Full title: **Inferring the neural basis of binaural phenomena with a modified autoencoder**

Short title: Inferring binaural mechanisms

Authors: Samuel S. Smith*[1], Joseph Sollini[1], Michael A. Akeroyd[1]

[1] Hearing Sciences, Mental Health and Clinical Neurosciences, School of Medicine, University of Nottingham, Nottingham, NG7 2RD, UK

* Corresponding author

Email: samuel.smith@nottingham.ac.uk

ABSTRACT

Broad organizational features of the auditory pathway have been replicated by artificial neural networks trained on diverse and naturalistic stimuli. It is credible that neural network analogues could also reveal specific neural computations underlying targeted auditory phenomena. This remains to be established. Here, we examined a modified autoencoder (a deep neural network akin to symbolic regression) trained to imitate an auditory behavior rooted in the binaural system. The behavior promoted the use of split-second interaural timing cues to improve the detection of a tone presented in noise. In the optimal network, we observed the emergence of specialized computations with prominent similarities to animal models. Artificial neurons developed a sensitivity to temporal delays that increased deeper into the network and were widely distributed in preference (extending to delays beyond the range permitted by head width), and the ensuing dynamics were consistent with a binaural cross-correlation mechanism. Our results attest to the generality of these solutions for performing signal detection at low frequencies. Moreover, this is a primary demonstration that deep learning methods can be used to infer tangible mechanisms underlying auditory perception.

AUTHOR SUMMARY

Do artificial intelligence (AI) systems process sounds in a similar way to the brain? If so, they represent a vital resource with which to investigate sensory systems. Here, we report an AI system that automatically developed auditory computations similar to those described in animal neurophysiology. The network learnt to use subtle discrepancies across its “two ears” to better detect a signal amongst a background noise. In the process, tuning to timing cues and established mechanisms of binaural hearing (relating to the use of both ears) developed internally. The results offer insight into computations occurring in deep brain structures that are difficult to measure with non-invasive techniques. Further, the results endorse informed associations made between AI systems and the auditory brain.

INTRODUCTION

The human-like 'behavior' achievable with deep neural networks (DNNs), combined with their unpremeditated network of computations, have seen them advocated as a new generation of model organisms [1]. These models are resource efficient, relatively easy to record from and perturb activity in, exempt to the usual ethical considerations of animal models, and potentially less restricted by species-specific ecology. However, the viability of DNN models as a means to reliably investigate the basis of sensory perception is of ongoing debate [2]. To evaluate this, evidence is required across multiple sensory modalities, methods, and scales of analysis.

A handful of recent studies have used DNNs to examine hierarchical organization in the auditory system [3–5]. These DNNs were found to replicate transformations in how information is broadly represented, such as a conversion from time-representations to rate-representations [5]. It remains to be established whether DNNs can also be used to uncover computations of the auditory brain on a finer scale. For example, can DNNs reveal the neural circuitry underlying specialized auditory behaviors of the type often targeted in animal studies? To address this lower level of analysis, we consider a modified autoencoder architecture [6,7] designed to promote mechanistic interpretation of a system by distilling its explanatory factors.

As a highly specialized sensory pathway [8], the binaural system represents an ideal context for evaluating the inferential abilities of this network architecture. We know that specialized binaural mechanisms are able to utilize interaural time differences (ITDs) less than one ten thousandth of a second to improve the detection of a signal in noise [9,10]. Yet, to date, the inner workings of DNN analogues configured to handle binaural audio have scarcely been considered [11–13]. Here, we trained a DNN, with a modified autoencoder architecture, to imitate the phenomena of binaural signal detection. The DNN was found to have striking similarities with neural dynamics (single- and population-level), and underlying mechanisms, described in animal models [14–17]. We find this to be encouraging evidence in support of DNNs as analogues of sensory processing.

RESULTS

Here, we examined whether a 5 layer DNN, with a modified variational autoencoder architecture (bottlenecking information through a small set of nodes in the central layer) [6,7], could be leveraged to learn about specific neural computations underlying audition. DNNs were trained to account for human behavior in a binaural signal detection task [9,10]. To augment the availability of training data, we generated data from a set of equations recognized as effective in predicting human binaural psychophysics [18–23]. DNNs were trained to mimic detection of a 500 Hz pure tone amongst a broadband noise, each with ITDs that varied trial-to-trial. The range of ITDs was restricted to fall within the human physiological range, i.e. as though they came from randomly chosen azimuthal locations in the real-world (**Fig 1A**). We found that the optimal DNN predicted unseen binaural detection data with a root mean square error of 2.5%. The predictions and internal dynamics of this DNN are the focus of this report (summary analytics across other DNN configurations are shown in **S1 Fig**).

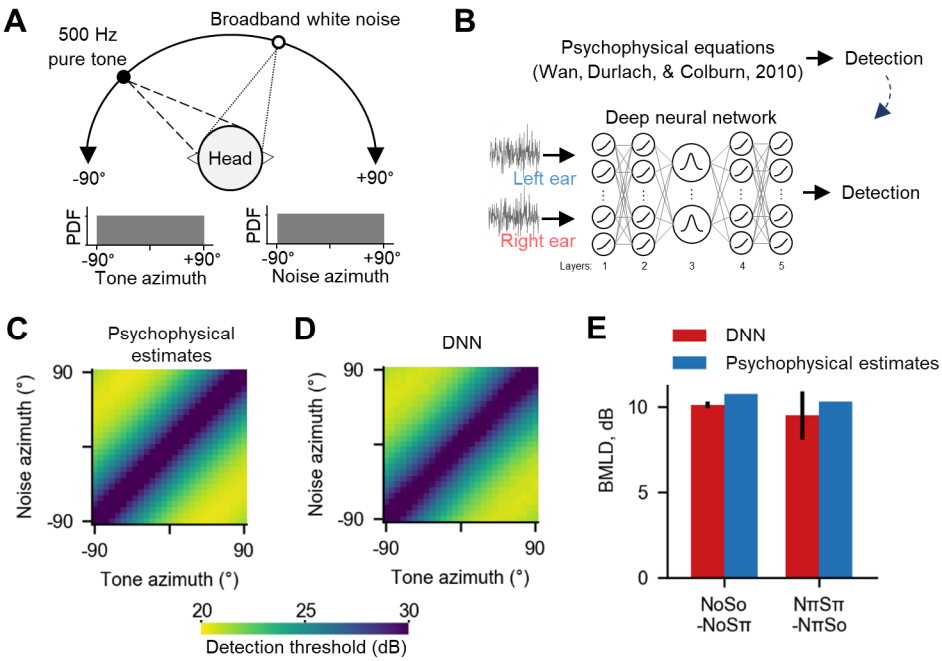


Fig 1. Modified autoencoder accounted for human binaural detection psychophysics.

