units in each layer transformed into principal components (Figure 6c–f) demonstrate partial clustering corresponding to unit categorisations based

on temporal response fields, suggesting that this relatively simple method of categorisation captures the principal modes of variance in unit behaviour.

3.5 | Interaction between superficial and deep layer activations with environmental salience of stimulus conditions

Hidden units that did not exhibit changes in activity had zero entropy, thus containing no information. Across model layers, the numbers of units that had zero entropy in response to different stimulus conditions are plotted in Figure 7a. The first and second layers had more units with zero entropy. From the units with nonzero entropy, layer average entropy during five input stimulus conditions are given in Figure 7b. This shows relatively high entropy in active units of the first layer across all stimulus conditions. There appears to be an interaction between having fewer zero-entropy units in layer 2 and higher mean entropy in layer 4 during frequency and increasing intensity deviant stimuli. Median sample entropies of units in layers 1 to 4 were .060, .070, .101 and .164, respectively; although this was clearly influenced by the numbers of zero-entropy units in each layer. Median sample entropy of stimulus conditions S, fD1, fD2, iD1 and iD2 were .090, .107, .110, .107 and .084, respectively. Average sample entropy of categorised units was also calculated as follows: zero units ($n = 23$, mean = .0, $\sigma = .0$), alpha units ($n = 31$, mean = .251, $\sigma = .283$), onset units ($n = 136$, mean = .192, $\sigma = .157$), offset units ($n = 44$, mean = .139, $\sigma = .134$), safety units ($n = 13$, mean = .123, $\sigma = .103$), danger units ($n = 8$, mean = .165, $\sigma = .058$) and other units ($n = 1$, mean = .052, $\sigma = .0$).

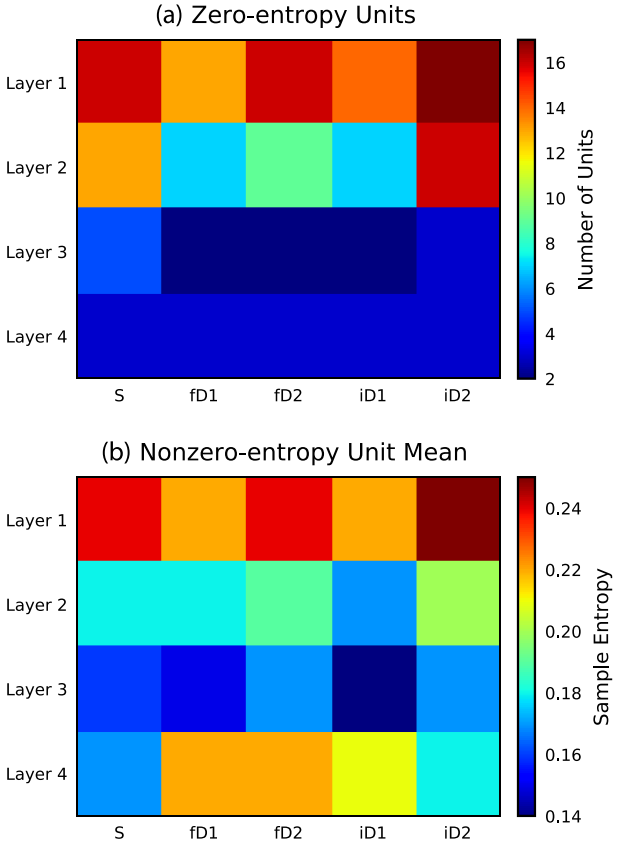

FIGURE 7 Interaction between superficial and deep layer activations with environmental salience of stimulus conditions. (a) The numbers of unresponsive units that showed zero entropy across layers and stimulus conditions. This data presents a layer-wise descending gradient, with fewer unresponsive units in deeper layers. There also appears to be a marginal stimulus-condition effect, with fD1, fD2 and iD1 inputs causing relatively fewer zero-entropy units in layers 2 and 3 than S or iD2 stimuli. (b) Average sample entropy from the remaining nonzero-entropy units in each layer in response to different stimulus conditions. Layer 1 units exhibit higher average sample entropy, with a tendency towards decreasing entropy in deeper layers, with the exception of a sharp increase in response to stimuli that evoked a positive amplitude long-latency response (i.e., fD1, fD2 and iD1). This analysis tentatively hints towards a possible link between the activity of fewer nonzero-entropy units in layers 2/3 and downstream initiation of context-dependent signals observed in the activity of layer 4 units

3.6 | Simulated auditory inputs elicit stereotypical responses to tone duration and intensity, but not frequency

Simulated auditory inputs not included among the set of training stimulus conditions were applied to the model to simulate experiments, the results of which are displayed in Figure 8. These were evaluated qualitatively based on knowledge from prior neurophysiological investigations (e.g., O'Reilly & Conway, 2021). Different duration stimuli produced stimulus offset responses with peak latency positively correlated with stimulus duration, in agreement with expectations. Different frequency tones influenced stimulus onset and offset response peak amplitudes, although not in the proportional manner expected based on prior results. Moreover, only deviant frequencies presented during model training elicited long-latency responses, defying the logical assumption that other frequencies which differ from the standard would also produce similar long-latency responses in vivo. Different intensity stimuli influenced stimulus onset and offset response peak amplitudes in a proportional manner, and both louder stimuli generated positive amplitude long-latency responses, meeting expectations.

