Emergence of Auditory Receptive Fields based on Surprise

Yashaswini[1]*, Sneha Dash[2]*, Sharba Bandyopadhyay[2]
[1] University of California, Berkeley, [2] Information Processing Laboratory, Indian Institute of Technology, Kharagpur
yashaswini@berkeley.edu, snehadash7441@gmail.com,
sharba@ece.iitkgp.ac.in

Understanding how sensory systems efficiently encode natural stimuli is a fundamental challenge in neuroscience. While the efficient coding hypothesis explains many aspects of sensory processing, its role in processing surprising auditory inputs remains unclear. We present two computational frameworks modeling the development of auditory neural receptive fields via unsupervised learning to address this challenge. In the first framework, a single-layer network's synaptic weights adapt to auditory inputs to maximize activations for surprising events while minimizing overall activity. The weights are adjusted using three factors (α, β, γ) and the gradient of the $l1$ norm of activations. An autoregressive generative model (CochleaNet), trained on LibriSpeech, provides the joint probability distribution to calculate surprise, defined as the negative log probability of time-frequency bin energy conditioned on previous time steps and other frequency channels. We find learning to be fast, with robust convergence of weights using random speech samples. This approach yields spectrotemporal receptive fields (STRFs) whose tuning properties closely match neurophysiological observations. Second, we propose a principled *Kalman-MI* formulation in which the generative prior, latent auditory state, and synaptic weights are jointly updated online. Mutual-information gradients, serving as a normative proxy for expected surprise reduction, drive adaptation in a linear-Gaussian state-space model, producing deviant-selective receptive fields in an oddball paradigm. Together, these complementary approaches aim to refine the interplay between sparse coding and surprise-driven learning, offering new insights into efficient sensory encoding.

1. Introduction

The remarkable efficiency of biological sensory systems in processing natural stimuli continues to inspire advances in computational neuroscience and artificial intelligence [1]. The complexity of temporal patterns in auditory signals, combined with the need to process information across multiple timescales simultaneously, makes the auditory system a particularly fascinating subject for studying neural computation and adaptation mechanisms. Beginning in the inner ear, acoustic signals undergo sophisticated processing through multiple stages, from basic feature extraction in the brainstem to complex pattern recognition in the auditory cortex [2]. A key distinguishing feature is the development of receptive fields along multiple timescales, allowing neurons to process both brief and extended acoustic events simultaneously [3]. Central to auditory processing is the concept of Bayesian surprise, which serves as a fundamental mechanism for learning and adaptation in the brain [4]. Surprise signals guide attention toward unexpected stimuli and drive synaptic adaptation, enabling the system to update its internal models continuously [5]. This surprise-driven learning is particularly relevant in auditory processing, where the temporal structure of sounds carries critical information, and the ability to detect and adapt to novel patterns is essential for survival. The principle of efficient coding, first proposed by [6] suggests that the auditory system optimizes its neural resources to maximize information transmission. This optimization manifests in the spectrotemporal receptive fields (STRFs) of auditory neurons, which are adapted to the statistical properties of natural sounds [7]. The sparse coding observed in auditory neural responses further supports this efficiency principle, with only a small fraction of neurons active at any given time, reducing noise and improving information transmission.

*Equal Contribution.

Third Conference on Parsimony and Learning (CPAL 2026).

We adopt two complementary computational frameworks to answer whether surprise-driven learning is key to the emergence of biologically meaningful auditory receptive fields . We first ask whether surprise estimates computed from real-world speech statistics drive the emergence of biologically realistic auditory receptive fields, motivating the CochleaNet-based framework. Having established this empirical feasibility, we then address its conceptual limitations - namely, the reliance on an external generative model that does not interact with neural activity, the treatment of surprise as an imposed modulatory signal rather than an emergent quantity, and the use of heuristic thresholds, by introducing a principled Kalman–MI formulation in which surprise minimization arises intrinsically from predictive inference.

This research makes three contributions. First, we propose a surprise-driven synaptic adaptation framework in which surprise estimates derived from a generative model of natural sounds guide learning under a sparsity constraint, yielding spectrotemporal receptive fields consistent with neurophysiological observations across multiple timescales. Second, through explicit ablations, we show that sparsity alone is insufficient to produce meaningful tuning, and that surprise provides a critical inductive signal. Third, we introduce a complementary normative framework based on Kalman filtering and mutual information maximization, in which surprise minimization emerges intrinsically from predictive inference, producing deviant-selective receptive fields in an oddball paradigm.

2. Methods

Our work employs two methodologies for understanding how auditory receptive fields and frequency selectivity emerge through surprise-driven adaptation.

1. **Approach 1: A data-driven deep generative model (CochleaNet)**, trained on natural speech to extract empirical surprise values and reproduce realistic tuning curves
2. **Approach 2: Kalman-Mutual Information Framework**, where gradients from mutual-information maximization (proxy for surprise minimization) drive synaptic adaptation and produce receptive fields through principled information-theoretic learning.

Both approaches capture different but converging aspects of auditory computation: the first provides empirical grounding using real audio signals, while the second provides a theoretical framework that accounts for the emergence of receptive fields through Kalman-style inference and information-maximizing plasticity. This framework was essential because the empirical model alone cannot reveal why adaptation occurs as a result of how uncertainty, prediction error, and sparsity interact to shape synaptic structure. In the following sections, we detail each methodology.

2.1. Approach 1: Deep Generative Model and Surprise-Based Adaptation

This section introduces a computational framework in which surprise estimates obtained from a deep generative model of natural speech modulate synaptic plasticity under a sparsity constraint. Building on work in efficient coding [8], [6] and surprise-based adaptation [9], we show that this combination of sparsity and surprise, rather than sparsity alone, leads to biologically plausible spectrotemporal receptive fields across multiple temporal scales.
We used the LibriSpeech dataset, comprising approximately 1,000 hours of English speech, to generate cochleagrams at various timescales. We applied 50 ERB-spaced gammatone filters, ranging from 75 to 4,800 Hz, across timescales of 4 ms, 8ms, 16 ms, 32 ms, and 64 ms. Leveraging the CochleaNet architecture, which we developed by drawing inspiration from MelNet [10] but specifically adapted it for cochleagram processing, we implemented an autoregressive model to predict the probability distributions of observed time-frequency bin energy. The network was trained across different timescales, experimenting with configurations that varied in hidden sizes from 16 to 64 neurons while maintaining a five-layer depth. We defined surprise as the negative log of the probability of observed time-frequency bin energy, conditioned on previous time steps and other frequency channels. The probability distribution of surprise values was calculated, with the minimum surprise threshold set at the first quartile and the saturation surprise threshold at the third quartile of the distribution. We designed a single-layer neural network whose synaptic weights were first modulated based on surprise, employing three adaptation factors: alpha, beta, gamma to modify synaptic

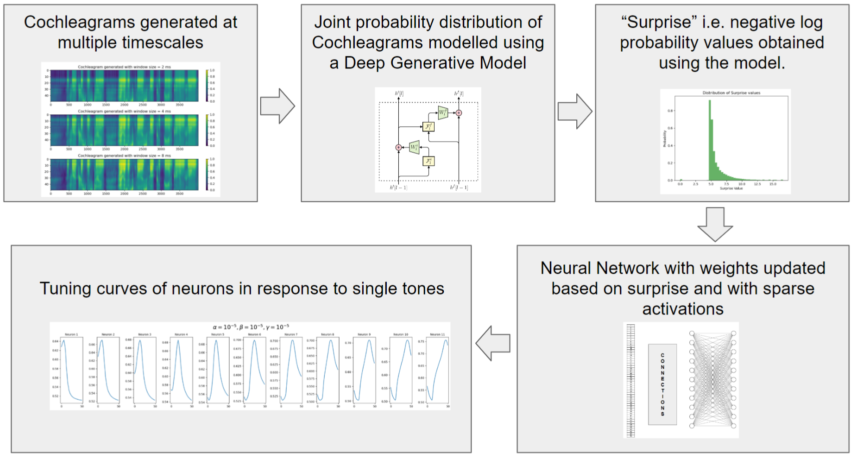

Cochleagrams generated at multiple timescales
Joint probability distribution of Cochleagrams modelled using a Deep Generative Model
"Surprise" i.e. negative log probability values obtained using the model.
Neural Network with weights updated based on surprise and with sparse activations
Tuning curves of neurons in response to single tones
CONNECTIONS

Figure 1: **Framework overview** Cochleagrams are generated at multiple timescales, with their joint probability distribution modeled using a deep generative model. Surprise values are calculated and used to update neural network weights. Sparsity is enforced by decreasing weights by the sparsity factor times the gradient of the l1 norm of activations with respect to the weights, resulting in tuning curves for neurons in response to single tones with frequencies corresponding to the best filters of gammatone filters

.

weights to mimic long term potentiation (alpha), synaptic depression (beta) and gamma to ensure neural stability and preventing runaway excitation. The synaptic weights were subsequently adjusted according to the gradient of the $l1$ norm of activations with respect to weights to enforce sparsity.

2.1.1. CochleaNet Autoregressive Model

We develop *CochleaNet*, a MelNet-inspired autoregressive generative model [10] adapted to cochleagram inputs. Let $x \in \mathbb{R}^{n \times T}$ denote a cochleagram with frequency index i and time index j. The model factorizes the joint distribution as

$$p(x) = \prod_j \prod_i p(x_{ij} \mid x_{<ij}; \theta_{ij}), \tag{1}$$

$$p(x_{ij} \mid x_{<ij}; \theta_{ij}) = \mathcal{N}(x_{ij}; \mu_{ij}, \sigma_{ij}), \tag{2}$$

where $x_{<ij}$ denotes all cochleagram bins preceding (i, j) under a row-major (time-first, low-to-high frequency) ordering, and $\theta_{ij} = (\mu_{ij}, \sigma_{ij})$ are the predicted mean and variance parameters.