(A) Data from a frontal field binaural detection task (generated using psychophysical equations) were used to train DNNs to detect a pure tone (black circle) in broadband noise (empty circle). Locations (and hence interaural time differences (ITDs)) of the tone and noise were chosen at random on each trial and were equally likely to come from each location (bottom panels). (B) The DNN was a 5-layer network with a low-dimensional central layer, i.e. layer 3, designed to promote simple representations of the input data. (C) Psychophysical estimates of human detection thresholds for a tone in noise, each with ITDs mapped from differing azimuths. (D) The DNN predictions for unseen testing data configured as in (C). (E) Binaural masking level differences (BMLDs) were derived for experimental stimulus configurations (NoSo/NoSπ, NπSπ/NπSo, Note: π is beyond the DNNs trained range). Error bars for DNN represent 95% confidence intervals.

Modified autoencoder accounted for human binaural detection psychophysics

As expected, we found that the DNN's detection thresholds (i.e. 69% correct performance) decreased as the difference between tone ITD and noise ITD increased (**Fig 1C,D**). For example, in diotic noise (noise ITD = 0) where the tone came from the far left, detection thresholds were significantly enhanced by 9 dB (from a maximum of 30.8 dB to 21.8 dB, two-sided unpaired t-test, $p \ll 0.0001$). To allow comparative assessment of the DNN and previously published data, we also tested the network on stimulus configurations typically employed to study binaural detection. These include tones and noise configurations which are either in-phase or completely out-of-phase across the ears. In the literature these stimuli are denoted as NoSo, NoSπ, NπSπ, NπSo, where N refers to the noise component, S the pure tone signal, with the successive subscripts denoting interaural phase difference (IPD) in radians. These stimuli incorporate ITDs that fall outside of the range permitted by head width. For example, a 500 Hz pure tone with an IPD of π corresponds to an ITD of 1000 μs, larger than that produced by the head size in the simulated training data (maximum of 655 μs [24]). This meant the DNN had no prior exposure to this size of ITD and so it was unclear how it would function over this range. We found that when the noise signal had zero IPD, the mean detection thresholds predicted by the DNN for corresponding homophasic (NoSo) and antiphasic (NoSπ) tones were 30.9 dB and 20.8 dB respectively. The gain in detection, commonly called the binaural masking level difference (BMLD), was 10.1 dB ($p \ll 0.0001$, two-sided unpaired t-test). Comparatively, when instead the noise signal was interaurally out-of-phase, the mean detection thresholds predicted for the corresponding homophasic (NπSπ) and antiphasic (NπSo) stimuli were 26.6 dB and 17.1 dB respectively. Their BMLD was 9.5 dB (two-sided unpaired t-test, $p \ll 0.0001$). These BMLDs are similar to those measured in people [25] and with estimates from psychophysical equations (10.7 dB and 10.3 dB respectively; **Fig 1E**).

Time delay tuning emerged early in the network

The DNN was able to account for key aspects of human binaural detection behavior. Given this, we next sought to understand the means by which the DNN derived this behavior,

i.e. its computations(s). A common property of animal binaural systems is ITD tuning (**Fig 2A**). We found that ITD tuning emerged within the early layers of the network. To demonstrate this we characterized noise delay functions in DNN nodes, i.e. their response to noises presented with varying ITDs as typically measured in animal physiology studies [26]. For nodes in the DNN's first layer, we observed significant ITD tuning in 63 out of 100 nodes (**Fig 2B**). Tuning was quantified by regressing a Gabor function [15], the combination of a cosine windowed by a Gaussian (overlaid in **Fig 2D**). By the DNN's second layer, significant ITD tuning had emerged in all 100 nodes (**Fig 2E**). Estimates of each nodes' best ITD (bITD) were inferred from Gabor fits (in order to account for nodes that were oscillatory in their noise delay responses, a form of phase-ambiguity noted in animal physiology [27], bITD was attributed to the most central tuning peak). In both the first and second layers of the DNN, we observed a wide distribution of bITDs, both within the simulated head-range, and beyond it.