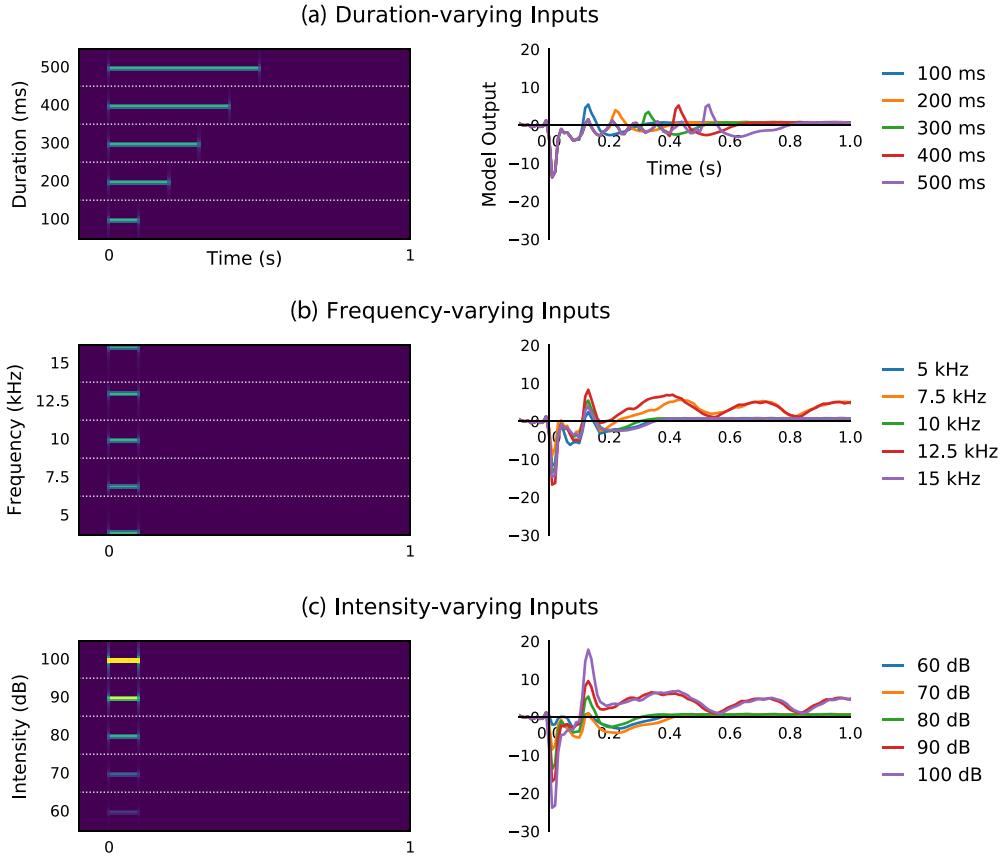

FIGURE 8 Simulated auditory inputs elicit stereotypical responses to tone duration and intensity, but not frequency. (a) Five different duration sounds were simulated and applied as input to the trained RNN (left side), producing simulated ERP waveforms (right side). These waveforms are comparable with those observed in O'Reilly and Conway (2021). (b) Responses to five different frequency sounds were examined. Here the expected relationship between stimulus onset response peak amplitude and tone frequency is not observed, and a long-latency response is only evoked by frequencies of deviant stimuli used for training the model. (c) Responses to five different intensity sounds were also examined. These exhibit correlation between sound intensity and stimulus onset and offset response peak amplitudes, as expected. Moreover, both of the louder tones evoked positive amplitude long-latency responses, whereas the quieter tones evoked more subtle negative amplitudes comparable with those of the standard stimulus. In left-hand side panels, horizontal white dotted lines separate different input stimuli, which are shown from 4.5 to 15.5 kHz.

4 | DISCUSSION

4.1 | Tone frequency changes and rising sound-level transitions trigger positive long-latency response

This discussion with concentrate on novel findings revealed from MVPA and analysis of the fitted hierarchical RNN model. However, to briefly summarise, previous analysis of these data using a double-epoch subtraction found that ascending and descending frequency, and louder but not quieter, novel sounds evoke a long-latency mismatch response from anaesthetised mice (O'Reilly, 2019a). As such, this long-latency mismatch response may be thought to reflect environmental salience, not an indiscriminate change-detection mechanism. Something overlooked previously, but brought to the fore by the results of MVPA in Figures 1 and 2, is that standard stimuli following quieter deviant stimuli (i.e., 70 to 80 dB) trigger a positive amplitude long-latency response resembling that evoked by increasing intensity deviant stimuli (i.e., 80 to 90 dB). Conversely, decreasing sound-level transitions (i.e., either 80 to 70 or 90 to 80 dB) appear to elicit a more subtle, slightly earlier, negative amplitude shift. These observations are consistent with findings from human EEG that demonstrate a prominent role of sound intensity level transition in modulating ERP component amplitudes during the time-window of MMN and P3a (Barry et al., 2022; O'Reilly, 2021a). This adds to evidence suggesting that neural responses to oddball stimulation depend on interactions between physiological adaptation and the physical properties of stimuli, which confound any potential role of prediction-error signalling

(Fishman & Steinschneider, 2012; Lazar & Metherate, 2003; May, 2021; May & Tiitinen, 2010; O'Reilly, 2021b; O'Reilly & Conway, 2021; O'Reilly & O'Reilly, 2021; Solomon et al., 2021; Umbricht et al., 2005).

An argument could be made that the effects observed in Figure 1 are primarily due to the physical properties of presented stimuli, not necessarily related to their probability or context. From this perspective, features of the auditory response evoked by different frequency or intensity stimuli are expected to be proportional to the respective physical quality of the stimulus. For this reason, the use of control paradigms in MMN research is encouraged to verify interpretations of resulting ERP components (Harms et al., 2014). It is apparent from the waveforms plotted in Figure 1a,b that both frequency and intensity have a modulatory influence over onset and offset response peak amplitudes, such that the standard stimulus elicits amplitudes that lie between those evoked by lower and upper deviant stimuli. These effects have been characterised in conscious and urethane-anaesthetised mice using the many-standards control sequence (O'Reilly & Conway, 2021). However, patterns of signals observed in the long-latency response window, from .2 to .5 s, are inconsistent with this view. In Figure 1a, both frequency deviant responses exhibit positive amplitude long-latency deflections that surpass the amplitude evoked by the standard frequency stimulus in this time-window. Furthermore, in Figure 1b, amplitudes evoked by 80-dB standard and 70-dB deviant stimuli essentially overlap during this time-window, whereas the 90-dB stimulus produces positive amplitude deflections. If these positive amplitude long-latency responses were simply due to the physical makeup of the stimuli, then standard stimulus responses should consistently lie between those of the two deviants, as they do for stimulus onset and offset peaks, but this is not the case. Results from other studies in urethane-anaesthetised mice have also demonstrated that low-probability stimuli in the oddball sequence, which violate sensory expectations, produce comparable long-latency features that were absent from responses to low-probability stimuli in the many-standards control sequence, which do not violate an established auditory regularity (Casado-Román et al., 2020; Kurkela et al., 2018; O'Reilly & Angsuwatanakul, 2021).