The network outputs (μ_{ij}, σ_{ij}) conditioned on $x_{<ij}$ and is trained by maximum-likelihood. Architectural and training details are provided in Appendix A.

2.1.2. Calculating Surprise Values from CochleaNet

To compute surprise values, we iterate over all elements of the cochleagram tensor. For each frequency-time bin x_{ij}, the corresponding context $x_{<ij}$ is passed through the CochleaNet autoregressive model, which outputs a predicted mean $\hat{\mu}_{ij}$ and standard deviation $\hat{\sigma}_{ij}$ for the Gaussian conditional distribution.

The predicted parameters are first constrained to ensure numerical stability. Then, for each bin we compute the log-likelihood. The *surprise* of that element is defined as the negative log-probability:

$$S_{ij} = -\log p(x_{ij} \mid x_{<ij}, \mu_{ij}, \sigma_{ij}).$$

The empirical distribution of surprise was estimated by maximum likelihood estimation. This was used to estimate the parameters of a few standard distribution. We used Residual Sum of Squares to test the goodness of fit of the distributions. The details regarding the best fit distribution are provided in Appendix B. We defined the low surprise threshold (m) at the first quartile of the distribution and the saturation threshold (s) at the third quartile of the distribution. These thresholds partition incoming stimuli into predictable, informative, and highly unexpected regions, mirroring biological observations of adaptation and gain control [5].

2.1.3. Surprise-based adaptation

Our model (Fig. 1) uses a single-layer neural network operating on auditory stimuli, employing three adaptation factors (alpha, beta, gamma) to modify synaptic weights. This tripartite mechanism aligns with neurobiological studies on synaptic homeostasis and plasticity [1] , indicating a principle for balancing excitability while enabling learning.

At a particular timescale, let Surprise $= S$, Min Threshold $= m$, and Saturation Threshold $= s$.

$$w_{ij}(t+1) = \begin{cases} w_{ij}(t) + \alpha(S), & \text{if } m < S < s \\ w_{ij}(t) - \beta(S), & \text{if } S < m \\ w_{ij}(t) + \gamma(s), & \text{otherwise} \end{cases}$$

2.1.4. Sparsity-Based Adaptation

After surprise-based synaptic adaptation, we apply adjustments according to the gradient of the $l1$ norm of activations with respect to weights to enforce sparsity. Here, we state the formulation needed to propose the synaptic weight update rule to promote sparsity. Detailed derivation can be found in Appendix F.

Let $g \in \mathbb{R}^n$ denote the CochleaNet-derived *surprise vector*, where g_j corresponds to surprise at frequency channel j. Let $W \in \mathbb{R}^{K\times n}$ be the synaptic weight matrix, with $K = \frac{n-b}{m} + 1$ neurons, filter size b, and stride m. Each neuron connects only to a local frequency band determined by (b, m). Activations are computed using a nonlinearity $f(\cdot)$ (sigmoid):

$$a = f(Wg), \qquad a \in \mathbb{R}^K.$$

The neural activations are computed as follows:

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_{b+1} \end{bmatrix} = f\left(\begin{pmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,b} & 0 & \cdots & 0 \\ 0 & w_{2,m+1} & \cdots & w_{2,m+b} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & w_{b+1,m+b+1} & \cdots & w_{b+1,n} \end{pmatrix} \begin{bmatrix} g_1 \\ g_2 \\ g_3 \\ \vdots \\ g_n \end{bmatrix}\right)$$

To enforce efficient coding, we penalize the L_1 norm of activations, $\|a\|_1 = \sum_i a_i$. The gradient of this objective with respect to W yields a structured update that respects the receptive field layout.

The masked input matrix is

$$G_{\text{active}} = \begin{bmatrix} g_1 & g_2 & \dots & g_b & 0 & \dots & 0 \\ 0 & \dots & 0 & g_{m+1} & \dots & g_{m+b} & \dots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & g_{mb+1} & \dots & g_n \end{bmatrix} \in \mathbb{R}^{K\times n}.$$

Let

$$F = \text{broadcast}(f'(Wg)) \in \mathbb{R}^{K\times n},$$

where each row contains the derivative of the corresponding neuron's activation. The resulting sparsity-driven weight update is

$$W_{\text{new}} = W - \eta\left(F \circ G_{\text{active}}\right), \tag{3}$$

where $\circ$ denotes element-wise multiplication and η is the learning rate. This update suppresses widespread activation while preserving localized, frequency-selective receptive fields.

2.1.5. Tuning Curve Generation

Receptive fields were evaluated using activation tuning curves. For each gammatone center frequency, a pure tone was presented, the resulting cochleagram was fed through the model, and neural responses were averaged over time, $A_i(f) = \frac{1}{T}\sum_t a_i(t; f)$, to obtain the tuning curve.

2.2. Approach 2: Learning the probability distribution through Mutual Information and Recursive Generative Modelling

Although the earlier framework successfully linked cochleagram-based surprise with sparsity-driven plasticity, and demonstrated that auditory neurons adapt their receptive fields to the statistical structure of natural sounds, yielding spectrotemporal receptive fields (STRFs) closely aligned with neurophysiological observations; it nonetheless exhibited important conceptual limitations. The approach relied on an external deep generative model that did not interact with neural activity, and surprise acted as an imposed modulatory signal rather than an emergent quantity arising from predictive inference. Here, in contrast to Approach 1, receptive field adaptation follows directly from minimizing predictive uncertainty, by formulating learning as mutual-information maximization within a Kalman filtering framework. Mutual information gradients arise intrinsically from an adaptive generative model that is continuously updated in tandem with synaptic plasticity.This approach therefore provides a principled account of deviant selectivity without requiring heuristic thresholds or externally imposed surprise signals. The use of mutual-information maximization as a learning objective is well established in sensory coding models; for example, [9] and other efficient-coding approaches have employed MI as a principled objective for driving neural adaptation.

We construct a dynamical auditory processing model that integrates an online generative predictor of future sensory input, a Kalman-style belief update over a latent state, and plasticity rules derived from a joint objective combining mutual-information maximization with sparsity regularisation. At each timestep the model receives an n-dimensional spectral vector from an oddball stimulus (90% standard, 10% deviant). The frequency axis is partitioned into overlapping patches, each driving one neuron through a learnable weight vector W_k. A global generative matrix H_{global} maps the latent history h_j to predictions of the upcoming spectral frame, and patch-specific predictions are extracted through selector matrices S_k. Predicted activation statistics are combined with realized activations to compute MI-based learning signals, which update both synaptic weights and the generative model. A Kalman correction updates the prior, ensuring tight coupling between inference of latent structure and adaptation of model parameters.

2.2.1. Mathematical Formulation of the Kalman-MI Framework

This section lays down the requisite mathematical formulation for the Kalman-MI Framework, without going into its detailed derivation (See Appendix C). At the end of the section, we present a formal algorithm for this framework. The system attempts to *maximize* how informative the neural activation is about the upcoming sensory input. A closed-form approximation of the mutual information between activation and sensory input is evaluated from the predicted activation statistics. Gradients of this MI measure modify the generative weights matrix H_{global} and the synaptic weights W_k.

Indices and basic notation

- Frequency-channel index: $i \in \{1, \dots, B\}$ (within a patch).
- Time index: $j \in \{1, \dots, T\}$.
- Patch index: $k \in \{1, \dots, K\}$.

Each patch k observes a contiguous set of frequency channels:

$$I_k = \{i_k,\, i_k + 1,\, \dots,\, i_k + B - 1\}.$$

Observed stimulus for a patch For each channel $i \in I_k$ at time j, let

$$x_{i,j}^{(k)}$$

denote the observed energy in that frequency bin for patch k.

Stack these into a vector:

$$x_{k,j} = \begin{bmatrix} x_{i_k,j}^{(k)} & x_{i_k+1,j}^{(k)} & \cdots & x_{i_k+B-1,j}^{(k)} \end{bmatrix}_{B\times 1}^{\top}$$

Global history upto a context window L The model maintains an nL-dimensional latent state $h_{j|j-1}$ containing the most recent L spectral frames .

$$h_j = [x_{1,j-1} \quad x_{2,j-1} \quad \cdots \quad x_{n,j-1} \quad x_{1,j-2} \quad \cdots \quad x_{n,j-L+1}]^{\top}_{nL\times 1}$$

It evolves under a fixed shift-register transition matrix F:

$$h_{j|j-1} = F h_{j-1|j-1}, \qquad P_{j|j-1} = F P_{j-1|j-1} F^{\top} + Q.$$

State-space generative model Conditioned on the hidden state $h_{k,j}$ and generative parameters $\theta = (F_k, H_{k,j}, Q_k, R_k)$, the observation model for patch k is:

$$x_{k,j} \mid \theta,\, h_{k,j} \sim \mathcal{N}(u_{k,j},\, \mathrm{Cov}_{k,j}),$$

Generative prediction. A global mapping $H_{\text{global}} \in \mathbb{R}^{n\times nL}$ predicts the next frame:

$$u_j = H_{\text{global}} h_{j|j-1}, \qquad u_{k,j} = S_k u_j, \qquad \mathrm{Cov}_{k,j} = S_k(H_{\text{global}} P_{j|j-1} H_{\text{global}}^{\top} + R_{\text{global}})S_k^{\top} + R_k.$$

Neural readout. Each neuron k has weights $W_k \in \mathbb{R}^B$:

$$z_{k,j} = W_k^{\top} x_{k,j}, \qquad a_{k,j} = f(z_{k,j}) + \varepsilon_{k,j}, \qquad m_{k,j} = W_k^{\top} u_{k,j}, \quad s_{k,j} = W_k^{\top} \mathrm{Cov}_{k,j} W_k.$$

Overall Objective. We use the (per-patch k, per-time j) loss:

$$\mathcal{L}_{k,j} \;=\; -\, I\big(a_{k,j},\, x_{k,j}\big) \;+\; \lambda\, \Phi\big(a_{k,j}\big),$$

where:

- $I(a, x)$ is the mutual information between the scalar activation of a patch $a_{k,j}$ and the vector of observed energies in the patch $x_{k,j}$.
- $a_{k,j} = f(z_{k,j}) + \varepsilon_{k,j}$ is the realized (noisy) activation . The sparsity penalty is computed from the instantaneous (noisy) activation.
- Φ is the sparsity penalty: $\Phi(a) = \log(1 + a^2)$.

