Functional Consequences of Synaptic Plasticity in Sensory Systems

Patrick Gill, Noopur Amin, Thane Fremouw, Sarah M. N. Woolley and Frédéric Theunissen

We present a model showing that relevant sensory information becomes efficiently encoded in sensory areas that exhibit two widespread spike timing-dependent phenomena: synaptic and excitability plasticity. Our model predicts neural networks represent surprising stimulus features relevant to post-sensory tasks. We show two supporting electrophysiological results in the urethane-anesthetized male zebra finch. First, neurons throughout the auditory pathway are excited more by surprising stimuli than by stimulus changes, suggesting that excitatory synapses in general are strengthened by surprise. Second, neurons in auditory areas immediately presynaptic to the song system represent only surprising features of birdsong, while auditory areas presynaptic to less specialized forebrain areas also represent surprising features of environmental sounds. Our model is not specific to birdsong, and may therefore illuminate how neurons in general specialize at encoding task-related information efficiently.

Neuroscientists characterize most general-purpose auditory[1], visual[2] and somatosensory[3] neurons as change detectors. Change-detecting sensory neurons are generally modeled using center-surround receptive fields[4] with balanced inhibitory and excitatory components[5] in order to be able to ignore broad, slow changes in the intensity of natural stimuli[6], which is efficient because there are ample opportunities to encode these extensive, persistent features. Difference-detecting, center-surround receptive fields are therefore seen as being useful in reducing the redundancy in the neural representation of natural stimuli[7], which tend to be highly correlated[8-10].

However, in a recent paper[11] we showed that neurons in Caudal Lateral Mesopallium (CLM, an auditory forebrain area specialized at representing conspecific song[12]) of the male zebra finch are best described not with a derivative code but with a surprise code. Surprise is defined as the minus logarithm of the stimulus probability given knowledge of the recent stimulus history and given knowledge of the statistics of the stimulus class (see Equation 3 in Methods and Figure 1 c). In surprise codes the firing probability is modeled using a surprise-Spectro-Temporal Receptive Field (surprise-STRF, see Figure 1 f) convolved with a representation of the stimulus surprise. Some changes in the stimulus are predictable, and predictable changes cause spikes in a derivative code but not a surprise code (e.g., compare the stronger representation of predictable stimulus changes such as those from 0.84 - 0.9 s in the derivative representation of Figure 1 b to their surprise representation in Figure 1 c).

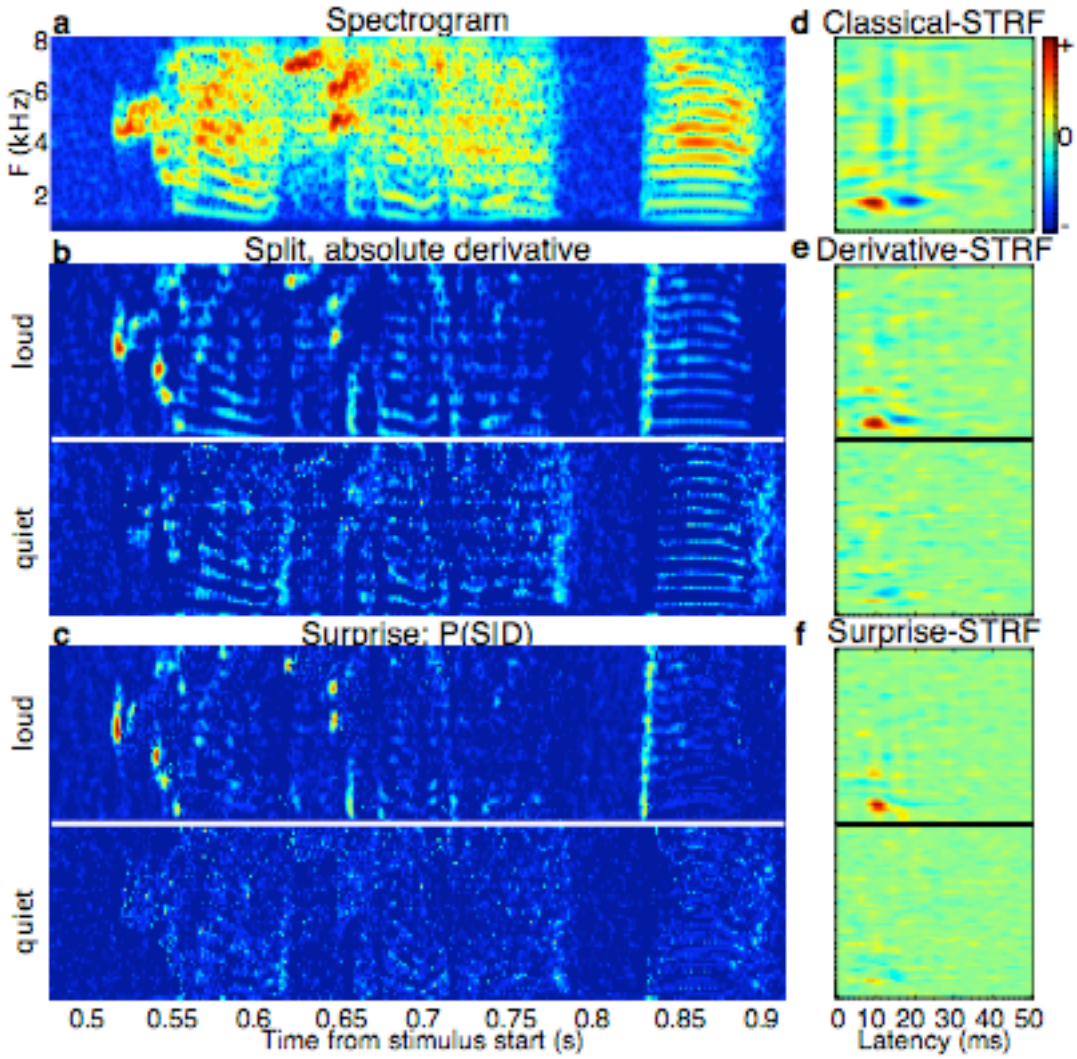

Figure 1: **Stimulus Representations and Associated STRFs** a) Spectrogram of a sample of zebra finch song. b) Split, absolute derivative of the spectrogram: louder-than-context entries appear in the top half of the representation labeled "loud" and quieter-than-context entries appear in the bottom half, labeled "quiet". See Methods and Equation 8 for more details. c) Surprise representation of the spectrogram: surprise of every stimulus element is proportional to its unexpectedness. See Methods and Equation 3. Quieter- and louder-than-expected elements are likewise separated. Predictable intra-syllable changes have a weaker representation than in b). d) Classical-STRF of a CLM neuron, obtained by reverse correlation to the spectrogram. e) Derivative-STRF of the same neuron, obtained from reverse correlation to the split, absolute derivative. f) Surprise-STRF of the same neuron, obtained from reverse correlation to stimulus surprise.

In our earlier study[11], we found PSTHs in CLM are described 40% better ($p = 5 * 10^{-7}$) using a surprise-STRF (which assumes firing probabilities are proportional to surprise) than using a derivative-STRF. This prediction improvement is due to

the fact that CLM neurons fire less to expected stimulus changes than to the intrinsically unpredictable elements of zebra finch song, yielding a sparser, more efficient neural code than the derivative code.

