Can Deep Learning Provide a Generalizable Model for Dynamic Sound Encoding in Auditory Cortex?

Jacob R. Pennington[1], Stephen V. David[2*]

[1]Washington State University, Vancouver, [2]Oregon Hearing Research Center, Oregon Health and Science University,

*Correspondence: SVD, 3181 SW Sam Jackson Park Road, MC L335A, Portland, OR 97239. Email: davids@ohsu.edu

JRP ORCID iD: https://orcid.org/0000-0001-9393-0474
SVD ORCID iD: https://orcid.org/0000-0003-4135-3104

Abstract

Recent advances in machine learning have shown that convolutional neural networks (CNNs) can provide powerful and flexible models of neural sensory processing. In the auditory system, the standard linear-nonlinear (LN) model and related variants are unable to account for high-order cortical representations. Thus far, it is unclear what additional insights can be provided by CNNs. CNNs can be difficult to fit with relatively small datasets, such as those available from single neuron recordings. In the current study, we developed a population encoding model: a CNN that simultaneously predicts activity of several hundred neurons recorded during presentation of a large, fixed set of natural sounds. This approach defines a spectro-temporal space that is shared by all neurons in the model and pools statistical power across the neural population. We tested a range of CNN architectures on data from primary and non-primary auditory cortex, varying the number and size of convolutional layers at the input and dense layers at the output. The population models performed consistently better than LN models. Moreover, performance was highly generalizable. The output layer of a model pre-fit using one population of neurons could be fit to novel single units, with performance close to that of neurons in the original fit data. These results indicate that population encoding models capture a general set of computations performed by auditory cortex. In this case, the model itself can be analyzed to provide a general characterization of auditory cortical representation.

Acknowledgments

The authors would like to thank Charles Heller and Mateo Lopez Espejo for their assistance with data collection for this study. We would also like to thank Nima Mesgarani and Menoua Keshishian for advice on fitting CNN models.

Introduction

A complete understanding of central neural sensory processing requires computational models that can account for responses to arbitrary natural stimuli (Wu et al., 2006). In the auditory system, encoding models such as the linear-nonlinear spectro-temporal receptive field (LN) model can predict some aspects of auditory neural tuning but often fail to predict time-varying responses to complex stimuli such as natural sounds (Machens et al., 2004; Espejo et al., 2019). Many variants of the LN model have been proposed that provide a more accurate characterization of auditory coding. Some of these variants build on traditional systems identification methods, accounting for second- and higher order nonlinearities (Eggermont, 1993; Atencio et al., 2008). Others have incorporated nonlinear elements derived directly from biological circuits, like short-term synaptic plasticity and gain control from local inhibitory populations (Rabinowitz et al., 2012; Espejo et al., 2019; Pennington and David, 2020). A third approach has been to use artificial neural networks, generalized models that combine large numbers of linear-nonlinear units (Harper et al., 2016). While theoretically appealing, large neural networks have proven challenging to fit, especially when data set size is limited. Direct comparisons between these different methods remain limited, as new models are typically compared only with the LN model as a baseline (but see (Harper et al., 2016; Keshishian et al., 2020; Pennington and David, 2020)). Thus far there is no clear consensus on what model architecture should replace the LN model as the standard for auditory neurons.

Recent advances in machine learning, in particular the development of backpropagation algorithms for convolutional neural networks (CNNs), have opened up the possibility of applying neural network analysis to electrophysiology datasets (Kell et al., 2018; Keshishian et al., 2020). CNNs have been adopted widely for signal processing problems, including speech recognition and other acoustic analysis (You et al., 2013; Chan et al., 2016). A small number of studies have indicated that CNNs can describe human BOLD fMRI data (Kell and McDermott, 2019) and ECoG data (Keshishian et al., 2020). In addition, CNNs have been applied to models of natural image representation in visual cortex (Cadieu et al., 2014). It remains an open question whether CNNs provide a useful characterization of single-neuron activity in auditory cortex.

One particular appeal of large CNNs is that they can serve as "foundation models" that can be applied to a wide range of new problems (Bommasani et al., 2021). Importantly, an effective foundation model can be applied to problems and data that were not included in the original dataset used for fitting. Such approaches are widely used for machine learning problems, including auditory (Devlin et al., 2019) and visual signal processing (Radford et al., 2021). The idea of pre-training has also been valuable for the analysis of neural data (Kell and McDermott, 2019). Motivated by the success of the foundation model approach, we argue that an effective CNN should be fully generalizable. A fully generalizable model should be able to account for the encoding properties of neurons that were not included in the original model fit. Moreover, it should be able to account for higher-order properties like sparse coding that are not captured by traditional LN models, but which have been hypothesized to be important for neural sensory coding (Barlow, 1961; Hromádka et al., 2008).

In the current study, we used CNNs to model the time-varying neural activity of a population of single-units in auditory cortex during presentation of a large natural sound library. To leverage the

statistical power across neurons, we developed a population encoding model in which the activity of many neurons was predicted by a single CNN with input layers shared across neurons. We found that population models were able to account for cortical activity substantially better than traditional LN models. In addition, the population models more closely reproduced the sparse coding properties of the neural data, and pre-trained versions of the models were able to successfully describe novel neural data. Thus, CNN models fit with large neural populations and diverse stimuli can provide a generalizable model of sound encoding by auditory cortex.

Methods

Data collection

All procedures were approved by the Oregon Health and Science University Institutional Animal Care and Use Committee and conform to standards of the Association for Assessment and Accreditation of Laboratory Animal Care (AAALAC).

Prior to experiments, ferrets (*Mustela putorius furo*, n = 5) were implanted with a custom steel head post to allow for stable recording. While under anesthesia (ketamine followed by isoflurane) and under sterile conditions, the skin and muscles on the top of the head were retracted from the central 3 cm diameter of skull. Several stainless-steel bone screws (Synthes, 6 mm) were attached to the skull, the head post was glued on the mid-line (Charisma), and the site was covered with bone cement (Charisma and/or Zimmer Palacos). After surgery, the skin around the implant was allowed to heal. Analgesics and antibiotics were administered under veterinary supervision until recovery.

After animals recovered from surgery and were habituated to a head-fixed posture, a small craniotomy (approximately 0.5 mm diameter) was opened over A1 or the secondary auditory field, PEG, immediately ventro-anterior to A1 (Fig. 1A, (Bizley et al., 2005; Atiani et al., 2014)). Neurophysiological activity was recorded using silicon multielectrode arrays (UCLA probes, Fig. 1B, (Shobe et al., 2015 p.2)). The array was inserted approximately normal to the cortical surface using a microdrive (Alpha-Omega Engineering EPS). Electrophysiological activity was amplified and digitized (Intan RHD-128) and recorded using open-source data acquisition software (OpenEphys). Recording site locations were confirmed as being in A1 or PEG based on tonotopy, frequency tuning and response latency (Kowalski et al., 1996; Bizley et al., 2005; Atiani et al., 2014).

Single- and multi-unit spiking events were extracted from the continuous, multichannel electrophysiological traces using Kilosort 2 (Pachitariu et al., 2016). Units were only kept for analysis if they maintained isolation and a stable firing rate over the course of the experiment. Unit isolation was quantified as the percent overlap of the spike waveform distribution with neighboring units and baseline activity. Isolation > 95% was considered a single unit, and isolation > 85% was considered a multi-unit (single units: A1, 567/849; PEG, 314/398). There was no significant difference in median prediction accuracy for any of the exemplar models between these groups in either brain area (U test, $p > 0.05$). Thus, we pooled single- and multi-unit data into a single population for this study, and we refer to these units as "neurons."

Stimulus presentation was controlled by custom software written in Matlab (Mathworks, R2017A). Digital acoustic signals were transformed to analog (National Instruments PCI6259) and amplified (Crown D-75a). Stimuli were presented through a flat-gain, free-field speaker (Manger) 80 cm distant, 0-deg elevation and 30-deg azimuth contralateral to the neurophysiological recording site. Prior to experiments, sound level was calibrated to a standard reference (Brüel & Kjær). Stimuli were presented at 60–65 dB SPL (peak-to-peak amplitude).

Natural sound stimuli

Data were collected during presentation of a library of natural sounds (595 1-sec samples, 0.5 sec ISI). Approximately 15% of these sounds were ferret vocalizations and environmental noises in the animal facility, recorded using a commercial digital recorder (44-KHz sampling, Tascam DR-400). Recordings included infant calls (1 week to 1 month of age), adult aggression calls, and adult play calls. No animals that produced the vocalizations in the stimulus library were used in the current study. The remaining 85% of sounds were drawn from a library of human speech, music and environmental noises developed to characterize natural sound statistics (McDermott et al., 2013). Activity was recorded during a single presentation of 577 samples and 20 repetitions of the remaining 18 samples. The low-repetition data were used for model estimation and the high-repetition data were used for model validation.

Modeling framework

For all analyses used in this study, spike data for each neuron was converted into a peristimulus time histogram (PSTH), $r(t)$, the time-varying spike rate, sampled at 100 Hz (10 ms bins). The input to each model consisted of a stimulus waveform converted into a log-spaced spectrogram, $s(f,t)$, using a second-order gammatone filter bank (Katsiamis et al., 2007). The filter bank included $F = 18$ filters with f_j spaced logarithmically from $f_{low} = 200$ to $f_{high} = 20{,}000$ Hz (approximately 1/3 octave per bin). The filter bank output was downsampled to 100 Hz to match the sampling of the neural PSTH. Log compression was also applied to the spectrogram, to further account for the action of the cochlea (Gill et al., 2006). Units in the subsequent model definitions correspond to these 1/3-octave frequency bins and 10 ms time bins.

<u>Linear-nonlinear models.</u> The linear-nonlinear spectro-temporal receptive field (LN) model is widely used in studies of neural auditory coding (Kowalski et al., 1996; Theunissen et al., 2001; Calabrese et al., 2011), and was used as a baseline for this study. The first stage of the LN model applied a finite impulse response (FIR) filter, h, to the stimulus spectrogram, to generate a linear firing rate prediction, r_{lin}:

$$r_{lin}(t) = \sum_{f}^{F} \sum_{u}^{U} h_{f,u} \; s(f,\, t-u)$$

The filter consisted of $F = 18$ spectral channels and $U = 15$ temporal bins (10 ms). In principle, this transformation can be applied to the spectrogram as a single 18x15 filter. In practice, the filter was implemented as a rank-4 factorization: projection onto an 18x4 spectral weighting matrix followed by convolution with a 4x15 temporal filter. This implementation substantially reduces the number of free parameters without sacrificing prediction accuracy (Thorson et al., 2015).