4.2 | Negative, 'safety', or positive, 'danger', long-latency responses can be interpreted to reflect environmentally inconspicuous or salient auditory stimuli

Informed by the results of MVPA and the RNN model, these long-latency responses are referred to using the term 'danger' for those that were initiated by an increase in environmental salience of auditory input, such as frequency deviant or increasing intensity deviant stimuli; contrasted with 'safety' responses, which were observed following relatively inconspicuous stimuli, such as quieter deviant stimuli or standard stimuli following a danger response. Danger and safety responses are suggested to be distinct from consecutive presentation of two standard stimuli, which reflect neutral environmental salience, and thereby evoke neither safety nor danger responses (top row of Figure 5). Choices of terminology for describing these observations are based on an interpretation of the evolutionary incentives placed on mammalian auditory systems, and are not presented as a rigorous taxonomy. Importantly, the model fails to reproduce the positive amplitude long-latency response evoked by 80-dB standards immediately following 70-dB deviants (Figures 1b and 1i, and the bottom row of Figure 4e), which probably should have engendered a response from the danger units. The relatively low number of trials containing this feature might explain why the model was unable to learn it.

Safety units were active in a relatively narrow time window (.15 to .25 s) following auditory stimulation, while danger units were active over a later, more extended period (.25 to .55 s). Interestingly, this agrees with observations that inconspicuous stimuli are normally processed more quickly than salient stimuli, which tend to produce larger, longer lasting auditory-evoked potentials (Heilbron & Chait, 2018; Horváth et al., 2008; Näätänen et al., 2007; Polich, 2007). Opposite polarity trajectories of ERP waveforms during the latency ranges associated with safety and danger unit activations (seen in Figures 4 and 5) may suggest that the underlying neurological processes are fulfilling separate roles. For example, perhaps activation of safety units lowers the threshold for detecting environmentally salient stimuli, whereas danger units are active when an auditory input exceeds this threshold, also causing it to rise. This interpretation shares elements of the model-adjustment theory of mismatch negativity (Garrido et al., 2009), although is distinct in that it places an emphasis on environmental salience, therefore differentiating between auditory novelty responses evoked by louder and softer deviant stimuli. Venturing any further towards attributing descriptors to these phenomena, such as inhibitory or excitatory, bottom-up or top-down, would be overly speculative, and require more elaborate recording methods to validate.

While these responses occur during the latency range of human MMN and P3a, it cannot be said definitively that they reflect the same underlying neurophysiological processes, given that the relationships among ERP components, neuroanatomy and complexity of perceptual processing are incompletely understood (Itoh et al., 2022;

Javitt et al., 1992; Komatsu et al., 2015). Nevertheless, the findings here replicate those found elsewhere in the animal model literature. Comparable long-latency responses to frequency oddball stimuli have been observed in anaesthetised rodents (Casado-Román et al., 2020; Chen et al., 2015; O'Reilly & Angsuwatanakul, 2021; Ruusuvirta et al., 1998). Curiously though, these have been absent from studies employing similar or near-identical stimulation and recording protocols in conscious mice (Harms et al., 2014; O'Reilly & Conway, 2021). To reconcile this apparent dichotomy between the presence of long-latency mismatch responses in anaesthetised rodents and absence thereof in conscious rodents, it may be speculated that the long-latency response to environmentally salient auditory stimuli quickly habituates if not reinforced in conscious animals, typical of the central orienting reflex associated with noradrenergic neurotransmission mediated via the locus coeruleus (Sara & Bouret, 2012; Shine, 2019). Future studies in anaesthetised and conscious animals may be designed to explore this hypothesis.

4.3 | Hidden units classified by their temporal response properties

Model hidden units were classified based on temporal response fields, essentially based on their latency of peak activation, into alpha, onset, offset, safety and danger units. Unresponsive units and units that did not fit into any of these categories were also considered. Time-domain patterns of unit activity in response to different stimulus conditions are shown in Figure 5, with traces coloured according to unit categorisation. The PCA results in Figure 6c–f demonstrate that these classifications reasonably-well capture the primary modes of variance in unit behaviour across all layers. It is evident that the patterns of activity from some of the hidden units spontaneously adopt an alpha frequency oscillation, presenting an emergent parallel with the prominent alpha rhythm observed in human EEG recordings (Quigley, 2021). The activity of these alpha units was suppressed during stimulus presentation, which tangentially lends support for the longstanding phase-reset theory of evoked potentials (Hanslmayr et al., 2007).

4.4 | Interaction between superficial and deep layer activations with environmental salience of stimulus conditions

Analysis of sample entropy calculated from hidden unit activations shown in Figure 7 hints towards a potential relationship between the behaviour of units in superficial layers with downstream manifestation of long-latency responses in deeper layers. Specifically, there were fewer layer 2 zero-entropy units and higher average sample entropy among layer 4 units during fD1, fD2 and iD1 stimulus conditions, all of which elicited a positive amplitude long-latency response, tentatively suggesting that the trigger signal for this long-latency response originates in earlier layers and feeds through the hierarchical model to influence its output. This offers a small glimpse into the potential structure underlying network responses to contextually different stimuli, however, development of more advanced analytical techniques will be required to explore these relationships more deeply (Barrett et al., 2019; Yang & Molano-Mazón, 2021).