The existence of surprise coding in CLM leaves three questions unanswered. First, if neurons encode unexpected events, then why is CLM not driven best by totally random artificial stimuli (such as white noise), which violate expectations maximally? Second, if having costs proportional to the logarithm of probabilities is the most efficient way to represent a stimulus[13], why is surprise coding found mostly in the secondary forebrain, but less in the primary auditory forebrain and not in the auditory midbrain[11]? Third, how do neural networks implement surprise codes?

We show here that these three questions can be answered simultaneously using a model incorporating two established spike timing dependent plasticity (STDP) phenomena. The first phenomenon is that if two synaptically-connected neurons fire action potentials within approximately 40 ms of each other, the strength of the synapse will increase if the presynaptic neuron fires first but decrease if the postsynaptic neuron fires first. This form of STDP has been demonstrated in many animal systems[14, 15] including the zebra finch forebrain[16]. The second phenomenon is that the presynaptic neuron becomes globally more excitable under conditions which cause the synapse to strengthen, but if the synapse is weakened through STDP the presynaptic neuron becomes less excitable[17-20]. While some STDP details are still not known (e.g. we have just begun to understand how synaptic strength and excitability saturate[21]), the qualitative aspects of these two STDP phenomena are sufficient to provide a model capable of answering the three questions posed above.

Results

1. Qualitative Model

In this section we explain how, under STDP, sensory neurons with surprising content that drives a postsynaptic area become more excitable than neurons with redundant sensory information. By definition of surprise, neurons with surprising content fire at the first possible indication of a stimulus feature (see Discussion). Suppose these surprising neurons also elicit action potentials in a downstream brain area. Then neurons with surprising, task-relevant (i.e. relevant to the postsynaptic area) content fire before both the postsynaptic neurons they drive and peer sensory neurons with un-surprising content. Under the two STDP rules mentioned above (synapse strength and excitability plasticity), surprising, task-relevant neurons become more excitable and have stronger excitatory synapses; neurons encoding redundant task-relevant information become less excitable and come to have weaker excitatory synapses. Here, being redundant to the

postsynaptic task is defined to mean that the postsynaptic targets consistently fire due to inputs from surprising neurons before the redundant neurons fire.

In summary, sensory areas that have neurons responding to both surprising and predictable content should find their surprising neurons become more excitable while their predictable neurons become less excitable. However, the trend to enhance surprise depends on action potentials in an area postsynaptic to (and driven by) that sensory area, without which STDP mechanisms would not apply, since there would be no stimulus-locked postsynaptic spikes. Section 4 in Results demonstrates reliance on postsynaptic activity *in vivo*.

Electrophysiological recordings in all sensory areas with these two forms of plasticity should therefore reveal mostly action potentials with (as much as possible given the scope of available neural computation) surprising content relevant to the tasks in postsynaptic areas. Whichever neurons have the most surprising relevant content will have the highest firing rates and therefore will contribute more action potentials to the neural code.

2. Quantitative Model

To double-check that these two STDP rules contribute to surprise coding, we simulated small sensory neural networks (see Methods) that had mixtures of surprise- derivative- and intensity-coding sensory neurons. All three types drove common postsynaptic targets. We found that the surprise-coding neurons were the only ones whose activity stays consistently strong under STDP (see Figure 2), which is expected, since neurons with surprising content fire before the other sensory neurons. Firing before other neurons engages the mechanisms of STDP, which cause neurons to become more excitable if they fire consistently before their postsynaptic targets.

As can be seen in Figure 2, our results are largely insensitive to noise, number of neurons, or synaptic connectivity; although in general, the more synapses connecting the sensory and post-sensory areas the faster surprise emerges as the dominant coding scheme.

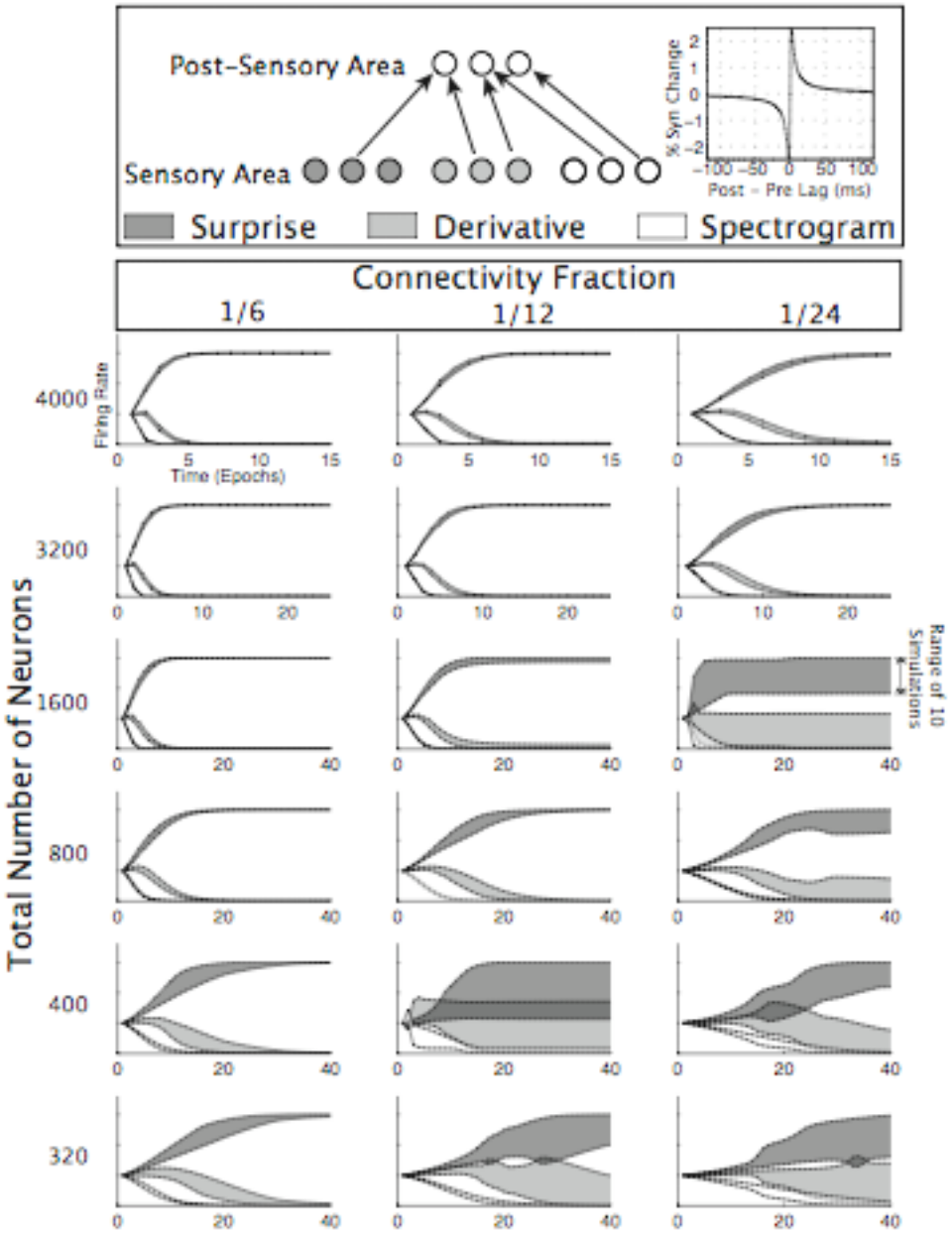

Figure 2: **Summary of Simulation Results** Top: Simulation schematics. Neurons from a sensory area have random classical-, derivative- or surprise-STRFs with fixed latency and are exposed to a zebra finch song. The sensory area drives a post-sensory area with neurons with STDP plasticity. The synapse strength and excitability in the sensory area are modified once per presentation epoch according to the synapse STDP rule (top right, see Equation 4 in Methods) and an excitability rule where excitability changes in an amount proportional to the mean synapse strength change and to a homeostatic factor which keeps the total number of spikes fired in the sensory area constant (see Methods). Main panel: 10-simulation ranges of firing rates of classical-, derivative- and surprise-neurons given different numbers of neurons (at left) and different connectivity fractions (top). In most cases,

surprise-bearing neurons eventually are the only active group, although in early stages often derivative-bearing neurons have slightly increased firing rates as well.