Mutual information objective. Using a second-order approximation of $I(a_{k,j}, x_{k,j})$, the predictive derivatives are

$$d_m = \partial I/\partial m_{k,j}, \qquad d_s = \partial I/\partial s_{k,j} \quad \text{(analytic expressions in Appendix C).}$$

Gradient lifts. Observation-space gradients:

$$g_{u,k,j} = d_m W_k, \qquad G_{Cov,k,j} = d_s(W_k W_k^{\top}).$$

Lifted to the global model:

$$\nabla_{H_{k,j}} I = g_{u,k,j} h_{j|j-1}^{\top} + 2 d_s (W_k W_k^{\top}) H_{k,j} P_{j|j-1}, \qquad H_{k,j} = S_k H_{\text{global}},$$

$$\nabla_{H_{\text{global}}} I = \sum_k S_k^{\top} \nabla_{H_{k,j}} I.$$

Synaptic plasticity.

$$\nabla_{W_k} I = d_m u_{k,j} + 2 d_s \mathrm{Cov}_{k,j} W_k, \qquad \nabla_{W_k} \Phi(a_{k,j}) = \frac{2a_{k,j}}{1 + a_{k,j}^2}\, f'(z_{k,j})\, x_{k,j}.$$

Kalman correction.

$$K_{k,j} = P_{j|j-1} H_{k,j}^{\top} \mathrm{Cov}_{k,j}^{-1}, \qquad h_{j|j} = h_{j|j-1} + \sum_k \big[K_{k,j}(x_{k,j} - u_{k,j}) + \eta_h H_{k,j}^{\top} g_{u,k,j}\big]\,.$$

Kalman-MI update (Algorithm). At time j:

$$
\begin{aligned}
&\text{Predict:} && h_{j|j-1} = F h_{j-1|j-1}, \quad P_{j|j-1} = F P_{j-1|j-1} F^\top + Q, \\
&\text{Observation prediction:} && u_{k,j} = S_k H_{\text{global}} h_{j|j-1}, \quad \mathrm{Cov}_{k,j} = S_k (H_{\text{global}} P_{j|j-1} H_{\text{global}}^\top + R_{\text{global}}) S_k^\top + R_k, \\
&\text{Activation:} && z_{k,j} = W_k^\top x_{k,j}, \quad a_{k,j} = f(z_{k,j}) + \varepsilon_{k,j}, \\
&\text{Predict scalar moments:} && m_{k,j} = W_k^\top u_{k,j}, \quad s_{k,j} = W_k^\top \mathrm{Cov}_{k,j} W_k, \\
&\text{MI gradients:} && g_{u,k,j} = d_m W_k, \quad G_{S,k,j} = d_s (W_k W_k^\top), \\
& && \nabla_{H_{k,j}} I = g_{u,k,j} h_{j|j-1}^\top + 2 d_s (W_k W_k^\top) H_{k,j} P_{j|j-1}, \\
& && \nabla_{W_k} I = d_m u_{k,j} + 2 d_s \mathrm{Cov}_{k,j} W_k, \quad \nabla_{W_k} \Phi = \frac{2 a_{k,j}}{1 + a_{k,j}^2} f'(z_{k,j}) x_{k,j}, \\
&\text{Kalman update:} && K_{k,j} = P_{j|j-1} H_{k,j}^\top \mathrm{Cov}_{k,j}^{-1}, \\
& && h_{j|j} = h_{j|j-1} + \sum_k \left(K_{k,j} (x_{k,j} - u_{k,j}) + \eta_h H_{k,j}^\top g_{u,k,j} \right), \\
&\text{Parameter update:} && W_k \leftarrow W_k - \eta_W \nabla_{W_k} I - \lambda \nabla_{W_k} \Phi, \\
& && H_{\text{global}} \leftarrow H_{\text{global}} - \eta_H \sum_k S_k^\top \nabla_{H_{k,j}} I.
\end{aligned}
$$

3. Results

3.1. Approach 1: Deep Generative Model and Surprise-Based Adaptation

A CochleaNet model pretrained on the full LibriSpeech corpus is used to compute surprise values for each stimulus. Using these fixed surprise signals, we train the downstream single-layer neural network on 10 recordings over 400 timesteps. To evaluate robustness, the learning procedure (for the synaptic weights of the single-layer network) is repeated five times using randomly sampled recordings. The hyperparameters used are listed in Table 1.

In Approach 1, we establish biological plausibility through a set of complementary analyses summarized in Fig. 2. The top three panels show normalized tuning curves across three integration timescales (4, 16, and 64 ms), illustrating the emergence of stable, frequency-selective responses that are consistent across independent runs. The fourth panel from the top plots a weight heatmap that visualizes the spectrotemporal evolution of synaptic weights, revealing gradual spectral localization of synaptic strengthening. Quantitatively, the resulting frequency selectivity (Q10dB; bottom right image) exhibits systematic broadening at faster timescales and sharpening at slower timescales, a hallmark of auditory cortical processing [11]. Notably, both the median values and percentile ranges of Q10dB closely match those reported for ferret A1 cortex [12], providing direct neurophysiological validation; detailed numerical comparisons are reported in Appendix D (Table 4 and Table 5). The bottom left image, shows smoothed percentage weight changes rapidly decaying to zero, indicating fast and robust convergence across all runs. While Fig. 2 presents representative examples for clarity, full convergence plots and weight dynamics across all timescales are also provided in Appendix D.

3.2. Approach 2: Kalman-MI Framework

Simulations were run for $T = 30{,}000$ timesteps (10 ms resolution) using a 22-channel cochleagram. A standard tone (channel 8) occurred with probability 0.9 and a deviant tone (channel 14) with probability 0.1, each generated as Gaussian spectral profiles ($\sigma_{\text{stim}} = 2.0$) with additive noise. The model used 6 neurons, each receiving a $B = 7$-channel patch (stride 3). The latent state stored $L = 20$ past frames (context window of 20 frames, $nL = 440$). H_{global} was initialized uniformly at 0.01, and W_k were initialized randomly; and the state prior was initialized as a broad uninformative Gaussian. The hyperparameters used are listed in Table 2. An ablation (Appendix E) sets $\eta_H = \eta_W = \eta_h = 0$ to isolate the contribution of MI gradients.

In Approach 2, we show how the model develops selective responses to the deviant stimulus (Fig. 3). The tuning curves (Fig. 3(a)) show that most neurons remain only weakly frequency-modulated, while Neuron 5