4.5 | Emergent computational neurophysiology observed from the model

The majority of hidden units were maximally active during the time-window of simulated auditory tones. This parallels biological neurons of the auditory cortex that are excited by sound stimulation (Bajo & King, 2012; King et al., 2018). Second in number were offset response units, for which the existence of biological counterparts is also supported by an extensive literature (Kopp-Scheinpflug et al., 2018; O'Reilly, 2019a; Solyga & Barkat, 2021). Relatively fewer units were grouped into more abstract categories of 'safety' and 'danger' units; safety units were located in layers 3 and 4, whereas danger units were situated only in layer 4. This somewhat anthropomorphic terminology has been selected to describe the relative environmental salience of stimuli that produced these responses. Danger units were activated by presentation of what are assumed to be salient changes in auditory input, specifically the first stimuli in fD1, fD2 and iD1 conditions. In contrast, comparatively inconspicuous changes in auditory input, such as the first stimulus in the iD2 condition, or second stimuli in fD1, fD2 or iD1 conditions, elicited activity from safety units. The existence of cortical neurons serving comparable functions is less firmly established than the aforementioned parallels between model behaviour and neurophysiological findings, although this seems a reasonable hypothesis derived from the model, particularly given there is evidence of neurons in subcortical structures that encode relative safety or danger attributed to incoming stimuli (Rogan et al., 2005; Sangha et al., 2013). Interestingly, these safety and danger units may also bear some of the hallmarks of negative and positive prediction error neurons that encode opposite direction mismatches

between input stimuli and expectations (Hertäg & Clopath, 2022).

An interesting wrinkle exposed by this aspect of model behaviour is that ERP component amplitudes accompanying the activity of safety and danger units are of opposite polarity. Were the biological existence of these units to be confirmed, this would challenge present formulations of the predictive coding theory of auditory processing by implying that prediction errors emerge differently based on stimulus context, at least with respect to environmental salience. Inclusion of this seemingly natural facet of sensory processing should not be at odds with a complete theory of auditory perception. Although, at a minimum this would require revision of the dominant view of predictive coding, which currently lacks an explanation for these observations (Casado-Román et al., 2020; Friston, 2005; Garrido et al., 2009; Lieder et al., 2013). There are alternative, albeit perhaps less exotic or conceptually appealing, explanations that have been proposed to account for differential responses to passive auditory oddball stimulation (Butler, 1968; May, 2021; May & Tiitinen, 2010; O'Reilly, 2021a, 2021b; O'Reilly & Conway, 2021; O'Reilly & O'Reilly, 2021). However, these are rarely considered preferable to the well-supported predictive coding framework, even when dutifully considered (Lacroix et al., 2022). As they currently stand, neither predictive coding nor alternative sensory processing theories succinctly encapsulate all of the physiological phenomena observed in response to sequences of sounds, which shall therefore remain the purview of future efforts directed towards improving our understanding these processes.

4.6 | Simulated auditory inputs elicit stereotypical responses to tone duration and intensity, but not frequency

Simulated experiments were conducted as a means of exploring the validity of model behaviour under conditions that were not exposed during training (Wacongne et al., 2012). Figure 8 displays the findings from three such experiments with the best-fitting hierarchical RNN. In these experiments, inputs representing five physically different audio stimuli were constructed and passed through the model to elicit output waveforms that were qualitatively evaluated against expectations based on established neurophysiology. Simulated duration-varying stimuli (Figure 8a) produced close agreement with results from a many-standards control sequence presented to the same cohort of anaesthetised mice (O'Reilly & Conway, 2021), presenting onset and offset responses also seen in anaesthetised rats (Nakamura et al., 2011). This suggests that the trained model reliably captures stimulus onset and offset responses, without producing unexpected activity, in response to simulated audio stimuli with different durations.

Results from simulated frequency-varying inputs (Figure 8b) are further from neurophysiological expectations. The relationships between tone frequency and obligatory stimulus-on and stimulus-off response peak amplitudes were not proportional, therefore not in agreement with previous findings (Nakamura et al., 2011; O'Reilly & Conway, 2021). Furthermore, long-latency responses were only observed from deviant frequency stimuli exposed to during training, defying the expectation that other frequencies that deviate from the standard would also produce comparable responses in vivo (Casado-Román et al., 2020; Chen et al., 2015; O'Reilly & Angsuwatanakul, 2021). This suggests that the model has not learned to trigger this response to frequencies that differ from the standard, rather it has learned to produce long-latency responses only to the specific deviant frequency stimuli used during training (7.5 and 12.5 kHz). Perhaps training the model across multiple datasets with different stimulus frequencies as standards and deviants would assist in overcoming this limitation (Yang & Molano-Mazón, 2021). Alternatively, training the model on longer sequences spanning several stimuli may be necessary to truly learn the underlying relationship between responses to frequent standard stimuli and infrequent deviant stimuli.

Model outputs in response to different intensity stimuli (Figure 8c) provide closer agreement with physiological expectations. These depict a proportional relationship between simulated stimulus intensity level and the magnitudes of onset and offset responses, which may be expected in vivo (Bajo & King, 2012; O'Reilly & Conway, 2021). Also, both louder stimuli produced positive amplitude long-latency responses, which appeals to the assumption that environmentally salient stimuli ought to generate similar responses. Quieter stimuli produced a more subtle, negative dip in output waveforms, corresponding to the safety signal already mentioned.

4.7 | Limitations

Data recorded from above the auditory cortices of urethane-anaesthetised mice during passive auditory oddball stimulation formed the basis for this analysis. Such experimental designs are intended to provide preclinical models with relevance to human auditory novelty responses. However, there are a few important limitations of this approach that should be considered to avoid over-interpreting the applicability of these findings to

research in humans. Firstly, while auditory novelty responses are observed in anaesthetised and non-responsive humans (Morlet & Fischer, 2014), it is unknown whether these would be present under urethane anaesthesia, which is not used in humans because of its carcinogenic properties. Secondly, the degree of similarity between mouse and human auditory neurophysiology is somewhat ambiguous. Siegel et al. (2003) suggested that the mouse auditory evoked response could be a time-compressed version of the human auditory evoked potential, given that both exhibit a series of alternating polarity deflections in their ERP waveform. Although it is unclear whether this hypothesis has been rigorously tested, perhaps due to difficulties in characterising the physiological origins of specific ERP features, which introduces a third limitation: the inverse source problem. Ultimately we wish to understand the neurophysiology inside the brain (of both mouse and human), by using signals recorded from above the skull that reflect an unknown function of this inner neural activity. Deciphering underlying neurophysiology from EEG signals is therefore fraught with uncertainty. As such, existence of hypothetical neural sources suggested by the model should be verified with higher resolution neuroimaging techniques, such as implanted electrodes, in vivo optical imaging, or fMRI.