3. Surprise Consistently Excites in Vivo

Aside from the prediction that surprising neurons should have their excitability increased, our model states that surprising content should cause neurons to form strong excitatory synapses, and thus surprising events should (more often than not) result in an increase in firing rates. This prediction is an alternative to the hypothesis that sensory neurons act as general change detectors, which would suggest that all stimulus changes (and not just surprising ones) should result in increasing firing rates. This second hypothesis, that sensory neurons detect either increases or decreases in the stimulus intensity compared to a spatially- or temporally-close reference, is implicit in the concept of the center-surround receptive field, which we wish to refine.

While changes and surprise are often highly correlated (compare Figures 1 b and 1 c), we can test whether sensory neurons tend to be more excited by changes or by surprises by quantifying the positive components of derivative-STRFs, (see Figure 1 e) and surprise-STRFs (see Figure 1 f). We estimated classical-, derivative- and surprise-STRFs to conspecific song in four auditory areas of the urethane-anesthetized male zebra finch (see Methods): 142 neurons in Mesencephallicus Lateralis Dorsalis (MLd – the auditory midbrain), 58 neurons in Ovoidalis (Ov – the auditory thalamus), 188 neurons in Field L (the primary, general-purpose forebrain auditory area) and 37 neurons in Caudal Lateral Mesopallium (CLM, a secondary auditory area specializing in the encoding of conspecific birdsong[12]). We then assessed the prevalence of negative coefficients in these classical-, derivative- and surprise-STRFs in two ways. First, we characterized the percentage of the filter that was positive (see Figure 3 a and Equation 5 in Methods; also compare the mostly positive surprise-STRF in Figure 1 f to the more negative derivative-STRF in Figure 1 e). Second, we compared the performance penalty for cross-validated predictions when negative coefficients are forbidden (see Figure 3 b and Equation 6 in Methods). The derivative representation used to generate Figure 3 was the same as that used in our earlier paper[11] and can be obtained from Equation 7 with d = 1 ms (see Methods).

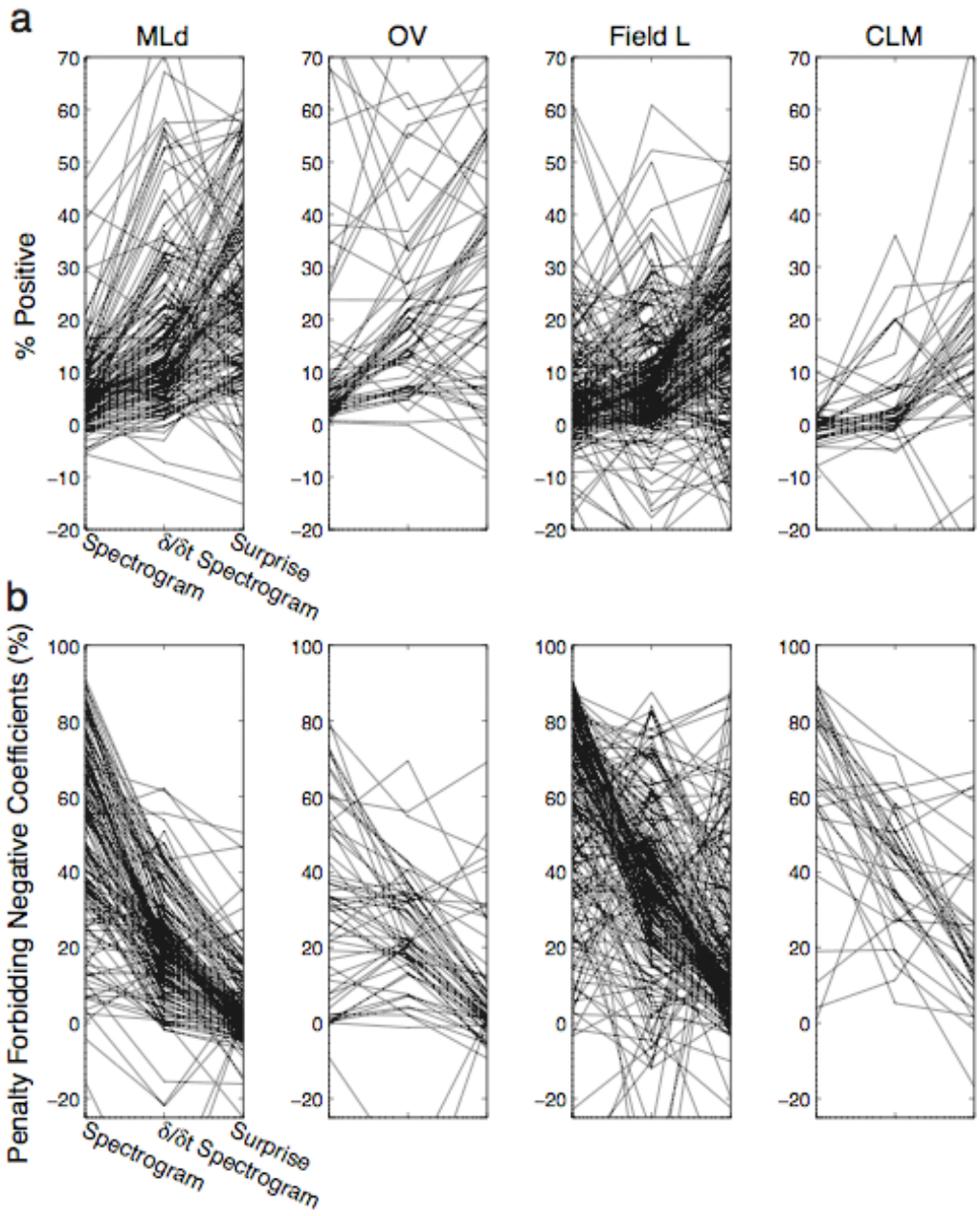

Figure 3: **Prevalence of Negative Coefficients** a) The % positive metric (see Equation 5) for classical-, derivative- and surprise-STRFs in all four sensory areas. Key: "Spectrogram" = classical-STRF, "δ/δt Spectrogram" = derivative-STRF (see Methods and Equation 7 with d = 1 ms), "Surprise" = surprise-STRF. Each neuron is represented with a line joining the % positive for the three STRF types. Positive slopes indicate surprise-STRFs with more positive coefficients than derivative-STRFs or derivative-STRFs with more positive coefficients than classical-STRFs. b) The % penalty metric (see Equation 6) for classical-, derivative- and surprise-STRFs in all four sensory areas. Negative slopes indicate smaller functional importance of negative coefficients in surprise-STRFs than in derivative-STRFs, or of derivative-

STRFs than in classical-STRFs. 12 derivative representations other than the temporal derivative were used, but surprise-STRFs always had significantly fewer and less functionally important negative coefficients in every area (see Methods).