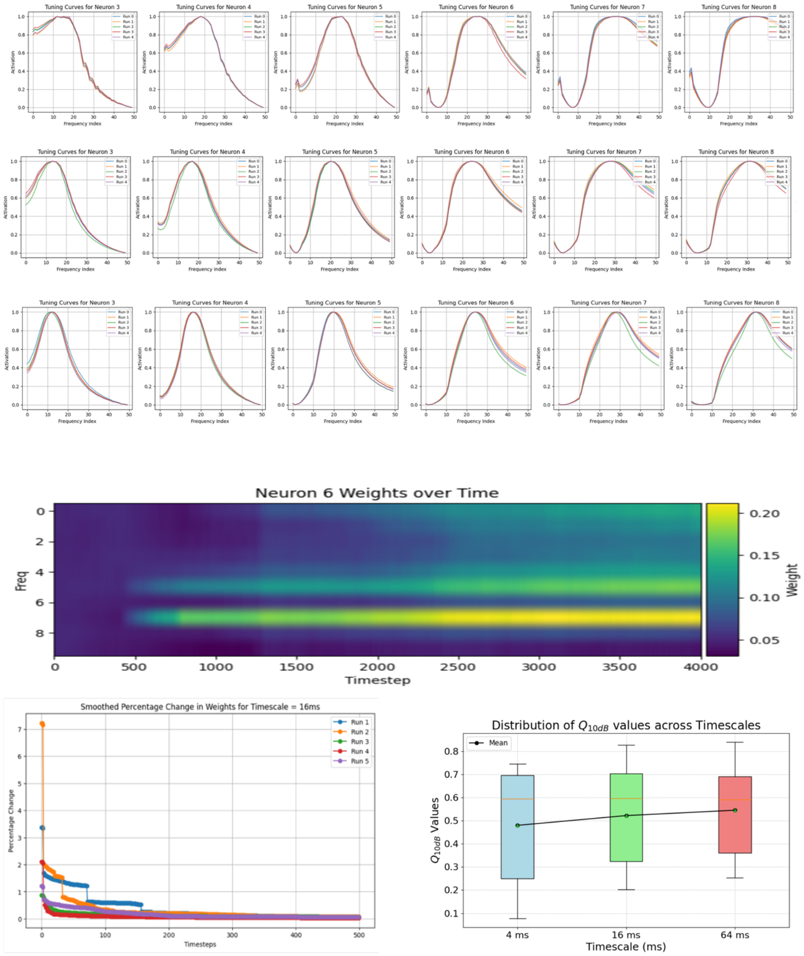


Figure 2: (a) (Top three rows) Normalized tuning curves for 4 ms, 16 ms, and 64 ms across five independent runs of each experiment, demonstrating consistent emergence of frequency selectivity by each neuron.(b) (Fourth row from the top) Weight update profile for a neuron showing frequency selectivity by synaptic weight strengthening over time. (c) (Bottom Left) Smoothed percentage changes in synaptic weights show rapid decay, indicating robust convergence for the representative timescale of 16ms (d) (Bottom Right) Q10dB values for 4 ms, 16 ms, and 64 ms, showing broader tuning at faster timescales and narrower tuning at slower timescales, consistent with findings in the auditory cortex [11].

develops a clear peak near the deviant channel and Neuron 4 shows a corresponding suppression, indicating an emergent division into excitatory and inhibitory deviant-sensitive units. The synaptic weight heatmaps (Fig. 3(b)) confirm this specialization: Neuron 5 displays a strong, systematic increase in weights around the deviant band, whereas other neurons remain largely stable. Finally, the evolution of the global generative matrix (Fig. 3(c)) reveals that the model allocates increasing predictive structure to the deviant frequency region.

Table 1: Hyperparameter values and initialization used in all experiments.

Hyperparameter	**Value**
Frequency input size (b)	50
Stride (m)	4
Learning rate (η)	10^{-4}
Surprise parameters (α, β, γ)	10^{-5}, 10^{-5}, 10^{-4}
Weight initialization	Uniform (0.05)

Table 2: Kalman–MI hyperparameter values and initialization

Hyperparameter	**Value**
Learning rates (η_H, η_W)	$10^{-4}, 5 \times 10^{-4}$
State update rate (η_h)	10^{-4}
Predictive-noise parameter (α)	0.05
Sparsity coefficient (λ)	5×10^{-6}
Weight Initialization	Random

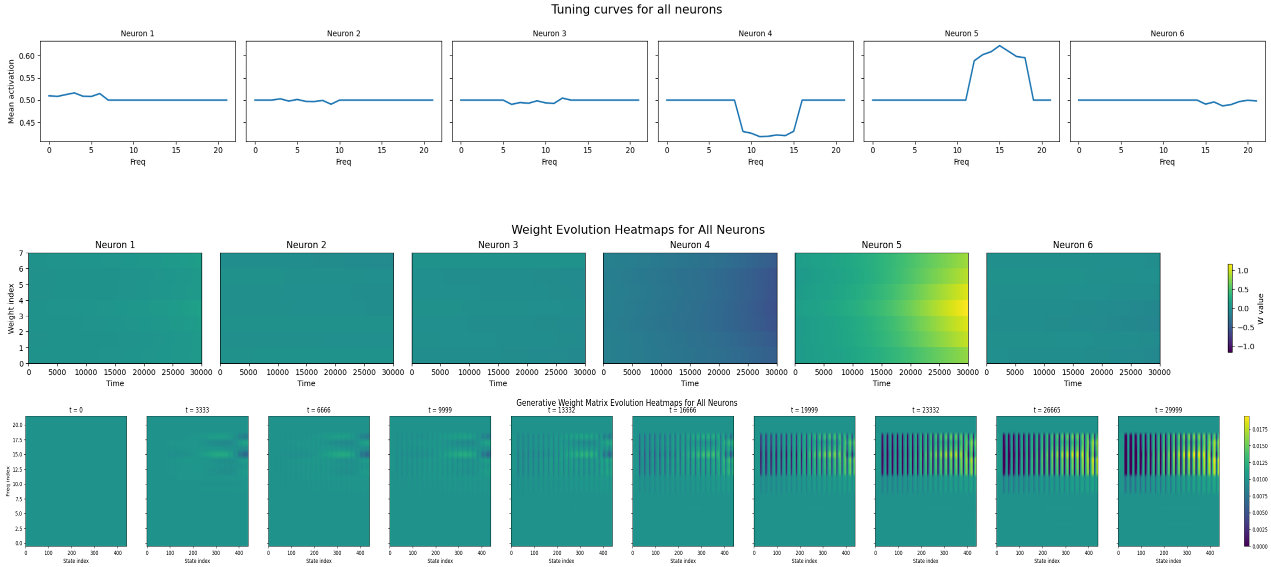


Figure 3: (a) (Top row) Normalized tuning curves for all neurons showing frequency selectivity for the deviant tone and sideband inhibition close to the deviant tone (b) (Middle Row) Weight update profile for all 6 neurons showing frequency selectivity for the deviant tone (c) (Bottom Row) Generative weight matrix update profile at 10 uniformly spaced timesteps across the total time period

4. Conclusion

This work introduced two complementary frameworks for understanding how auditory receptive fields and deviant sensitivity emerge from the interaction of prediction, surprise, and efficient coding. The first framework, *CochleaNet*, demonstrated that a deep generative model trained on natural speech can provide empirically grounded surprise estimates which, when coupled with sparsity-driven plasticity, produce STRFs whose tuning widths, frequency selectivity, and temporal scaling closely match neurophysiological observations. These results show that surprise-guided learning on real acoustic statistics is key to reproduce auditory cortical phenomena. The second framework, *Kalman-MI*, addressed conceptual limitations of purely data-driven heuristic approaches by offering a normative, fully adaptive formulation in which synaptic weights, latent states, and the generative prior are updated online through mutual-information maximization. Although evaluated here in a simplified oddball setting, the model exhibits hallmark auditory features such as deviant selectivity, inhibitory sidebands, and structured generative weights, providing mechanistic insight into how prediction and uncertainty shape auditory representations over time. Taken together, CochleaNet establishes *what* receptive-field structures emerge when learning is driven by realistic speech statistics, while the Kalman-MI framework explains *why* such structures arise from first-principles optimization under uncertainty.

Extending the Kalman-MI formulation to real-world speech would require a multi-layer architecture in which intermediate representations define the evolving input distribution for downstream layers, introducing substantial mathematical and computational complexity. Further, it would require the formulation of a non-linear generative model, as opposed to the linear-Gaussian generative model assumed here. We view this as an important direction for future work.

By unifying empirical modeling with theoretical parsimony, this work provides a foundation for future work combining large-scale generative models with online Bayesian learning to build biologically grounded yet computationally scalable models of auditory cortex. This understanding could have significant implications for various applications, from improved hearing aids to more sophisticated speech recognition systems and noise cancellation technologies [13], alongside informing the development of more effective and biologically-inspired artificial neural networks, particularly for processing temporal sequences and complex data streams.

References

[1] Nikolaus Kriegeskorte and Pamela K. Douglas. Cognitive computational neuroscience. *Nature Neuroscience*, 21(9):1148–1160, August 2018. ISSN 1546-1726. doi: 10.1038/s41593-018-0210-5. URL http://dx.doi.org/10.1038/s41593-018-0210-5.

[2] Jennifer K. Bizley and Yale E. Cohen. The what, where and how of auditory-object perception. *Nature Reviews Neuroscience*, 14(10):693–707, September 2013. ISSN 1471-0048. doi: 10.1038/nrn3565. URL http://dx.doi.org/10.1038/nrn3565.

[3] Anne-Lise Giraud and David Poeppel. Cortical oscillations and speech processing: emerging computational principles and operations. *Nature Neuroscience*, 15(4):511–517, March 2012. ISSN 1546-1726. doi: 10.1038/nn.3063. URL http://dx.doi.org/10.1038/nn.3063.

[4] Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. *Vision Research*, 49(10): 1295–1306, 2009.

[5] Gloria G. Parras, Javier Nieto-Diego, Guillermo V. Carbajal, Catalina Valdés-Baizabal, Carles Escera, and Manuel S. Malmierca. Neurons along the auditory pathway exhibit a hierarchical organization of prediction error. *Nature Communications*, 8(1), December 2017. ISSN 2041-1723. doi: 10.1038/s41467-017-02038-6. URL http://dx.doi.org/10.1038/s41467-017-02038-6.

[6] M. S. Lewicki. Efficient coding of natural sounds. *Nature Neuroscience*, 5(4):356–363, 2002.

[7] Frédéric E. Theunissen and Julie E. Elie. Neural processing of natural sounds. *Nature Reviews Neuroscience*, 15(6):355–366, May 2014. ISSN 1471-0048. doi: 10.1038/nrn3731. URL http://dx.doi.org/10.1038/nrn3731.

[8] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties. *Nature*, 381 (6583):607–609, 1996.

[9] Muneshwar Mehra, Adarsh Mukesh, and Sharba Bandyopadhyay. Earliest experience of a relatively rare sound but not a frequent sound causes long-term changes in the adult auditory cortex. *The Journal of Neuroscience*, 42(8):1454–1476, December 2021. ISSN 1529-2401. doi: 10.1523/jneurosci.0431-21.2021. URL http://dx.doi.org/10.1523/JNEUROSCI.0431-21.2021.