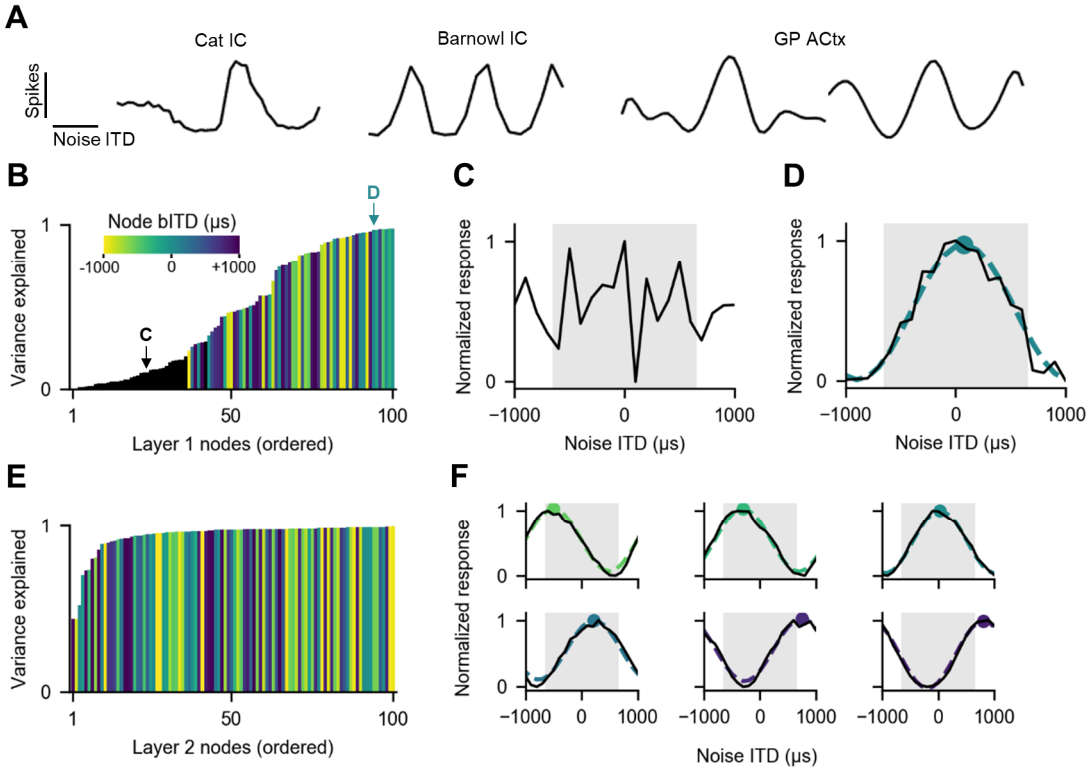


Fig 2. Time delay tuning emerged early in the network. (A) Neural ITD tuning curves have been observed in a number of animal species, including cat [26] (inferior colliculus, IC), barnowl [17] (IC), and guinea pig [14] (auditory cortex, GP ACtx). (B) ITD tuning emerged as a property of nodes within layer 1 of the DNN. ITD tuning was defined as the proportion of variance explained (R^2) by fits (Gabor functions) regressed to noise delay responses for nodes in the DNN's layers. Bars are color-coded by the nodes' best ITD (bITD; black indicates the Gabor fit was not significantly better than a linear fit). (C) Example of responses to noise ITDs in a layer 1 node with no clear tuning. The gray box underlays represent the ITD-limit for our simulation (modelled on the human head). (D) Example of ITD responses in a layer 1 node tuned near 0 μs. (E) ITD tuning was a dominant characteristic of nodes in layer 2. (F) Individual examples of ITD tuning within a selection of nodes in layer 2.

Network dynamics matched those of a cross-correlation mechanism

ITD tuning is an axiomatic property in many explanations of binaural signal detection. Therefore, in previous animal studies [14–17], clarity regarding the mechanism underlying binaural detection has been obtained by examining neural responses to a set of commonly employed tone-in-noise conditions (i.e. NoSo, NoSπ, NπSπ, NπSo). To examine the mechanism inferred by the DNN, we measured node activations in responses to the same stimuli. We found that in the DNN's second layer, node responses varied by bITD and the interaural phase of the noise presented (two-way ANOVA, $F[98,202]=12.5$ for main effect of bITD, $F[1,202]=31.2$ for main effect of noise phase, $F[98,202]=2.9$ for their interaction, $p \ll 0.0001$ for all). When a signal was presented amongst an in-phase noise (No), responses were largest for nodes with bITDs near 0 μs and decreased as bITDs were increasingly non-zero. Conversely, amongst an out-of-phase noise (Nπ), responses were lowest for nodes with bITDs near 0 μs and increased as bITDs deviated away from this (**Fig 3A**). The effects of tone phase on node dynamics were more subtle, although these dynamics were also in accordance with a nodes' tuning properties. Nodes tuned to smaller ITDs responded most to in-phase tones (So) and least to out-of-phase tones (Sπ), and vice-versa for nodes tuned to larger ITDs (**Fig 3B,C**).