The absence of functionally-important negative coefficients in a derivative-STRF would have indicated that neurons generally increase their firing rate in response to both predictable and surprising stimulus changes. Both "On" and "Off" derivative-detecting neurons (as well as neurons with mixtures of "On" and "Off" character) should all have exclusively non-negative derivative-STRF coefficients; "On" activity is represented by positive coefficients in the top panel of the derivative-STRF and "Off" activity is represented by positive coefficients in the bottom panel of the derivative-STRF (see Figure 1 e).

Instead, as shown in Figure 3 a, surprise-STRFs have more positive coefficients (see Equation 5 in Methods) than derivative- or classical-STRFs in all four auditory regions. Therefore, it is more correct to say these neurons fire in response to surprise than in response to "On" or "Off" stimulus changes. This difference between surprise- and derivative-STRFs is statistically significant (p values are $3 * 10^{-6}$, 0.02, $1 * 10^{-9}$, and $4 * 10^{-6}$, in MLd, Ov, Field L and CLM, respectively, one-tailed binomial test). These negative coefficients are also less functionally important in surprise-STRFs than in derivative-STRFs (see Equation 6 in Methods) as shown in Figure 3 b; these differences are highly significant (p values are $3 * 10^{-24}$, $4 * 10^{-7}$, $4 * 10^{-15}$, and $6 * 10^{-6}$, in MLd, Ov, Field L and CLM, respectively). We also tested 12 other stimulus derivative representations, such as the one shown in Figure 1 b, to investigate the possibility that some other stimulus derivative representation might result in STRFs with as many positive coefficients as surprise-STRFs (see Methods). However, we did not find any stimulus derivative representation that has fewer (or less functionally important) negative coefficients than the surprise representation.

4. Evidence of Post-Sensory Influence on the Neural Code in Vivo

An essential component in our model is activity of neurons postsynaptic to the sensory area in question. An irrelevant stimulus (i.e. one that does not evoke a stimulus-related response in any area postsynaptic to the sensory area) will engage neither synaptic nor excitability plasticity in a stimulus-dependent way, therefore the excitability and synaptic strengths of sensory neurons will not be influenced by this stimulus' statistics.

The zebra finch auditory system provides a rare opportunity to test the importance of post-sensory activity in the modification of the neural code. While the areas MLd, Ov and Field L encode all auditory stimuli used by higher areas, CLM is presynaptic to areas leading to the song production system, which are activated

primarily by conspecific song[12]. (See Figure 4, right column for a diagram of the connectivity of the four auditory areas discussed here.) Since the postsynaptic targets of CLM neurons are activated by birdsong, the surprise code CLM uses should reflect expectations of birdsong. In contrast, every auditory response in the forebrain passes through Field L, so we expect the surprise code in Field L reflects the statistics of all sounds that have an impact on any zebra finch behavior mediated by forebrain neurons.

Using our definition of relevance, only birdsong is relevant to CLM, while any sound that can influence behavior is relevant to Field L. Thus the surprise code in Field L should reflect the statistics not only of song, but also of natural environmental sounds. Since Ov and MLd are both presynaptic to Field L, these areas too should treat environmental sounds as being relevant, as Field L is postsynaptic to Ov, which is in turn postsynaptic to MLd (see Backpropagation in Discussion).

Our prediction therefore is that the expectations implicit in the surprise code (i.e. the corpus of stimuli used to estimate $P(S|D)$ of Equation 3, Methods) of MLd, Ov and Field L come from both birdsong and environmental sounds since both these classes of stimuli modulate firing in postsynaptic areas, whereas CLM's expectations are derived from birdsong alone. To test this prediction, we quantified the surprise of a group of random ripple stimuli ("modulation-limited noise" or ML noise[22]) given either expectations of conspecific song alone, or expectations of song and a standard corpus of environmental sounds[23] (see Methods for how surprise was calculated based on expectations of song & environmental sounds).

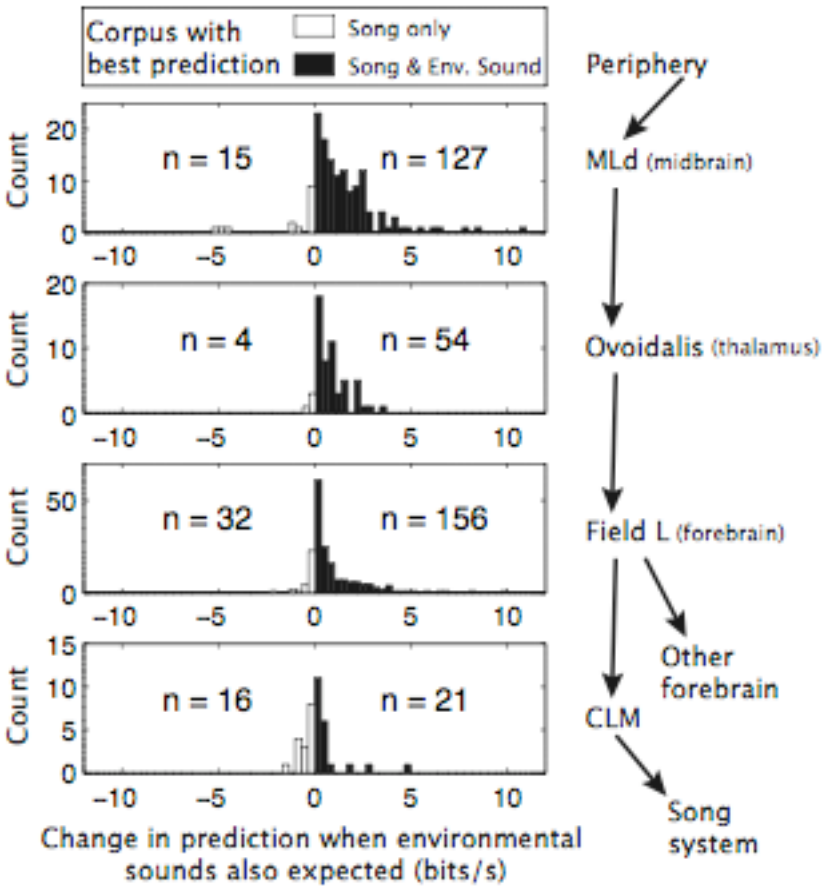

Figure 4: **Effects of Postsynaptic Modification** Histograms of the prediction improvement of surprise-STRFs in response to ML noise when expectations are based on a corpus of song & environmental sounds instead of being based on song alone. Counts of the negative and positive histogram entries are also shown. At right is a connectivity diagram of relevant zebra finch auditory areas. In the three areas feeding into general auditory processing (MLd, Ovoidalis and Field L), there is significant evidence that the surprise code had been tuned to represent environmental sounds as well as birdsong, as is seen in the bias of the histogram towards positive values. In CLM (the area presynaptic to song-specific areas), no evidence exists for modifications from environmental sounds.

