Predictive coding for natural vocal signals in the songbird auditory forebrain

Srihita Rudraraju[1], Michael E. Turvey[1*], Bradley H. Theilman[2*], Timothy Q. Gentner[1,2,3,4]

[1]Department of Psychology, University of California San Diego, 9500 Gilman Dr., La Jolla, CA 92093.
[2]Neurosciences Graduate Program, University of California San Diego, 9500 Gilman Dr., La Jolla, CA 92093.
[3]Department of Neurobiology, University of California San Diego, 9500 Gilman Dr., La Jolla, CA 92093.
[4]Kavli Institute for Brain and Mind, University of California San Diego, 9500 Gilman Dr., La Jolla, CA 92093.

[*]To whom correspondence should be addressed; E-mail: srrudrar@ucsd.edu
[†]To whom correspondence should be addressed; E-mail: tgentner@ucsd.edu

Predictive coding posits that incoming sensory signals are compared to an internal generative model with resulting error signals carried in the responses of single neurons. Empirical support for predictive coding in individual neurons, particularly in the auditory system and for natural stimuli, has proven difficult to observe. Here, we developed a neural network that uses current sensory context to predict future spectral-temporal features in a natural communication signal, birdsong. Using this model, we represent the waveform of any birdsong as either a set of weighted "latent" predictive features evolving in time, or a corresponding error representation that reflects the difference between the predicted and actual song. We then recorded responses of single neurons in caudomedial nidopallium (NCM), caudal mesopallium (CMM) and Field L, analogs of mammalian auditory cortex, in anesthetized European starlings listening to conspecific songs, and computed the linear/non-linear receptive fields for each neuron fit separately to the spectro-temporal, predictive, and error representations of song. Comparisons between the quality of each receptive field model reveal that NCM spiking responses are best modeled by the predictive spectrotemporal features of song, while CMM and Field L responses capture both predictive and error features. Neural activity is selective for carrying information explicitly

about prediction and prediction errors, and their preferences vary across the auditory forebrain. We conclude that this provides strong support for the notion that individual neurons in songbirds encode information related to multiple stimulus representations guided by predictive coding simultaneously.

Introduction

In sensory processing, especially of temporally patterned signals such as speech and audio, existing knowledge includes an internal model for the expectation of future events. Animals capable of vocal learning and communication produce learned complex sounds with predictive structures in the course of their natural behavior. Exploring the dynamic structure of learned vocalizations has been a subject of enduring interest. Additionally, evidence indicates that the sequential organization of learned signals is not uniform or random; rather, it exhibits discernable patterns. The brain's ability to predict temporal sensory events, hypothesized within predictive coding theory, is believed to enhance sensory processing by encoding expectations and errors arising from disparities between those expectations and actual sensory input through neural activity. Predictive coding offers a conceptual framework to capture the congruential notion of perceptual inference.

Predictive coding is modulated by feedback pathways that adjust feedforward signals in a context-dependent manner (Asilador & Llano, 2021; A. M. Lesicko et al., 2022; A. M. H. Lesicko & Geffen, 2022). There are other lines of literature that support the theory that predictive coding is implemented in the auditory cortex, where neural activity is motivated by learned associations between stimuli either in different sensory modalities or motor-related

regions. Inputs from the secondary motor cortex suppress neural responses in deep and superficial layers in the mouse auditory cortex during movements (Audette et al., 2022; Holey & Schneider, 2023). Similarly, when an auditory stimulus becomes associated with a visual stimulus in a behavior relevant context, it results in the experience-dependent suppression of visual responses within the primary visual cortex (Bimbard et al., 2023; Garner & Keller, 2022). Expectation based on olfactory cues modulate neuronal responses in mice primary sensory areas to learned auditory stimuli (Gilday & Mizrahi, 2023). However, in the structure of communication systems, predictive coding is concerned with the real-time prediction of familiar stimuli, a phenomenon that cannot be entirely explained by associative learning.

Previous studies while useful with understanding the relationships between stimuli in different sensory modalities, ignores the fact that natural stimuli in general have a range of temporal and spatial correlations built into them, and do not focus on the predictive elements within the statistical patterns of natural stimuli (Narula & Hahnloser, 2021). Songbirds produce songs characterized by rich spectrotemporal features with predictive structures. Furthermore, findings highlight structure in both long and short term dependencies between vocal elements across time in birdsong and speech are ubiquitous in vocal communication (Sainburg et al., 2019). Predictive coding offers a conceptual framework for signals where predictions occur at both short term and long term timescales, extending beyond learned sequences of synthetic tones or arbitrary pairings of sounds in simulated environments.

Predictive coding serves as the means to encapsulate the idea that perception involves an inferential process. The auditory landscape is predominantly characterized by patterns in time

and frequency, and integrating these specific prior probabilities into our perceptual systems is crucial. This incorporation is expected to significantly influence our subsequent cognitive processes. In this paper, we have introduced a methodology capable of unveiling both the feedforward signals and top-down error processing of natural signals at the level of individual neurons within the framework of predictive coding. We first introduce a deep learning model that generates supervised statistical approximation of the stimulus, and subsequently measure the putative error relative to an unbiased stimulus estimate. To explore these features of predictive coding directly in the auditory cortex, we examined individual neural responses in anesthetized European starlings in the caudomedial nidopallium (NCM), caudal mesopallium (CMM) and Field L, analogs of human higher-order auditory cortex, listening to conspecific songs. We then applied logistic regression models to study single neuron coding to discover statistics of the stimulus encoded in the response corresponding to general, prediction and prediction error spectrotemporal auditory features. Auditory neurons exhibit responses to multiple representations of the stimulus. Nevertheless, the specific information preferred by each neuron appears to be influenced by the organization of the auditory pathway. These findings support the notion of predictive coding being implemented in the context of natural signals where individual auditory neurons carry information about the signal, the predictive features and the prediction error features in the signal at any point in time.

RESULTS

Stimulus representations obtained by the generative model

Predictive coding theories hypothesize the brain must be equipped with a generative model that maps from hidden causes to predict incoming sensory information (or events). To determine how the perceptual process of mapping from sensory events to causes is implemented

mechanistically through predictive coding in individual neurons, here we proposed a neural network model that illustrates the generative model and consecutively defined representations of the stimulus corresponding to prediction, signal and error as given in the predictive coding framework (Rao & Ballard, 1999).

We performed Fourier transformation on a library of bandpass-filtered natural birdsong waveforms to generate spectrograms of auditory stimulus, and removed low-amplitude background noise. We trained neural networks on the library to generate a feature space that is predictive of future auditory input. Each birdsong spectrogram was then segmented such that the networks trained to predict the immediate next temporal segment (about 340 ms) based on the most recent temporal spectrogram segment (about 340 ms) (Figure 1A).

We propose a deep convolutional network (Kingma & Welling, 2022) Temporal Convolutional Model (TCM) trained to predict future segments of song as a proxy to the generative model in the brain (Figure 1B). The architecture of this model used a combination of convolutional layers and fully connected layers in the encoder and decoder; and has been shown to resemble neuronal processing in the visual system (LeCun & Bengio, 1995; Carlson et al., 2012; Klein et al., 2003; Yamins et al., 2014; Zhao & Zhaoping, 2011). The encoder compresses the past spectrographic segment taken as inputs (32x32 units) into a latent representation (256 hidden units), which is then passed through the decoder to predict the output future spectrogram segment (32x32 units). We proposed a loss function that is a combination of prediction loss and multidimensional scaling (MDS) distance loss. Prediction loss aims to minimize the mean squared error between predicted future spectrogram segment and ground-truth future

spectrogram segment. The goal of multi-dimensional scaling (MDS) algorithms is to keep information loss to a minimum when projecting data into lower dimensions. It is a non-linear optimization problem which tries to optimize the mapping in the target dimension based on the original pairwise distance information, thus preserving the global structure (Wickramasinghe & Sice, 2021).

For training, we analyzed datasets consisting of over 30 hours of vocalizations and over discrete vocalization units (see Methods; these datasets can be obtained from zenodo; Arneodo et al., 2019). This dataset was chosen because it contains large repertoires of vocalizations from relatively acoustically isolated individuals that can be cleanly segmented. Each vocalization dataset was temporally segmented as described above and used as the training set for TCM. A library of natural starling songs comprising a diverse set of vocalization units were used for testing.

Using TCM trained weights, we projected the past song segments of the test set as inputs to obtain compressed latent representations trained to make predictions of future song segments of the test set. To follow a convention throughout the paper, we labeled computations on spectrograms as segments, and representations of latent space as features. Under this model, the latent space containing predictive spectrotemporal features of song represents the prediction component of the predictive coding framework (prediction features), compared to the more generalized class of spectrotemporal features comprising the whole of incoming song representing the signal component (signal-fft segment). We captured the prediction error component (prediction error segment) by computing the squared difference between the output

of TCM (or predicted future song segment) and the ground-truth future song segment (Figure 1C). As a control, we trained an autoencoder model (Sainburg et al., 2020) that shares a similar architecture to TCM, but instead reconstructs the input past spectrogram segment. The compressed latent space of the autoencoder (signal-ae features) also comprises generalized spectrotemporal song features representing the signal component of the predictive coding framework. The resulting acoustic spaces obtained by the generative model are illustrated in Figure S1.