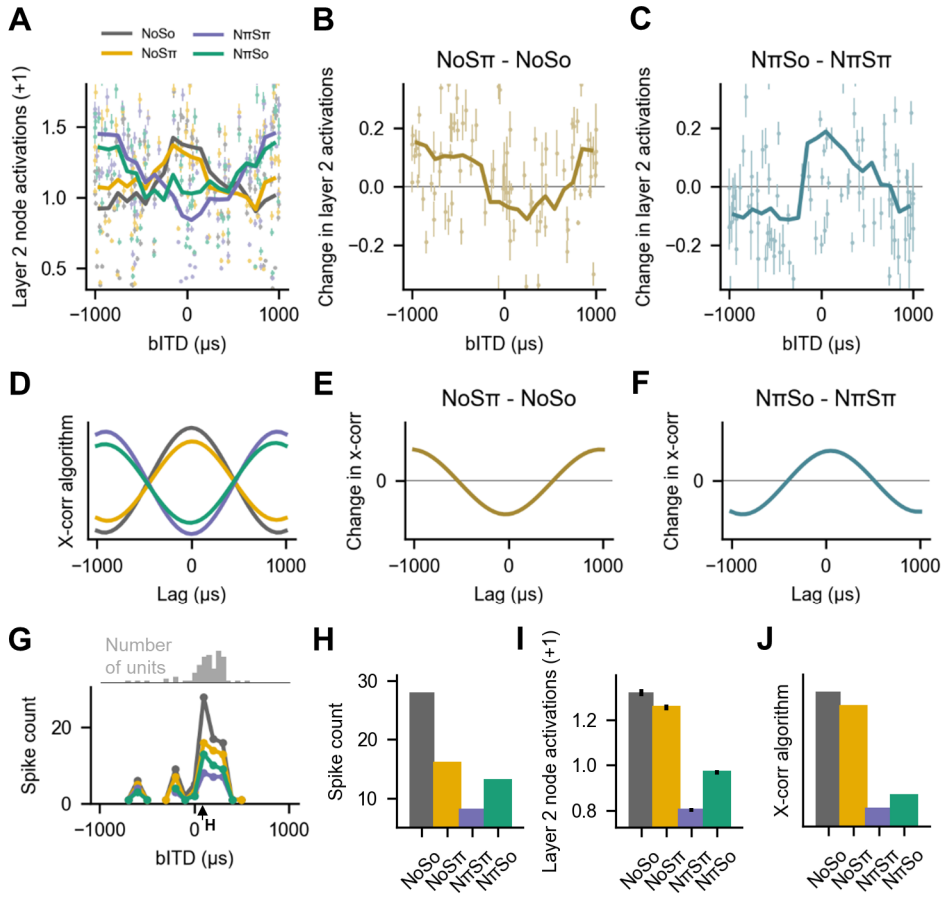


Fig 3. Network dynamics matched those of a cross-correlation mechanism of the type suggested in animal physiology. (A) The activations of layer 2 nodes were recorded for a set of binaural detection stimuli (NoSo, NoSπ, NπSπ, NπSo; color-coded). Single node data in lighter colors, where error bars represent 95% confidence intervals. Moving averages are overlaid and color-coded. (B) Change in layer 2 node activation for paired stimulus configuration NoSo/NoSπ (Error bars represent 95% confidence intervals. Moving averages are overlaid and color-coded). (C) Equivalent to (B) for the paired stimulus configuration NπSπ/NπSo. (D,E,F) These profiles matched that of a simple cross-correlation algorithm. (G) In addition, similar qualitative profiles have been observed in units sampled in the guinea pig animal model [14]. Unit histogram is presented at the top of the panel. (H) Spike counts, (I) node activations (moving average) and (J) cross-correlation algorithm for bITDs/lags near 100 μs (i.e. the bITD for which the most spikes overall were recorded).

These responses are commensurate with a binaural cross-correlation mechanism. The concept of binaural cross-correlation is predicated on the existence of coincidence detectors that encode temporally offset signals (similar to the dynamics already described in layers 1 and 2) [28]. Computationally, a binaural cross-product can be calculated by summing the point-by-point product of two temporally offset signals. Comparative outputs from a simple binaural cross-correlation algorithm (namely for signals in noise passed through narrow-band filters centered at 500 Hz) are shown in **Fig 3D-F**, and were found to resemble responses across the DNN's layer 2 nodes (Pearson's $r=0.36$, $p \ll 0.0001$). A number of animal studies have reported neural responses consistent with a cross-correlation mechanism [14–17], which were consistent with the DNN node responses (**Fig 3G-J**).

Early network ablation was detrimental to binaural detection

The functional importance of early layers in the DNN was further interrogated by inflicting targeted damage and observing knock-on effects to BMLDs. We set to zero the weights of a fixed proportion of nodes in a specific layer of the network, i.e. ablating them. Our ablation range varied from 0% (none) to 50% (lots). We then observed the corresponding NoSo-NoSπ BMLDs (**Fig 4**). We found that ablation to nodes in the DNN's first layer were most detrimental to BMLDs, to such an extent that with as little as 5% ablation the DNN failed to predict a significant BMLD (criterion set at $p<0.05$ with Bonferonni correction, Student's two-tailed t-test). When ablating the second layer, the DNN initially failed to predict a BMLD when 35% of nodes were ablated. Layers later in the network were more robust to the effects of ablation. BMLDs withstood ablations up to and including 50% (maximum tested) of the nodes in fourth layer, and significant BMLDs were not exhibited when 40% or more of the nodes in the fifth layer were ablated.

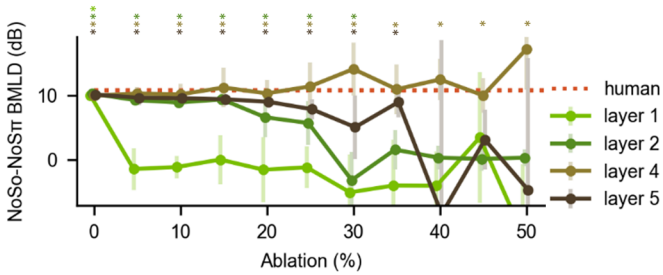


Fig 4. Early network ablation was detrimental to binaural detection. BMLDs (NoSo-NoSπ) were predicted following varying levels of random ablation (setting to zero) to nodes in the DNN's layers. Error bars represent 95% confidence interval. Color-coded asterisks indicate significant BMLDs ($p<0.05$ following Bonferonni correction).

Low-dimensional network representations imitated neural signature of population-level cortical activity

Responses in the DNN's second layer combine together to form the low-dimensional representations encoded in the DNN's central nodes (i.e. layer 3). This bottleneck (i.e. autoencoder) architecture has proven successful in extracting key explanatory variables in other systems [7], and our interest was for something analogous for binaural detection. In the DNN, six central nodes were deemed operational (**S1 Fig**), and we found noteworthy similarities between these nodes and population level cortical responses in the guinea pig [14] (comparable observations have also been reported in other animals [15,17] and in the guinea pig inferior colliculus [16]). The central nodes could take the value of any real number, positive or negative (a necessary limitation imposed by the network autoencoder architecture). Although not essential for our main conclusions, node responses are presented polarity-corrected to best correspond with noise ITD tuning in cortex (**Fig 5A,C**; guinea pig auditory cortex).