As seen in the histograms of Figure 4, a surprise code based on song alone predicts worse than a code based on song and environmental sounds in MLd, Ov and Field L, but there is no significant difference in CLM. Mean differences in prediction scores between the song & environmental expectations and the song-alone expectations are MLd: 18% ($p = 2 * 10^{-18}$, two tailed Wilcoxon signed rank test), Ov: 20% ($p = 2 * 10^{-10}$), Field L: 19% ($p = 1 * 10^{-21}$) and CLM: 4.9% ($p = 0.6$). The only area whose expectations are not significantly modified by environmental sounds is CLM; not only is the distribution of prediction improvements statistically indistinguishable from 0 (see Figure 4, bottom

histogram), the CLM distribution has a significantly lower median than Field L, Ov or MLd (two sided Wilcoxon rank sum test, $p = 2 * 10^{-4}$, $p = 2 * 10^{-5}$, $p = 6 * 10^{-8}$ respectively). Therefore, in every area except CLM, expectations formed from environmental sounds significantly influence the surprise-based neural code.

For completeness, we also compared expectations of environmental sounds alone to expectations of the mixture of song and environmental sounds above. We found that, omitting song from the training corpus, predictions to ML noise went down in MLd, Ov, Field L and CLM by 7.1% ($p = 3 * 10^{-17}$), 1.9% ($p = 0.3$), 4.3% ($p = 1 * 10^{-7}$) and 10% ($p = 2 * 10^{-4}$) respectively. The prediction decrease in CLM was significantly larger than that in Field L ($p = 0.007$, two-tailed Wilcoxon rank sum test), indicating that song expectations are more important in CLM than in Field L.

In summary, there is significant influence of birdsong seen in CLM, Field L and MLd while Ov shows a smaller, statistically insignificant influence from song. The surprise codes of MLd, Ov and Field L (but not CLM) show influence from environmental sounds, consistent with our prediction that since no auditory area downstream of CLM encodes environmental sounds, CLM's surprise code should not be influenced by environmental sound statistics.

Discussion

In the Introduction, three questions were raised. First, why do areas successfully modeled with a surprise code not respond robustly to random (and thus unpredictable), artificial, behaviorally-irrelevant stimuli? Second, if surprise-coding is a good strategy, why is there a variety of neural codes in the avian auditory processing stream? Third, how might networks of neurons come to use a surprise code?

To answer the first question, we showed that synaptic and excitability plasticity together can increase the firing rates of neurons which happen to carry the most surprising information relevant to a postsynaptic task. Stimulus features irrelevant to any task in postsynaptic areas will not engage these plasticity mechanisms.

To answer the second question, our method of surprise selection can act only on the variety of stimulus-response functions already present in the sensory area. The earliest-firing neurons with content that happens to be task relevant have their excitability increased, and these surprising-as-possible neurons become postsynaptic targets for sensory areas further upstream. From metabolic studies[24, 25], we know that only a small fraction of sensory neurons contribute a significant number of spikes to the neural code. From synaptic surveys, we know that a small subset of forebrain neurons have exceptionally strong synapses, and this subset tends to connect to other neurons with strong synapses[26]. If the small fraction of neurons that are functionally active and strongly interconnected

coincides with the small fraction of neurons that are most surprising (which is likely under our model, since surprising neurons alone have their excitability consistently increased and their synapses strengthened), then we have uncovered a possible explanation for why so many sensory neurons do not spike often[24, 25, 27]. They have stimulus-response relationships which do not contribute novel or needed information to a postsynaptic area, and thus they had their excitability decreased (and did not become targets of strong excitatory synapses) to the point that now they are nearly silent.

To answer the third question, we provided four main points in support of a mechanism whereby STDP causes sensory areas to contain spikes with mostly task-relevant, surprising stimulus information. First, we described qualitatively that (by definition) surprising, relevant neurons will be the first to contain indicators of stimulus features relevant to postsynaptic areas, so under STDP surprising neurons should have stronger synapses and should become more excitable. Second, we showed in a numerical model that a sensory network with mixed surprising and un-surprising content has all but its surprising neurons become less excitable, meaning that the spikes in a sensory area with STDP will tend to be allocated to neurons with stimulus-response relationships that contain relevant surprising information. Third, we showed evidence *in vivo* of the consistent excitation associated with surprising features by showing that the surprise-STRF has more positive coefficients than any derivative-STRF we found. Fourth, we showed that *in vivo* sensory neurons upstream of song-specific regions have a surprise code based on statistics of song, while areas presynaptic to general forebrain areas (in which both song and environmental sounds are important) have a surprise code based on both song and environmental sounds.

Backpropagation

Our model implies a form of backpropagation[28] in that information about what is relevant to a downstream task is propagated upstream. Presynaptic neurons with useful content increase their firing rates, thus becoming more relevant to neurons even further upstream. This form of backpropagation requires no unknown physiological mechanism, and may help reconcile the gulf between multi-layered perceptrons, which require backpropagation for efficient training, and multi-layered biological neurons, where no suitable correspondence to backpropagation had until now been found[29].

Backpropagation might be critically important to the maintenance of relevant surprise detection. Excitability plasticity does not last longer than one week *in vivo* even though the changes in behavior produced by excitability plasticity can last much longer[19]. We propose relevant surprise detection is initially mediated in part by excitability plasticity, which in turn helps strengthen excitatory synapses from presynaptic neurons with surprising content, seen in Section 3 of Results. Synaptic plasticity can last years, spanning multiple time scales and physical

mechanisms[21, 30], therefore this type of backpropagation explains how behavior initially mediated by excitability plasticity outlives excitability plasticity.

BOS Selectivity

An unexpected benefit of our model is that we may be able to explain the sensitivity of area HVC to the bird's own song (BOS). HVC is a songbird area noted for its selectivity to BOS[31]. HVC is postsynaptic to the rest of the auditory system and immediately presynaptic to the motor areas used in song generation. Since the postsynaptic targets of HVC are activated nearly exclusively during the production of song, our model would suggest that HVC should fire predominantly to stimuli that predict firing in the song system, and thus should have selectivity to BOS.

Surprise and Overall Firing Rates

Our analysis relies on the relationship between surprise and primacy. The first reliable indicators of any stimulus feature must be surprising (in terms of having low probability given the stimulus history and given knowledge of the stimulus class' typical behavior); otherwise they are *ipso facto* not the earliest reliable indicator. STDP has been shown in simulation to cause neurons to fire as early as possible[32]. Let us examine the equivalence of surprise and primacy with more mathematical precision. Suppose that the excitability of a sensory neuron follows this rule:

$$act = k_1 + k_2 \log\left(\frac{P_{Inc}}{P_{Dec}}\right) \tag{1}$$

Here *act* is the activity, or average firing rate, of the neuron, P_{Inc} is the near-past averaged probability of an event occurring which increases the excitability of the neuron (such as firing before a postsynaptic neuron) and P_{Dec} is the probability of an event which decreases the excitability of the neuron, *i.e.*, the probability that the sensory neuron fires a spike shortly after a postsynaptic feature detector. Since *act* in Equation 1 is governed by the ratio of P_{Inc} to P_{Dec}, *act* will tend towards a stable asymptote based on the neuron's function in its circuit, and will be stable both to multiplicative increases to both P_{Inc} and P_{Dec} and to external perturbations (as has been demonstrated in the tadpole retinotectal system[33]).