[10] Sean Vasquez and Mike Lewis. Melnet: A generative model for audio in the frequency domain. *ArXiv*, abs/1906.01083, 2019. URL https://api.semanticscholar.org/CorpusID:174798083.

[11] F. A. Rodriguez, C. Chen, H. L. Read, and M. A. Escabí. Spectral and temporal modulation tradeoff in the inferior colliculus. *Journal of Neurophysiology*, 103(2), 2010. doi: 10.1152/jn.00813.2009. URL https://doi.org/10.1152/jn.00813.2009.

[12] Jennifer K. Bizley, Fernando R. Nodal, Israel Nelken, and Andrew J. King. Functional organization of ferret auditory cortex. *Cerebral Cortex*, 15(10):1637–1653, October 2005. doi: 10.1093/cercor/bhi042. URL https://doi.org/10.1093/cercor/bhi042.

[13] Enrique A. Lopez-Poveda and Almudena Eustaquio-Martín. Objective speech transmission improvements with a binaural cochlear implant sound-coding strategy inspired by the contralateral medial olivocochlear reflex. *The Journal of the Acoustical Society of America*, 143(4):2217–2231, April 2018. ISSN 1520-8524. doi: 10.1121/1.5031028. URL http://dx.doi.org/10.1121/1.5031028.

Appendix

A. CochleaNet Architecture and Training

A.1. CochleaNet Architecture

Inspired by image autoregressive models that predict distributions pixel-by-pixel, this architecture estimates a distribution for each element in the spectrogram's time-frequency domain. Unlike images, spectrograms show frequency-dependent variations. To address this, the network utilizes a fully recurrent architecture with multiple processing stacks. One stack captures information from previous frames (time-delayed), while the other considers all past elements within a frame and the time-delayed stack's output (frequency-delayed). These stacks are interconnected at each layer, and both employ residual connections for deeper network training. Finally, the final layer's output from the frequency-delayed stack feeds into the calculation of unconstrained parameters.

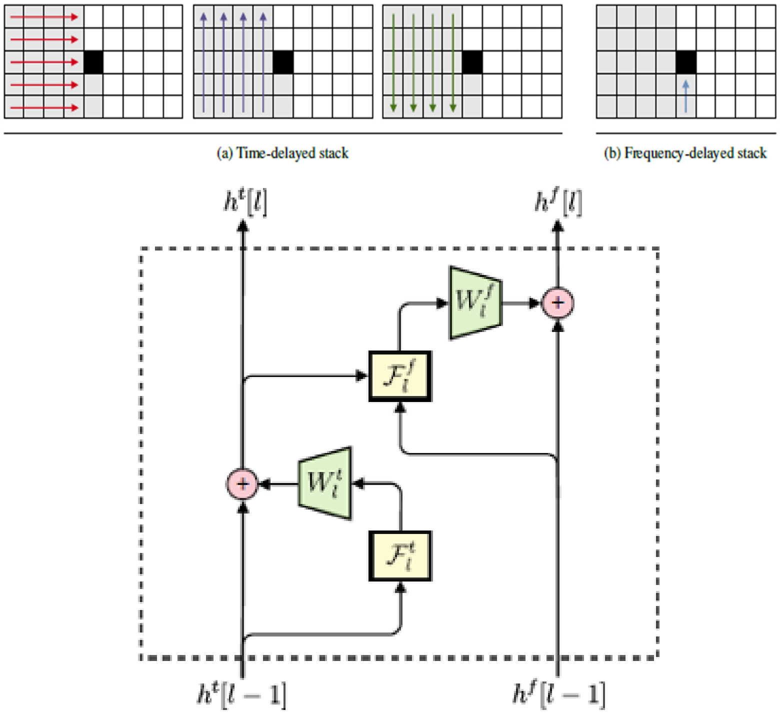


Figure 4: **CochleaNet architecture.** *Top:* Visualization of the time-delayed and frequency-delayed stacks. *Bottom:* Computational graph for a single network layer. F_l^t and F_l^f denote the transformations computed by the time-delayed and frequency-delayed stacks, respectively, at layer l. Their outputs are projected via weight matrices W_l^t and W_l^f and summed with the layer input to form residual connections. At the final layer, the network predicts the parameters of the conditional probability distribution governing each frequency-time bin (see main text). These parameters are obtained via a linear readout:

$$\hat{\theta}_{ij} = W_\theta \, h_{ij}^f[L],$$

where $\hat{\theta}_{ij}$ denotes the predicted distribution parameters, $h_{ij}^f[L]$ is the frequency-delayed representation at layer L, and W_θ is a learned projection matrix.

A.2. CochleaNet Training Details

We trained the above model on the Cochleagrams generated across 5 different timescales. We trained the model using Log-Likelihood Loss and Adam optimizer. The Cochleagrams are chunked to have dimension

= [50, 200]. We obtain 5 different models for 5 different timescales. The table below highlights the hyperparameters used for training that gave the best results.

Table 3: CochleaNet training hyperparameters for different temporal resolutions.

Timescale (ms)	Learning Rate	Hidden Size	Number of Layers
4	1×10^{-5}	16	5
8	1×10^{-5}	32	5
16	1×10^{-5}	32	5
32	1×10^{-5}	32	5
64	1×10^{-5}	64	5

A.3. CochleNet Results

A.3.1. Sampled cochleagrams from the 5 different models

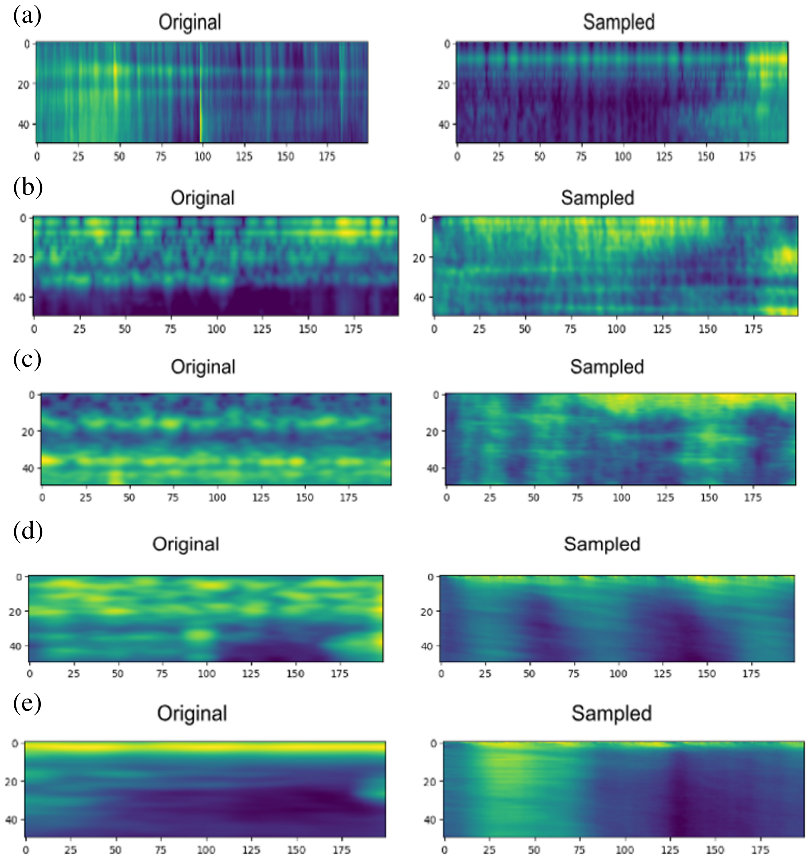


Figure 5: Sampled cochleagrams from CochleaNet models trained at different timescales: (a) 4 ms, (b) 8 ms, (c) 16 ms, (d) 32 ms, and (e) 64 ms. Visual inspection confirms that samples are similar to the corresponding real cochleagrams at each timescale.

B. Analysis of Surprise Values Distribution

This appendix summarizes the statistical analysis used to identify an appropriate probability distribution for surprise values computed from the generative model. Figure 6 reports the residual sum of squares (RSS) for several candidate families, while Figure 7 shows that the Johnson-Su distribution provided the best overall fit.

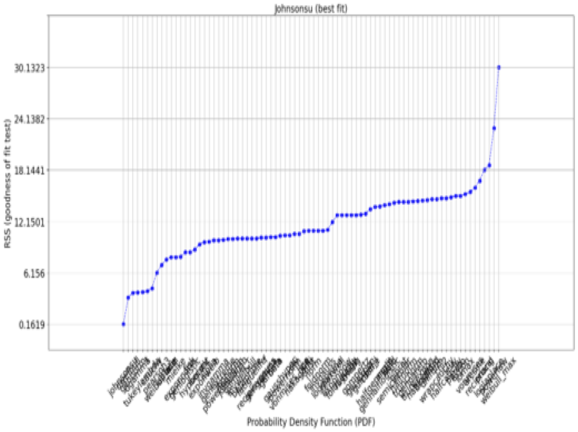


Figure 6: Residual sum of squares (RSS) for each candidate distribution tested.