The output of the filtering operation was then used as the input to a static sigmoid nonlinearity that mimicked spike threshold and firing rate saturation to produce the final model prediction. For this study, we used a double exponential nonlinearity:

$$\mathrm{r(t)} = \mathrm{b} + \mathrm{a}\ \exp[-exp(k(y_{lin}(t) - s)]$$

where the baseline spike rate, saturated firing rate, firing threshold, and gain are represented by b, a, s and k, respectively (Thorson et al., 2015). This model predicts the activity of each neuron independently from the rest of the recorded data, but we implemented it in Tensorflow using custom layers such that we could run a single "model" for the full neural population (Abadi et al., 2016). This approach enabled a dramatic speedup in fit time compared to a one-fit-at-a-time strategy and allowed us to use the same optimization routine as the population models.

Single-neuron CNN models. The single-neuron CNN models used in this study have two layers (Fig. 1D). The first layer is a 1D convolutional layer, composed of the following operations: apply spectral weights to the input, convolve the weighted inputs with a 1x25 temporal filter, then apply an activation function (rectified linearity, ReLU) to the output of the convolution. The filter length allows the model to account for context effects over a period of up to 250 ms. The number of filters used determined the number of model architectures, with a total of six variations ranging from 4 to 18 filters. The second layer is a dense layer with a single unit, which forms the output.

For both the single-neuron and population versions of our neural network models, we used two kinds of activation functions. For all intermediate dense and convolutional layers, we used a ReLU activation function with an added parameter for input offset. For the final (output) dense layer, we used a double exponential nonlinearity as described for the LN models (Thorson et al., 2015).

Population CNN models. We focused on four major architectures of population CNN model (Fig. 1E): a single 1D convolutional layer with no activation function followed by one dense layer (pop LN), a single 1D convolutional layer followed by two dense layers (1D CNN), two 1D convolutional layers followed by two dense layers (1D CNNx2), or three 2D convolutional layers followed by two dense layers (2D CNN).

The pop LN model is similar to the LN model, in that the encoding properties of a single neuron can be collapsed into an LN model. The only difference is that the entire population shares a subspace defined by the 1D convolutional filters applied to the input, and weights for individual neurons are only computed in the final layer. We consider this the minimal change needed to convert the single-neuron LN model into a population model. The number of filters in the first layer ranged from 4 to 300, with 15 total variations. The number of units in the output layer is equal to the number of neurons in the population used for fitting (n = 849 for A1, n = 398 for PEG). Also note that while the pop LN model is technically a CNN, we will often refer to it as the LN model when contrasting its properties with the more complex CNN models.

The 1D CNN models have a similar architecture to the CNN single models, except there is a hidden dense layer after the 1D convolutional layer and the number of units in the output layer matches the number of neural responses in the dataset. We used sixteen model variations with between 5 and 230 filters, and between 10 and 300 hidden units.

The 1D CNNx2 models are akin to the 1D CNN models, but with two consecutive convolutional layers at the input instead of a single layer. The first layer used 1x15 filters while the second used 1x10 filters, yielding the same 250ms total "context memory" as a single 250 ms filter. The number

of filters in the first and second layers varied from 5 to 150 and 10 to 200, respectively, and the number of units in the hidden dense layer ranged from 20 to 250. A total of 14 parameter combinations was tested.

For the 2D CNN models, 1D convolutional layers are replaced with three consecutive 2D convolutional layers each using ten 3x8 filters, providing a cumulative 240 ms of context memory and approximately one octave of spectral information. The number of units in the hidden dense layer ranged from 4 to 300, with a total of 15 sizes tested.

To choose the structure of each model, we built progressively larger models out from the canonical single-neuron LN model by hand-selecting models with increasing size and number of network layers. For each model type and size, we evaluated a small number of wide-ranging hyperparameter combinations like learning rate and early stopping criteria, selected a combination that provided stable fits and good prediction accuracy, and then varied convolutional filter count and/or dense unit count to produce a continuum of model sizes. This was not an exhaustive exploration of the possible hyperparameters for the model architectures chosen, nor did we dive deeply into the full range of possible architectures.

A more extensive exploration of network hyperparameters like number and type of layers, dropout, and regularization should be beneficial. It remains unclear whether our population models can account for known response nonlinearities like short-term synaptic plasticity or gain control that can impact sensory activity in cortex (Ahveninen et al., 2011; Rabinowitz et al., 2011; David and Shamma, 2013; Harper et al., 2016; Lohse et al., 2020). Explicitly modeling effects like these within the neural network could lead to improved performance if the current CNN architecture is not able to account for them. The stimuli used for this study were not designed to probe these response dynamics, and thus additional experimental data are required to study these possibilities.

Model optimization

Model parameters were fit using Tensorflow's implementation of the Adam algorithm for stochastic gradient descent, using a mean squared error (MSE) loss function (Abadi et al., 2016; Kingma and Ba, 2017). Post-fitting performance was measured on a held-out validation dataset using prediction correlation, defined as the correlation coefficient (Pearson's R) between prediction and actual response, adjusted for uncertainty in the actual response due to the finite sampling of validation data (Hsu et al., 2004).

For each fit, model parameters were randomly initialized nine times, with a tenth initialization set to the mean of the parameter distributions. Each initialization was used to fit a subset of the model with the output nonlinearity excluded. The best submodel fit was then used to parameterize a subsequent full model fit. In past studies we found that this heuristic approach improved single-neuron model performance, and the same proved to be true for this study's population models (Thorson et al., 2015; Pennington and David, 2020). To mitigate overfitting, an early stopping criterion was set using twenty percent of the estimation set as temporary validation data. We also tested dropout and L2 regularization (Krogh and Hertz, 1991; Srivastava et al., 2014), but we found little to no benefit and chose to exclude them from this study for simplicity.

This fit procedure was repeated in two stages:

- First stage: Models were fit to all units from one brain region simultaneously (n = 849 A1; n = 398 PEG).
- Second stage: The output layer was re-fit to one unit at a time while parameters for earlier layers were kept fixed.

The second stage served two purposes. First, it provided a modest increase in model performance by optimizing the final layer parameters for each unit separately rather than optimizing a single loss function, produced by a weighted sum of the loss across neurons. Second, it ensured there were no differences in fitting process for the models in the generalization analyses.

For the test of generalization, the same two-stage procedure was used. However, in held-out models for each site S, the K neurons from that site were excluded from the first stage fit. The fit resulting from the first stage was then used to initialize fits for those K neurons in the second stage. In the case of the matched model, K neurons from other sites with prediction scores similar to those in S (for a single-neuron LN model) were excluded during the first stages. Thus, the matched model provided a control for the generalization test, in that the number of neurons used for fitting this model was the same as for the held-out model.

Exemplar models

After our initial exploration of a wide range of model sizes, subsequent analyses focused on exemplar models from each architecture. These exemplars were chosen such that each had similar complexity, and that complexity was as low as possible while keeping any reduction in prediction accuracy to a minimum. The architectural hyperparameters were as follows, in layer-order, where N represents the number of neural responses in the data:

- CNN single: 6 1D convolutional units, 1 output unit.
- pop LN: 120 1D convolutional units (250 ms), N output units.
- 1D CNN: 100 1D convolutional units (250 ms), 120 hidden dense units, N output units.
- 1D CNNx2: 70 1D convolutional units (150 ms), 80 1D convolutional units (100 ms), 100 hidden dense units, N output units.
- 2D CNN: 3 layers of 10 2D convolutional units each, in series (each 80 ms x 3 spectral bins, approximately 1 octave), 90 hidden dense units, N output units.

Equivalence analysis

Equivalence provides a quantitative measure of prediction similarity by computing the Pearson correlation coefficient between PSTHs for a pair of models for each neuron. For this study, we computed equivalence between a pair of 1D CNNx2 and pop LN models and for a pair of 1D CNNx2 and 2D CNN models. We then tested the significance of the median difference between distributions using the Wilcoxon signed-rank test. This is a relative comparison: if a distribution is shifted farther to the right, then that pair of models is more similar to each other than the other pair of models. This analysis was a variant of a previous method that computed the partial

correlation between PSTHs, relative to the prediction of a baseline model (Pennington and David, 2020). The simpler comparison used in this study introduces the limitation that it cannot be meaningfully applied to models that explain a high amount of variance in the data. In that case, the model predictions *must* be highly correlated with each other so a third, baseline model would be required to measure their differences.

Analysis of lifetime sparseness

Previous studies have argued that the brain encodes sensory stimuli with a sparse code, which provides a useful scheme for reading out sensory information downstream (Barlow, 1961; Vinje and Gallant, 2000; Hromádka et al., 2008; Willmore et al., 2011). We used a lifetime sparseness metric to determine how well different models were able to predict the sparseness properties of single units. The high repetition-count validation data provided a measure of sound evoked activity at each point in time, $r(t)$, and sparseness was computed from the distribution of these responses (Willmore et al., 2011):

$$S = 1 - \frac{[\sum_{t=1}^{T} r(t)]^2}{\sum_{t=1}^{T} r(t)^2}$$

According to this metric, neurons with zero response at most times and a non-zero response for only a few times will have high sparseness ($S = 1$), while neurons with many non-zero responses will have low sparseness ($S = 0$). To evaluate predicted sparseness, the time-varying model prediction was subjected to the same metric.