Auditory single neurons respond to multiple distinct features of song representations

We explored the neurophysiological basis of predictive coding implementation by relating stimulus representations to neural activity. We recorded spiking activity in anesthetized European starlings from populations of neurons in the primary auditory forebrain region Field L, and two secondary auditory forebrain regions (the caudal mesopallium CMM and caudal medial nidopallium NCM) (Brainard & Doupe, 2000; Nottebohm, 2005; Woolley et al., 2005) while they were listening to natural birdsongs consisting of a diverse set of motifs (Figure S2A). Figure S2B shows all the recording sites. Neural recordings were sorted using Kilosort2 followed by manually curation. Single units were identified using our rating criteria (n=1749 neurons, 8 birds, 14 penetration sites; see Methods). Neurons in all three regions respond to a variety of motifs. NCM and CMM neurons display precise, sparse activity and spiking selectively to specific features of conspecific song (Amin et al., 2007; Gentner & Margoliash, 2003; M ller & Leppelsack, 1985; Theunissen et al., 2004). Field L neurons in contrast respond broadly to natural auditory stimuli (Bonke et al., 1979; Grace et al., 2003; Leppelsack & Vogt, 1976; Lewicki & Arthur, 1996; Margoliash, 1986). Example song-aligned sample spike trains for an

individual neuron (across 20 presentations of stimulus) and spiking activity in a population of units in NCM (n=150 neurons), CMM (n=79 neurons) and Field L (n=43 neurons) respectively are shown in Figure 2A, 2B and 2C . To relate stimulus representations to neural activity, we computed composite receptive fields (CRFs) for auditory cortical neurons using the Maximum Noise Entropy (MNE) method (Fitzgerald et al., 2011). Previous work has shown that the MNE model well-describes the sensitivity to higher-order features exhibited by NCM neurons (Kozlov & Gentner, 2016; Vahidi, 2021).

For each single unit in simultaneously recorded populations, we independently computed the MNE model parameters a, h and J that optimally relate the neuron's response to each of the three different stimulus representations: either the short-time Fourier transform spectrogram or the reduced dimensionality representation of the spectrogram; the projection of the spectrogram into a TCM latent space; or the squared error between TCM predicted future spectrogram and the true spectrogram. 80% of the song stimuli was used for training. This yields a version of each neuron's CRF fit to either: 1) all spectrotemporal features of conspecific song (signal-fft-CRF and signal-ae-CRF) or 2) only the predictive spectrotemporal features of song (prediction-CRF) or 3) spectrotemporal features of song corresponding to prediction error (prediction error-CRF). Example MNE features and CRFs for an individual NCM neuron are shown in Figure 2D.

The eigenvalue spectrum of J contained significant eigenvalues, indicating multiple covariant acoustic features that drive spiking of the neuron. We obtained significant eigenvectors of the MNE model's J matrix that along with parameter h define the receptive field of each neuron; non-linear and linear features of the receptive field respectively. The parameters obtained for each neuron from the trained MNE model can be used to predict their spiking response to novel stimuli. We modeled the response of each neuron to the held out portion (20%)

of song stimuli not used for MNE training. We evaluated the prediction by computing the correlation coefficient between the modeled response obtained from each stimulus representation and the empirical response for that neuron (Figure 2E). The features of the signal-fft-CRF (Pearson's correlation coefficient, r=0.71) combined spectrotemporal features representing tones, harmonic stacks. The features of the prediction-CRF (r=0.75) and signal-ae-CRF (r=0.48) are obtained by nonlinear transformation (latent weights) of the TCM and autoencoder respectively, and thus are not visually decipherable. The prediction error-CRF (r=0.80) features are obtained by fitting the TCM residual spectrogram segment. These CRF structures resemble spectrotemporal features, but are less robust as they have been filtered through the model.

We found multiple, distinct receptive-field features in individual high-level auditory neurons in response to all four stimulus representations. Although the prediction-CRF and signal-ae-CRF do not display spectrotemporal structure, their MNE features produced high predictive quality of the empirical response variance. The quality of signal-ae-CRF (mean Pearson's correlation coefficients, r=0.79±0.005) modeled responses were similar to those produced from signal-fft-CRFs (r=0.75±0.005; t=-5.6, p=2.09e-8; paired t-test; Figure S3) conveying that in addition to spectrographic representation as in previous studies (Fitzgerald et al., 2011; Kozlov & Gentner, 2016), latent spaces can be used as an alternative representation of the stimulus.

Single auditory neurons encode for predictive and prediction error acoustic features

We modeled the response of each neuron to held out portion (20%) of song stimuli not used for MNE training. We repeated this procedure to generate a null distribution of MNE predictions using randomly shuffled versions of the spike trains for each neuron. The mean

correlation coefficient across all the unshuffled MNEs (r=0.73 ± 0.004, range=0.70 − 0.75 over three regions) was significantly higher than that for the shuffled MNEs (r=-5.73e-4 ± 4.98e-4, range=-4.84e-3 - -1.54e-6) for each of the separate stimulus representations across NCM (signal-fft, t=155.8, p=0.0; prediction, t=170.8, p=0.0; prediction error, t=116.2, p=0.0; signal-ae, t=162.8, p=0.0) CMM (signal-fft, t=124.3, p=0.0; prediction, t=134.1, p=0.0; prediction error, t=83.9, p=0.0; signal-ae, t=130.6, p=0.0) and Field L (signal-fft, t=96.2, p=0.0; prediction, t=96.6, p=0.0; prediction error, t=91.7, p=0.0; signal-ae, t=94.1, p=0.0) (Figure S4A-L).

The receptive field for a neuron allows the experimenter to know which features in the stimulus are associated with modulations in the neuron's response. The MNE receptive fields allow for quantitative comparison of distinct predictive coding guided stimulus representations by response modeling. Across all recorded NCM populations (n=568 neurons, 3 birds), the prediction-CRFs (mean r±SEM=0.77±0.004) yielded excellent models of of each neuron's empirical spiking response to novel song, accounting for 59.81% of the response variance in comparison with signal-fft-CRF (mean r±SEM=0.75±0.005; t=-7.7, p=6.76e-14 paired t-test; Figure 3A). The same pattern is observed across all the neural populations in CMM and Field L (CMM, t=-12.1, p<1e-23 paired t-test; Field L, t=-11.2, p<1e-23 paired t-test; Figure 3B and 3C), with the empirical correlation for prediction stronger than that for signal. In contrast, for all NCM neurons, the empirical correlation coefficients computed from error-CRFs are much weaker (mean r=0.48±0.004, t=75.8, p<1e-23 paired t-test; Figure 3D) compared to signal-fft-CRFs.

However, CMM (mean r=0.64±0.008; Figure 3E) and Field L (mean r=0.70±0.008; Figure 3F) recorded populations show significantly higher error-CRF correlation contrary to

NCM (t=-19.1, p<1e-23 t-test CMM vs NCM; t=-24.9, p<1e-23 t-test Field L vs NCM). In the context of CMM, we observed that certain blocks exhibited lower correlation coefficients for error-CRFs when compared to the remaining blocks.

To assess whether the predictive coding hypothesis is supported with an alternative representation of the signal, we tested the comparison between prediction-CRFs and error-CRF, in this instance with respect to a latent representation of the spectrotemporal features (signal-ae-CRFs). In populations recorded across all three regions, we observed that prediction-CRFs explained equally significant portions of the variance of empirical neural response compared to signal-ae-CRFs (NCM, t=14.9, p<1e-23 paired t-test; CMM, t=-7.3, p=9.7e-13 paired t-test; Field L, t=-6.7, p=4.7e-11 paired t-test; Figure 3G, 3H and 3I). The same pattern is observed for error-CRFs across all neuron populations, with the NCM empirical correlation significantly weaker than that for the signal-ae-CRFs (t=86.2, p=0.0 paired t-test; Figure 3J). In CMM and Field L, the empirical correlation was slightly weaker than that for signal-ae-CRFs (CMM, t=22.2, p<1e-23 paired t-test; Field L, t=20.5, p<1e-23 paired t-test; Figure 3K and 3L), but followed a similar pattern to that of signal-fft-CRFs (NCM, t=7.4, p=2.3e-13; CMM, t=6.3 , p=5.1e-10 ; Field L, t=11.4, p<1e-23 t-test). In contrast, for modeling responses from only the linear component of the MNE receptive field (Figure S4A-L), the correlation with the empirical response is much weaker (signal-fft, NCM, r=0.21±0.005, t=100.1, p=0.0; CMM, r=0.29±0.007, t=72.3, p<1e-23; Field L, r=0.26±0.006, t=73.4, p<1e-23 paired t-test). The relationship between signal, prediction and prediction error CRFs for the linear model (Figure S5) parallels a familiar progression to that of the full model although predictive coding features are more notable in the full model (NCM, prediction, t=0.9, p=0.36; prediction error, t=-30.2, p<1e-23; CMM,

prediction, t=0.7, p=0.48; prediction error, t=-2.5, p=0.01; Field L, prediction, t=0.8, p=0.41; prediction error, t=8.5, p=1.3e-16 paired t-test).