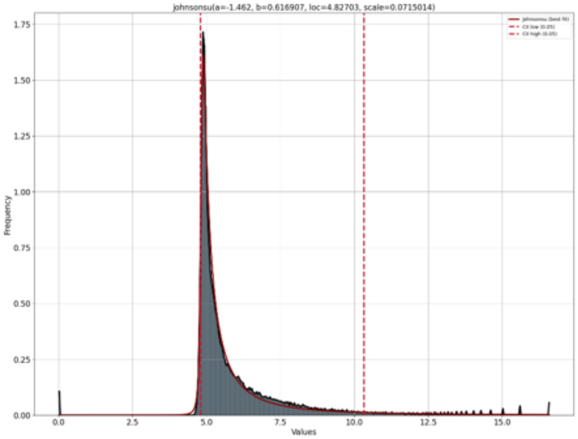


Figure 7: Best-fitting distribution for surprise values (Johnson–Su).

C. Mathematical Derivation of the Kalman-MI Framework

This appendix contains the complete mathematical specification of the Kalman-MI model, including indexing conventions, generative assumptions, mutual-information gradients, lifted state-space derivatives, and full parameter updates. The expressions below correspond exactly to the implementation used in the experiments and are provided for completeness and reproducibility.

C.1. Indices and basic notation

- Frequency-channel index: $i \in \{1, \dots, B\}$ (within a patch).
- Time index: $j \in \{1, \dots, T\}$.
- Patch index: $k \in \{1, \dots, K\}$.

Each patch k observes a contiguous set of frequency channels:

$$I_k = \{i_k,\, i_k + 1,\, \dots,\, i_k + B - 1\}.$$

C.2. Observed stimulus for a patch

For each channel $i \in I_k$ at time j, let

$$x_{i,j}^{(k)}$$

denote the observed energy in that frequency bin for patch k.

Stack these into a vector:

$$x_{k,j} = \begin{bmatrix} x^{(k)}_{i_k,j} \\ x^{(k)}_{i_k+1,j} \\ \vdots \\ x^{(k)}_{i_k+B-1,j} \end{bmatrix} \in \mathbb{R}^B.$$

C.3. Observed stimulus and global history

$$h_j = \begin{bmatrix} x_{1,j-1} \\ \vdots \\ x_{n,j-L+1} \end{bmatrix} \in \mathbb{R}^{nL},$$

and the patch observation vector is the slice $x_{k,j} = S_k x_j$.

C.4. State-space generative model

Conditioned on the hidden state $h_{k,j}$ and generative parameters $\theta = (F_k, H_{k,j}, Q_k, R_k)$, the observation model for patch k is:

$$x_{k,j} \mid \theta,\, h_{k,j} \sim \mathcal{N}(u_{k,j},\, \mathrm{Cov}_{k,j}),$$

C.5. Global observation mapping (single H)

We propose a single global generative mapping

$$H_{\mathrm{global},j} \in \mathbb{R}^{n \times nL},$$

which maps the full spectral history $h_j \in \mathbb{R}^{nL}$ to a global predicted mean $u_j \in \mathbb{R}^n$:

$$u_j = H_{\mathrm{global},j}\, h_{j|j-1}.$$

Each patch k then reads a contiguous block of B rows from u_j. Let $I_k = \{i_k, \ldots, i_k + B - 1\}$ denote the channel indices of patch k (as above) and define the binary row-selector matrix $S_k \in \{0,1\}^{B \times n}$ that extracts those rows:

$$u_{k,j} = S_k\, u_j \in \mathbb{R}^B, \qquad H_{k,j} \equiv S_k\, H_{\mathrm{global},j} \in \mathbb{R}^{B \times nL}.$$

$H_{k,j}$ is the corresponding row-block of $H_{\mathrm{global},j}$.

C.6. State-space generative model (with H_{global})

Conditioned on the full history $h_{j|j-1}$ and parameters $\theta = (F, H_{\mathrm{global},j}, Q, R_k)$, the global predicted mean and per-patch predictive covariance are

$$u_j = H_{\mathrm{global},j}\, h_{j|j-1} \in \mathbb{R}^n, \qquad u_{k,j} = S_k u_j \in \mathbb{R}^B,$$

The predicted global covariance (for all channels) is

$$\mathrm{Cov}_j \;=\; H_{\mathrm{global},j}\, P_{j|j-1}\, H_{\mathrm{global},j}^{\top} \;+\; R_{\mathrm{global}},$$

and the per-patch predictive covariance is obtained by selection:

$$\mathrm{Cov}_{k,j} \;=\; S_k\, \mathrm{Cov}_j\, S_k^{\top} \;+\; R_k \;=\; (S_k H_{\mathrm{global},j})\, P_{j|j-1}\, (S_k H_{\mathrm{global},j})^{\top} \;+\; R_k.$$

C.7. Shift register structure of F.

The state-transition matrix F shifts the history downward by one frame:

$$F = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ I_n & 0 & \cdots & 0 \\ 0 & I_n & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & I_n \end{bmatrix} \in \mathbb{R}^{nL \times nL}.$$

where n is the total number of frequency channels. After each update, the newest spectrum x_j overwrites the top block of $h_{j|j}$.This gives F_k a fixed structure, and we do not need to learn it.

C.8. Complete per-patch, per-time loss

We adopt the (per-patch k, per-time j) loss used in the code:

$$\mathcal{L}_{k,j} \;=\; -\,I\big(a_{k,j},\, x_{k,j}\big) \;+\; \lambda\,\Phi\big(a_{k,j}\big),$$

where:

- $I(a,x)$ is the mutual information between the scalar activation of a patch $a_{k,j}$ and the vector of observed energies in the patch,$x_{k,j}$.
- $a_{k,j} = f(z_{k,j}) + \varepsilon_{k,j}$ is the realized (noisy) activation . In code the penalty is computed from the instantaneous (noisy) activation.
- Φ is the sparsity penalty: $\Phi(a) = \log(1 + a^2)$. Its derivative is

  $$\Phi'(a) = \frac{2a}{1 + a^2}.$$

- The idea is to jointly maximize the Mutual Information (MI) between the stimulus and the response and minimize overall neuronal activity. (promote sparsity)

Total loss over patches and time is

$$\mathcal{L} \;=\; \sum_{j=1}^{T}\sum_{k=1}^{K} \mathcal{L}_{k,j}.$$

C.9. Gradient of the loss w.r.t. the synaptic weights W_k

Using the chain rule, the gradient of $\mathcal{L}_{k,j}$ w.r.t W_k comprises two contributions:

$$\nabla_{W_k}\mathcal{L}_{k,j} = -\underbrace{\nabla_{W_k} I(a_{k,j}, x_{k,j})}_{\text{MI-part}} + \lambda\,\underbrace{\nabla_{W_k}\Phi\big(a_{k,j}\big)}_{\text{sparsity-part}}.$$

C.10. Kalman prediction

$$h_{k,j|j-1} = F_k\, h_{k,j-1|j-1}, \tag{4}$$

$$P_{k,j|j-1} = F_k\, P_{k,j-1|j-1}\, F_k^\top + Q_k. \tag{5}$$

C.11. Predictive drive and synpatic weights

The linear readout is:

$$z_{k,j} = W_k^\top x_{k,j}.$$

Assuming a Gaussian predictive law:

$$z_{k,j} \mid h_{k,j} \sim \mathcal{N}(m_{k,j},\, s_{k,j}),$$

with

$$m_{k,j} = W_k^\top\, u_{k,j}, \tag{6}$$

$$s_{k,j} = W_k^\top\, \mathrm{Cov}_{k,j}\, W_k. \tag{7}$$

C.12. Activation nonlinearity

The neural activation is:

$$a_{k,j} = f(z_{k,j}) + \varepsilon_{k,j}, \qquad f(z) = \sigma(z) = \tfrac{1}{1+e^{-z}},$$

with heteroscedastic noise:

$$\varepsilon_{k,j} \sim \mathcal{N}\big(0,\, \sigma^2_{\varepsilon,k,j}\big)\,, \qquad \sigma^2_{\varepsilon,k,j} = \alpha\, \mu_{a,k,j}.$$

Under a second-order Delta approximation,

$$\mu_{a,k,j} = f(m_{k,j}) + \tfrac{1}{2} f''(m_{k,j})\, s_{k,j}, \tag{8}$$

$$\sigma^2_{a,k,j} = f'(m_{k,j})^2\, s_{k,j} \;+\; \alpha\,\big(f(m_{k,j}) + \tfrac{1}{2} f''(m_{k,j})\, s_{k,j}\big)\,. \tag{9}$$

Thus

$$a_{k,j} \sim \mathcal{N}(\mu_{a,k,j},\, \sigma^2_{a,k,j}).$$

C.13. Predictive MI derivatives

From the MI objective $I(a_{k,j}, x_{k,j})$, recall that mutual information is defined as:

$$I(a_{k,j}, x_{k,j}) = H(a_{k,j}) - H(a_{k,j} \mid x_{k,j}),$$

where the predictive activation is Gaussian with variance $\sigma^2_{a,k,j}$ and the conditional distribution (given the stimulus) has variance $\alpha\, \mu_{a,k,j}$. Thus,

$$H(a_{k,j}) = \frac{1}{2}\log\big(2\pi e\, \sigma^2_{a,k,j}\big)\,, \qquad H(a_{k,j} \mid x_{k,j}) = \frac{1}{2}\log(2\pi e\, \alpha\, \mu_{a,k,j})\,.$$