Calculation of neural responsiveness (SNR)

To generate subsets of data from A1 and PEG with balanced levels of auditory responsiveness, we assessed the reliability of auditory responses across repeated presentations of the same stimulus, using an approach modified from (Pennington and David, 2020). We refer to this reliability quantification as SNR and define it as follows. For each neuron, we quantified *total power*, T, as the variance of the time-varying spike rate, r_j, averaged across presentations of the same stimulus, $j = 1 \ldots m$,

$$\mathrm{T} = \frac{1}{m} \sum_{k=1}^{m} \langle r_j, r_j \rangle$$

where $\langle \ldots \rangle$ denotes a dot product. We then quantified *signal power*, S, as the average correlation between the time varying spike rate on different presentations of the same stimulus:

$$\mathrm{S} = \frac{1}{m} \sum_{k=1}^{m} \langle r_j, r_k \rangle, \ j \neq k$$

This process was repeated for each stimulus $i = 1...n$. SNR was then measured as the average ratio of signal power to total power across stimuli,

$$\mathrm{SNR} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\frac{S_i}{T_i}$$

Statistical methods

For all comparisons, we used non-parametric tests since the distributions of the relevant variables were non-normal and no other distribution was apparent. Statistical significance was assigned for $p < 0.05$. For paired tests, like comparing prediction correlations between different models on the same set of neurons, we used a two-sided Wilcoxon signed-rank test (referred to below as "signed-rank test"). For non-paired tests, we used a two-sided Mann-Whitney U test ("U test"). To determine whether a model's prediction was above chance for a given neuron, or if one model's prediction was significantly more accurate than another's for that neuron, we used a jack-knifed t-test. For the population analyses reported in Figs. 3-10, we only included data from neurons for which the 1D CNNx2, pop LN, and CNN single exemplar models all performed above chance. We considered neurons that met this criterion to be auditory-responsive (n = 777/849 A1 neurons, n = 339/398 PEG neurons) and assumed the excluded neurons were non-auditory.

A toolbox for systematic comparison of encoding models

All the models in this study were fit using the Neural Encoding Model System (NEMS, https://github.com/LBHB/NEMS). We have developed this open-source software package to be a flexible and extensible tool for fitting models to sensory neurophysiology data. Originally NEMS was focused on single-neuron encoding models, but through the course of this research we expanded it to include population models and additional computational backends like Tensorflow. Implementing all our models in a common framework helped us to eliminate unhelpful differences in our programming pipeline, like optimization routine and cost function evaluation, that otherwise might arise. We hope to continue adding new model paradigms to NEMS and invite contributions and suggestions from other researchers to help us make this software helpful and accessible for the broader neuroscience community.

Results

To comprehensively characterize the neural encoding of natural sounds, we recorded spiking activity from auditory cortex of awake, passively listening ferrets during presentation of a large natural sound library. Neural data was collected with 64-128 channel linear silicon arrays that recorded simultaneous activity of 10-95 single- and multi-units during each experiment. Recordings were performed in primary auditory cortex (A1) and a secondary auditory field (PEG) anterior-ventral to A1 (Bizley et al., 2005; Atiani et al., 2014). The same stimuli were used in all recordings (A1: 22 sites, 849 units, PEG: 11 sites, 398 units).

Convolutional neural networks with a shared tuning space for neural populations.

We used a convolutional neural network (CNN) to describe the functional relationship between the natural sound spectrogram and the time-varying neural activity. While deep learning tools have proven effective at a wide range of analytically similar problems, fitting these models requires large datasets. The amount of data available from many neurophysiological studies is limited, and this limitation is compounded by the fact that sensory-evoked neural activity is not reliable, varying substantially across repeated stimulus presentations. The current study took advantage of the fact that the same stimuli were presented during multiple experiments to fit a single CNN that simultaneously modeled the entire population of neurons being studied. In this framework, the input stimulus passed through a series of shared layers for all neurons. Only the last, fully connected layer computed a separate weighted sum to predict the activity of each neuron individually (Fig. 1E). Thus, the earlier layers comprised a shared, general model of sound processing in auditory cortex.

Within the population model framework, we evaluated several CNN architectures. CNNs used in visual processing problems typically apply multiple layers of small, two-dimensional (2D) convolutional kernels to an input image. We adapted this architecture to an auditory model, in which 2D filters are convolved along the time and frequency dimensions of the sound spectrogram (2D CNN, (Keshishian et al., 2020)). To link the CNN architecture directly to LN models typically used in the auditory system, we also developed a one-dimensional CNN model (1D CNN), in which filters were convolved only in time. These filters integrated with separate weights across frequency, specified by either a Gaussian function or a dense linear weighting (Thorson et al., 2015). Thus, the first layer of 1D CNN models closely resembled spectro-temporal receptive fields that define the LN model (Fig. 2A, E).

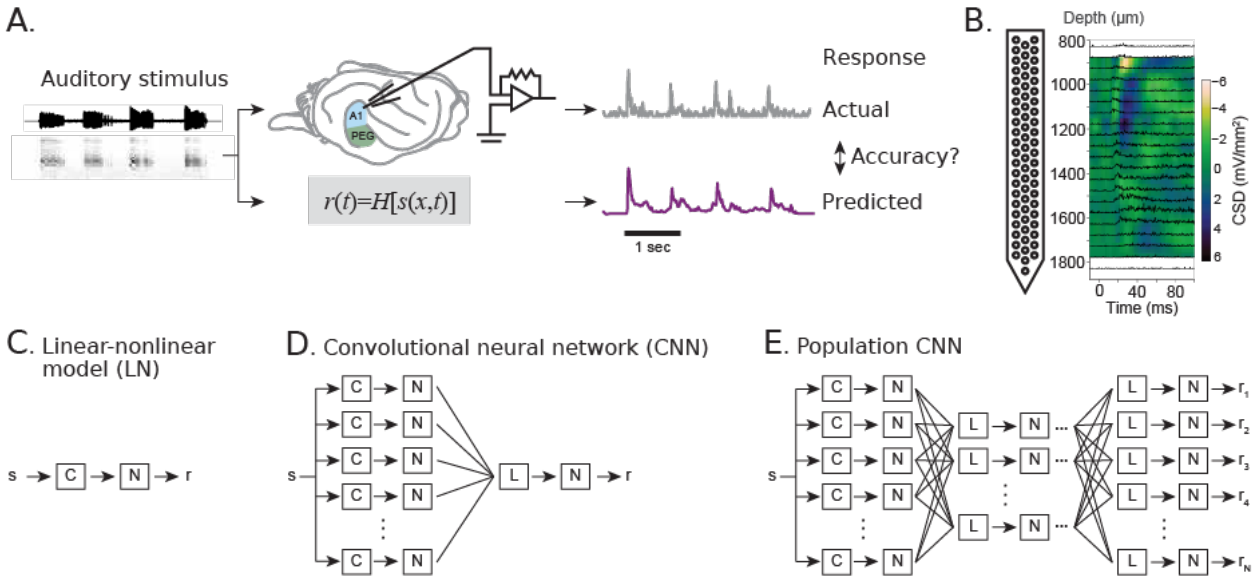

Figure 1. A. Schematic of experimental design. Encoding models were optimized to predict the actual time-varying neural spiking response to a library of natural sounds using the stimulus spectrograms as input. Neural data were collected from ferret primary auditory cortex (A1) and a secondary field (PEG) during passive listening. **B.** Multi-channel linear electrode arrays were used to record from 64-128 channels simultaneously. Current source density (CSD) plot from one recording illustrates sampling across layers of auditory cortex. **C.** Standard linear-nonlinear (LN) model consists of a single convolutional filter (C) followed by a static nonlinearity (N). As is the case for a traditional spectro-temporal receptive field, this convolution is one-dimensional: a separate filter is convolved in time for each frequency channel and then summed across frequency. **D.** Single-neuron convolutional neural network (CNN single) consists of a bank of LN units that are then combined linearly by a subsequent fully connected layer (L), followed by another static nonlinearity. **E.** Population CNN models consisted of one or more convolutional layers and one or model fully connected layers. Fitting the activity of many neurons simultaneously permitted much larger models than CNN single. Convolutional units were either one-dimensional (1D CNN) or two-dimensional (2D CNN).

We used the same dataset to fit several baseline models for comparison with the population CNNs. These included a standard linear-nonlinear model (LN, Fig 1C), which is widely used to describe the encoding properties of single neurons (Aertsen and Johannesma, 1981; Theunissen et al., 2001; Pennington and David, 2020). To control for possible benefits of increased statistical power from pooling data across neurons in a single model, we also fit a population LN model (pop LN), in which the activity of the recorded neural population was modeled simultaneously as the weighted sum of a shared bank of 1D convolutional filters. This model resembled the 1D CNN model, but it did not contain the nonlinear activation functions between layers. Finally, to distinguish population effects from improvements due to the neural network architecture, we fit a single-neuron CNN model (CNN single, Fig. 1D). Due to the sampling limitations of single unit data, these models contained substantially fewer units than the larger population models but maintained an architecture similar to the 1D CNN. In all cases, models were fit using standard back-propagation methods (Kingma and Ba, 2017), minimizing the mean-squared error between predicted and actual neural time-varying activity.

An example of a 1D CNN fit to the A1 dataset illustrates the comprehensive nature of the CNN models (Fig. 2). The filters in the first, convolutional layer are diverse and resemble spectro-temporal filters fit by a traditional LN model. These filters define a space of spectro-temporal

channels upon which processing by the subsequent dense layers depends. The positive and negative weights connecting the dense layers produce the time-varying output specific to each unit (Fig. 2C), which is compared to the actual neural activity (Fig. 2D). Prediction correlation varies between units but is close to a perfect score of 1.0 for many units (Fig. 2F).