The analysis revealed that our proxy for the internal model encapsulates significant prediction and prediction error acoustic structure in the generated representations. The MNE features verily capture statistical relationships between individual neuron activity and stimulus characterized by predictive coding. Our data suggests that auditory neurons respond to multiple representations of the stimulus. However, the preference for the specific type of information encoded in each neuron is influenced by the organization of the auditory pathway. In summary, these first findings point at the evidence of predictive coding hypothesis implemented in single auditory neurons during song listening. We substantiated our claims by selecting a different architecture for our internal model, this time using a simple feedforward neural network (Temporal Predictive Model or TPM; Figure S6A-C). The TPM generated prediction features followed a similar response modeling quality compared to that of our proxy internal model (Figure S6D).

Neurons along the auditory hierarchy selective for carrying information about prediction and prediction errors

Response modeling by the MNE method enables quantification of specific aspects of the internal model during song listening. Since prediction error is a function of the signal input to the TCM and the output of the TCM derived from decoding the learnt latent predictive weights, we hypothesize that prediction, prediction error and the signal are correlated with each other. We evaluated their respective contributions following a partitioning strategy (multiple regression

model of response probabilities), separating the explained response variance into what was explained uniquely by signal, prediction and prediction error and what was shared between the feature spaces. In the partitioning strategy, a full regression model was used, treating the empirical neural response as a composition of modeled response probabilities. We compared the unique and shared variance partitions across the three brain regions (NCM, CMM and Field L). Consistent with our hypothesis, the shared contribution of prediction and signal was the largest of the variance partitions within each region (NCM, partial r=0.51±0.004; CMM, partial r=0.72±0.006; Field L, partial r=0.68±0.007). Additionally, the shared variance between prediction error and signal was also significantly larger in comparison to the unique contribution of prediction error (NCM, t=11.5, p<1e-23; CMM, t=38.2, p<1e-23; Field L, t=57.4, p<1e-23 paired t-test; Figure 4A, 4B and 4C).

Prediction error features encode significantly higher portions of the empirical response in CMM and Field L in comparison to NCM (Figure 3D, 3E and 3F), however, we observed that most of this variance is shared between prediction and prediction error (CMM, t=-28.6, p<1e-23; Field L, t=-34.8, p<1e-23 t-test vs NCM). In order to determine a neuron's selectivity to encode distinct acoustic features, we compared the unique contributions of the components of the predictive coding framework. Remarkably, unique partial correlations for prediction features are the strongest predictor of neuron responses in NCM (partial r=0.39±0.004), CMM (partial r=0.18±0.006) and Field L (partial r=0.20±0.006) significantly higher than partial correlations for signal (NCM, t=11.3, p<1e-23; CMM, t=10.5, p=1.2e-23; Field L, t=15.6, p<1e-23 paired t-test) and prediction error (NCM, t=72.2, p<1e-23; CMM, t=19.0, p<1e-23; Field L, t=23.9, p<1e-23 paired t-test; Figure 4D, 4E, 4F and S7). However, both the signal and prediction error

contribute a significant portion of the unique response variance that do not overlap with the prediction determined response variance. Consequently, predictive coding features combinedly describe the activity of any neuron better than an alternative consisting of only signal acoustic features (NCM, 65.5±0.006% variance, t=57.8, p<1e-23; CMM, 48.4±0.012% variance, t=15.8, p<1e-23; Field L, 52.9±0.010% variance, t=17.4, p<1e-23 paired t-test vs signal; Figure 4G, 4H, 4I and S8).

In addition, the unique partial correlations of prediction features in Figure 4D, 4E and 4F show the differences in variability of overall distribution between the three regions (NCM σ=0.08; CMM σ=0.15; Field L σ=0.16). The higher variance in CMM and Field L variance suggests that neurons in these regions have varied encoding preferences. To better understand the variability of encoded acoustic features within a population, we quantified the propensity of neurons based on their unique response variance partitions. We applied HDBSCAN clustering (McInnes et al., 2017) to unique signal, prediction and prediction error partial correlations of neurons within a simultaneously recorded population. We found that prediction features modulated response encoding of all neurons in an NCM population (n=194 neurons, silhouette score=0.18; Figure 5A). We also found that CMM (n=82 neurons, silhouette score=0.38; Figure 5B) and Field L (n=96 neurons, silhouette score=0.46; Figure 5C) populations code for predictive features, however, a subpopulation of neurons had enhanced response encoding attributed to spectrotemporal features that were not predictive or composed of error. We found that clusterability increased continually from NCM (recorded populations n=3, mean silhouette score=0.20) to CMM (n=5, mean silhouette score=0.29) and Field L (recorded populations n=6, mean silhouette score=0.44; one-way ANOVA F = 5.83, p=0.02; Figure 5D and S9). We found

distinct predictive and non-predictive clusters within a population in CMM and FIeld L that subsequently converge in NCM, which is consistent with literature finding prediction or prediction error specific neurons (Attinger et al., 2017; Eliades & Wang, 2008; Huang et al., 2020; Keller et al., 2012; Keller & Hahnloser, 2009; Schultz & Dickinson, 2000). Our data therefore suggests that sensory information relevant for update of the internal model is realized through multiple representations in individual neurons with different dependencies determined by their organization in the auditory pathway. The reason for these differences in relative contributions is that we are conducting an unsupervised exploration across all putative single units we are recording from.

Discussion

We combined computational modeling with electrophysiological activity to study internal representations of sensory information under the framework of predictive coding in the auditory domain in songbirds. Our findings bridge the gap between qualitative evidence of predictive coding theory and a mechanistic understanding of the implementation of this theory. We have presented a set of computational methods for quantitative implementation and systemwide evaluation of the predictive coding hypothesis. We demonstrate that during song listening, individual neuron responses in primary and secondary auditory regions are modeled by expectations of future song and uncertainty of song.

Our results suggest hierarchical organization in the sensory cortex determined by internal models that are thought to generate predictions of incoming inputs yielding efficient neural encoding. This result was enabled by a generative model representative of the bird's internal model. We developed a neural network consisting of convolutional layers (Temporal

Convolutional Model; TCM) to extract features of the acoustic stimulus corresponding to (1) overall song features or signal obtained from spectrogram segments (2) expectation of future song or prediction obtained from latent space weights of the TCM and (3) uncertainty of song or prediction error obtained from the difference between actual and TCM predicted future spectrogram segment. We examined neurophysiological responses of individual neurons fit to separate acoustic components of the generative model using the MNE model. This approach provides quantitative estimations of response variance captured in their receptive fields driven by separate acoustic features.

Evidence of predictive coding in the form of prediction errors have been found throughout the brain. Subsets of neurons in visual cortex (Attinger et al., 2017; Keller et al., 2012; Zmarz & Keller, 2016), auditory cortex (Eliades & Wang, 2008; Keller & Hahnloser, 2009) and barrel cortex (Ayaz et al., 2019) code for a mismatch between feedback and feedforward information. Neuronal responses in various regions of the auditory cortex could be explained by sensory predictions of future scenes and responses to deviant stimuli or omissions or surprise or adaptation to repetitive standard sounds (Garrido et al., 2009; Gill et al., 2008; Näätänen et al., 2004, p. 1995, 2007; Netser et al., 2011; Singer et al., 2018; Ulanovsky et al., 2003). Our results, in corroboration with previous literature, show that in anesthetized songbirds, single neuron auditory responses in NCM, CMM and Field L are modeled collectively by stimulus representations that capture covariant structure in the predictive spectrotemporal acoustic features of song and spectrotemporal features capturing uncertainty between actual and expected song. Thus, individual neurons are not only summing their inputs, but also predict

future events and minimize error in their predictions, which we propose is a crucial component of the brain's process of efficient coding.