From the MI objective $I(a_{k,j}, x_{k,j})$:

$$d_{m,k,j} = \frac{\partial I}{\partial m_{k,j}} = \frac{1}{2\sigma^2_{a,k,j}}\left(2ff''s_{k,j} + \alpha(f' + \tfrac{1}{2}f'''s_{k,j})\right) - \frac{1}{2\mu_{a,k,j}}\left(f' + \tfrac{1}{2}f'''s_{k,j}\right),$$

$$d_{s,k,j} = \frac{\partial I}{\partial s_{k,j}} = \frac{1}{2\sigma^2_{a,k,j}}(f^2 + \alpha f'') - \frac{1}{2\mu_{a,k,j}} f''.$$

C.14. Final combined gradient:

$$\boxed{\nabla_{W_k}\mathcal{L}_{k,j} = -\big(d_{m,k,j}\, u_{k,j} + 2 d_{s,k,j}\, \mathrm{Cov}_{k,j}\, W_k\big) \;+\; \lambda\, \frac{2a_{k,j}}{1 + a^2_{k,j}}\, f'\big(z_{k,j}\big)\, x_{k,j}.}$$

C.15. Per-patch lifts to observation and global state

Per-patch observation-space gradients are computed with $H_{k,j} = S_k H_{\mathrm{global},j}$: The MI derivatives lift to observation-space for each patch as

$$g_{u,k,j} \;=\; \frac{\partial I}{\partial u_{k,j}} \;=\; d_{m,k,j}\, W_k \in \mathbb{R}^B,$$

$$G_{S,k,j} \;=\; \frac{\partial I}{\partial \mathrm{Cov}_{k,j}} \;=\; d_{s,k,j}\,(W_k W_k^\top) \in \mathbb{R}^{B\times B}.$$

Lifted to the *global* state-space (same $h_{j|j-1}$ for all patches):

$$g_{h,k,j} = H_{k,j}^\top g_{u,k,j} = (S_k H_{\mathrm{global},j})^\top g_{u,k,j} = H_{\mathrm{global},j}^\top S_k^\top g_{u,k,j} \in \mathbb{R}^{nL},$$
$$G_{P,k,j} = H_{k,j}^\top G_{S,k,j} H_{k,j} = H_{\mathrm{global},j}^\top S_k^\top G_{S,k,j} S_k H_{\mathrm{global},j} \in \mathbb{R}^{nL\times nL}.$$

Because the global latent state $h_{j|j-1}$ is shared across patches, the total state-space derivatives sum over patches:

$$g_h^{(\mathrm{total})} = \sum_{k=1}^{K} g_{h,k,j}, \qquad G_P^{(\mathrm{total})} = \sum_{k=1}^{K} G_{P,k,j}.$$

C.16. Parameter gradients

C.16.1. Parameter gradients: global H

Per-patch gradient of I w.r.t. the local observation block $H_{k,j}$ (row-block of the global matrix) is:

$$\nabla_{H_{k,j}} I = g_{u,k,j}\, h_{j|j-1}^{\top} \; + \; 2d_{s,k,j}\, (W_k W_k^{\top})\, H_{k,j}\, P_{j|j-1} \in \mathbb{R}^{B \times nL}.$$

Accumulate these into the gradient of the global observation matrix by inserting each per-patch row-block into the appropriate rows:

$$\nabla_{H_{\text{global},j}} I \;=\; \sum_{k=1}^{K} S_k^{\top}\, \nabla_{H_{k,j}} I \;=\; \text{row-block-accumulate}\big\{\nabla_{H_{k,j}} I\big\}_{k=1}^{K} \in \mathbb{R}^{n \times nL}.$$

Equivalently in index form:

$$\nabla_{H_{\text{global},j}} I\big[i_k : i_k + B - 1,\; :\big] \mathrel{+}= \; \nabla_{H_{k,j}} I.$$

C.16.2. Parameter gradients: per-patch W_k

The per-patch readout gradient for W_k is unchanged in form:

$$\nabla_{W_k} I = d_{m,k,j}\, u_{k,j} \; + \; 2d_{s,k,j}\, \mathrm{Cov}_{k,j}\, W_k,$$

and the sparsity (realized-activation) contribution is

$$\nabla_{W_k} \Phi(a_{k,j}) = \frac{2a_{k,j}}{1 + a_{k,j}^2}\, f'(z_{k,j})\, x_{k,j}.$$

C.17. Combined loss gradients and parameter updates

Per-patch loss is $\mathcal{L}_{k,j} = -I(a_{k,j}, x_{k,j}) + \lambda \Phi(a_{k,j})$. Thus the full gradients are

$$\nabla_{W_k} \mathcal{L}_{k,j} = -\big(d_{m,k,j}\, u_{k,j} + 2d_{s,k,j}\, \mathrm{Cov}_{k,j}\, W_k\big) \; + \; \lambda\, \frac{2a_{k,j}}{1 + a_{k,j}^2}\, f'(z_{k,j})\, x_{k,j},$$

and the global H gradient is the accumulation

$$\nabla_{H_{\text{global},j}} \mathcal{L} = -\sum_{k=1}^{K} S_k^{\top} \nabla_{H_{k,j}} I$$

Apply gradient steps (learning rates η_W, η_H):

$$W_k \leftarrow W_k - \eta_W\, \nabla_{W_k} \mathcal{L}_{k,j}, \qquad H_{\text{global},j} \leftarrow H_{\text{global},j} - \eta_H\, \nabla_{H_{\text{global},j}} \mathcal{L}.$$

C.18. Kalman correction and posterior update (global state)

For each patch compute the innovation $\nu_{k,j} = x_{k,j} - u_{k,j}$ and the patch Kalman gain using $H_{k,j} = S_k H_{\text{global},j}$:

$$K_{k,j} = P_{j|j-1} H_{k,j}^{\top} \mathrm{Cov}_{k,j}^{-1}.$$

Update the shared posterior latent state by summing patch contributions (and the small predictive-gradient push):

$$h_{j|j} = h_{j|j-1} + \sum_{k=1}^{K} \Big(K_{k,j}\, \nu_{k,j} + \eta_h\, H_{k,j}^{\top} g_{u,k,j}\Big).$$

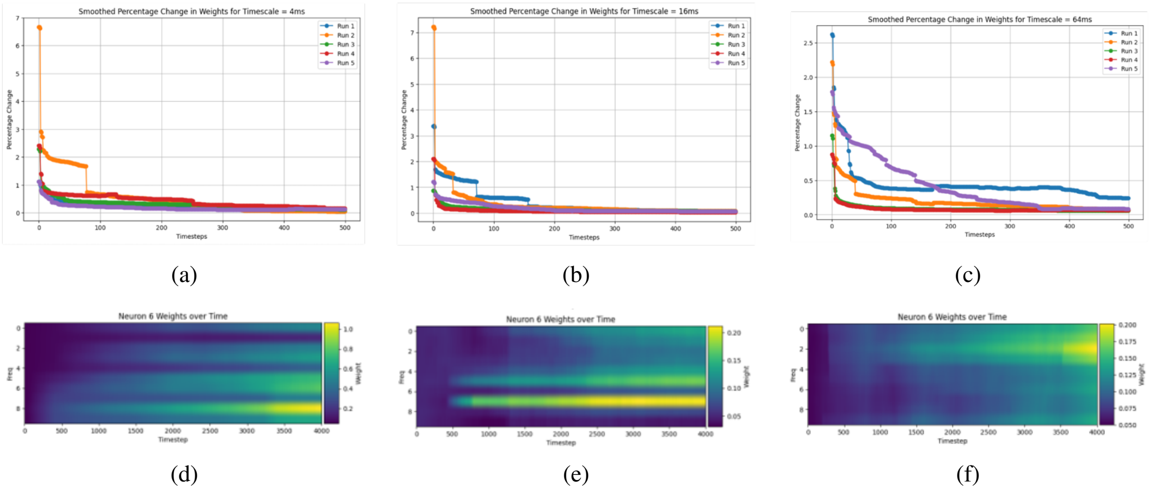
Posterior covariance $P_{j|j}$ may be updated using the Joseph form (summing patch contributions appropriately), ensuring numerical stability:

$$P_{j|j} = (I - \sum_k K_{k,j} H_{k,j})\, P_{j|j-1}\, (I - \sum_k K_{k,j} H_{k,j})^{\top} + \sum_k K_{k,j} R_k K_{k,j}^{\top},$$

D. Additional Results

Here, we present detailed results corresponding to Section 3.1 in the main manuscript.

D.1. Fast convergence of the Weights



(a) (b) (c)

(d) (e) (f)

Figure 8: (a-c) Smoothed percentage change in synaptic weights across five runs shows fast and robust convergence at (a) 4 ms, (b) 16 ms, and (c) 64 ms timescales. Lower panels: Weight matrices of a neuron over time illustrate stable adaptation and frequency-specific structure at each timescale.

D.2. Alignment of Q10dB values with neurophysiological data

Table 4: Q_{10dB} values (median, 10th, and 90th percentiles) from our model across timescales. Highlighted cells show close correspondence with A1 cortex measurements (Table 5).

Parameter	4 ms	16 ms	64 ms
Q_{10dB} (median)	0.594	0.596	0.590
10th percentile	0.146	0.234	0.330
90th percentile	0.738	0.778	0.760

Table 5: Q_{10dB} values (median and 10th–90th percentile) for different ferret auditory cortical areas, adapted from [12]. A1 cortex values are highlighted.