4.8 | Further considerations of the model

We ought to acknowledge that any links between model behaviour and neurobiology are necessarily speculative. The present model portrays interesting parallels with experimental neurophysiology in terms of its spontaneous alpha-frequency rhythms, onset and offset units, and potentially also, although yet to be confirmed, danger and safety units, which encode differential responses to environmentally salient and inconspicuous stimuli, respectively. Nevertheless, computational models are treated as tools for studying the potential mechanisms employed within the brain and are not assumed to be directly homologous. Specifically in relation to artificial neural networks, these are known to be biologically implausible, but can still be valuable for neuroscience research (Barak, 2017; Barrett et al., 2019; Yang & Molano-Mazón, 2021).

One of the particularly appealing elements of artificial neural networks, despite their biological implausibility, is that they develop latent states that may be considered roughly analogous to the behaviour of neural sources, thereby enabling a richer analysis of the potential dynamics of neural activity underlying patterns observed in electrophysiology data. However, it is important to avoid over-interpretation. For example, numbers used to refer to model layers do not imply any hypothetical relationship with cortical layers that are conventionally numbered I to VI. Moreover, the terms superficial and deep, used in the context of artificial neural networks, refer to the relative proximity of layers to the input signal, not to be confused with how these terms are used to discuss cortical layers. It may be possible to apply a vast array of established and novel analytical methods to study the hierarchical RNN, which will provide plenty avenues for future research, although the present article's scope is limited in an effort to keep it concise.

As a final cautionary note, the present application of an RNN to model 'idealised experiment' data is subject to some of the same limitations as conventional ERP analysis, such as assuming that evoked components are relatively stationary across subjects and trials. In many cases this may be an invalid assumption. However, MVPA is partially immune to this limitation, being applied to individual subject data separately, and its application has confirmed (as shown in Figures 1 and 2) that the long-latency components evoked by frequency and rising intensity level sounds are not spurious phenomena.

5 | CONCLUSION

This reanalysis of cortical auditory-evoked potentials from anaesthetised mice in response to frequency and intensity oddball paradigms offers two related insights that were undiscovered in previous examinations of these data. Firstly, asymmetric sound level transitions in the intensity oddball paradigm modulate cortical electrophysiology in different respects, such that rising sound level transitions, which may be perceived as reflecting greater environmental salience, akin to frequency deviant stimuli, produce a positive amplitude long-latency feature; whereas relatively inconspicuous, falling sound level transitions produce a subtle negative amplitude long-latency feature in the evoked response waveform. Secondly, through the mechanisms depicted by a computational model, these features correlating with relative changes in environmental salience can be conceptualised as reflecting potential 'danger' or 'safety' responses, respectively, which are encoded by the activity of distinct sources that influence the cortical auditory-evoked response. The hypothesised biological existence of these separate sources may be evaluated in future animal or human neuroimaging experiments. This study also highlights the potential synergy between MVPA and computational modelling approaches, and the myriad intriguing parallels between established auditory neurophysiology

and the emergent behaviour of a hierarchical recurrent neural network fitted to cortical auditory evoked potential data.

ACKNOWLEDGEMENTS

Data used in this study were obtained during a doctoral training award funded by the United Kingdom Engineering and Physical Sciences Research Council (grant no. EP/F50036X/1). The computational model was trained using a computer system and graphics processing unit (NVIDIA Titan RTX) purchased with financial support from the Research Institute of Rangsit University (grant no. 90/2561). Open access publishing facilitated by The University of Sydney, as part of the Wiley - The University of Sydney agreement via the Council of Australian University Librarians.

CONFLICT OF INTEREST

The authors declare to have no competing interests concerning the faithful conduct of this research.

AUTHOR CONTRIBUTION

Jamie A. O'Reilly: Conceptualization, methodology, software, formal analysis, data curation, writing–original draft, writing–review & editing, visualization. **Thanate Angsuwatanakul:** Formal analysis, writing–review & editing. **Jordan Wehrman:** Conceptualization, methodology, software, formal analysis, writing–review & editing.

DATA AVAILABILITY STATEMENT

Data and code associated with this article can be accessed from https://osf.io/hkmut/.

ORCID

Jamie A. O'Reilly https://orcid.org/0000-0002-2250-3077

Jordan Wehrman https://orcid.org/0000-0002-9358-6092

REFERENCES

Angsuwatanakul, T., O'Reilly, J., Ounjai, K., Kaewkamnerdpong, B., & Iramina, K. (2020). Multiscale entropy as a new feature for EEG and fNIRS analysis. *Entropy*, *22*, 189–212. https://doi.org/10.3390/e22020189

Avissar, M., & Javitt, D. C. (2018). Mismatch negativity: A simple and useful biomarker of N-methyl-D-aspartate receptor (NMDAR)-type glutamate dysfunction in schizophrenia. *Schizophrenia Research*, *191*, 1–4. https://doi.org/10.1016/j.schres.2017.11.006

Bae, G. Y., Leonard, C. J., Hahn, B., Gold, J. M., & Luck, S. J. (2020). Assessing the information content of ERP signals in schizophrenia using multivariate decoding methods. *NeuroImage Clin.*, *25*, 102179. https://doi.org/10.1016/j.nicl.2020.102179

Bajo, V. M., & King, A. J. (2012). Cortical modulation of auditory processing in the midbrain. *Frontiers in Neural Circuits*, *6*, 1–12. https://doi.org/10.3389/FNCIR.2012.00114

Barak, O. (2017). Recurrent neural networks as versatile tools of neuroscience research. *Current Opinion in Neurobiology*, *46*, 1–6. https://doi.org/10.1016/j.conb.2017.06.003

Barrett, D. G., Morcos, A. S., & Macke, J. H. (2019). Analyzing biological and artificial neural networks: Challenges with opportunities for synergy? *Current Opinion in Neurobiology*, *55*, 55–64. https://doi.org/10.1016/J.CONB.2019.01.007

Barry, R. J., De Blasio, F. M., Rushby, J. A., MacDonald, B., Fogarty, J. S., & Cave, A. E. (2022). Stimulus intensity effects and sequential processing in the passive auditory ERP. *International Journal of Psychophysiology*, *176*, 149–163. https://doi.org/10.1016/j.ijpsycho.2022.03.005