If P_{Inc} is equal to a constant C_1 (related to synaptic efficiency) times the firing rate R, and P_{Dec} is equal to a similar constant C_2 times the firing rate R times $P(S|D)$ (which is the probability that the event is predictable given the stimulus history, thus the probability that the sensory neuron fires after a stimulus-savvy postsynaptic neuron, see Methods near Equation 3), then Equation 1 reduces to the following:

$$act = k_1 + k_2 \log\left(\frac{C_1 R}{C_2 R\, P(S \mid D)}\right)$$

$$act = k_1 + k_2\left(\log(C_1) - \log(C_2)\right) - k_2 \log\left(P(S \mid D)\right)$$

$$act = k_3 - k_2 \log\left(P(S \mid D)\right) \qquad (2)$$

Here, k_3 is the sum of all constants on the second line. Equation 2 shows that neural activity under the above assumptions should become linearly related to -$P(S|D)$: the minus log of the stimulus probability given the recent stimulus history. In other words, it is predicted that under STDP if the excitability of a neuron follows the form of Equation 1, its firing rate will become proportional to $-\log(P(S|D))$ which is surprise[11], and downstream neurons should also develop firing rates proportional to surprise.

The Role of Feedback

The connectivity diagram at the right of Figure 4 lacks feedback connections (axons traveling upstream from their presynaptic stimulating sensory areas). We did not include them in our model, but under STDP they should also perform the role of training more peripheral neurons to ignore sensory content that is predictable, as follows. If a stimulus feature can be predicted in a less-peripheral area but not in a more-peripheral area, then the firing of the less-peripheral neuron becomes the earliest indication of the presence of that stimulus feature, and upstream neurons encoding for the predictable feature become redundant and therefore less excitable. Thus feedback connections should help squelch the representation of predictable stimulus elements in more peripheral areas, a cortical function that has been hypothesized before[34].

Application to Other Fields

We have shown that sensory systems with two known STDP rules will tend to represent surprising sensory content that is relevant to postsensory tasks. Not only do we see our model as a general method of how plastic sensory brain areas come to process sensory data efficiently, but also the same principles may apply to Pavlovian conditioning, cognitive functions and motor planning tasks: the first neurons able to accomplish the desired computations become more active, and their synapses become more excitatory.

If the evolutionary purpose of STDP is indeed to enhance the activity of the fastest possible neural mechanisms capable of performing a task, then our finding should help the following areas of study. Searches for specific synaptic biological phenomena could be guided by their presumed function; for example if Equation 1 held then emphasizing surprise would be the result. Neuromimetic computing

could make use of the abstraction that STDP results in relevant surprise coding, and possibly artificial STDP neural networks could be trained more quickly than by modeling every spike. Finally, since the role of plastic neural tissue seems to be determined by presynaptic and postsynaptic activity, it may be possible to use a paired stimulation paradigm to re-train pieces of human cortex to recover lost cognitive abilities or to develop new ones.

Methods

Quantifying Surprise

To estimate the surprise in conspecific song, we used an identical quantification of surprise to that in an earlier paper[11].

$$Surprise = \begin{bmatrix} -\log(P(S \mid D)) + \log(P(S_{ML} \mid D)) & if \quad S > S_{ML}, \; 0 \quad otherwise \\ -\log(P(S \mid D)) + \log(P(S_{ML} \mid D)) & if \quad S < S_{ML}, \; 0 \quad otherwise \end{bmatrix} \tag{3}$$

Here D is the Domain (the recent stimulus history relevant to determining the likelihood of S), S is the stimulus intensity whose surprise is being considered, S_{ML} is the most likely stimulus given D, and $P(S|D)$ stands for the conditional probability of S given D and depends on the knowledge of statistical dependencies in the sounds from a representative corpus of 59 unfamiliar conspecific songs. *Surprise*, S_{ML}, S, and D are all functions of frequency and time. As long as $P(S|D)$ is unimodal, S can be recovered from *Surprise*. D includes latencies between 4 and 7 ms prior to S and frequencies within 625 Hz of S.

We quantified the surprise of the random modulation-limited noise (ML noise) stimulus in three ways: once with $P(S|D)$ obtained from the corpus of 59 zebra finch songs, once from both the song corpus and a selection from the Pittsburgh Natural Sounds[23], and once with the environmental sounds alone. We included the shortest 36 environmental sounds so that the total duration of the environmental sounds (241 s) was approximately twice as long as the duration of the 59 zebra finch songs (117 s) because cochlear filters have been shown to be optimized for mixtures of animal vocalizations and environmental sounds in a 1:2 ratio[35]. The names of the environmental sounds we used were: breaking_branches, breaking_branches2, breaking_twigs, Chewing, Doorclose, fire, footsteps, Glass break, Hammer, Horse Galloping, Knocking, Ocean, Pourwatr, rain, Rain2, Rattle Snake, ripples, River, rubbing_twings, rustle, Sand Paper, scratching_branch, shaking_tree, Shuffling Cards, squeeky_floor, stream, Tear Paper, Thunder, Turnpage, walking_snow+twigs, walking_snow, Water Bubbling, Water Draining, Water Drippng, Water Drops, and Wind.

Quantitative Model

For section 2 of Results, we set up a "sensory layer" of neurons whose activity is driven by a zebra finch song and a "post-sensory" area whose activity is driven by excitatory synapses from neurons in the sensory layer. Neurons in the sensory layer had inputs proportional to the outputs of random classical-STRFs, derivative-STRFs (see Equation 7, with d = 1 ms and Figure 1 e) or surprise-STRFs (see Equation 3 and Figure 1 f), all with a constant latency. Since we believe timing is critical to determining which neurons become active, we eliminated timing discrepancies by forcing all model neurons to have identical latency. We allowed the excitability and synaptic strengths of the sensory layer to become modified by STDP, based on the expected firing times of sensory neurons relative to their postsynaptic targets. Rather than simulate spikes (which was computationally impractical for the noise levels we desired) we calculated analytically the firing probability over time for every neuron, both in the sensory area and in the post-sensory area, and then used a mean-field STDP approximation in discreet-time epochs.

Epochs consisted of the following steps. The firing probability of sensory neurons was set to be proportional to the random STRF convolved with the stimulus (one representative birdsong) in the form of a spectrogram, derivative spectrogram or surprise spectrogram. These firing rates where then rectified, and the average firing rate was set to a constant. The time-dependent firing rates of the post-sensory neurons were then determined by taking the sum of the firing rates of sensory neurons weighted by synaptic strengths, and adding a 1 ms latency to account for synaptic delay and integration. For each synapse, the net "mean field" effect of STDP was estimated by the dot product of the cross-correlation of presynaptic and postsynaptic firing rates with the following time-dependent STDP rule (plotted in the upper-right of Figure 2):

$$Change(t) = \frac{t(0.1ms)}{\left(t^2 + (2ms)^2\right)} \tag{4}$$

The actual change applied to each synapse was multiplying the current synaptic strength by $1 + \varepsilon(Change \bullet CC)$ where ε is a learning rate, *Change* is as in Equation 4 and *CC* is the cross-correlation of pre- and postsynaptic firing rates. The excitability of each sensory neuron was also multiplied by the product of all synaptic change factors, then adjusted so that no neuron had a firing rate above 50 Hz. All excitabilities were scaled to keep the mean firing rate in the sensory area constant of 0.5 Hz, our estimate of a typical mean firing rate[24]. The mean firing rate of each class of sensory neuron (Surprise, Derivative or Spectrogram) was stored at the end of every epoch.

We did not use our postulated method of weighting exciting and depressing events (see Equation 1) for two reasons: it has not yet been experimentally

demonstrated, and we did not wish to imply that our finding that our main finding (relevant surprise detection increases excitability) requires this specific, untested assertion.