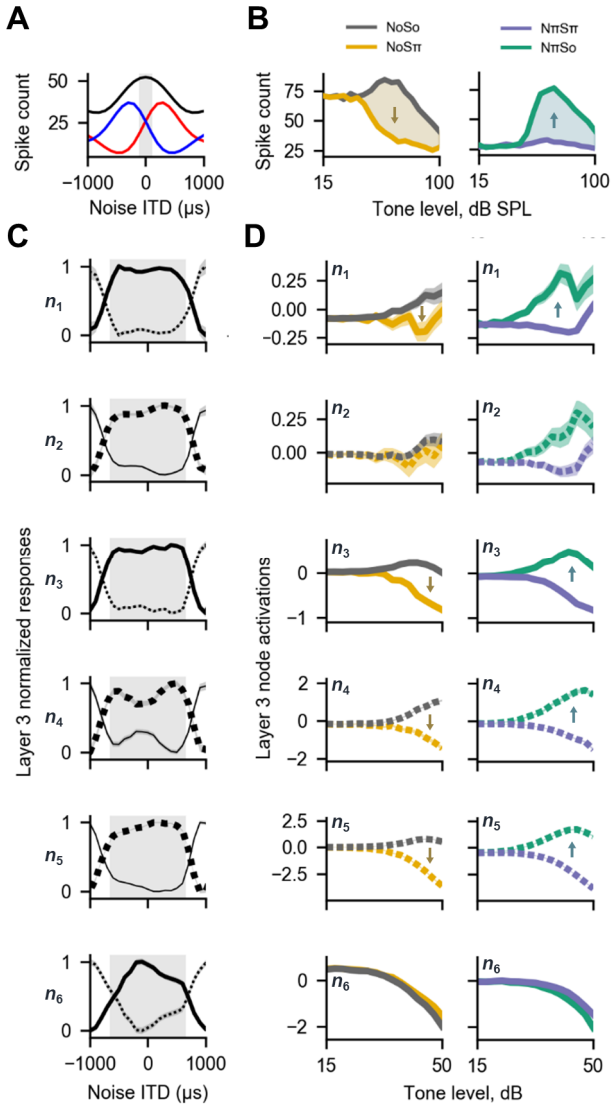


Fig 5. Low-dimensional network representations imitated neural signature of population-level cortical activity. (A) A noise delay function from a representative neuron in right guinea pig auditory cortex [14] (red line) alongside its reflection representative of left cortex (blue line) and their sum (black line). The gray box underlays represent the ITD-limit for the guinea pig. (B) Population masked rate-level functions recorded from guinea pig auditory cortex [14], in response to experimental binaural detection stimuli (NoSo, NoSπ, NπSπ, NπSo) are shown amongst arrows indicating changes as stimuli become more easily detectable. (C) Noise delay functions from operational nodes in the DNN's layer 3 (n_1-n_6). The gray box underlays represent the ITD-limit for an average human head. Dashed lines represent polarity corrected responses (such that noise delay functions in (**C**) are peaked as in (A)). (D). Masked rate-level functions in response to the same stimuli configurations as in (B, i.e. NoSo, NoSπ, NπSπ, NπSo). Dashed lines represent polarity corrected responses. Lightly shaded regions represent 95% confidence intervals.

An interesting feature of binaural processing can be seen when comparing detection behavior and animal physiological data. Similar improvements in behavioral performance can be attributed to completely different alterations in neural dynamics. For example, BMLDs across NoSo/NoSπ conditions are similar to those found across NπSπ/NπSo conditions (**Fig 1E**). However, guinea pig neural data [14] suggests different neural dynamics underlie these similar BMLDs. In cortical recordings, population spike counts dropped amongst an No signal, as a pure tone went from So to Sπ (left panel in **Fig 5B**). Conversely, amongst an Nπ signal, as a pure tone transitioned from Sπ to So, population spike counts increased (right panel in **Fig 5B**). These opposing dynamics therefore represent a unique signature of processing

The majority of layer 3 nodes displayed the same opposing dynamics in response to homophasic/antiphasic stimuli as observed in the guinea pig auditory cortex. Threshold responses in four latent nodes (n_1, n_3, n_4 and n_5, in left panels of **Fig 5D**) were found to be significantly lower in response to NoSπ relative to NoSo ($p<0.05$, two-sided unpaired t-test). Conversely, threshold responses in the same nodes to NπSo at a threshold level were significantly higher in comparison to NπSπ ($p<0.05$, two-sided unpaired t-test, right panels in **Fig 5D**). One latent node was qualitatively different to the others (node 6 of **Fig 5D**), and seemed to encode offsets related to the interaural phase of noise (No v. Npi, $p\ll0.0001$, two-sided unpaired t-test).

DISCUSSION

Deep neural networks as analogues of the auditory system

Recent DNN studies have described correlates with broad organizational principles in the auditory system [3–5]. Here, we set out to discover the efficacy of DNNs as a means to infer proximate mechanisms of audition underlying a targeted auditory phenomena. We modelled binaural hearing (namely, in the context of signal detection), representative of a highly specialized system for which the computations within DNN analogues have barely been examined [11–13]. We found that a relatively shallow DNN was able to successfully utilize binaural discrepancies in auditory stimuli, as opposed to seeking an alternative strategy [29] and/or failing to exhibit binaural detection behavior. Notably, similarities with animal single-unit and population-level neural responses were observed. ITD tuning, a well-known characteristic of binaural neurons, more often discussed in the context of sound localization [26], emerged early in the DNN. Additionally, the dynamics of second layer nodes, in response to tones presented in broadband noise, resembled a binaural cross-correlator [30]. Latent nodes, central within the DNN's autoencoder architecture, were compatible with the downstream dynamics of a cross-correlation mechanism and resembled guinea pig cortical neural recordings [14]. These dynamics were not hard-coded into the network and together highlight insights obtainable by treating DNNs as model organisms [1].

However, comparison between DNNs and neural biology come accompanied by an asterisk. The network was not constructed with the goal of accurately mimicking neuronal biophysics or hierarchical complexity, but instead a trade-off was made in which a modified autoencoder architecture was applied to facilitate interpretation and optimization. The inclusion of additional structural priors (e.g. hemispherity) and biologically inspired processes (e.g. spiking neural networks [3,31]) could well have merit, but any impact on interpretability should be weighed.

Human binaural detection

In addition to binaural detection serving as an ideal topic to interrogate DNN capabilities, it is a key function for successful communication in real-life environments. Consequently, binaural deficits have real-world consequences [8,32]. However, understanding of binaural detection in humans has been mired due to ambiguity regarding whether animal neurophysiology data satisfactorily accounts for human psychophysics. There is ongoing debate as to which of a number of theoretical frameworks best relate to human binaural detection [18,30,33,34]. For example, animal neural data lend support to a theory of binaural cross-correlation [14–17]. Whereas, human behavior appears to be equally well, if not better, described by a noise equalization and subtraction scheme [18,21,22]. These discrepancies have not been resolved with human imaging data [35–38], for which resolution and response variability are key limitations [39]. Our results indicate that, in a highly connected network, binaural cross-correlation functionality is selected to explain psychophysics.

Conclusion

In conclusion, our results demonstrate that a DNN, with a modified autoencoder architecture, can discover neural computations underlying auditory phenomena. These computations (namely, temporal delay tuning followed by a binaural cross-correlation mechanism) corroborate observations made in animal physiology. The work demonstrates the potential for deep learning methods, in bridging the interdisciplinary gap between psychophysics and neurophysiology, to clarify understanding of auditory perception.