Brønnick, K. S., Nordby, H., Larsen, J. P., & Aarsland, D. (2010). Disturbance of automatic auditory change detection in dementia associated with Parkinson's disease: A mismatch negativity study. *Neurobiology of Aging*, *31*, 104–113. https://doi.org/10.1016/J.NEUROBIOLAGING.2008.02.021

Butler, R. A. (1968). Effect of changes in stimulus frequency and intensity on habituation of the human vertex potential. *The Journal of the Acoustical Society of America*, *44*, 945–950. https://doi.org/10.1121/1.1911233

Casado-Román, L., Carbajal, G. V., Pérez-González, D., & Malmierca, M. S. (2020). Prediction error signaling explains neuronal mismatch responses in the medial prefrontal cortex. *PLoS Biology*, *18*(12), e3001019. https://doi.org/10.1371/journal.pbio.3001019

Chen, I.-W. W., Helmchen, F., & Lütcke, H. (2015). Specific early and late oddball-evoked responses in excitatory and inhibitory neurons of mouse auditory cortex. *The Journal of Neuroscience*, *35*(36), 12560–12573. https://doi.org/10.1523/JNEUROSCI.2240-15.2015

De-Wit, L., Machilsen, B., & Putzeys, T. (2010). Predictive coding and the neural response to predictable stimuli. *The Journal of Neuroscience*, *30*(26), 8702–8703. https://doi.org/10.1523/JNEUROSCI.2248-10.2010

Fishman, Y. I., & Steinschneider, M. (2012). Searching for the mismatch negativity in primary auditory cortex of the awake monkey: Deviance detection or stimulus specific adaptation? *The Journal of Neuroscience*, *32*, 15747–15758. https://doi.org/10.1523/JNEUROSCI.2835-12.2012

Friedman, D., Cycowicz, Y. M., & Gaeta, H. (2001). The novelty P3: An event-related brain potential (ERP) sign of the brains evaluation of novelty. *Neuroscience and Biobehavioral Reviews*, *25*(4), 355–373. https://doi.org/10.1016/S0149-7634(01)00019-7

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

Friston, K., & Kiebel, S. (2009). 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

Garrido, M. I., Kilner, J. M., Stephan, K. E., & Friston, K. J. (2009). The mismatch negativity: A review of underlying mechanisms. *Clinical Neurophysiology*, *120*, 453–463. https://doi.org/10.1016/j.clinph.2008.11.029

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (pp. 249–256). PMLR.

Grootswagers, T., Wardle, S. G., & Carlson, T. A. (2017). Decoding dynamic brain patterns from evoked responses: A tutorial on multivariate pattern analysis applied to time series neuroimaging data. *Journal of Cognitive Neuroscience*, *29*(4), 677–697. https://doi.org/10.1162/jocn_a_01068

Hanslmayr, S., Klimesch, W., Sauseng, P., Gruber, W., Doppelmayr, M., Freunberger, R., Pecherstorfer, T., & Birbaumer, N. (2007). Alpha phase reset contributes to the generation of ERPs. *Cerebral Cortex*, *17*, 1–8. https://doi.org/10.1093/CERCOR/BHJ129

Harms, L., Fulham, W. R., Todd, J., Budd, T. W., Hunter, M., Meehan, C., Penttonen, M., Schall, U., Zavitsanou, K., Hodgson, D. M., & Michie, P. T. (2014). Mismatch negativity (MMN) in freely-moving rats with several experimental controls. *PLoS One*, *9*, e110892. https://doi.org/10.1371/journal.pone.0110892

Heilbron, M., & Chait, M. (2018). Great expectations: Is there evidence for predictive coding in auditory cortex? *Neuroscience*, *389*, 54–73. https://doi.org/10.1016/J.NEUROSCIENCE.2017.07.061

Hertäg, L., & Clopath, C. (2022). Prediction-error neurons in circuits with multiple neuron types: Formation, refinement, and functional implications. *Proceedings of the National Academy of Sciences*, *119*(13), e2115699119. https://doi.org/10.1073/PNAS.2115699119

Higuchi, Y., Seo, T., Miyanishi, T., Kawasaki, Y., Suzuki, M., & Sumiyoshi, T. (2014). Mismatch negativity and P3a/reorienting complex in subjects with schizophrenia or at-risk mental state. *Frontiers in Behavioral Neuroscience*, *8*, 172. https://doi.org/10.3389/FNBEH.2014.00172/BIBTEX

Horváth, J., Winkler, I., & Bendixen, A. (2008). Do N1/MMN, P3a, and RON form a strongly coupled chain reflecting the three stages of auditory distraction? *Biological Psychology*, *79*, 139–147. https://doi.org/10.1016/j.biopsycho.2008.04.001

Itoh, K., Konoike, N., Nejime, M., Iwaoki, H., Igarashi, H., Hirata, S., & Nakamura, K. (2022). Cerebral cortical processing time is elongated in human brain evolution. *Scientific Reports*, *12*, 1–13. https://doi.org/10.1038/s41598-022-05053-w

Javitt, D. C., Schroeder, C. E., Steinschneider, M., Arezzo, J. C., & Vaughan, H. G. (1992). Demonstration of mismatch negativity in the monkey. *Electroencephalography and Clinical Neurophysiology*, *83*, 87–90. https://doi.org/10.1016/0013-4694(92)90137-7

King, A. J., Teki, S., Willmore, B. D., Schreiner San Francisco, C., & Francisco, S. (2018). Recent advances in understanding the auditory cortex. F1000Research *7*, 1000–1555. https://doi.org/10.12688/f1000research.15580.1

King, J. R., & Dehaene, S. (2014). Characterizing the dynamics of mental representations: The temporal generalization method. *Trends in Cognitive Sciences*, *18*(4), 203–210. https://doi.org/10.1016/J.TICS.2014.01.002

Komatsu, M., Takaura, K., & Fujii, N. (2015). Mismatch negativity in common marmosets: Whole-cortical recordings with multichannel electrocorticograms. *Scientific Reports*, *5*, 15006. https://doi.org/10.1038/srep15006