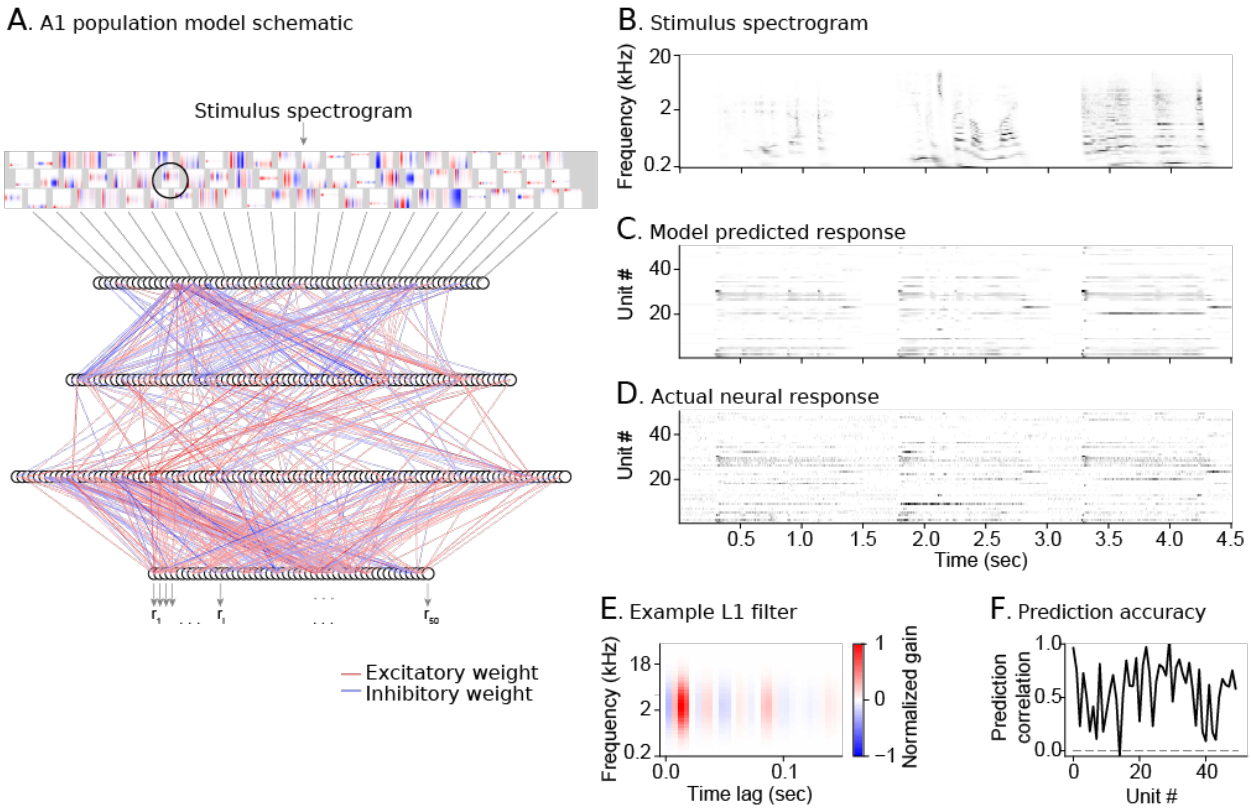

Figure 2. Example population CNN model fit. **A.** Schematic of fitted weights between each layer. Line color indicates the weight, with red indicating positive (excitatory) weights and blue indicating negative (inhibitory) weights. For compactness, data are shown for 50 out of the n = 849 A1 units fit in the output layer. **B**. Example spectrogram of three sounds in the validation set used to measure prediction accuracy. **C**. Heat map shows the predicted time-varying response of a unit in each row. **D**. Heat map of actual time-varying activity of the same neurons, plotted as in C. **E.** One filter from the first 1D convolutional layer of the population CNN model (circled in A). The filter resembles one typically observed in a traditional LN model fit to a single neuron. **F.** Noise-corrected prediction correlation for each unit, measured from the validation data.

CNN models consistently outperform standard LN models

For each model architecture, we fit variants of differing complexity by changing the number of units in one or more layers, which in turn varied the number of fit parameters. Note that a CNN single model with one filter in the first layer reduces to a standard LN model. Thus, by varying layer size, we explored a continuous space of models ranging from the LN model to much larger CNNs. For the LN model, we varied complexity by changing the rank of the convolutional filter (Thorson et al., 2015). For the 1D CNN models, filters were constrained to have rank 1, which consistently produced the best-performing models.

A Pareto plot compares the performance of models with different architectures and sizes for the A1 and PEG data (Fig. 3). For each fitted model, we assessed performance by measuring the prediction correlation—the noise-corrected Pearson correlation coefficient (Hsu et al., 2004)—between predicted and actual time-varying responses in a held-out validation set that was not used for fitting (Hsu et al., 2004; Thorson et al., 2015). A separate prediction correlation was computed for each neuron in the model. A value of 1.0 indicates that the neuron's activity was predicted as accurately as possible given the uncertainty in the actual response, and a correlation of 0 indicates chance performance. We identified exemplar models from each broad architectural group with approximately the same number of free parameters per unit and near-asymptotic performance (circled points, Fig. 3). Unless otherwise noted, subsequent analysis focuses on these models and the subset of auditory-responsive neurons, defined as those whose activity could be predicted above chance by the 1D CNNx2, pop LN and CNN single models ($p < 0.05$; A1: n = 777/849, PEG: n = 339/398).

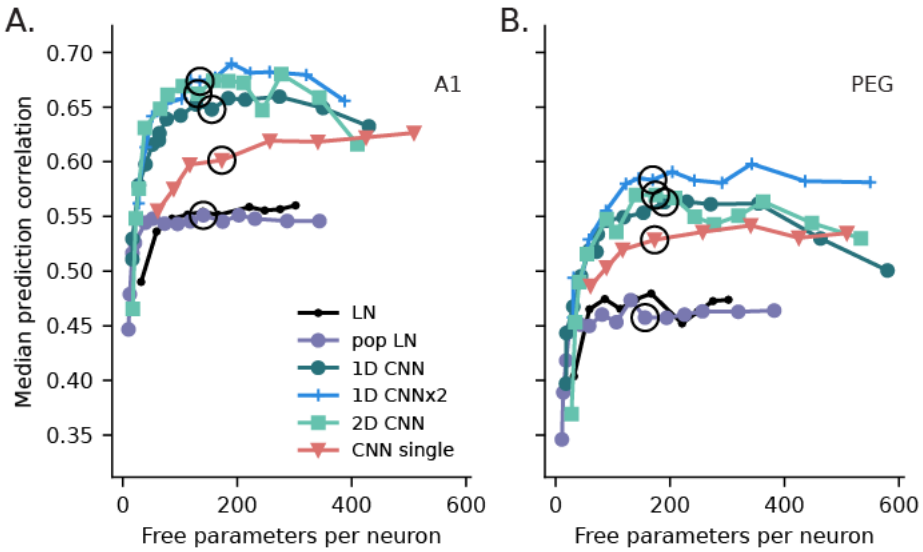

Figure 3. Pareto plots comparing performance of different model architectures of varying size. **A**. Parameter count versus median noise-corrected prediction correlation for each model in A1 (n = 777 auditory-responsive neurons). Lines connect models of different size for each of the six broad model architectures. Circles indicate exemplar models from each architecture with similar complexity that are examined in more detail below. For population models, parameter count is normalized by the number of neurons that were simultaneously fit. The CNN models maintain consistently higher prediction accuracy across a wide range of parameter counts. **B**. Pareto plot comparing model performance in PEG, plotted as in A (n = 339). Relative performance differences across model types are comparable to A1, but median prediction accuracy is consistently lower for all models.

First, we consider the results from A1 data. Prediction accuracy improved as model size increased and approached asymptotic performance at around 150-200 free parameters per neuron for most architectures. The CNN models tended to decrease in accuracy for the largest model sizes, presumably due to overfitting. When model complexity was matched, prediction correlation was higher for population CNN models than for other architectures. This difference can be seen by comparing performance of models requiring the same number of parameters per neuron (Fig. 3).

The greater accuracy of CNN models is consistent across the neural population: the 1D CNNx2 model predicted responses more accurately than the pop LN model ($p < 0.05$, jackknifed t-test) for almost half of auditory-responsive A1 neurons (380/777) while the opposite was true for only a handful of neurons (Fig. 4A). On average, the best 1D CNNx2 model accounted for 47% of the explainable variance in the A1 data, compared to 31% for the best LN model and 39% for the best CNN single model. The explanatory power of the 1D CNNx2 model is also greater than that of a nonlinear model that explicitly incorporates short-term plasticity and gain control, which accounted for 37% of explainable variance in a similar natural sound dataset (Pennington and David, 2020).

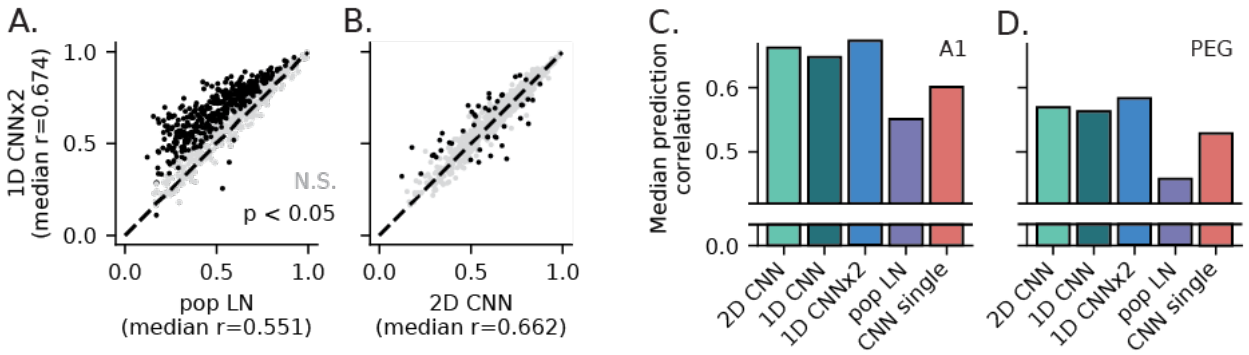

Figure 4. Performance summary for exemplar models. **A**. Scatter plot comparing prediction correlation for the pop LN and 1D CNNx2 models, for each neuron in the A1 dataset. The 1D CNNx2 model had a significantly higher prediction correlation for 380 of the n = 777 auditory-responsive neurons in the recording. **B**. Scatter plot comparing the 2D CNN and 1D CNNx2 models, plotted as in A. Prediction correlations were comparable in this case, but the 1D CNNx2 model still represents a small overall improvement (signed-rank test, $p = 1.10 \times 10^{-8}$). **C**. Median prediction accuracy in A1 for exemplar models representing each architecture. All differences were statistically significant (signed-rank test, 1D CNN vs. 2D CNN: $p = 9.21 \times 10^{-3}$, other comparisons: $p < 10^{-7}$). **D**. Prediction accuracy for PEG data, plotted as in C. In this case, the difference between 1D CNNx2 and 2D CNN was not significant (signed-rank test, $p = 0.883$), but all other differences were ($p < 10^{-4}$). Although overall prediction accuracy is lower for PEG than A1, the relative difference in performance between models is the same for both areas, and the 1D CNNx2 model is the best-performing in both.

In addition to their increased accuracy relative to LN models, the population CNN models performed better than the CNN single model. This result indicates that the improved prediction accuracy of the population models is not solely due to their neural network architecture but also reflects the gain in statistical power from pooling data across neurons. At the same time, the CNN single model did outperform both the LN and pop LN models. Thus, even with less data, the CNN architecture shows an improvement over traditional LN models. There is no theoretical constraint on the performance of the CNN single model relative to the population CNN models, and one would expect it to match their performance if more data were available. However, for the experimental datasets considered here, the population CNN models benefited from pooling across population data.