METHODS AND MATERIALS

Training stimuli

Stimuli parameters were selected to maximize comparative opportunities with published experimental data. Pure tones were produced at a frequency of 500 Hz, and presented at levels between 0 and 50 dB. Pure tones were 20 ms long (10 periods) and produced with a sample rate of 20 kHz. Pure tones were masked by randomly distributed broadband noise (50-5000 Hz, limited by 6th order Butterworth bandpass filter) with an overall level of 60 dB. The tone and noise were gated simultaneously. Horizontal perception of space at low frequencies is largely based upon interaural time differences (ITDs) [40]. Given this, tones and noises were simulated with ITDs mapped from two independent angles in the azimuth between -90° (far left) and 90° (far right). ITDs were derived from Woodworth's equation [24], assuming a head radius of 0.0875 m.

Binaural detection rates and thresholds

The theory of equalization and cancellation [18] has wide human psychophysical support, successful in predicting binaural masking level difference (BMLD) data [18,33] and binaural pitch phenomena [18–20], underpinning other models of binaural hearing [21], and proven psychophysically favourable relative to other prominent theories [22]. Detection thresholds were calculated from phenomenological equations derived from this theory [18,23]:

$$D(\tau_S, \tau_N) = 31 - 10\log_{10}\max\left\{\frac{k-\cos(\omega_0\tau_S-\omega_0\tau_N)}{k-\gamma(\tau_N-\tau_0)}, 1\right\} \text{ dB} \quad (1)$$

where τ_s and τ_N are the interaural time lags of the signal and noise, ω_0 is the angular frequency of the pure tone signal, $k = (1+\sigma_\varepsilon^2)e^{\omega_0^2\sigma_\delta^2}$ where σ_ε^2 and σ_δ^2 are jitter (internal noise) parameters, γ is the normalized envelope of the autocorrelation of the narrow-band noise output of a filter centred at the target tone frequency, and τ_0 is an optimal time equalization parameter. The parameters were chosen according to Durlach's [18] original formulation, e.g. γ assumes a filter with triangular gain characteristics. Psychometric functions

were derived under the assumption that detection thresholds represented a d' of 1 in a yes-no experiment [41,42]:

$$R(a, D) = \Phi\left(0.501611 \times 10^{0.1(a-D)}\right) \quad (2)$$

where a is pure tone amplitude and D the detection threshold (**Equation 1**).

Deep neural network

We trained autoencoder-based DNNs [6,7] to predict the detection rates of tones presented amongst noise, each with varying ITDs. We based the architecture meta-parameters upon the DNNs described in [7]. Networks took 800 input values comprised of 400 samples from the left-ear waveform and 400 samples from the right-ear waveform. These inputs were passed through two 100 neuron exponential linear unit layers (ELU), referred to as the "decoder" portion of the network. This was followed by a layer of 10 latent Gaussian nodes ($\gg$ than the parameters varied in the generation of training stimuli) with minimal uncorrelated representations, constrained by a parameter β. This was followed by another two 100 neuron exponential linear unit layers, referred to as the "decision" portion of the network. All layers were fully connected and feed-forward. DNNs were trained and validated (95%/5% split respectively) on one million instances of a random phase tone at a random level (0-50 dB) in randomly generated white noise, both presented with ITDs mapped from random angles in the azimuth, and the corresponding detection rates (Equation 2).

The Adam optimization algorithm was used to minimize the cost function:

$$C_\beta(\hat{x}, x, \sigma, \mu) = \|\hat{x} - x\|_2^2 - \frac{\beta}{2}\sum_i \log\left(\sigma_i^2\right) - \mu_i^2 - \sigma_i^2 \quad (3)$$

where $\hat{x}$ and x are predicted and true detection rates respectively, and σ and μ are the standard deviation and mean of latent Gaussian nodes respectively. Batch size (number of training instances employed in each iterative update of network parameters) was set to 256. The learning rate (training hyperparameter) was set to 5×10^{-4} for 1000 epochs (total passes of entire training dataset). Ten DNNs were trained for each value of β, namely 0, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, and 10^{-2} (60 in total). The DNN with the least root mean square error, between predicted detection rates and ground truth, for the validation dataset, was selected for further analysis.

Central nodes were considered operational if the Kullback–Leibler divergence between their individual responses and a unit Gaussian was larger than 0.1 bits (**S1 Fig**).

Network detection thresholds

For a given stimulus configuration, DNN detection thresholds were obtained by calculating the mean of 10 detection rates across tone levels set between 0 and 50 dB in 2.5 dB steps and regressing a psychometric curve (Equation 2). This was repeated 10 times for a given stimulus configuration. Stimuli for which detection thresholds were derived included:

- random phase tones amongst randomly generated broadband noise with ITDs each mapped from fixed azimuthal locations,
- and random phase tones and randomly generated broadband noise each either in- or out-of-phase (i.e. NoSo, NoSπ, NπSπ, and NπSo).

Detection thresholds were also derived following ablations, where a set percentage of a given layers nodes were randomly nullified.

Node representations

Node activations were measured as a function of ITD for broadband noise (50-5000 Hz, 60 dB). ITDs ranged from -2000 μs to 2000 μs in steps of 100 μs. Node activations were also measured in response to So (in-phase) or Sπ (out-of-phase) signals masked by either No (in-phase) or Nπ (out-of-phase) broadband noise. Activations in layer 2 nodes were measured in response to a tone level of 35 dB amongst a noise level of 60 dB (in **Fig 3A**, activations were displayed with +1 added to their value because the minimum value of the ELU activation function is -1) [5]. In latent Gaussian nodes (in the central layer 3), masked rate-level functions to NoSo, NoSπ, NπSπ, and NπSo were measured amongst pure tone levels varied between 0 and 50 dB in 2.5 dB steps and 60 dB broadband noise. For all response measurements, stimuli were 20 ms long with a sample rate of 20 kHz. For a given stimulus configuration, activations were measured in response to 5000 random generations.

Binaural cross-correlation algorithm

For comparative purposes, outputs from a binaural cross-correlation algorithm were calculated [43]. The stimuli NoSo, NoSπ, NπSπ, and NπSo were generated for a 35 dB tone and a 60 dB randomly distributed broadband noise. Stimuli were sampled at 20 kHz and were 1 s in duration. Signals were passed through gammatone filters centered at 500 Hz and passed through a model of neural transduction [44]. The outputs were then delayed relative to one another, and the cross-products calculated and summated.