Kopp-Scheinpflug, C., Sinclair, J. L., & Linden, J. F. (2018). When sound stops: Offset responses in the auditory system. *Trends in Neurosciences*, *41*, 712–728. https://doi.org/10.1016/j.tins.2018.08.009

Kurkela, J. L. O., Lipponen, A., Kyläheiko, I., & Astikainen, P. (2018). Electrophysiological evidence of memory-based detection of auditory regularity violations in anesthetized mice. *Scientific Reports*, *8*, 3027. https://doi.org/10.1038/s41598-018-21411-z

Lacroix, A., Harquel, S., Mermillod, M., Vercueil, L., Alleysson, D., Dutheil, F., Kovarski, K., & Gomot, M. (2022). The predictive role of low spatial frequencies in automatic face processing: A visual mismatch negativity investigation. *Frontiers in Human Neuroscience*, *0*, 101. https://doi.org/10.3389/FNHUM.2022.838454

Lazar, R., & Metherate, R. (2003). Spectral interactions, but no mismatch negativity, in auditory cortex of anesthetized rat. *Hearing Research*, *181*(1–2), 51–56. https://doi.org/10.1016/S0378-5955(03)00166-7

Lieder, F., Daunizeau, J., Garrido, M. I., Friston, K. J., & Stephan, K. E. (2013). Modelling trial-by-trial changes in the mismatch negativity. *PLoS Computational Biology*, *9*, e1002911. https://doi.org/10.1371/journal.pcbi.1002911

Mathys, C., Daunizeau, J., Friston, K. J., & Stephan, K. E. (2011). A Bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*, *0*, 1–20, 9. https://doi.org/10.3389/FNHUM.2011.00039/BIBTEX

May, P. J. C. (2021). The adaptation model offers a challenge for the predictive coding account of mismatch negativity. *Frontiers in Human Neuroscience*, *15*, 1–10. https://doi.org/10.3389/fnhum.2021.721574

May, P. J. C., & Tiitinen, H. (2010). Mismatch negativity (MMN), the deviance-elicited auditory deflection, explained. *Psychophysiology*, *47*, 66–122. https://doi.org/10.1111/j.1469-8986.2009.00856.x

Morlet, D., & Fischer, C. (2014). MMN and novelty P3 in coma and other altered states of consciousness: A review. *Brain Topography*, *27*, 467–479. https://doi.org/10.1007/S10548-013-0335-5

Näätänen, R., Gaillard, A. W. K., & Mäntysalo, S. (1978). Early selective-attention effect on evoked potential reinterpreted. *Acta Psychologica (Amst)*, *42*, 313–329. https://doi.org/10.1016/0001-6918(78)90006-9

Näätänen, R., Jacobsen, T., & Winkler, I. (2005). Memory-based or afferent processes in mismatch negativity (MMN): A review of the evidence. *Psychophysiology*, *42*, 25–32. https://doi.org/10.1111/j.1469-8986.2005.00256.x

Näätänen, R., Paavilainen, P., Rinne, T., & Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: A review. *Clinical Neurophysiology*, *118*, 2544–2590. https://doi.org/10.1016/j.clinph.2007.04.026

Nakamura, T., Michie, P. T., Fulham, W. R., Todd, J., Budd, T. W., Schall, U., Hunter, M., & Hodgson, D. M. (2011). Epidural auditory event-related potentials in the rat to frequency and duration deviants: Evidence of mismatch negativity? *Frontiers in Psychology*, *2*, 367. https://doi.org/10.3389/fpsyg.2011.00367

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational*

Intelligence and Neuroscience, *2011*, 1–9. https://doi.org/10.1155/2011/156869

O'Reilly, J. A. (2019a). Double-epoch subtraction reveals long-latency mismatch response in urethane-anaesthetized mice. *Journal of Neuroscience Methods*, *326*, 108375. https://doi.org/10.1016/j.jneumeth.2019.108375

O'Reilly, J. A. (2019b). Event-related potential arithmetic to analyze offset potentials from conscious mice. *Journal of Neuroscience Methods*, *318*, 78–83. https://doi.org/10.1016/j.jneumeth.2019.01.018

O'Reilly, J. A. (2021a). Can intensity modulation of the auditory response explain intensity-decrement mismatch negativity? *Neuroscience Letters*, *764*, 136199. https://doi.org/10.1016/j.neulet.2021.136199

O'Reilly, J. A. (2021b). Roving oddball paradigm elicits sensory gating, frequency sensitivity, and long-latency response in common marmosets. *IBRO Neuroscience Reports*, *11*, 128–136. https://doi.org/10.1016/j.ibneur.2021.09.003

O'Reilly, J. A., & Angsuwatanakul, T. (2021). More evidence for a long-latency mismatch response in urethane-anaesthetised mice. *Hearing Research*, *408*, 108296. https://doi.org/10.1016/j.heares.2021.108296

O'Reilly, J. A., & Conway, B. A. (2021). Classical and controlled auditory mismatch responses to multiple physical deviances in anaesthetised and conscious mice. *The European Journal of Neuroscience*, *53*, 1839–1854. https://doi.org/10.1111/ejn.15072

O'Reilly, J. A., & O'Reilly, A. (2021). A critical review of the deviance detection theory of mismatch negativity. *NeuroScience*, *2*, 151–165. https://doi.org/10.3390/neurosci2020011

Otten, L. J., Alain, C., & Picton, T. W. (2000). Effects of visual attentional load on auditory processing. *Neuroreport*, *11*, 875–880. https://doi.org/10.1097/00001756-200003200-00043

Parras, G. G., Nieto-Diego, J., Carbajal, G. V., Valdés-Baizabal, C., Escera, C., & Malmierca, M. S. (2017). Neurons along the auditory pathway exhibit a hierarchical organization of prediction error. *Nature Communications*, *8*, 2148. https://doi.org/10.1038/s41467-017-02038-6

Phukhachee, T., Maneewongvatana, S., Angsuwatanakul, T., Iramina, K., & Kaewkamnerdpong, B. (2019). Investigating the effect of intrinsic motivation on alpha desynchronization using sample entropy. *Entropy* 2019, *21*, 237–252. https://doi.org/10.3390/E21030237

Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology*, *118*, 2128–2148. https://doi.org/10.1016/J.CLINPH.2007.04.019