Among population models of similar complexity, prediction accuracy was quite similar, suggesting that the specific architecture was not critical to performance. However, adding an extra convolutional layer (1D CNNx2) did significantly improve the performance of the 1D CNN model

(signed-rank test, $p = 3.89 \times 10^{-18}$, Figs. 3, 4B). This modest but significant difference suggests that further improvements can be achieved by exploring CNN hyperparameters in more detail.

Prediction correlations for the LN and pop LN models were nearly identical, indicating that pooling data across neurons did not benefit prediction accuracy for this less complex model (signed-rank test between exemplar models; A1: $p = 0.133$, PEG: $p = 0.167$). We confirmed that the two models also produced highly equivalent PSTHs (data not shown). For subsequent analyses, we focused only on the pop LN model, since there are fewer differences between its architecture and the population CNNs.

The relationships described for A1 also held true for the PEG data. However, we observed two major differences between areas. First, median prediction correlations were consistently lower for PEG across all architectures and model sizes. We explored one possible reason for this difference in a subsequent analysis (see Fig. 10).

Population models generalize to novel datasets

We hypothesized that, unlike single-neuron models, population models capture an encoding subspace that is shared across neurons in the brain area that they model. Whether constrained by information bottlenecks in upstream neural pathways or by developmental plasticity from learning of behaviorally important sound features, this population subspace is much smaller than the space of all possible stimuli. We reasoned that if the models were indeed capturing this subspace, then they should be able generalize to novel data that were not included in the original model fit.

To explore this hypothesis, we re-fit the 1D CNNx2, pop LN, and CNN single models using two alternate approaches (Fig. 5A). In a "held-out" model, all data from one recording site was excluded from the first stage of the model fit. The output layer was then re-fit to each excluded neuron in the second stage while model parameters were kept fixed for all other layers. Thus, for the held-out model, all parameters except those in the final output layer were fit only using data from experiments that did not include the neuron being predicted. The "matched" model provided a control in which an equal number of neurons from other sites were excluded from the first stage of fitting. This ensured that the amount of fit data was matched to that of the held-out model, but in this case data from the predicted neuron's recording site was used to fit all layers (see Methods for more details).

When we compared the performance of the held-out and matched versions of the three models (Fig. 5B, C), we found that both population models generalized well: the difference in prediction accuracy between the two fitting strategies was not statistically significant (held-out vs matched, U test, A1; 1D CNNx2: $p = 0.482$; pop LN: $p = 0.994$; PEG; 1D CNNx2: $p = 0.440$; pop LN: $p = 0.947$). In contrast, the CNN single model generalized poorly: the matched fitting strategy performed substantially better in this case (A1: $p = 6.73 \times 10^{-91}$, PEG: $p = 6.49 \times 10^{-33}$).

This last outcome provided a control for the generalization test. The CNN single model is only fit to one neuron at a time, so it is unlikely to capture the sensory space encoded by a novel neuron. However, the single-unit model did generalize above chance, so it provided a useful baseline to

contrast with the ability of the population models to generalize. The finding that the pop LN model generalized to new data illustrates a broader value of the population modeling approach. The LN model spans a smaller domain of functional representations than the CNN models and thus does not predict activity as accurately. Still, the population model captures the auditory space spanned by this simpler model, permitting it to generalize to data from new neurons. This observation suggests that the population model approach can benefit analysis using any model architecture. Thus, it should also be able to identify encoding spaces for architectures that outperform those considered in the current study.

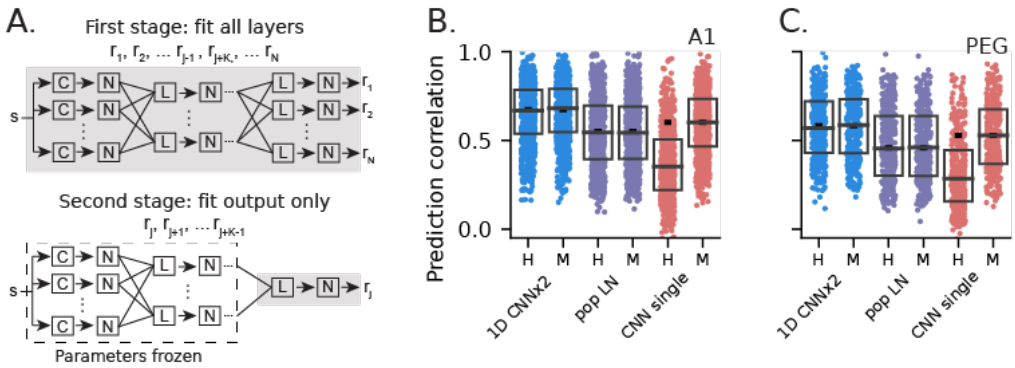

Figure 5. Generalization of model performance to novel neurons. **A.** Schematic of two-stage fitting process for held-out and matched models. For the held-out model, all K neurons from a recording site are excluded from the first stage fit. All parameters except those in the final layer are frozen (dashed blue box) while fitting the K individual responses in the second stage. For the matched model, K neurons from other sites with similar prediction accuracy are excluded during the first stage. **B**. Prediction accuracy of A1 held-out (H) and matched (M) models. Boxes show the 1st, 2nd and 3rd quartiles, and the small horizontal dash shows the median performance for the full model (Fig. 3C). The difference in prediction accuracy between held-out and matched was insignificant for the population models (U test, 1D CNNx2: $p = 0.482$; pop LN: $p = 0.994$), indicating that they generalize well to novel data. In contrast, there was a substantial decrease in prediction accuracy for the CNN single model ($p = 6.73 \times 10^{-91}$). **C**. Generalization for PEG neurons, plotted as in B. Once again, performance was the same for held-out and matched population models but significantly decreased for the held-out CNN single model (U test, 1D CNNx2: $p = 0.440$; pop LN: $p = 0.947$; CNN single: $p = 6.49 \times 10^{-33}$).

Given the ability of a pre-trained model to generalize to new data, we hypothesized that such a model should also be beneficial to the analysis of smaller datasets that measure neural responses to fewer auditory stimuli. The amount of data available from typical neurophysiological recordings is often limited, relative to the large datasets typically required for CNN models. This problem is especially acute in studies of animal behavior, where data is limited by the number of trials an animal is motivated to perform during a single recording session. To evaluate the utility of the population CNN for smaller datasets, we used an approach related to the held-out versus matched analysis (see Fig. 5).

For a given neuron, spiking data was subsampled over a range of 10-100% of the original dataset. A 1D CNNx2 held-out model was pre-trained on 100% of the data from all recording sites, excluding as above the site of the current neuron. The output layer was then fit using only the subsampled data. In other words, the model drew on a much larger dataset for fitting the initial

layers but only used the smaller temporal sample of data from the current neuron for fitting the final layer. We compared this model to a standard model for which both stages of fitting used data from the entire neural population, including data from the current neuron's recording site, but only using the fraction of data included in the subsample. Performance of models fit with the smaller datasets is quite variable, but the pre-trained held-out models performed better than the standard model on average (Fig. 6A, signed-rank test, $p = 2.91 \times 10^{-21}$). The benefit of pre-training extended across all subsamples tested (Fig. 6B, signed-rank test, $p < 0.05$).

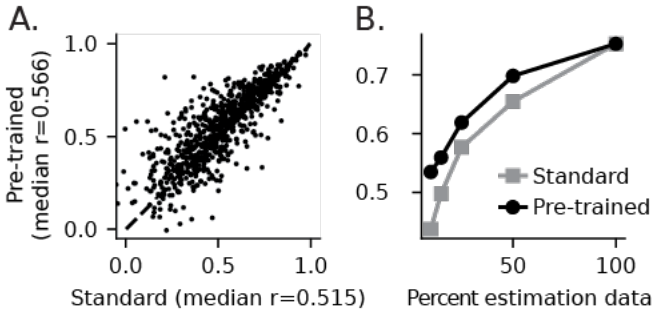

Figure 6. Generalization of a pre-trained 1D CNNx2 model to smaller datasets. The pre-trained model was fit to every stimulus but excluded neurons from one site. The output layer was then re-fit to an excluded neuron using a fraction of the available data. The standard model was fit to all neurons but only a subset of stimuli was used for the entire fit. **A.** Scatter plot compares prediction correlation between the pre-trained and standard models using 10% of available data. On average, the pre-trained model more accurately predicted the subsampled data (signed-rank test, $p = 2.91 \times 10^{-21}$). **B.** Curves compare median prediction correlation for pre-trained and standard models fit to subsampled data. Improved accuracy of the pre-trained model was consistent across all subsample sizes (signed-rank test, $p < 0.05$). Performance converges when all stimuli are used to fit both models.

Population CNN models show a high degree of equivalence

A striking observation in our performance analysis is that the 1D CNNx2 and 2D CNN models predicted activity with nearly equal accuracy (Fig. 4B). Given the differences in architecture, it was not immediately clear whether both models captured the same functional properties or if their improvements over the LN model instead reflected their ability to capture distinct aspects of the neurons' function. Qualitative inspection of PSTHs predicted by the 1D CNNx2 and 2D CNN models suggested that they produced similar predictions, which were visibly distinct from the pop LN model prediction (Fig. 7).