Statistical Analysis

ITD tuning was quantified by fitting a Gabor function [15] to noise delay responses. The parametric expression for a Gabor function is:

$$G = Ae^{-(ITD-bITD)^2/2s^2}\cos\big(2\pi F(ITD - bITD)\big) + C \tag{4}$$

in which we characterized a nodes' best ITD as the parameter $bITD$, F is the tuning curve frequency, A is a scaling factor (constrained to be positive), C is a constant offset, and s is a decay constant. These parameters were fit with the non-linear least squares algorithm *curve_fit* (a SciPy function[45]). An F-test was used to assess whether a Gabor function was a significantly better fit to noise delay responses than a linear function of ITD.

We performed Student's two-tailed t-tests (assuming unequal variance) to assess BMLDs and differences in node responses at threshold tone levels. We also used Student's two-tailed t-tests (assuming unequal variance) to assess BMLDs following network ablations, for which a Bonferonni correction was applied to offset the impact of testing multiple ablation rates. Pearson product-moment correlation was calculated between the average responses of nodes to NoSo, NoSπ, NπSπ, and NπSo, and the delay matched outputs of a binaural cross-correlation algorithm. A two-way ANOVA was also run for these nodes responses, with main effects of best ITD and noise phase. For the outlined statistical analyses, the criterion for significance (following multiple comparison corrections, when applied) was set to p=0.05. Error bars and lightly shaded underlays in figures are 95% confidence intervals.

Code Availability

Code generated during this study is available at https://github.com/HearingSciences/BinauralDetection_DNN.

ACKNOWLEDGEMENTS

S.S.S. and M.A.A. are supported by the Medical Research Council (grant number MR/S002898/1). J.S. is funded by a Nottingham Research Fellowship from the University of Nottingham. We are grateful for access to the University of Nottingham's Augusta High Performance Computer service. Thanks to Alan Palmer for assistance in sharing previously published physiology data. Thanks to Wilten Nicola for his valuable comments during the writing phase.

REFERENCES

1. Scholte HS. Fantastic DNimals and where to find them. NeuroImage. Academic Press Inc.; 2018. pp. 112–113. doi:10.1016/j.neuroimage.2017.12.077
2. Cichy RM, Kaiser D. Deep Neural Networks as Scientific Models. Trends in Cognitive Sciences. 2019. pp. 305–317. doi:10.1016/j.tics.2019.01.009
3. Khatami F, Escabí MA. Spiking network optimized for word recognition in noise predicts auditory system hierarchy. Cohen YE, editor. PLOS Comput Biol. 2020;16: e1007558. doi:10.1371/journal.pcbi.1007558
4. Kell AJE, Yamins DLK, Shook EN, Norman-Haignere S V., McDermott JH. A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. Neuron. 2018;98: 630-644.e16. doi:10.1016/j.neuron.2018.03.044
5. Koumura T, Terashima H, Furukawa S. Cascaded tuning to amplitude modulation for natural sound recognition. J Neurosci. 2019;39: 5517–5533. doi:10.1523/JNEUROSCI.2914-18.2019
6. Higgins I, Matthey L, Pal A, Burgess C, Glorot X, Botvinick M, et al. B-VAE: Learning basic visual concepts with a constrained variational framework. 5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings. 2017.
7. Iten R, Metger T, Wilming H, del Rio LL, Renner R. Discovering physical concepts with neural networks. Phys Rev Lett. 2018;124: 010508. doi:10.1103/PhysRevLett.124.010508
8. Schnupp JWH, Carr CE. On hearing with more than one ear: Lessons from evolution. Nature Neuroscience. 2009. pp. 692–697. doi:10.1038/nn.2325
9. Hirsh IJ, Burgeat M. Binaural Effects in Remote Masking. J Acoust Soc Am. 1958;30: 827–832. doi:10.1121/1.1909781
10. Hirsh IJ. The Influence of Interaural Phase on Interaural Summation and Inhibition. J Acoust Soc Am. 1948;20: 536–544. doi:10.1121/1.1906407

11. Adavanne S, Politis A, Nikunen J, Virtanen T. Sound Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Networks. IEEE J Sel Top Signal Process. 2018;13: 34–48. doi:10.1109/JSTSP.2018.2885636

12. Vecchiotti P, Ma N, Squartini S, Brown GJ. End-to-end Binaural Sound Localisation from the Raw Waveform. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. Institute of Electrical and Electronics Engineers Inc.; 2019. pp. 451–455. doi:10.1109/icassp.2019.8683732

13. Francl A, McDermott J. Deep neural network models of sound localization reveal how perception is adapted to real-world environments. bioRxiv. 2020; 2020.07.21.214486. doi:10.1101/2020.07.21.214486

14. Gilbert HJ, Shackleton TM, Krumbholz K, Palmer AR. The neural substrate for binaural masking level differences in the auditory cortex. J Neurosci. 2015;35: 209–220. doi:10.1523/JNEUROSCI.1131-14.2015

15. Lane CC, Delgutte B. Neural correlates and mechanisms of spatial release from masking: Single-unit and population responses in the inferior colliculus. J Neurophysiol. 2005;94: 1180–1198. doi:10.1152/jn.01112.2004

16. McAlpine D, Jiang D, Palmer AR. Binaural masking level differences in the inferior colliculus of the guinea pig. J Acoust Soc Am. 1996;100: 490–503. doi:10.1121/1.415862

17. Asadollahi A, Endler F, Nelken I, Wagner H. Neural correlates of binaural masking level difference in the inferior colliculus of the barn owl (Tyto alba). Eur J Neurosci. 2010;32: 606–618. doi:10.1111/j.1460-9568.2010.07313.x

18. Durlach NI. Equalization and Cancellation Theory of Binaural Masking-Level Differences. J Acoust Soc Am. 1963;35: 1206–1218. doi:10.1121/1.1918675

19. Klein MA, Hartmann WM. Binaural edge pitch. J Acoust Soc Am. 1981;70: 51–61. doi:10.1121/1.386581

20. Hartmann WM, McMillon CD. Binaural coherence edge pitch. J Acoust Soc Am. 2001;109: 294–305. doi:10.1121/1.1331680

21. Breebaart J, van de Par S, Kohlrausch A. Binaural processing model based on

contralateral inhibition. I. Model structure. J Acoust Soc Am. 2001;110: 1074–1088. doi:10.1121/1.1383297