Quigley, C. (2021). Forgotten rhythms? Revisiting the first evidence for rhythms in cognition. *The European Journal of Neuroscience*, 1–9. https://doi.org/10.1111/ejn.15450

Rahman, M., Willmore, B. D. B., King, A. J., & Harper, N. S. (2020). Simple transformations capture auditory input to cortex. *Proceedings of the National Academy of Sciences*, *117*, 28442–28451. https://doi.org/10.1073/PNAS.1922033117

Richman, J. S., & Moorman, J. R. (2000). Physiological time-series analysis using approximate and sample entropy. *American Journal of Physiology-Heart and Circulatory Physiology*, *278*, H2039–H2049. https://doi.org/10.1152/ajpheart.2000.278.6.h2039

Rogan, M. T., Leon, K. S., Perez, D. L., & Kandel, E. R. (2005). Distinct neural signatures for safety and danger in the amygdala and striatum of the mouse. *Neuron*, *46*, 309–320. https://doi.org/10.1016/J.NEURON.2005.02.017/ATTACHMENT/019D5727-F78E-4AA3-A852-9650968F6AAC/MMC1.PDF

Rosburg, T., & Kreitschmann-Andermahr, I. (2016). The effects of ketamine on the mismatch negativity (MMN) in humans - a meta-analysis. *Clinical Neurophysiology*, *127*, 1387–1394. https://doi.org/10.1016/j.clinph.2015.10.062

Ruusuvirta, T., Penttonen, M., & Korhonen, T. (1998). Auditory cortical event-related potentials to pitch deviances in rats. *Neuroscience Letters*, *248*, 45–48. https://doi.org/10.1016/S0304-3940(98)00330-9

Sangha, S., Chadick, J. Z., & Janak, P. H. (2013). Safety encoding in the basal amygdala. *The Journal of Neuroscience*, *33*, 3744–3751. https://doi.org/10.1523/JNEUROSCI.3302-12.2013

Sara, S. J., & Bouret, S. (2012). Orienting and reorienting: The locus Coeruleus mediates cognition through arousal. *Neuron*, *76*, 130–141. https://doi.org/10.1016/J.NEURON.2012.09.011

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2013). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. 2nd Int. Conf. Learn. Represent. ICLR 2014 - Conf. Track proc.

Schröger, E., & Wolff, C. (1998). Attentional orienting and reorienting is indicated by human event-related brain potentials. *Neuroreport*, *9*, 3355–3358. https://doi.org/10.1097/00001756-199810260-00003

Shine, J. M. (2019). Neuromodulatory influences on integration and segregation in the brain. *Trends in Cognitive Sciences*, *23*, 572–583. https://doi.org/10.1016/j.tics.2019.04.002

Siegel, S. J., Connolly, P., Liang, Y., Lenox, R. H., Gur, R. E., Bilker, W. B., Kanes, S. J., & Turetsky, B. I. (2003). Effects of strain, novelty, and NMDA blockade on auditory-evoked potentials in mice. *Neuropsychopharmacology*, *28*, 675–682. https://doi.org/10.1038/sj.npp.1300087

Solomon, S. S., Tang, H., Sussman, E., & Kohn, A. (2021). Limited evidence for sensory prediction error responses in visual cortex of macaques and humans. *Cerebral Cortex*, *31*, 3136–3152. https://doi.org/10.1093/cercor/bhab014

Solyga, M., & Barkat, T. R. (2021). Emergence and function of cortical offset responses in sound termination detection. *eLife*, *10*, e72240. https://doi.org/10.7554/eLife.72240

Taaseh, N., Yaron, A., & Nelken, I. (2011). Stimulus-specific adaptation and deviance detection in the rat auditory cortex. *PLoS ONE*, *6*, e23369. https://doi.org/10.1371/JOURNAL.PONE.0023369

Takegata, R., Tervaniemi, M., Alku, P., Ylinen, S., & Näätänen, R. (2008). Parameter-specific modulation of the mismatch negativity to duration decrement and increment: Evidence for asymmetric processes. *Clinical Neurophysiology*, *119*, 1515–1523. https://doi.org/10.1016/j.clinph.2008.03.025

Todd, J., Michie, P. T., Schall, U., Karayanidis, F., Yabe, H., & Näätänen, R. (2008). Deviant matters: Duration, frequency, and intensity deviants reveal different patterns of mismatch negativity reduction in early and late schizophrenia. *Biological Psychiatry*, *63*, 58–64. https://doi.org/10.1016/j.biopsych.2007.02.016

Treder, M. S. (2020). MVPA-light: A classification and regression toolbox for multi-dimensional data. *Frontiers in Neuroscience*, *14*, 289. https://doi.org/10.3389/FNINS.2020.00289/BIBTEX

Umbricht, D., & Krljes, S. (2005). Mismatch negativity in schizophrenia: A meta-analysis. *Schizophrenia Research*, *76*, 1–23. https://doi.org/10.1016/j.schres.2004.12.002

Umbricht, D., Vyssotki, D., Latanov, A., Nitsch, R., & Lipp, H. P. (2005). Deviance-related electrophysiological activity in mice: Is there mismatch negativity in mice? Clin. *Neurophysiology*, *116*, 353–363. https://doi.org/10.1016/j.clinph.2004.08.015

Wacongne, C., Changeux, J. P., & Dehaene, S. (2012). A neuronal model of predictive coding accounting for the mismatch negativity. *The Journal of Neuroscience*, *32*, 3665–3678. https://doi.org/10.1523/JNEUROSCI.5003-11.2012

Yang, G. R., & Molano-Mazón, M. (2021). Towards the next generation of recurrent network models for cognitive neuroscience. *Current Opinion in Neurobiology*, *70*, 182–192. https://doi.org/10.1016/J.CONB.2021.10.015

How to cite this article: O'Reilly, J. A., Angsuwatanakul, T., & Wehrman, J. (2022). Decoding violated sensory expectations from the auditory cortex of anaesthetised mice: Hierarchical recurrent neural network depicts separate 'danger' and 'safety' units. *European Journal of Neuroscience*, *56*(3), 4154–4175. https://doi.org/10.1111/ejn.15736