Parameter	All areas	A1	AAF	PSF	PPF	ADF	AVF
Q_{10dB} (median)	1.37	0.59	0.27	0.45	0.67	0.35	0.25
10th-90th percentile	0.84-3.1	0.41-0.74	0.19-0.88	0-2.31	0-0	0.18-1.96	0.21-1.5

Model Tuning Curves Q10dB values show close correspondence with A1 cortex reference data from [12], particularly in median values (0.594-0.596 vs. 0.59) and 90th percentiles (0.738-0.778 vs 0.74) as highlighted in Table 2 and Table 3.

E. Ablation Study

A key claim of the main paper is that *surprise-driven learning*, rather than sparsity alone, is critical for the emergence of biologically meaningful auditory receptive fields and deviant selectivity. To address the concern that standard sparse coding objectives can also produce structured receptive fields, we perform explicit ablations comparing **MI + Sparsity** against a **Sparsity-only** baseline under otherwise identical conditions.

For the Kalman-MI framework introduced in the paper, the sparsity-only baseline is obtained by removing the mutual-information-driven terms while retaining the sparsity based learning, learning rates, architectures, and stimulus statistics.

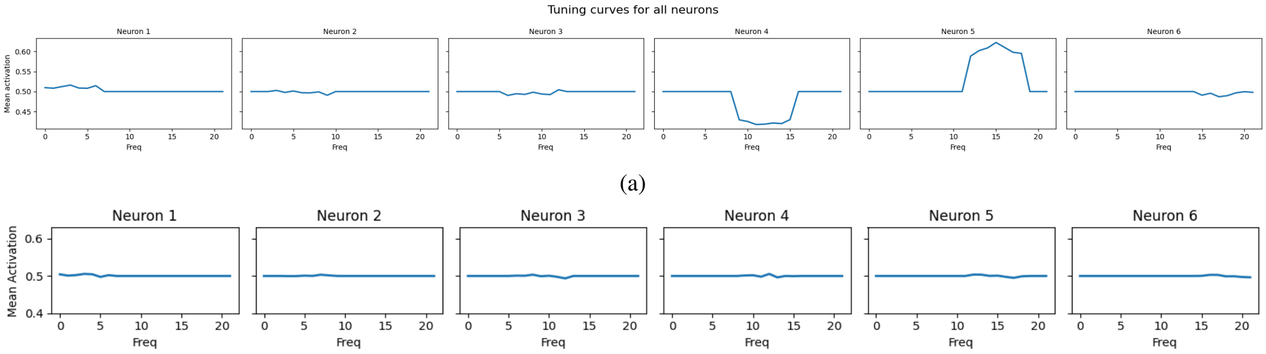
Simulations were run for $T = 30{,}000$ timesteps (10 ms resolution) using a 22-channel cochleagram. A standard tone (channel 8) occurred with probability 0.9 and a deviant tone (channel 14) with probability 0.1, each generated as Gaussian spectral profiles ($\sigma_{\text{stim}} = 2.0$) with additive noise. The model used 6 neurons, each receiving a $B = 7$-channel patch (stride 3). The latent state stored $L = 20$ past frames (context window of 20 frames, $nL = 440$). H_{global} was initialized uniformly at 0.01, and W_k were initialized randomly; and the state prior was initialized as a broad uninformative Gaussian.

E.1. Kalman-Mutual Information Framework: MI + Sparsity vs. Sparsity Only

The hyperparameters used for the MI + Sparsity experiments are already listed in Table 2. The sparsity-only ablation (Appendix E) sets $\eta_H = \eta_W = \eta_h = 0$, thus disbling MI-based learning. Figure 9 contrasts the two cases under an oddball stimulation paradigm.

In the **MI + Sparsity** condition, neurons exhibit clear deviant selectivity and structured sideband inhibition, including inhibitory responses in neurons adjacent to the deviant-selective neuron.

In contrast, the **Sparsity-only** baseline fails to develop any frequency tuning or deviant selectivity.



(a)

(b)

Figure 9: Kalman-MI ablation study. **Top:** Frequency tuning curves for Sparsity + Kalman-MI paradigm. **Bottom:** Frequency tuning curves for Sparsity-only paradigm . Mutual-information-driven learning yields deviant selectivity and sideband inhibition, which are absent under sparsity-only learning.

Summary of ablation results. In the Kalman-MI framework, sparsity alone is insufficient to produce stable, biologically meaningful receptive fields or deviant selectivity. Using only sparsity results in flat, non-selective responses. Surprise or mutual information-driven learning thus provides a critical inductive signal that shapes receptive field structure beyond what can be achieved by sparse coding objectives alone.

F. Derivation for the Sparsity Based Adaptation learning rule

After surprise-based synaptic adaptation, we apply adjustments according to the gradient of the l_1 norm of activations with respect to weights to enforce sparsity. Here, we derive the synaptic weight update rule to promote sparsity. The following parameters are used in the computation below:

b = Frequency input size (or filter size), n = Total number of frequencies, m = Stride.

η is the learning rate. The neural activations are computed as follows:

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_{b+1} \end{bmatrix} = f\left(\begin{pmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,b} & 0 & \cdots & 0 \\ 0 & w_{2,m+1} & \cdots & w_{2,m+b} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & w_{b+1,m+b+1} & \cdots & w_{b+1,n} \end{pmatrix} \begin{bmatrix} g_1 \\ g_2 \\ g_3 \\ \vdots \\ g_n \end{bmatrix} \right)$$

The l_1 norm of activations is calculated as:

$$l_1 = \sum a_i = a_1 + a_2 + \cdots + a_{(b+1)}$$

This can be expressed in matrix form as:

$$l_1 = [1 \quad 1 \quad \ldots \quad 1]_{1\times(b+1)} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{(b+1)} \end{bmatrix}$$

Individual neural activations are computed using an activation function f (sigmoid in our case):

$$a_1 = f(w_{1,1}g_1 + w_{1,2}g_2 + \cdots + w_{1,b}g_b)$$

The gradient of the l_1 norm with respect to weights is:

$$\frac{\partial l_1}{\partial w} = \begin{bmatrix} \frac{\partial l_1}{\partial w_{1,1}} & \frac{\partial l_1}{\partial w_{1,2}} & \cdots & \frac{\partial l_1}{\partial w_{1,b}} & 0 & \cdots & 0 \\ 0 & \frac{\partial l_1}{\partial w_{2,m+1}} & \cdots & \frac{\partial l_1}{\partial w_{2,m+b}} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0\ldots & \frac{\partial l_1}{\partial w_{b+1,mb+1}} & \cdots & \frac{\partial l_1}{\partial w_{b+1,n}} \end{bmatrix}_{(b+1)\times n}$$

Where each partial derivative is:

$$\frac{\partial l_1}{\partial w_{i,j}} = f'(\sum_{j=mi-m+1}^{mi+b-m} w_{i,j}g_j)g_j$$

Expanding this:

$$\frac{\partial l_1}{\partial w} = \begin{bmatrix} f'(\sum_{j=1}^{b} w_{1,j}g_j)g_1 & f'(\sum_{j=1}^{b} w_{1,j}g_j)g_2 & \cdots & 0 \\ 0 & f'(\sum_{j=m+1}^{m+b} w_{2,j}g_j)g_{m+1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & f'(\sum_{j=mb+1}^{n} w_{b+1,j}g_j)g_n \end{bmatrix}_{(b+1)\times n}$$

For computational efficiency, we can express this using matrices and vectors:

$$G_{\text{active}} = \begin{bmatrix} g_1 & g_2 & g_3 & \ldots & g_b & 0 & \ldots & 0 & \\ 0 & \ldots & 0 & g_{m+1} & g_{m+2} & \ldots & g_{m+b} & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \\ 0 & 0 & \ldots & 0 & g_{mb+1} & g_{mb+2} & \ldots & g_n & \end{bmatrix}_{(b+1)\times n} \qquad g = [g_1 \quad g_2 \quad \ldots \quad g_n]^{\top}_{n\times 1}.$$

Compute:

$$Q = W \cdot g = \begin{pmatrix} \sum_{j=1}^{b} w_{1,j} g_j \\ \sum_{j=m+1}^{m+b} w_{2,j} g_j \\ \vdots \\ \sum_{j=mb+1}^{n} w_{b+1,j} g_j \end{pmatrix}$$

$$R = f'(Q)$$

Define F, which is a $(b+1) \times n$ matrix obtained by broadcasting the $(b+1) \times 1$ vector R across n columns.

$$F = Broadcast(f'(W{\cdot}g)) = \begin{bmatrix} f'(\sum_{j=1}^{b} w_{1,j} g_j) & f'(\sum_{j=1}^{b} w_{1,j} g_j) & \cdots & 0 \\ 0 & f'(\sum_{j=m+1}^{m+b} w_{2,j} g_j) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & f'(\sum_{j=mb+1}^{n} w_{b+1,j} g_j) \end{bmatrix}_{(b+1)\times n}$$

$$F{\circ}G_{active} = \begin{bmatrix} f'(\sum_{j=1}^{b} w_{1,j} g_j) g_1 & f'(\sum_{j=1}^{b} w_{1,j} g_j) g_2 & \cdots & 0 \\ 0 & f'(\sum_{j=m+1}^{m+b} w_{2,j} g_j) g_{m+1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & f'(\sum_{j=mb+1}^{n} w_{b+1,j} g_j) g_n \end{bmatrix}_{(b+1)\times n}$$

(required gradient matrix)

The sparsity update rule is:

$$W_{\text{new}} = W - \eta \left(F \circ G_{\text{active}} \right),$$

where $F = \mathrm{broadcast}(f'(Wg))$ and G_{active} masks inactive weight entries. This enforces sparse neural code properties consistent with efficient coding theory.