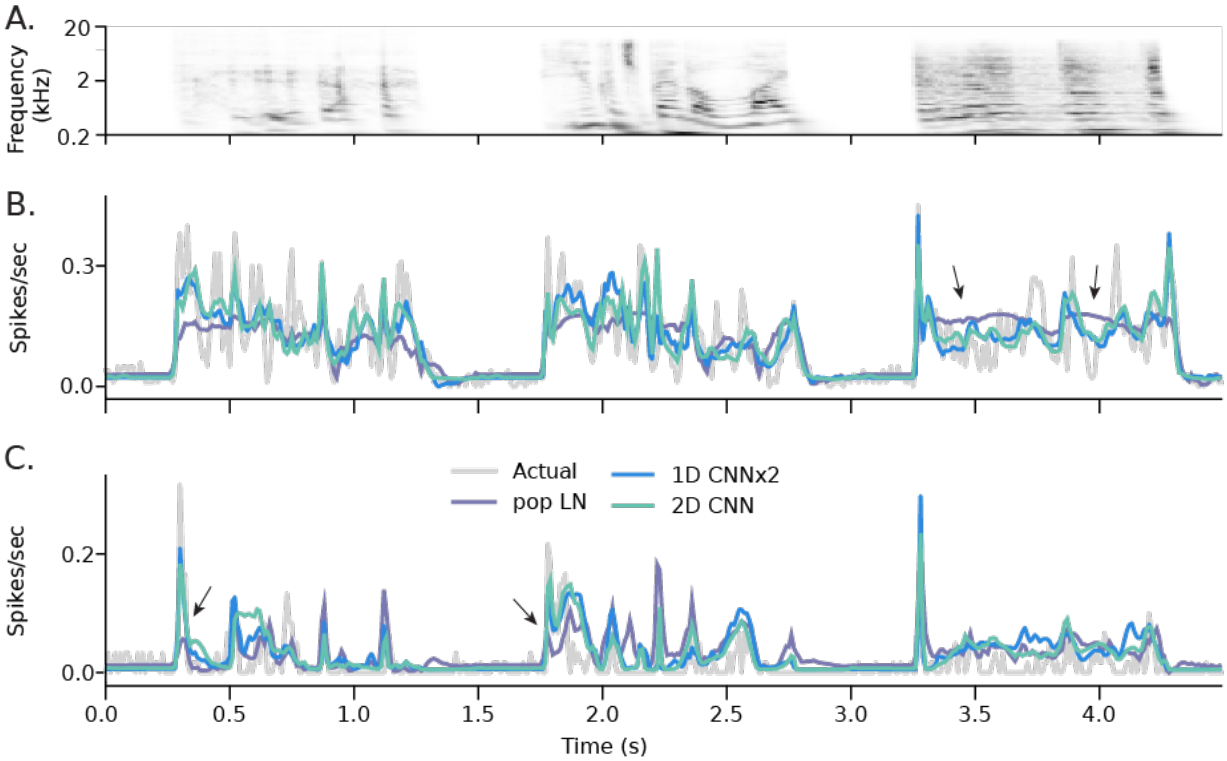

Figure 7. Examples of actual versus predicted PSTHs for two neurons from Figure 2. **A.** Segment of stimulus spectrogram from validation dataset used for testing prediction accuracy. **B.** Actual PSTH (gray) overlaid with the PSTHs predicted by the 1D CNNx2 (light blue), 2D CNN (teal), and pop LN (purple) models. Both CNN models have similar predictions that correlate well with the actual response (1D CNNx2: r = 0.812, 2D CNN: r = 0.800). The pop LN model fails to predict many of the features in the actual PSTH (r = 0.590, arrows, 3-5 sec). C. Actual and predicted PSTH for a second neuron. Again, both CNN models have higher prediction accuracy and can predict large transient responses better than the LN model (1D CNNx2: r = 0.729, 2D CNN: r = 0.748, pop LN: r = 0.535; arrows, 0.25, 1.75 sec).

To quantify prediction similarity between models, we computed model "equivalence" as the Pearson correlation coefficient between PSTHs predicted for each neuron (Fig. 8, (Pennington and David, 2020)). The 1D CNNx2 model had much closer equivalence to the 2D CNN model than to the pop LN model (signed-rank test, A1: $p = 1.47 \times 10^{-128}$, PEG: $p = 1.64 \times 10^{-55}$). This similarity of time-varying predictions indicates that there is indeed substantial overlap in the computations each CNN uses to predict neural responses.

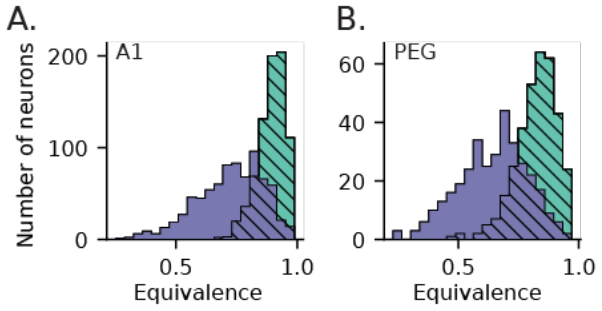

Figure 8. Quantification of equivalence between CNN and pop LN exemplar models. **A**. Histogram of equivalence (correlation between predicted PSTHs) between the 2D CNN and 1D CNNx2 models (teal, hatched) and between the 1D CNNx2 and pop LN models (purple) for A1 neurons. Equivalence is significantly greater between the two CNN models than between the 1D CNNx2 and pop LN models (signed-rank test, $p = 1.47 \times 10^{-128}$). This result indicates that CNN models achieve higher prediction accuracy over the LN architectures in similar ways. **B.** As A, but for PEG neurons. Here again we observed higher median equivalence between the CNN models ($p = 1.64 \times 10^{-55}$).

CNN models predict sparser responses

A model that provides a fully general characterization of neural encoding should account for higher order statistical properties beyond basic selectivity for spectro-temporal features. One characteristic property of auditory cortex and other sensory brain areas is lifetime sparseness, which measures the frequency at which stimuli evoke changes in spike rate from spontaneous levels (Vinje and Gallant, 2000; Hromádka et al., 2008; Willmore et al., 2011). As sparseness increases, neurons become selective for progressively smaller subsets of stimuli. A sparse code has been proposed as an efficient strategy for exposing important features in complex natural stimuli (Barlow, 1961), but linear filters tend to predict much less sparsity than is observed in actual neural responses.

To evaluate the ability of CNN models to account for sparse activity in auditory cortex, we computed lifetime sparseness for predictions by the pop LN and 1D CNNx2 models, as well as for the actual PSTH response of the same neurons. We excluded neurons with low prediction accuracy to ensure an accurate measure of sparseness ($r < 0.2$; A1: $n = 750$, PEG: $n = 316$). Histograms of spike rate for an example neuron illustrate that the CNN model matches the distribution of actual spiking more closely than the LN model (Fig. 9A-B). In this example, the CNN prediction histogram peaks at 0 spikes/sec, while the LN model has a peak at 1 spike/sec. On average, the CNN model is less sparse than the actual data but predicts consistently sparser responses than the pop LN model in both A1 and PEG (Fig. 9C, D, signed-rank test, A1: $p = 1.72 \times 10^{-98}$, PEG: $p = 1.81 \times 10^{-41}$). Thus, in addition to accounting for variance in evoked activity, the CNN model is better able to capture a higher order property of neurons in auditory cortex.

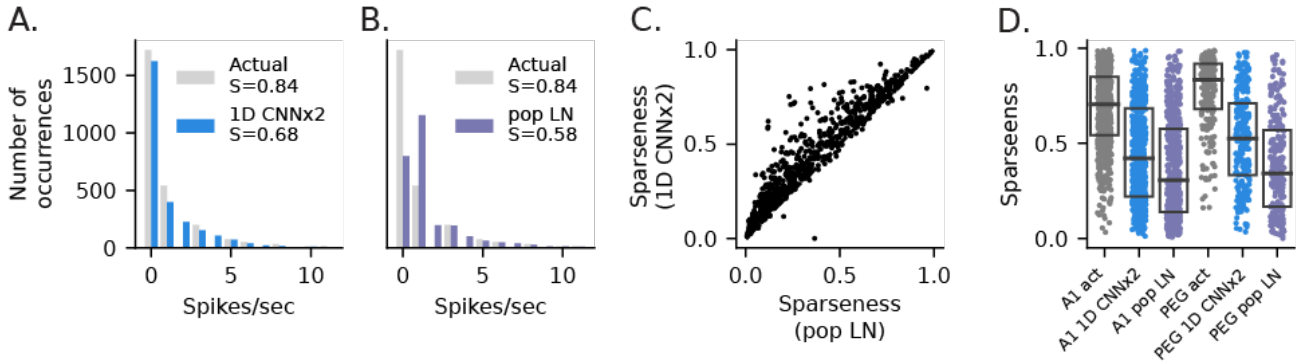

Figure 9. CNN models predict sparse auditory selectivity. **A.** Histogram of actual spike rate for one unit during the entire course of natural sound presentation in the validation data (gray) and rate predicted by the 1D CNNx2 model (light blue) reveals a similar distribution of responses, peaking at zero spikes/sec. **B.** Histogram of actual response and prediction by the pop LN model, plotted as in A. The LN distribution peaks at 1 spike/sec, and the distribution is not as well matched to the actual data. **C.** Scatter plot compares lifetime sparseness predicted by the pop LN and 1D CNNx2 models for 750 A1 neurons. CNN predictions are consistently sparser. **D.** Distribution of actual sparseness and sparseness predicted by the two models for A1 and PEG data. The 1D CNNx2 model does not match the sparseness of the actual data but predicts greater sparseness than the pop LN model (U test, A1: $p = 1.72 \times 10^{-98}$, n = 750 PEG: $p = 1.81 \times 10^{-41}$, n = 316).

Reduced SNR in PEG only partially explains decreased model performance.

Sensory responses of PEG neurons tend to be noisier, or less reliable, than A1 responses (Atiani et al., 2014), which could explain the difference in prediction accuracy between the two brain regions (Fig. 4). To test this possibility, we measured the reliability of neural responses to repeated presentations of identical stimuli (SNR) and selected the largest possible subset of units from each brain region for which binned distributions of SNR scores for the two subsets were identical (Fig. 10). Median SNR was indeed higher for A1 (md = 0.0989, U test: $p = 2.49 \times 10^{-6}$) than for PEG (md = 0.0713), and the A1 subset had a lower median SNR (md = 0.0709) while the median for the PEG subset was mostly unchanged (md = 0.0710).

When we compared prediction accuracy between these SNR-matched subsets, we found that all models still predicted A1 responses better than PEG responses (U test; 1D CNNx2: $p = 6.69 \times 10^{-4}$; pop LN: $p = 6.74 \times 10^{-4}$; CNN single: $p = 4.78 \times 10^{-4}$). However, the gap in median prediction correlation was smaller (median difference, 1D CNNx2: 0.0535, pop LN: 0.0420, CNN single: 0.0276) than for the full datasets (1D CNNx2: 0.0907, pop LN: 0.0936, CNN single: 0.0728), indicating that the decreased SNR of PEG neural responses does account for some of the reduced prediction accuracy. However, the persistence of the difference in performance after matching SNR indicates that a greater portion of PEG dynamics is not captured by the models in this study.

Generalizability extends across brain regions.