We observe variations in the type of information carried by neurons across auditory regions. NCM neurons mostly carry information uniquely about the expectation of future events, while CMM and Field L neurons carry information shared by uncertainty and expectation. These regional differences are an indication that there is something predictive going on. The predictive coding theory is potentially a helpful way to explain these differences as the theory refers to hierarchical organization of the sensory cortex. Understanding the contributions different regions may make within the PC framework is useful moving forward. The pattern of unique variance across every region is qualitatively similar but what changes across regions is the relative proportion of variance that is uniquely explained by these different components. Thus we believe that there is not an explicitly predictive or error region, but what the system is doing is differentiating between unique and shared variance. We also observe that the proportion of neurons within a population that are uniquely predictive also varies across regions. These results point to a novel claim that there is evidence of a computational gradient along the sensory hierarchy.

Prediction and prediction error neuronal populations are believed to be distinct (Bastos et al., 2012; Hertäg & Clopath, 2022). Previous literature has mapped explicit computations onto individual neurons (Keller 2008; refs). Our findings reveal that a clear distinction between signal, prediction and error neurons is possible maybe for a small subset of the population, but most neurons carry information that can be explicitly tied to more than one or all representations

of the stimulus. One possibility of this inconsistency could be in the construction of external representations that we relate to neural activity. The generative model we used in this paper to obtain acoustic representation may be correlated with each other or the lower dimensional latent spaces obtained might have captured both signal and error components that we have failed to isolate. The other possibility is that the scale for relative computation conceptualized by the interaction between internal model and externally driven representation is correlated but not uniquely dissociable at the level of an individual neuron.

We believe our study shows, using natural stimulus or birdsong and with robusting modeling approaches, neurons can carry information about multiple computations in the predictive coding framework. There are two ways to think about this problem (1) We view signal, prediction and error features of predictive coding as being defined by externally measured variables, whereas these features should be defined by patterns of spiking activity. These different ways of defining representations of the stimulus might be correlated with one another, but not at a level that allows resolution of explicit contributions about the internal model or incoming sensory signal within individual neurons. In that respect, the question remains whether these findings are imposed by the implementation of predictive coding models or by our understanding of the relationship between external variables, that we measure to represent the stimulus, and the internal model, and the geometries of internal spiking representations that carry sensory information.

METHODS

Subjects

Under a protocol approved by the Institutional Animal Care and Use Committee of the University of California, San Diego, we collected electrophysiology data collected from n=8 adult (>1 yr old) European starlings (*Sturnus vulgaris*). All birds were wild-caught in southern California. We did not control for the sex of the subject.

Stimuli

European starlings (both male and female) produce long, spectrotemporally rich and individualized songs composed of repeating, shorter acoustic segments known as 'motifs' that are learned over the bird's lifespan. Motifs are on the order of 0.5 to 1.5 s in duration, and largely unique to each individual. To create all the stimuli for these experiments, we used a large library of previously recorded natural starling songs as stimuli (Arneodo et al., 2019). Original song samples were recorded from fourteen European starlings at either 44.1 or 48 kHz (16-bit) over the course of several days to weeks, at various points throughout the year in sound-isolated chambers. Some birds were administered testosterone before audio recordings to increase singing behavior.

Training stimuli.

To construct training stimuli for the Temporal Convolutional Model (TCM) neural network, we downsampled each song waveform in the zenodo database to 24 kHz and computed its corresponding STFT spectrogram with NFFT=128, a Hanning window of length 128, and 50% window overlap. We excluded the DC component and log scaled the spectrogram magnitudes. To reduce their dimensionality, we averaged pairwise neighboring time-bins twice and frequency bins once as in (Kozlov & Gentner, 2016)yielding spectrograms with 32 frequency bins and approximately 10.5 ms time bins. We then parsed spectrograms into segments of 32 time bins in length (ca. 336 ms), with neighboring segments offset by a single time step (10.5 ms). The set of resulting 32 x 32 (1024) dimensional spectrogram segments served as inputs for the TCM network training. We paired each input spectrogram segment with the immediately following, non-overlapping segment from the same song, which became the output training target of the TCM network (see Figure 1). We refer to the input and output segments as the "current" and "future" segments, respectively. We reserved 10% of the training set as a validation subset.

Physiological test stimuli.

Starlings produce long, spectrotemporally rich, and individualized songs composed of repeating, shorter acoustic segments known as "motifs" that are learned over the bird's lifespan. Motifs are on the order of 0.5–1.5 s in duration. During electrophysiological recording sessions, we presented five one-minute long natural birdsongs and short synthetic songs each consisting of 1-5 natural birdsong motifs. The songs and motifs were manually extracted from a larger library of natural starling songs produced by several singers. The synthetic songs were created using motifs from eleven singers. The singers featured in our test stimuli are different from the ones

employed during the training phase. Original song samples were recorded from European starlings at either 44.1 or 48 kHz (16-bit).

Temporal Convolutional Model

We trained a neural network model to generate predictions of song on the timescale of the approximate average length of starling song motifs. Specifically, the network takes a 336ms song segment as input and predicts the next immediate future song segment of the same length.

Architecture.

The TCM neural network (Figure 1) comprises both an encoder and a decoder. The encoder is a combination of 5 convolutional layers and a fully connected layer, projected to a 256 dimensional latent space, and a mirror image decoder with a fully connected layer and 5 deconvolutional layers (Table 1). Convolutional layers used different numbers of 3´3 filters and rectified linear unit (ReLU) as the nonlinear activation function. Batch normalization was not used. The TCM network served as our proxy for the internal generative model, and we used the 256-d latent space to represent the features of any given input (current segment) that are most predictive of the output (future segment). As a control representation of all the spectro-temporal features in song, we trained a convolutional autoencoder to the TCM network, except that it learned to reconstruct the current (input) segment rather than the future.

We implemented, trained, and tested the TCM and autoencoder networks described above using Tensorflow. The convolutional and linear weights were initialized to be uniformly random (the default setting in Tensorflow). The loss function for TCM is a sum of a "prediction

loss" and "distance loss". We defined prediction loss as the mean squared error between the true and predicted output, and distance loss using a Multi-dimensional Scaling (MDS) function that maximally preserves the pairwise distances between input and output data points, and favors the preservation of global dataset structure. The architectural setting of our autoencoder model was similar to the TCM with stacked convolutional layers. The autoencoder loss function is a sum of "reconstruction loss", defined as the mean squared error between true and reconstructed output, and MDS distance loss. The models were trained with the ADAM optimizer with a learning rate 0.001, and using mini-batches of size 128 and without using dropout regularization. All additional architectural details can be found in the GitHub repository. The autoencoder we used followed those in our AVGN repository (http://github.com/tsainb/avgn_paper).

Signal, prediction, and prediction error representations.

Each current spectrogram segment was sampled from the physiological test stimulus dataset for each block and passed through the TCM to produce the corresponding future spectrogram segment. The current spectrogram segment corresponds to *signal-fft* segment; we considered TCM latent space as *prediction* features; and the *prediction error* segment is computed by the squared difference between the predicted future spectrogram segment and true future spectrogram segment at each pixel. The autoencoder latent space obtained using the same current spectrogram segment was considered as *signal-ae* features.

Electrophysiology

Naive starlings were anesthetized with urethane (0.7 mg/kg) and head-fixed to a stereotaxic apparatus in an acoustic isolation chamber (Acoustic Systems). A small craniotomy

was opened over the region NCM, CMM or field L. We placed multi-channel silicon electrodes (Masmanidis 128DN; Yang et al., 2020) in the region of interest until auditory-evoked activity was observed on the majority of channels. Songs were played to the subjects at 60-dB mean-level while we recorded action potentials extracellularly. After placing the electrode, the bird was left in silence for 30-60 minutes prior to starting trials. Each stimulus was repeated 20 times in a random order with a random inter-trials interval of 2-5 seconds. Recording blocks were obtained from independent populations of cells in both hemispheres and at different depths within the region of interest (Table 2)

Spike Sorting

Simultaneously recorded blocks were spike sorted with Kilosort2 (Stringer et al., 2019). Automatic sorting was performed in Matlab (Mathworks), and putative units were manually curated using the phy interface. Single units were identified by clustering principal components of the spike waveforms, only when fewer than 1% violations of the refractory period (assumed to equal 2 ms) occurred, and only from recordings with an excellent signal-to-noise ratio (large-amplitude extracellular action-potential waveforms). Rest of the units were classified as multiunit activity or noise based on signal-to-noise ratio. Only putative single units were used for further analyses.

Maximum Noise Entropy (MNE) Models

MNE Inputs.

MNE models were fit for each of the obtained acoustic feature spaces: 1) *signal-fft* segments pairwise averaged across both frequency and temporal axes to form 16x16 (256)

dimensional spectrogram segments that serve as the stimuli for the *signal-fft* MNE model; 2) latent spaces obtained from the TCM (256 dimensions) served as the stimuli for *prediction* MNE model; 3) the *prediction error* segments were pairwise averaged across both frequency and temporal axes to form 16x16 (256) dimensional spectrogram segments to serve as stimuli for the *prediction error* MNE model and 4) autoencoder latent spaces (256 dimensions) served as stimuli for *signal-ae* MNE model. The stimulus samples were z-scored before inputting into their corresponding MNE models.