We explored a large parameter range to show that our main finding is robust. We investigated networks with 320, 400, 800, 1600, 3200 and 4000 total neurons. For all our simulations there were three times as many neurons in the sensory layer as in the post-sensory layer. Initial synapse strength was set to be random, and the chance that any particular synapse existed between a given sensory neuron and a given post-sensory neuron in any given simulation was 1 in 6 (a relatively dense interconnectivity), 1 in 12 or 1 in 24 (a relatively sparse connectivity). We ran each simulation 10 times with different initial random synapse connectivity and strength, and reported the entire 10-simulation range of firing rates as a function of number of elapsed epochs in Figure 2.

Surprise and Excitation

Positive Metrics

STRFPAK version 5.2 (available for download at http://strfpak.berkeley.edu) was used to calculate classical-, derivative- and surprise-STRFs. We used two methods of assessing the degree to which surprise-STRFs are more positive than derivative-STRFs: a % positive measure (see Figure 3 a) and a functional importance measure (see Figure 3 b). The % positive measure was computed as follows:

$$\%\ Positive = \frac{\sum_{ft} S_{ft}}{\sum_{ft} |S_{ft}|} \tag{5}$$

where S_{ft} is STRF coefficient at latency t and frequency band f. The % Positive is thus restricted to being between -100% (for a STRF composed entirely of nonpositive coefficients) and 100% (for a STRF composed entirely of nonnegative coefficients).

Our second method of assessing the importance of negative coefficients is to determine the decrease in prediction scores to validation data when negative coefficients are disallowed. STRFPAK version 5.2 includes the option of setting negative STRF coefficients to 0 for predictions on validation data. To assess the functional importance of negative coefficients, we used a % penalty metric as follows:

$$\%\ Penalty = \frac{Fit_{Norm} - Fit_{NoNeg}}{Fit_{Norm}} \tag{6}$$

Fit_{Norm} is the regular predicted information fit score[36] and Fit_{NoNeg} is the prediction score once negative coefficients have been zeroed out.

Other Derivative Representations

To control for the possibility that the form of the spectrogram derivative used was not the one which generates minimal negative coefficients, we searched for (but could not find) a different stimulus derivative representation in which derivative-STRFs have fewer negative coefficients (by either metric, see Equations 5 and 6) than surprise-STRFs. The remainder of this section documents the specifics of the derivative representations which all had significantly more negative coefficients than the surprise-STRF using either metric.

We investigated the following 12 other derivative representations: 9 different temporal derivative representations where the time step of the derivative ranged from 2ms to 10 ms (the default derivative representation has a 1 ms time step), a spectral derivative, a joint spectro-temporal derivative, and a difference-from-mean-domain derivative.

The 10 total time derivatives were calculated as follows:

$$Der = \begin{bmatrix} |S(t) - S(t-d)| & if & S(t) - S(t-d) > 0, & 0 & otherwise \\ |S(t) - S(t-d)| & if & S(t) - S(t-d) < 0, & 0 & otherwise \end{bmatrix} \quad (7)$$

Der is the split absolute derivative, with louder-than-context entries in the top half and quieter-than-context entries in the bottom half (see Figure 1 b for an example); *t* is time. The parameter *d* controls the duration of the derivative baseline, and varied from 1 ms to 10 ms.

Similarly, the spectral derivative was the difference in intensity between a given spectrogram frequency band and the band of immediately lower frequency, split so that increases and decreases in intensity as a function of frequency appear as positive entries in separate sections of the double-tall stimulus representation.

The joint spectro-temporal derivative was formed by taking the split spectral derivative of the split temporal derivative with a 1 ms baseline (see Equation 7 with d = 1 ms), and results in a quadruple-tall stimulus representation where 75% of the entries are 0: e.g. the top section has nonzero entries only when the present spectrogram band is more intense both than the spectrogram band of immediately lower frequency and the spectrogram band of the same frequency 1 ms in the past.

The difference-from-mean-domain derivative (shown in Figure 1 b) was calculated as follows:

$$\begin{bmatrix} |S-D| \;\; if \;\; S-D>0, \;\; 0 \;\; otherwise \\ |S-D| \;\; if \;\; S-D<0, \;\; 0 \;\; otherwise \end{bmatrix} \quad (8)$$

Here, S is the stimulus intensity at a particular time and frequency and D is the mean value of the stimulus in the domain used to predict how surprising S is for the surprise representation (a rectangle with latency range 4 – 7 ms and frequency range within 625 Hz of S, see above near Equation 3).

In all areas the surprise-STRF was more positive than all of the derivative-STRFs, both in terms of positive percentage (Equation 5, p values are less than or equal to $3 * 10^{-6}$, 0.04, $1 * 10^{-4}$, and 0.001 in MLd, Ov, Field L and CLM, respectively, one-tailed binomial test) and in terms of prediction penalty when negative coefficients are removed (Equation 6, p values are less than or equal to $1 * 10^{-21}$, $2 * 10^{-5}$, $2 * 10^{-7}$, and 0.01 in MLd, Ov, Field L and CLM, respectively).

We were justified in using a one-tailed binomial test because in all 104 independent tests (2 metrics * 4 areas * 13 derivative representations), negative coefficients are more prevalent in derivative- than in surprise-STRFs, almost always with a high degree of significance, thus there is an *a priori* reason to expect surprise-STRFs to be more positive.

As an aside, we double-checked our earlier finding that in CLM surprise-STRFs outperform derivative-STRFs[11] by comparing the prediction accuracy of the best derivative-STRF for CLM each neuron (determined *post hoc*, outside of any regularization framework) to the prediction accuracy of the surprise-STRF. This comparison is statistically unfair against the surprise-STRF for two reasons. First, selecting the best from 13 different derivative representations gives significantly more flexibility to the computations the derivative-STRFs could perform (thus matching the idiosyncrasies of each neuron). Second, one noisy measurement (the prediction score of the surprise-STRF) is compared to the maximum, not the mean, of 13 noisy measurements (the 13 different derivative-STRF prediction scores), so noise works in favor of high derivative-STRF scores. Nonetheless, in CLM the surprise-STRF still out-performs the best derivative-STRF consistently, by an average of 10%, p = 0.005 (Wilcoxon signed rank test).

Neurophysiological Recordings

Neural data were obtained from 57 adult male zebra finches. All subjects were reared in a colony in natural family groups, and were not exposed to any of the songs used as a stimulus prior to the neurophysiological recordings session. Single-unit responses were obtained with extracellular tungsten electrodes in urethane-anesthetized birds. The location of the recordings was verified with standard histological techniques: for CLM, n = 37; Field L, n = 188; Ovoidalis, n=58 and MLd, n = 142. 46 subjects underwent simultaneous recording in either

in both Field L and MLd (n = 29) or in both CLM and Field L (n = 17); the remaining 11 underwent recording in Ovoidalis exclusively. (*** MORE OVOIDALIS METHODS HERE IF THIS PUBLISHED BEFORE NOOPUR"S PAPER ***)

Sounds were played from a loudspeaker placed 15 cm in front of the animal, and sound levels had peak intensity of 70 dB SPL. All neurons in CLM were in the lateral subdivision. Neurons in Field L were sampled from all sub-regions (L1, L2a, L2b, and L3). Data from 46 (*** ALL 57 IF NOOPUR'S PAPER PUBLISHED FIRST ***) of these birds were also used in previously published work[11], and additional information on stimulus design, neurophysiological recordings and histological techniques can be found in previous studies[22, 37, 38] (*** Insert ref to Noopur's paper here too once it's published ***). All experimental procedures were approved by the Animal Care and Use Committee of UC Berkeley.