22. Culling JF. Evidence specifically favoring the equalization-cancellation theory of binaural unmasking. J Acoust Soc Am. 2007;122: 2803. doi:10.1121/1.2785035
23. Wan R, Durlach NI, Colburn HS. Application of an extended equalization-cancellation model to speech intelligibility with spatially distributed maskers. J Acoust Soc Am. 2010;128: 3678–3690. doi:10.1121/1.3502458
24. Woodworth R, Barber B, Schlosberg H. Experimental psychology. 1954.
25. Durlach NI, Colburn HS. Binaural phenomena. In: Carterette EC, Friedman MP, editors. Handbook of perception, Vol IV, Hearing. New York: Academic; 1978. pp. 365–447.
26. Joris P, Yin TCT. A matter of time: internal delays in binaural processing. Trends in Neurosciences. Elsevier Current Trends; 2007. pp. 70–78. doi:10.1016/j.tins.2006.12.004
27. Singheiser M, Gutfreund Y, Wagner H. The representation of sound localization cues in the barn owl's inferior colliculus. Front Neural Circuits. 2012;6: 45. doi:10.3389/fncir.2012.00045
28. Jeffress LA. A place theory of sound localization. J Comp Physiol Psychol. 1948;41: 35–39. doi:10.1037/h0061495
29. Malhotra G, Evans BD, Bowers JS. Hiding a plane with a pixel: examining shape-bias in CNNs and the benefit of building in biological constraints. Vision Res. 2020;174: 57–68. doi:10.1016/j.visres.2020.04.013
30. Colburn HS. Theory of binaural interaction based on auditory-nerve data. II. Detection of tones in noise. Cit J Acoust Soc Am. 1977;61: 525. doi:10.1121/1.381294
31. Goodman DFM, Brette R. Learning to localise sounds with spiking neural networks. Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010, NIPS 2010. 2010. Available: http://staff.aist.go.jp/m.goto/RWC-MDB/
32. Ching TYC, van Wanrooy E, Dillon H, Carter L. Spatial release from masking in normal-hearing children and children who use hearing aids. J Acoust Soc Am. 2011;129: 368–

375. doi:10.1121/1.3523295

33. Domnitz RH, Colburn HS. Analysis of binaural detection models for dependence on interaural target parameters. J Acoust Soc Am. 1976;59: 598–601. doi:10.1121/1.380904

34. Dietz M, Lestang JH, Majdak P, Stern RM, Marquardt T, Ewert SD, et al. A framework for testing and comparing binaural models. Hear Res. 2018;360: 92–106. doi:10.1016/j.heares.2017.11.010

35. Wack DS, Cox JL, Schirda C V., Magnano CR, Sussman JE, Henderson D, et al. Functional Anatomy of the Masking Level Difference, an fMRI Study. Zhan W, editor. PLoS One. 2012;7: e41263. doi:10.1371/journal.pone.0041263

36. Wack DS, Polak P, Furuyama J, Burkard RF. Masking Level Differences – A Diffusion Tensor Imaging and Functional MRI Study. Zhan W, editor. PLoS One. 2014;9: e88466. doi:10.1371/journal.pone.0088466

37. Sasaki T, Kawase T, Nakasato N, Kanno A, Ogura M, Tominaga T, et al. Neuromagnetic evaluation of binaural unmasking. Neuroimage. 2005;25: 684–689. doi:10.1016/j.neuroimage.2004.11.030

38. Fowler CG. Electrophysiological evidence for the sources of the masking level difference. Journal of Speech, Language, and Hearing Research. 2017. pp. 2364–2374. doi:10.1044/2017_JSLHR-H-16-0251

39. Jiang D, McAlpine D, Palmer AR. Detectability index measures of binaural masking level difference across populations of inferior colliculus neurons. J Neurosci. 1997;17: 9331–9339. doi:10.1523/jneurosci.17-23-09331.1997

40. Rayleigh, Lord. XII. On our perception of sound direction . London, Edinburgh, Dublin Philos Mag J Sci. 1907;13: 214–232. doi:10.1080/14786440709463595

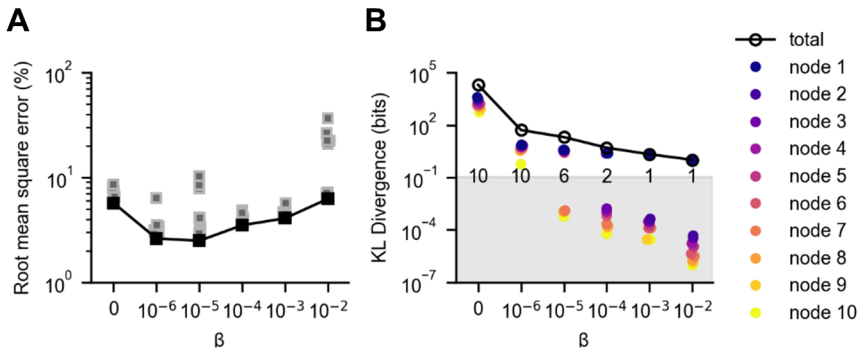
41. Egan JP, Lindner WA, Mcfadden4 D. Masking-level differences and the form of the psychometric functton'. 1969.

42. Ingleby JD. Signal detection theory and psychophysics. J Sound Vib. 1967;5: 519–521. doi:10.1016/0022-460x(67)90197-6

43. Akeroyd M. A binaural cross-correlogram toolbox for MATLAB. 2017.

44. Meddis R, Hewitt MJ, Shackleton TM. Implementation details of a computation model of the inner hair-cell/auditory-nerve synapse. J Acoust Soc Am. 1990;87: 1813–1816. doi:10.1121/1.399379

45. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17: 261–272.



S1 Fig. Meta-parameter search. (A) Prediction error for 60 DNNs (10 for each value of β, see Methods) tested on the validation dataset. The DNNs with the minimum error for each value of β are represented with large black squares. (B) Information transmission was investigated for the most accurate DNNs for each value of β. Displayed is the total Kullback–Leibler (KL) divergence between latent nodes in layer 3 and an isotropic Gaussian (empty black circles). The KL divergence between each individual node and a unit Gaussian is also shown (color coded in order). The gray region represents nodes deemed to be suppressed during training. The number of nodes with KL divergences above this region (out of 10) are typed on the upper border of this region.