For each single unit and for each trial, the unit's spike train was binned in time to correspond with the outputs of the Temporal Convolutional Model described above. Likewise, the temporal spike bins were added pairwise along with the four acoustic feature spaces to yield bins of about 10.5 ms in length indicating the number of spikes in from this neuron in that bin. Trial average spike count was computed in each bin and treated as the response vector. MNE inputs were split into 80-20% train and test set.

Shuffling spiking responses.

To isolate the effects of temporal organization, we computed shuffled versions of responses of individual neurons. We computed the "permuted shuffle" by shuffling each time bin across the entire stimulus set, while preserving the spiking coactivity across all cells within a single time bin and trial. This shuffle destroys stimulus-locked responses but preserves the temporal coactivation structure of the population response.

Fitting MNE Models.

MNE models fit a logistic function to capture the spiking response of each neuron, characterizing it as a linear combination of first- and second-order features of the stimulus. Parameters *a*, *h* and *J* were optimized to constrain stimulus-response relationships representing mean firing rate, spike-triggered average and spike-triggered covariance respectively. The model optimization process adopts a four-fold jackknife approach, where acoustic inputs and response data were split into four batches. In each jackknife iteration, three of the batches were used as training data, and the remaining batch was reserved for testing. Using a nonlinear conjugate gradient algorithm, the log loss between the response and the weighted input was minimized. To prevent overfitting, early stopping with a ten-epoch criterion on the testing data log loss is enforced as a regularization measure. The optimized weights from each jackknife were then averaged, yielding a set of mean weights used in all subsequent analyses. A detailed description of the method is given in Fitzgerald et al. 2011. The optimized MNE model takes the form:

$$P(y|s) = \frac{1}{1 + e^{-(a+hs+s^T Js)}} \qquad (1)$$

where $P(y|s)$ is a time series of predicted spiking probability.

Modeling MNE responses to novel stimuli.

Trained MNE models were used to generate predicted neural responses. To predict the response of a unit, the parameters computed from the trained model were applied to the testing set. Response predictions were generated using either the complete model or only the trained linear features (parameters *a* and *h*). To assess prediction accuracy, Pearson's correlation coefficient (*r*) was calculated to quantify the degree of correlation between the predicted spiking probability and the empirical response given by the averaged trial-average spike counts for each

unit (Figure 2E). These values were used to compare prediction quality across MNE models across different acoustic feature spaces.

Variance partitioning of MNE modeled responses

Regressors.

The regressors we used to model empirical neuronal spiking activity in response to birdsong are organized into three feature spaces: the general spectrotemporal features of song, features predictive of incoming song outputted by TCM and residual prediction error. The regressors are based on the response probabilities modeled on MNE testing set derived using their respective CRFs (Equation 1).

Variance Partitioning.

The regression analysis adapted from multiple linear regression was performed using custom code written in R. The empirical neural response for each unit is predicted as a linear combination of the *signal-fft*, *prediction* and *prediction error* MNE modeled responses. Specifically, the following models were fit containing feature spaces:

1) signal
2) prediction
3) prediction error
4) signal and prediction
5) signal and error
6) prediction and error
7) signal, prediction and error

Adjusted R-squared score was computed as a performance measure. Partial Pearson's correlation coefficients calculated using ppcor library in R were used to measure unique and shared variances of signal, prediction and prediction error.

Categorality metric for unique variances

HDBSCAN clustering was performed on partial Pearson's correlation coefficients of signal, prediction and prediction error of individual units within each simultaneously recorded population. Each clustering used default parameterization of HDBSCAN, setting the minimum samples at 1. The minimum cluster size parameter is varied to achieve the best possible clusterability quantified by silhouette score (Equation 2 and 3). The silhouette score is the mean silhouette coefficient across all of the samples in a dataset, where the silhouette coefficient measures how distant each point is to points in its own category, relative to its distance from the nearest point in another category. It is therefore taken as a measure of how well clustered together elements are that belong to the same category. Silhouette scores range from -1 to 1, with 1 being more clustered.

The silhouette score S, is computed as the mean of the silhouette coefficients for each data point. For each data point (i), the silhouette coefficient s_i is the mean distance between the data point and all other data points in the same cluster (a_i), minus the distance to that points nearest neighbor belonging to a different cluster (b_i), divided by the maximum of a_i and b_i, which can be written as:

$$s_i = \begin{cases} 1 - \dfrac{a_i}{b_i}, & if\ a_i < b_i \\ 0, & if\ a_i = b_i \\ \dfrac{b_i}{a_i} - 1, & if\ a_i > b_i \end{cases} \qquad (2)$$

$$S = \frac{1}{n}\sum_{i=1}^{n} s_i \qquad (3)$$

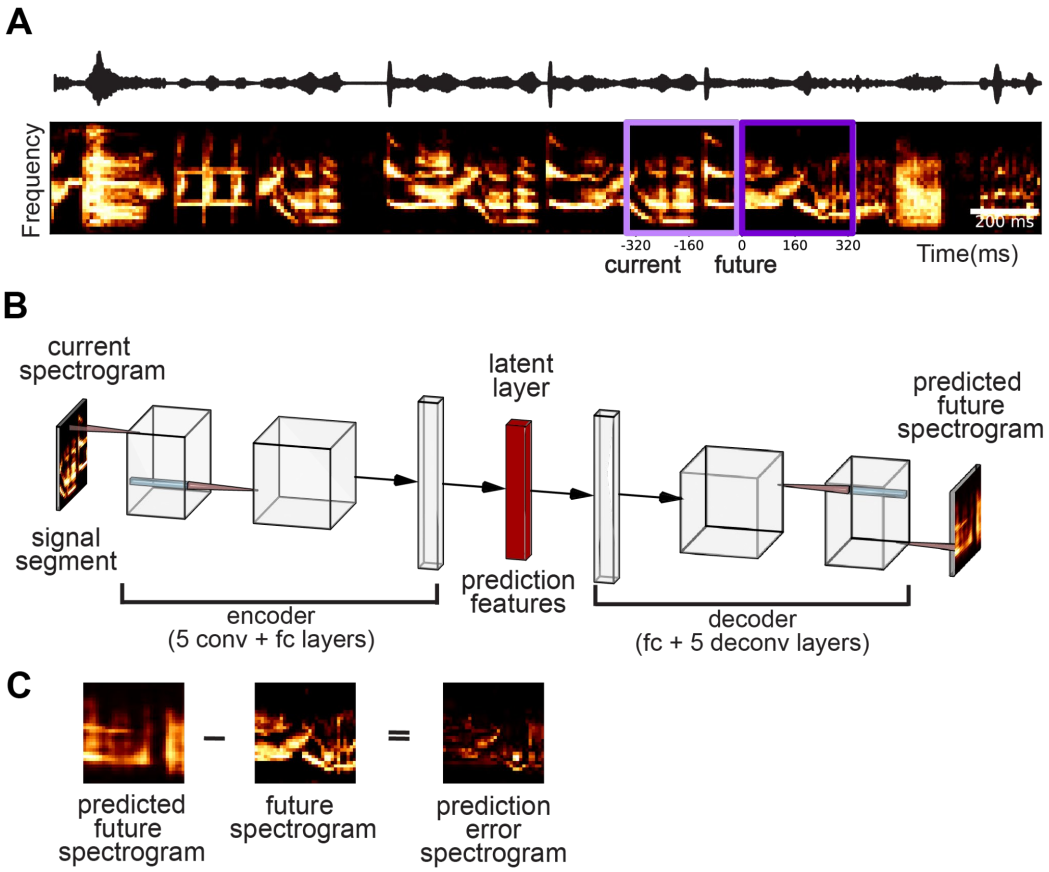

Figure 1. Complex acoustic feature spaces of European starling song obtained by a representative internal generative model. A) Top: Segment of an example starling song showing the time-varying pressure waveform (top) and spectrogram (bottom; frequency range=0.3-12 kHz, scale bar=200 ms). For predictive neural network (NN) training, we divided spectrograms into contiguous "current" and "future" segments each roughly 340 ms in length (32 freq channels x 32 time bins). B) Schematic of the Temporal Convolutional Model (TCM) trained on the starling song library to predict future segments from current segments of song. The NN architecture shows convolutional and fully connected layers with current and future spectro-temporal segments of 32 frequencies and 32 time bins used as inputs and outputs, respectively, and the latent space comprising 256 hidden units. TCM model loss is a combination of prediction loss (computed as mean squared error between true and predicted outputs) and distance loss (computed using multi-dimensional scaling). Spectrogram-driven activation of the trained latent space is considered to be the *prediction* representation of the stimulus, i.e. the latent features of the input that optimally predict the upcoming future spectrogram segment. C) Example of a *prediction error segment* determined by the squared difference between true future spectrogram segment and TCM predicted spectrogram segment.