Acknowledgements

We appreciate Alyosha Molnar's suggestion to add the numerical simulation in section 2 of the Results. None of this work would have been possible without the thoughtful, expert help of our two laboratory technicians and animal care specialists B. R. and Y. M.

This work was supported by NIH grants DC007293, MH66990 and MH59189 to FET.

References

1. Sutton, S., Tueting, P., Zubin, J. & John, E.R. Information Delivery and the Sensory Evoked Potential. *Science* **155**, 1436-1439 (1967).
2. Kuffler, S.W. Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology* **16**, 37-68 (1953).
3. Yamaguchi, S. & Knight, R.T. P300 generation by novel somatosensory stimuli. *Electroencephalography and Clinical Neurophysiology* **78**, 50-55 (1991).
4. Knudsen, E. & Konishi, M. Center-surround organization of auditory receptive fields in the owl. *Science* **202**, 778-780 (1978).
5. Wehr, M. & Zador, A.M. Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex. *Nature* **426**, 442-446 (2003).
6. van Hateren, J.H. Theoretical predictions of spatiotemporal receptive fields of fly LMCs, and experimental validation. *Journal of Comparative Physiology A* **171**, 157-170 (1992).

7. Barlow, H.B. Sensory mechanisms, the reduction of redundancy, and intelligence. *NPL Symposium on the Mechanization of Thought Process* **10**, 535-539 (1959).
8. Dong, D.W. & Atick, J.J. Statistics of natural time-varying images. *Network: Computation in Neural Systems* **6**, 345-358 (1995).
9. Field, D.J. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A* **4**, 2379-2394 (1987).
10. Singh, N.C. & Theunissen, F. Modulation spectra of natural sounds and ethological theories of auditory processing. *Journal of the Acoustical Society of America* **114**, 3394-3411 (2003).
11. Gill, P., Woolley, S.M.N., Fremouw, T. & Theunissen, F.E. What's That Sound? Auditory Area CLM Encodes Stimulus Surprise, Not Intensity or Intensity Changes. *Journal of Neurophysiology* **99**, 2809-2820 (2008).
12. Gentner, T.Q. & Margoliash, D. Neuronal populations and single cells representing learned auditory objects. *Nature* **424**, 669-674 (2003).
13. Shannon, C.E. & Weaver, W. *The Mathematical Theory of Communication* (University of Illinois Press, Illinois, Indiana, 1998).
14. Dan, Y. & Poo, M.-M. Spike Timing-Dependent Plasticity: From Synapse to Perception. *Physiological Reviews* **86**, 1033-1048 (2006).
15. Cassenaer, S. & Laurent, G. Hebbian STDP in mushroom bodies facilitates the synchronous flow of olfactory information in locusts. *Nature* **448**, 709-713 (2007).
16. Boettiger, C.A. & Doupe, A.J. Developmentally Restricted Synaptic Plasticity in a Songbird Nucleus Required for Song Learning. *Neuron* **31**, 809-818 (2001).
17. Li, C.-y., Lu, J.-t., Wu, C.-p., Duan, S.-m. & Poo, M.-M. Bidirectional Modification of Presynaptic Neuronal Excitability Accompanying Spike Timing-Dependent Synaptic Plasticity. *Neuron* **41**, 257-268 (2004).
18. Daoudal, G. & Debanne, D. Long-Term Plasticity of Intrinsic Excitability: Learning Rules and Mechanisms. *Learning & Memory* **10**, 456-465 (2003).
19. Moyer, J.R., Jr., Thompson, L.T. & Disterhoft, J.F. Trace Eyeblink Conditioning Increases CA1 Excitability in a Transient and Learning-Specific Manner. *Journal of Neuroscience* **16**, 5536-5546 (1996).
20. Zhang, W. & Linden, D.J. The other side of the engram: experience-driven changes in neuronal intrinsic excitability. *Nat Rev Neurosci* **4**, 885-900 (2003).
21. Clem, R.L., Celikel, T. & Barth, A.L. Ongoing in vivo experience triggers synaptic metaplasticity in the neocortex. *Science* **319**, 101-104 (2008).
22. Hsu, A., Woolley, S.M.N., Fremouw, T.E. & Theunissen, F.E. Modulation Power and Phase Spectrum of Natural Sounds Enhance Neural Encoding Performed by Single Auditory Neurons. *Journal of Neuroscience* **24**, 9201-9211 (2004).
23. Smith, E.C. & Lewicki, M.S. Efficient auditory coding. *Nature* **439**, 978-982 (2006).
24. Lennie, P. The Cost of Cortical Computation. *Current Biology* **13**, 493-497 (2003).

25. Shoham, S., O'Connor, D.H. & Segev, R. How silent is the brain: is there a "dark matter" problem in neuroscience? *J Comp Physiol A Neuroethol Sens Neural Behav Physiol* **192**, 777-784 (2006).
26. Song, S., Sjostrom, P.J., Reigl, M., Nelson, S. & Chklovskii, D.B. Highly nonrandom features of synaptic connectivity in local cortical circuits. *PLoS Biol* **3**, e68 (2005).
27. Olshausen, B.A. & Field, D.J. Sparse coding of sensory inputs. *Curr Opin Neurobiol* **14**, 481-487 (2004).
28. Werbos, P.J. Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences. in *Applied Matematics* (Harvard University, Cambridge, Massachusetts, 1974).
29. Stork, D.G. Is backpropagation biologically plausible? *Proceedings of the International Joint Conference on Neural Networks, Washington DC* **2**, 241-246 (1989).
30. Fusi, S., Drew, P.J. & Abbott, L.F. Cascade Models of Synaptically Stored Memories. *Neuron* **45**, 599-611 (2005).
31. Margoliash, D. Acoustic Parameters Underlying the Responses of Song-Specific Neurons in the White-Crowned Sparrow. *Journal of Neuroscience* **3**, 1039-1057 (1983).
32. Guyonneau, R., VanRullen, R. & Thorpe, S.J. Neurons Tune to the Earliest Spikes Through STDP. *Neural Computation* **17**, 859-879 (2005).
33. Zhou, Q., Tao, H.W. & Poo, M.-M. Reversal and Stabilization of Synaptic Modifications in a Developing Visual System. *Science* **300**, 1953-1957 (2003).
34. Rao, R.P.N. & Ballard, D.H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience* **2**, 79-87 (1999).
35. Lewicki, M.S. Efficient coding of natural sounds. *Nature Neuroscience* **5**, 356-363 (2002).
36. Hsu, A., Borst, A. & Theunissen, F. Quantifying variability in neural responses and its application for the validation of model predictions. *Network: Computation in Neural Systems* **15**, 91-109 (2004).
37. Woolley, S.M.N., Fremouw, T.E., Hsu, A. & Theunissen, F. Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. *Nature Neuroscience* **8**, 1371-1379 (2005).
38. Woolley, S.M.N., Gill, P.R. & Theunissen, F. Stimulus-Dependent Auditory Tuning Results in Synchronous Population Coding of Vocalizations in the Songbird Midbrain. *Journal of Neuroscience* **26**, 2499-2512 (2006).