Based on the observation that population models generalize well to new data within cortical region, we also asked how well a model pre-trained on A1 data could generalize to data from PEG. We used the held-out approach to pre-train a 1D CNNx2 model on A1 data, and then fit the output

layer using data from PEG neurons in the matched-SNR subset. We then compared the prediction accuracy of this model ("cross-area held-out") to the PEG held-out 1D CNNx2 model from above (Fig. 5). The important distinction is that the output layer of each model was always fit to data from a PEG neuron, but the earlier layers were fit either using A1 responses or PEG responses.

We found no significant difference in prediction accuracy between the cross-area model (median r = 0.581) and the standard PEG held-out model (md = 0.569, signed-rank test, p = 0.119). In other words, a model pre-trained on A1 data performed just as well as one pre-trained on PEG data for predicting the activity of PEG neurons. This similarity in performance suggests that CNN models can generalize well across areas in addition to new neurons in the same area.

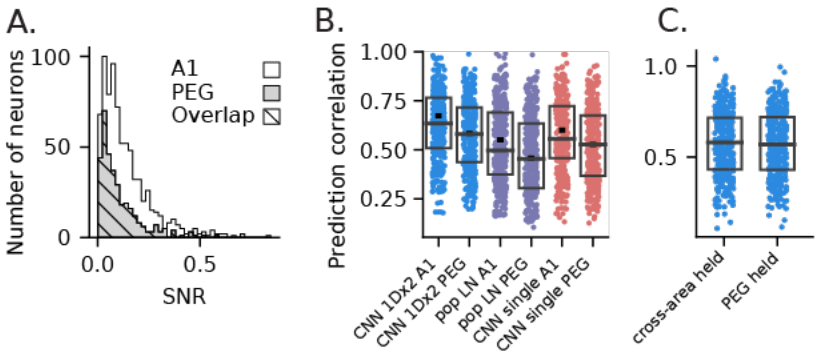

Figure 10. Comparison of model performance between primary (A1) and secondary (PEG) auditory cortical regions after controlling for differences in SNR. **A**. SNR scores for all neurons from A1 (white), PEG (gray), and an overlapping subset (hatched) of neurons with matched distributions. The subset was formed primarily by excluding high-SNR neurons from A1 since PEG neural responses were less reliable overall. **B**. Prediction accuracy of three exemplar models for neurons in the A1 and PEG subsets. Boxes show the 1st, 2nd and 3rd quartile for the selected subset, and the small horizontal dash indicates median performance across all neurons from that brain region. The increase in prediction accuracy for A1 is smaller for this subset, but still significant (U test; 1D CNNx2: $p = 6.685 \times 10^{-4}$; pop LN: $p = 6.743 \times 10^{-4}$; CNN single: $p = 4.78 \times 10^{-4}$), indicating that the increased SNR of A1 responses only partially accounts for our models' higher prediction accuracy for A1 neurons. **C.** Prediction correlations for the 1D CNNx2 PEG held-out model (Fig. 5C, median r = 0.569) and a "cross-area" held-out model (md = 0.581) that used A1 neural responses for the first stage of fitting and corresponding PEG responses from the matched-SNR subset for the second stage. Pre-fitting to A1 data proved to be just as effective as pre-training on PEG data for predicting PEG neural responses (signed-rank test, p = 0.119).

Discussion

We developed a convolutional neural network (CNN) architecture that simultaneously models the responses of several hundred neurons to dynamic natural stimuli. These neural population CNNs consistently outperformed traditional LN models of auditory coding, as well as CNNs fit to the activity of individual neurons. Moreover, population models were generalizable: the output layer of a pre-trained model could be re-fit to novel data without a reduction in model performance. The improved prediction accuracy and generalizability of the population CNN models is consistent with the hypothesis that they describe a comprehensive spectro-temporal space that encompasses sound encoding by any neuron in the auditory field being analyzed.

Population network models provide a natural expansion of linear-nonlinear models

Many alternatives to the LN model have been proposed to better characterize auditory encoding. Polynomial expansions, based on Taylor and Volterra series, are the classic way of extending linear models to better account for sensory activity (Marmarelis, 1989; Eggermont, 1993; Williamson et al., 2015). However, while theoretically well motivated, in practice these models provide only modest improvements over the linear model. This shortcoming is likely due to two factors. First, second order models require qualitatively more data for fitting and are thus more prone to estimation noise. Second, even with adequate fit data, the expanded functional space of a second order model may not align well to the actual nonlinear properties of auditory neurons. Other studies have attempted to address the latter limitation by developing models that account for specific biological nonlinearities, such as contrast gain control and short-term plasticity (Williamson et al., 2016; Espejo et al., 2019; Pennington and David, 2020). With a more empirically targeted design, these models tend to require less data for fitting, but still only provide modest improvements to performance. These models capture specific functional properties, but they may be limited in their ability to account for the full complement of nonlinearities present in sensory neural responses. Neural network models hold the promise of being able to account for a broader range of nonlinear properties, but have traditionally required complex fitting procedures to account for data limitations (Harper et al., 2016).

In the current study, we showed that the linear-nonlinear STRF architecture can be extended smoothly into small convolutional networks, with units identical to the traditional linear filters in the STRF. Moreover, we leveraged statistical power across neurons to fit machine learning models that performed consistently better than the LN model and other nonlinear models (Pennington and David, 2020). Prediction accuracy improved as model complexity increased, reaching an asymptote at complexities of about 150-200 free parameters per neuron for all CNN architectures tested (1D CNN, 1D CNNx2, 2D CNN). This number reflects an increase in model complexity relative to efficiently parameterized LN models, but still falls far short of the approximately 450 free parameters per neuron required for a full-rank LN model (Thorson et al., 2015).

Surprisingly, there was not a clear "winner" among the population CNNs we developed: these models were highly equivalent and differences in prediction accuracy were small. Despite theoretical guarantees that some CNN exists that *can* approximate a given neural encoding function, there is no guarantee that any specific CNN of a fixed size can meet that goal. There is

also no guarantee that a network's parameters can be optimized for a particular function algorithmically, even if the chosen architecture is sufficiently large (Colbrook et al., 2022). Accordingly, we had no expectation of equivalence between the CNNs we built despite their similarities. In practice, for example, one might expect idiosyncrasies of neural activity, such as non-Gaussian noise and a small fit dataset, to make some models more susceptible than others to fit errors or local minima. However, our results demonstrate that population CNN models can be fit robustly, regardless of the details of model architecture.

Toward a generalized model of auditory cortex

The goal of a fully generalizable brain model that can simulate neural activity across novel experimental conditions is not new (Amunts et al., 2016). However, attempts to build detailed models from the ground up have met with limited success, and it remains unclear which details of neural circuits and biophysical mechanisms are required to mimic natural function. In the current study, we take a more top-down approach to building a generalizable model. We use nonlinear regression to capture functional properties of the auditory system, without accounting for detailed biological circuitry. In this sense, the population CNN models we developed bear a closer resemblance to foundation models, which have grown increasingly valuable across the field of machine learning (Bommasani et al., 2021). In the domain of language processing, for example, models such as BERT and GPT-3 can be applied to a wide range of problems with little additional training (Devlin et al., 2019; Brown et al., 2020). A foundation model for biological auditory coding could, in the same way, be applied to a wide range of brain processes. Previous studies of auditory and visual neurophysiology have developed a similar idea (Cadieu et al., 2014; Kell et al., 2018). Intermediate representations in large CNN models fit initially for a visual or image processing problem can be used in a generalized linear model to predict neurophysiological responses to sensory stimuli. In the current study, we found such an approach was able to account not just for sensory selectivity but also for the time-varying response to dynamic natural stimuli.

A general model for biological auditory processing can have wider applications than explaining selectivity for sound features. For example, in this study, the population CNN model was also able to account for sparse coding properties in auditory cortex better than the LN model. Thus, it could also account for higher order properties of the system that may be important for biological signal processing, even though they were not imposed by the original fit. Following this logic, more comprehensive models, trained on larger and more diverse datasets, could account for a wide range of functional properties, in the same way that foundation models generalize to applications outside their original training.

While maximal performance was reached for a complexity of 150-200 free parameters per neuron in the current study, this should not be interpreted as a hard limit on complexity of auditory cortical representations. Our analysis of the relationship between fit dataset size and performance shows that prediction accuracy is limited by the amount of fit data available (Fig. 6). While we did present a large and diverse set of natural stimuli in the current study, these results indicate that further increasing the amount of fit data should lead to even more accurate model predictions on validation data. A more general model fit to a larger sample of natural sounds will likely require more free parameters and greater complexity.

Functional differences across the cortical hierarchy

One open question in studies of the auditory system is how sound representations evolve across the cortical processing hierarchy. Traditional LN models have been effective at characterizing sound encoding in the brainstem and midbrain (Miller et al., 2002; Versnel et al., 2009; Typlt et al., 2012), but their accuracy is limited in A1 (Machens et al., 2004; Espejo et al., 2019). Studies in non-primary auditory cortex are much less common. Sound-evoked activity in these brain areas is generally less reliable across repeated stimulus presentations and can undergo complex, long-lasting sensory adaptation, limiting the efficacy of encoding model analysis (Atiani et al., 2014; Lu et al., 2018). In this study, CNN models were able to achieve a substantial improvement in performance over the LN model in PEG. However, prediction accuracy was lower for PEG than for A1, even after accounting for differences in response reliability between areas. This difference is consistent with the idea that PEG neurons exhibit more selective, nonlinear response properties than A1. In the absence of data limitations, a CNN or related model should be able to account for this activity, but the current results indicate that a larger fit dataset is required to characterize PEG neurons as accurately as A1 neurons.

Neural population dynamics and the space of sensory representation

Population models leverage statistical power across neural populations and achieve greater accuracy than models fit to individual neurons. The fact that population CNNs generalize readily to novel neural data suggests that, embedded in their parameters, is a complete representation of the spectro-temporal features encoded across the entire cortical field. The dimensionality of the model's final hidden layer determines how many channels are recombined to predict the activity of every neuron. Here, the best-performing 1D CNNx2 model relied on 100 channels in this layer. The performance of the held-out models indicated that the same space could account equally well for the activity of any neuron in A1. Questions around the dimensionality of cortical sensory coding spaces are an active area of research (Stringer et al., 2019). This study suggests that much of the sensory activity of the many thousands of neurons in auditory cortex can be accounted for by a relatively low-dimensional space.