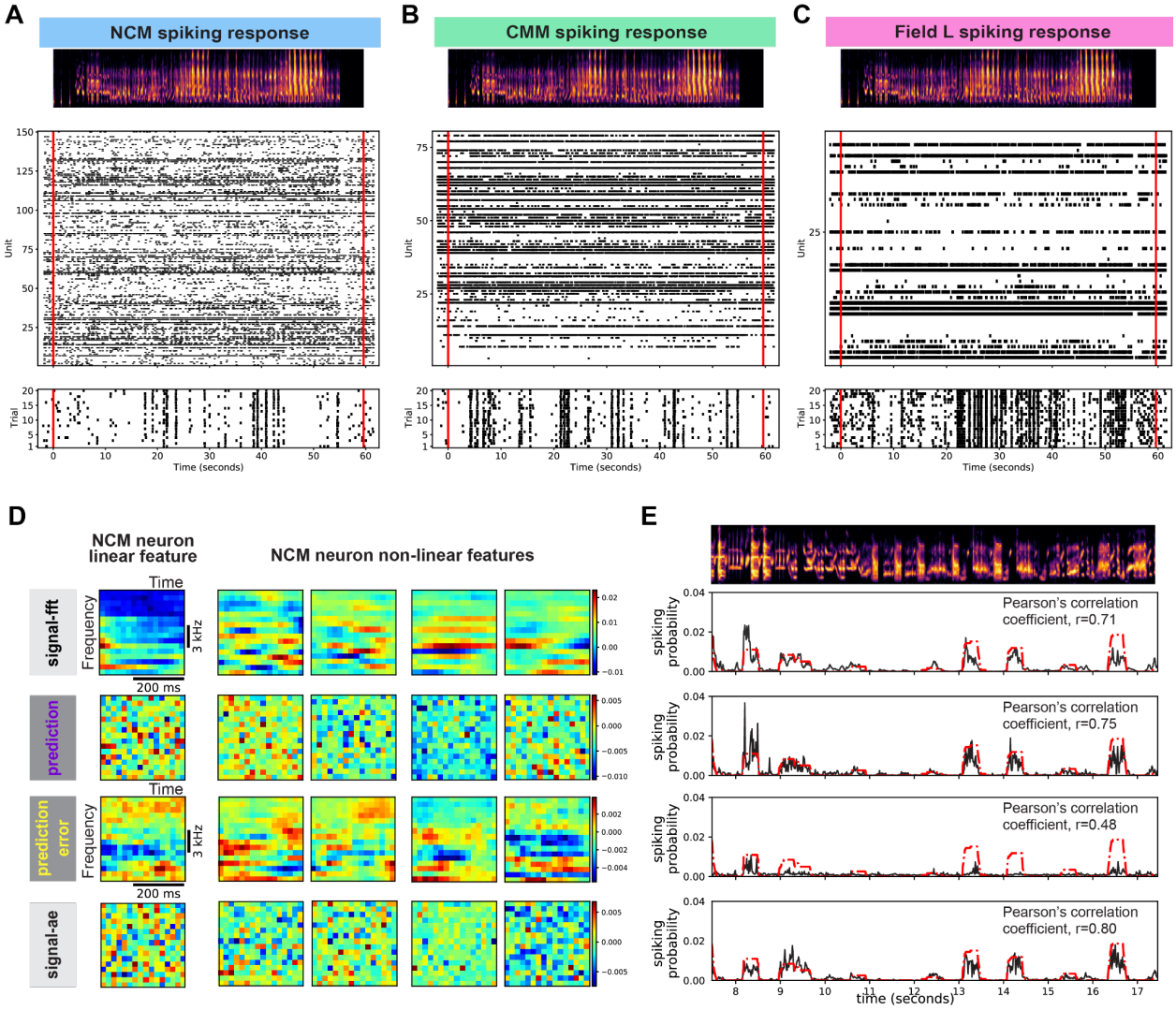

Figure 2. Composite receptive fields of auditory neurons fit to different stimulus representations. Extracellular population response of auditory neurons recorded simultaneously in (A) NCM (n=150 units), showing spectrogram of example stimulus (top), population spike rasters where each row shows the response of a unique unit to the same presentation of the example stimulus (middle), and a single unit spike raster showing the response to 20 repetitions of the example stimulus (bottom). Red horizontal lines denote the start and end of the stimulus. B, C) as in A) but for simultaneously recorded units in CMM (n=79 units) and Field L (n=43 units), respectively. D) Composite receptive fields (CRF; columns) of one NCM neuron fit to different stimulus representations (rows) showing linear component (left-most column, obtained from parameter *h* of the MNE model) and four most significant non-linear components (right-most columns, obtained from MNE parameter *J* for each representation of the stimulus. *signal-fft*: spiking response is fit to spectrogram coefficients. *prediction*: spiking response is fit to only predictive features of song from TCM latent space. *prediction error*: spiking response of fit to error representation of stimulus. *signal-ae*: spiking response of fit to 256-d representation of song from the latent space of the autoencoder. Note that the TCM and autoencoder latent weight matrices and their corresponding CRFs do not have a 2-d spectro-temporal structure. E) CRF

response predictions for novel stimuli not used to compute the CRF. Response probabilities modeled by *signal-fft-CRFs* (black) correlated with empirical neuron response (red) (Pearson's correlation coefficient r=0.71), *prediction-CRF* modeled response probabilities (purple; r=0.75), *prediction error-CRF* modeled response probabilities (yellow; r=0.48) and *signal-ae-CRF* modeled response probabilities (gray; r=0.80).

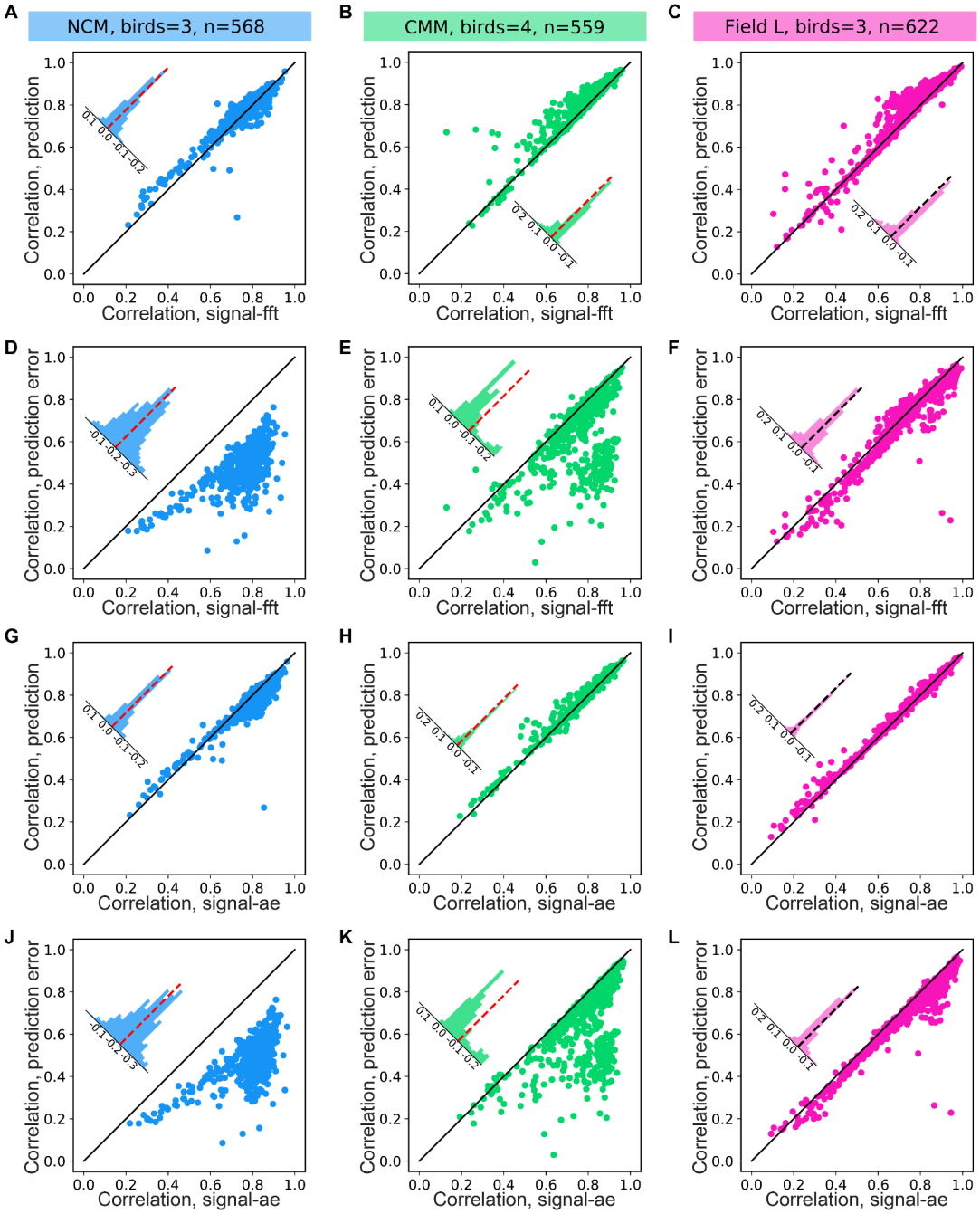

Figure 3. Single neuron responses in auditory regions are best modeled by acoustic features constituting predictive and prediction error features. Distribution of Pearson's correlation coefficient values of *prediction-CRF* modeled responses with empirical responses when compared with correlation coefficient values of *signal-fft-CRF* modeled responses with empirical responses in A) NCM (birds=3, n=568 units), B) CMM (birds=4, n=559 units) and C) Field L (birds=3, n=622 units). Distribution of Pearson's correlation coefficient values of prediction *error-CRF* modeled responses with empirical responses when compared with correlation coefficient values of *signal-fft-CRF* modeled responses with empirical responses in D) NCM, E) CMM and F) Field L. Scatterplot of response correlations between CRF modeled and empirical