As high channel-count recordings in neurophysiological research continue to become more prevalent and grow in scale, the feasibility and value of population models will also increase. We expect these benefits to be particularly strong for studies where dataset duration is constrained by experimental design, such as studies of behavior. When motivation or other behavioral factors limit the number of trials during recordings, statistical power can be leveraged from the large number of neurons. In addition, sensory coding properties can be considered in a constrained sensory subspace identified by analysis of larger datasets.

References

Abadi M et al. (2016) TensorFlow: A System for Large-Scale Machine Learning. In, pp 265–283 Available at: https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi [Accessed April 30, 2022].

Aertsen AM, Johannesma PI (1981) The spectro-temporal receptive field. A functional characteristic of auditory neurons. Biol Cybern 42:133–143.

Ahveninen J, Hämäläinen M, Jääskeläinen IP, Ahlfors SP, Huang S, Lin F-H, Raij T, Sams M, Vasios CE, Belliveau JW (2011) Attention-driven auditory cortex short-term plasticity helps segregate relevant sounds from noise. Proceedings of the National Academy of Sciences 108:4182–4187.

Amunts K, Ebell C, Muller J, Telefont M, Knoll A, Lippert T (2016) The Human Brain Project: Creating a European Research Infrastructure to Decode the Human Brain. Neuron 92:574–581.

Atencio CA, Sharpee TO, Schreiner CE (2008) Cooperative nonlinearities in auditory cortical neurons. Neuron 58:956–966.

Atiani S, David SV, Elgueda D, Locastro M, Radtke-Schuller S, Shamma SA, Fritz JB (2014) Emergent selectivity for task-relevant stimuli in higher-order auditory cortex. Neuron 82:486–499.

Barlow HB (1961) Possible principles underlying the transformation of sensory messages. In: Sensory Communication (Rosenbluth WA, ed), pp 217–234. Cambridge, MA: MIT Press.

Bizley JK, Nodal FR, Nelken I, King AJ (2005) Functional organization of ferret auditory cortex. Cerebral Cortex 15:1637–1653.

Bommasani R et al. (2021) On the Opportunities and Risks of Foundation Models. arXiv:210807258 [cs] Available at: http://arxiv.org/abs/2108.07258 [Accessed April 18, 2022].

Brown TB et al. (2020) Language Models are Few-Shot Learners. arXiv:200514165 [cs] Available at: http://arxiv.org/abs/2005.14165 [Accessed May 3, 2022].

Cadieu CF, Hong H, Yamins DLK, Pinto N, Ardila D, Solomon EA, Majaj NJ, DiCarlo JJ (2014) Deep neural networks rival the representation of primate IT cortex for core visual object recognition. PLoS computational biology 10:e1003963.

Calabrese A, Schumacher JW, Schneider DM, Paninski L, Woolley SMN (2011) A generalized linear model for estimating spectrotemporal receptive fields from responses to natural sounds. PloS one 6:e16104.

Chan W, Jaitly N, Le Q, Vinyals O (2016) Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 4960–4964.

Colbrook MJ, Antun V, Hansen AC (2022) The difficulty of computing stable and accurate neural networks: On the barriers of deep learning and Smale's 18th problem. Proceedings of the National Academy of Sciences 119:e2107151119.

David SV, Shamma SA (2013) Integration over Multiple Timescales in Primary Auditory Cortex. J Neurosci 33:19154–19166.

Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:181004805 [cs] Available at: http://arxiv.org/abs/1810.04805 [Accessed May 3, 2022].

Eggermont JJ (1993) Wiener and Volterra analysis applied to the auditory system. Hearing Research 66:177–201.

Espejo ML, Schwartz ZP, David SV (2019) Spectral tuning of adaptation supports coding of sensory context in auditory cortex. PLoS Computational Biology 15:e1007430.

Gill P, Zhang J, Woolley SM, Fremouw T, Theunissen FE (2006) Sound representation methods for spectro-temporal receptive field estimation. Journal of Computational Neuroscience 21:5–20.

Harper NS, Schoppe O, Willmore BDB, Cui Z, Schnupp JWH, King AJ (2016) Network Receptive Field Modeling Reveals Extensive Integration and Multi-feature Selectivity in Auditory Cortical Neurons. PLoS Computational Biology 12:e1005113.

Hromádka T, DeWeese MR, Zador AM (2008) Sparse Representation of Sounds in the Unanesthetized Auditory Cortex. PLOS Biology 6:e16.

Hsu A, Borst A, Theunissen FE (2004) Quantifying variability in neural responses and its application for the validation of model predictions. Network: Computation and Neural Systems 15:91–109.

Katsiamis AG, Drakakis EM, Lyon RF (2007) Practical Gammatone-Like Filters for Auditory Processing. EURASIP Journal on Audio, Speech, and Music Processing 2007:1–15.

Kell AJ, McDermott JH (2019) Deep neural network models of sensory systems: windows onto the role of task constraints. Elsevier Ltd.

Kell AJE, Yamins DLK, Shook EN, Norman-Haignere SV, McDermott JH (2018) A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. Neuron 98:630-644.e16.

Keshishian M, Akbari H, Khalighinejad B, Herrero JL, Mehta AD, Mesgarani N (2020) Estimating and interpreting nonlinear receptive field of sensory neural responses with deep neural network models Serre T, Frank MJ, Englitz B, Lesica NA, eds. eLife 9:e53445.

Kingma DP, Ba J (2017) Adam: A Method for Stochastic Optimization. arXiv:14126980 [cs] Available at: http://arxiv.org/abs/1412.6980 [Accessed April 30, 2022].

Kowalski N, Depireux DA, Shamma SA (1996) Analysis of dynamic spectra in ferret primary auditory cortex. I. Characteristics of single-unit responses to moving ripple spectra. Journal of Neurophysiology 76:3503–3523.

Krogh A, Hertz J (1991) A Simple Weight Decay Can Improve Generalization. In: Advances in Neural Information Processing Systems. Morgan-Kaufmann. Available at: https://papers.nips.cc/paper/1991/hash/8eefcfdf5990e441f0fb6f3fad709e21-Abstract.html [Accessed April 30, 2022].

Lohse M, Bajo VM, King AJ, Willmore BDB (2020) Neural circuits underlying auditory contrast gain control and their perceptual implications. Nat Commun 11:324.

Lu K, Liu W, Zan P, David SV, Fritz JB, Shamma SA, Zan XP, Stephen X, David V, Fritz JB, Shihab X, Shamma A (2018) Implicit memory for complex sounds in higher auditory cortex of the ferret. The Journal of Neuroscience 38:9955–9966.

Machens CK, Wehr MS, Zador AM (2004) Linearity of cortical receptive fields measured with natural sounds. Journal of Neuroscience 24:1089–1100.

Marmarelis VZ ed. (1989) Advanced Methods of Physiological System Modeling. Boston, MA: Springer US. Available at: http://www.springerlink.com/index/10.1007/978-1-4613-9789-2.

McDermott JH, Schemitsch M, Simoncelli EP (2013) Summary statistics in auditory perception. Nature neuroscience 16:493–498.

Miller LM, Escabi MA, Read HL, Schreiner CE, Escabí MA (2002) Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. J Neurophysiol 87:516–527.

Pachitariu M, Steinmetz N, Kadir S, Carandini M, Harris KD (2016) Kilosort: realtime spike-sorting for extracellular electrophysiology with hundreds of channels. Available at: http://biorxiv.org/content/early/2016/06/30/061481.abstract http://biorxiv.org/lookup/doi/10.1101/061481.

Pennington JR, David SV (2020) Complementary Effects of Adaptation and Gain Control on Sound Encoding in Primary Auditory Cortex. eNeuro 7 Available at: https://www.eneuro.org/content/7/6/ENEURO.0205-20.2020 [Accessed April 21, 2022].

Rabinowitz NC, Willmore BDB, Schnupp JWH, King AJ (2011) Contrast gain control in auditory cortex. Neuron 70:1178–1191.

Rabinowitz NC, Willmore BDBB, Schnupp JWHH, King AJ (2012) Spectrotemporal contrast kernels for neurons in primary auditory cortex. Journal of Neuroscience 32:11271–11284.

Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J (2021) Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp 8748–8763. PMLR.

Shobe JL, Claar LD, Parhami S, Bakhurin KI, Masmanidis SC (2015) Brain activity mapping at multiple scales with silicon microprobes containing 1,024 electrodes. Journal of neurophysiology 114:2043–2052.

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research 15:1929–1958.

Stringer C, Pachitariu M, Steinmetz N, Carandini M, Harris KD (2019) High-dimensional geometry of population responses in visual cortex. Nature 571:361–365.

Theunissen FE, David SV, Singh NC, Hsu A, Vinje WE, Gallant JL (2001) Estimating spatial temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. Network: Computation in Neural Systems 12:289–316.

Thorson IL, Liénard J, David SV (2015) The essential complexity of auditory receptive fields Theunissen FE, ed. PLoS Computational Biology 11:e1004628.

Typlt M, Englitz B, Sonntag M, Dehmel S, Kopp-Scheinpflug C, Ruebsamen R (2012) Multidimensional characterization and differentiation of neurons in the anteroventral cochlear nucleus. PloS one 7:e29965.

Versnel H, Zwiers MP, Opstal AJ van (2009) Spectrotemporal Response Properties of Inferior Colliculus Neurons in Alert Monkey. J Neurosci 29:9725–9739.

Vinje WE, Gallant JL (2000) Sparse coding and decorrelation in primary visual cortex during natural vision. Science 287:1273–1276.

Williamson RS, Ahrens MB, Linden JF, Sahani M (2016) Input-Specific Gain Modulation by Local Sensory Context Shapes Cortical and Thalamic Responses to Complex Sounds. Neuron 91:467–481.

Williamson RS, Sahani M, Pillow JW (2015) The equivalence of information-theoretic and likelihood-based methods for neural dimensionality reduction. Bethge M, ed. PLoS computational biology 11:e1004141.

Willmore BDB, Mazer JA, Gallant JL (2011) Sparse coding in striate and extrastriate visual cortex. Journal of neurophysiology 105:2907–2919.

Wu MC-K, David SV, Gallant JL (2006) Complete functional characterization of sensory neurons by system identification. Annual Review of Neuroscience 29:477–505.

You Z, Wang X, Xu B (2013) Investigation of deep Boltzmann machines for phone recognition. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp 7600–7603.