neural responses using *prediction-CRF* in G) NCM, H) CMM and I) Field L, and *prediction-error-CRF* in J) NCM, K) CMM and L) Field L compared to response correlations modeled using *signal-ae-CRFs*. Dashed lines in residual histograms indicate the position of means in each distribution. In all plots, the diagonal line indicates unity. A distribution plot of residuals is shown with the dashed line indicating the mean residual.

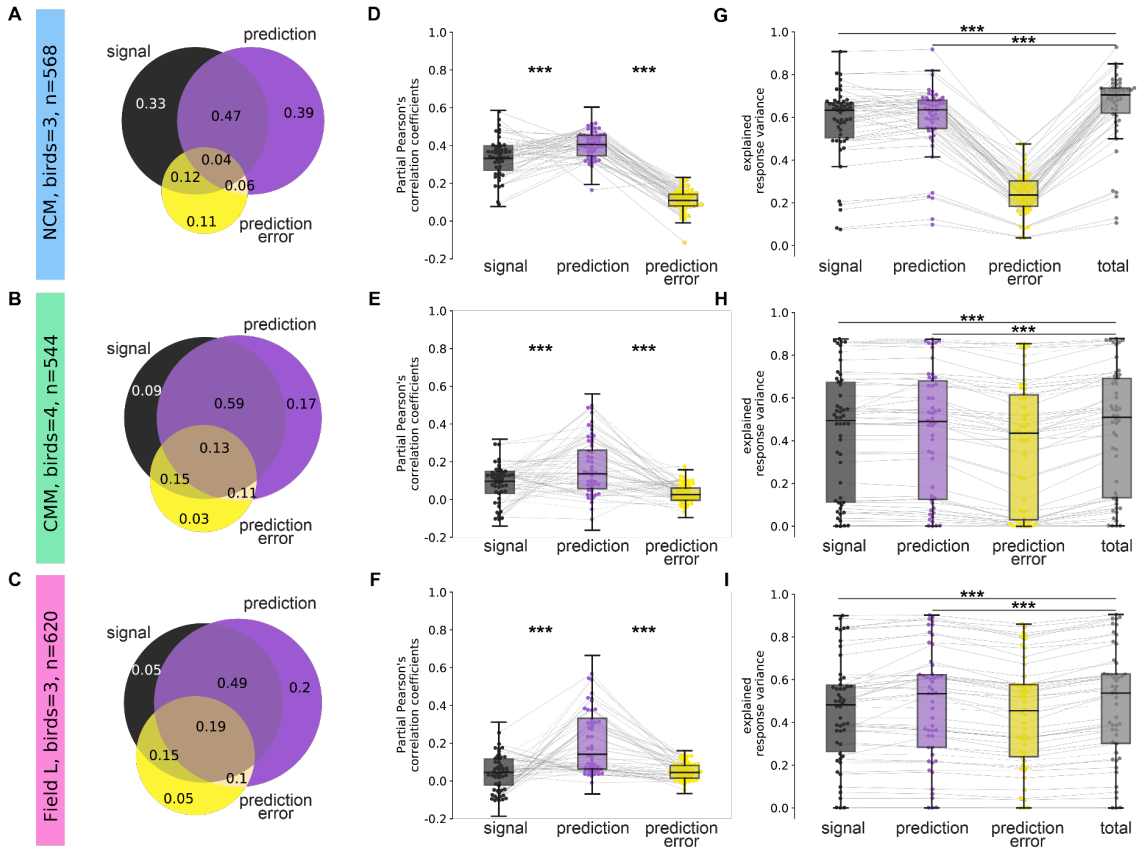

Figure 4. Relative contributions of signal, prediction and prediction error features vary across auditory regions. Visualization of the mean unique and shared contributions (partial Pearson's correlation coefficients) of signal (black), prediction (purple) and prediction error (yellow) acoustic features in A) NCM (birds=3, n=568 units), B) CMM (birds=4, n=544 units) and C) Field L (birds=3, n=620 units). Partial Pearson's correlation coefficients representing the unique contributions of (i) signal, (ii) prediction, (iii) prediction error, for a sample of n=50 neurons in D) NCM, E) CMM and F) Field L. We compared performances determined by variance in the neural signal of signal-prediction-error model to those of signal, prediction and prediction error in G) NCM H) CMM I) Field L. Barplots represent the overall distribution of all neurons with a swarmplot for a sample of n=50 neurons. * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

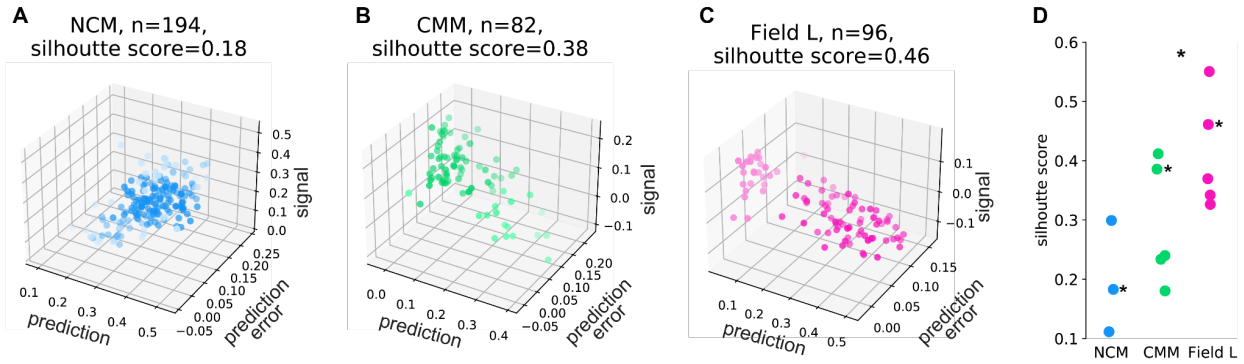

Figure 5. Clusterability of auditory neurons based on unique contributions of signal, prediction and prediction error. HDBSCAN clustering of partial Pearson's correlation coefficients corresponding to unique signal, prediction and error variances of neurons within a simultaneously recorded population in A) NCM (n=194) B) CMM (n=82) and C) Field L (n=96). D) Distribution of silhouette scores for all recorded populations showing differing clusterability in NCM (recorded populations n=3), CMM (n=5) and Field L (n=6; one-way ANOVA F = 5.83, p=0.02). Higher silhouette scores indicative of higher clusterability. * p<0.05, ** p<0.01, *** p<0.001.

References

Amin, N., Doupe, A., & Theunissen, F. E. (2007). Development of Selectivity for Natural Sounds in the Songbird Auditory Forebrain. *Journal of Neurophysiology*, *97*(5), 3517–3531.

Arneodo, Z., Sainburg, T., Jeanne, J., & Gentner, T. (2019). An acoustically isolated European starling song library (Version v1) [dataset]. Zenodo.

Asilador, A., & Llano, D. A. (2021). Top-Down Inference in the Auditory System: Potential Roles for Corticofugal Projections. *Frontiers in Neural Circuits*, *14*, 615259.

Attinger, A., Wang, B., & Keller, G. B. (2017). Visuomotor Coupling Shapes the Functional Development of Mouse Visual Cortex. *Cell*, *169*(7), 1291-1302.e14.

Audette, N. J., Zhou, W., La Chioma, A., & Schneider, D. M. (2022). Precise movement-based predictions in the mouse auditory cortex. *Current Biology*, *32*(22), 4925-4940.e6.

Ayaz, A., Stäuble, A., Hamada, M., Wulf, M.-A., Saleem, A. B., & Helmchen, F. (2019). Layer-specific integration of locomotion and sensory information in mouse barrel cortex. *Nature Communications*, *10*(1), 2585.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical Microcircuits for Predictive Coding. *Neuron*, *76*(4), 695–711.

Bimbard, C., Sit, T. P. H., Lebedeva, A., Reddy, C. B., Harris, K. D., & Carandini, M. (2023). Behavioral origin of sound-evoked activity in mouse visual cortex. *Nature Neuroscience*, *26*(2), 251–258.

Bonke D., Scheich H., Langner G. Responsiveness of units in the auditory neostriatum of the guinea fowl (Numida meleagris) to species-specific calls and synthetic stimuli. *J Comp Physiol [A]* 132: 243–255, 1979.

Brainard, M. S., & Doupe, A. J. (2000). Auditory feedback in learning and maintenance of vocal behaviour. *Nature Reviews Neuroscience*, *1*(1), 31–40.

Carlson, N. L., Ming, V. L., & DeWeese, M. R. (2012). Sparse Codes for Speech Predict Spectrotemporal Receptive Fields in the Inferior Colliculus. *PLoS Computational Biology*, *8*(7), e1002594.

Eliades, S. J., & Wang, X. (2008). Neural substrates of vocalization feedback monitoring in primate auditory cortex. *Nature*, *453*(7198), 1102–1106.

Fitzgerald, J. D., Sincich, L. C., & Sharpee, T. O. (2011). Minimal Models of Multidimensional Computations. *PLoS Computational Biology*, *7*(3), e1001111.

Garner, A. R., & Keller, G. B. (2022). A cortical circuit for audio-visual predictions. *Nature Neuroscience*, *25*(1), 98–105.

Garrido, M. I., Kilner, J. M., Stephan, K. E., & Friston, K. J. (2009). The mismatch negativity: A review of underlying mechanisms. *Clinical Neurophysiology*, *120*(3), 453–463.

Gentner, T. Q., & Margoliash, D. (2003). Neuronal populations and single cells representing learned auditory objects. *Nature*, *424*(6949), 669–674.

Gilday, O. D., & Mizrahi, A. (2023). Learning-Induced Odor Modulation of Neuronal Activity in Auditory Cortex. *The Journal of Neuroscience*, *43*(8), 1375–1386.

Gill, P., Woolley, S. M. N., Fremouw, T., & Theunissen, F. E. (2008). What's That Sound? Auditory Area CLM Encodes Stimulus Surprise, Not Intensity or Intensity Changes. *Journal of Neurophysiology*, *99*(6), 2809–2820.

Grace, J. A., Amin, N., Singh, N. C., & Theunissen, F. E. (2003). Selectivity for Conspecific Song in the Zebra Finch Auditory Forebrain. *Journal of Neurophysiology*, *89*(1), 472–487.

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

Holey, B. E., & Schneider, D. M. (2023). *Sensation and expectation are embedded in mouse motor cortical activity* [Preprint]. Neuroscience.

Huang, K.-H., Rupprecht, P., Frank, T., Kawakami, K., Bouwmeester, T., & Friedrich, R. W. (2020). A virtual reality system to analyze neural activity and behavior in adult zebrafish. *Nature Methods*, *17*(3), 343–351.

Keller, G. B., Bonhoeffer, T., & Hübener, M. (2012). Sensorimotor Mismatch Signals in Primary Visual Cortex of the Behaving Mouse. *Neuron*, *74*(5), 809–815.

Keller, G. B., & Hahnloser, R. H. R. (2009). Neural processing of auditory feedback during vocal practice in a songbird. *Nature*, *457*(7226), 187–190.

Kingma, D. P., & Welling, M. (2022). *Auto-Encoding Variational Bayes* (arXiv:1312.6114). arXiv. http://arxiv.org/abs/1312.6114

Klein, D. J., König, P., & Körding, K. P. (2003). Sparse Spectrotemporal Coding of Sounds. *EURASIP Journal on Advances in Signal Processing*, *2003*(7), 902061.

Kozlov, A. S., & Gentner, T. Q. (2016). Central auditory neurons have composite receptive fields. *Proceedings of the National Academy of Sciences*, *113*(5), 1441–1446.

Lecun, Y. and Bengio, Y (1995). Convolutional Networks for Images, Speech and Time-Series.

Leppelsack, H. J., & Vogt, M. (1976). Responses of auditory neurons in the forebrain of a songbird to stimulation with species-specific sounds. *Journal of Comparative Physiology ? A*, *107*(3), 263–274.

Lesicko, A. M., Angeloni, C. F., Blackwell, J. M., De Biasi, M., & Geffen, M. N. (2022). Corticofugal regulation of predictive coding. *eLife*, *11*, e73289.

Lesicko, A. M. H., & Geffen, M. N. (2022). Diverse functions of the auditory cortico-collicular pathway. *Hearing Research*, *425*, 108488.

Lewicki, M. S., & Arthur, B. J. (1996). Hierarchical Organization of Auditory Temporal Context Sensitivity. *The Journal of Neuroscience*, *16*(21), 6987–6998.

Margoliash, D. (1986). Preference for autogenous song by auditory neurons in a song system nucleus of the white-crowned sparrow. *The Journal of Neuroscience*, *6*(6), 1643–1661.

McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, *2*(11), 205.

Müller, C. M., & Leppelsack, H.-J. (1985). Feature extraction and tonotopic organization in the avian auditory forebrain. *Experimental Brain Research*, *59*(3).

Näätänen, R., Paavilainen, P., Rinne, T., & Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: A review. *Clinical Neurophysiology*, *118*(12), 2544–2590.

Näätänen, R., Pakarinen, S., Rinne, T., & Takegata, R. (2004). The mismatch negativity (MMN): Towards the optimal paradigm. *Clinical Neurophysiology*, *115*(1), 140–144.

Narula, G., & Hahnloser, R. H. R. (2021). Songbirds are excellent auditory discriminators, irrespective of age and experience. *Animal Behaviour*, *175*, 123–135.

Netser, S., Zahar, Y., & Gutfreund, Y. (2011). Stimulus-Specific Adaptation: Can It Be a Neural Correlate of Behavioral Habituation? *The Journal of Neuroscience*, *31*(49), 17811–17820.

Nottebohm, F. (2005). The Neural Basis of Birdsong. *PLoS Biology*, *3*(5), e164.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87.

Sainburg, T., Theilman, B., Thielk, M., & Gentner, T. Q. (2019). Parallels in the sequential organization of birdsong and human speech. *Nature Communications*, *10*(1), 3636.

Sainburg, T., Thielk, M., & Gentner, T. Q. (2020). Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLOS Computational Biology*, *16*(10), e1008228.

Schultz, W., & Dickinson, A. (2000). Neuronal Coding of Prediction Errors. *Annual Review of Neuroscience*, *23*(1), 473–500.

Singer, Y., Teramoto, Y., Willmore, B. D., Schnupp, J. W., King, A. J., & Harper, N. S. (2018). Sensory cortex is optimized for prediction of future input. *eLife*, *7*, e31557.

Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M., & Harris, K. D. (2019). High-dimensional geometry of population responses in visual cortex. *Nature*, *571*(7765), 361–365.

Theunissen, F. E., Amin, N., Shaevitz, S. S., Woolley, S. M. N., Fremouw, T., & Hauber, M. E. (2004). Song Selectivity in the Song System and in the Auditory Forebrain. *Annals of the New York Academy of Sciences*, *1016*(1), 222–245.

Ulanovsky, N., Las, L., & Nelken, I. (2003). Processing of low-probability sounds by cortical neurons. *Nature Neuroscience*, *6*(4), 391–398.

Vahidi, N. W. (2021). *Spatial and Temporal Organization of Composite Receptive Fields in the Songbird Auditory Forebrain* [Preprint]. Neuroscience.

Wickramasinghe, P., & Sice, G. F. (2021). Multidimensional Scaling for Gene Sequence Data with Autoencoders. *2021 2nd International Conference on Computing and Data Science (CDS)*, 516–523.

Woolley, S. M. N., Fremouw, T. E., Hsu, A., & Theunissen, F. E. (2005). Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. *Nature Neuroscience*, *8*(10), 1371–1379.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.

Yang, L., Lee, K., Villagracia, J., & Masmanidis, S. C. (2020). Open source silicon microprobes for high throughput neural recording. *Journal of Neural Engineering*, *17*(1), 016036.

Zhao, L., & Zhaoping, L. (2011). Understanding Auditory Spectro-Temporal Receptive Fields and Their Changes with Input Statistics by Efficient Coding Principles. *PLoS Computational Biology*, *7*(8), e1002123.

Zmarz, P., & Keller, G. B. (2016). Mismatch Receptive Fields in Mouse Visual Cortex. *Neuron*, *92*(4), 766–772.