Rare Events, Real Signals: Functional Ensembles as Units of Computation in Deep Spiking Networks

Aditi Aravind[1,2,*] **Konstantinos Ladakis**[1,2,*] **Mario Alexios Savaglio**[1,2]
Stelios M. Smirnakis[4,5] **Maria Papadopouli**[1,2,3,†]

[1]University of Crete, Greece
[2]Foundation of Research & Technology - Hellas (FORTH), Greece
[3] Archimedes Research Unit, Athena Research Center, Athens, Greece
[4]Harvard Medical School, USA [5]Brigham and Women's Hospital, USA

† Corresponding author: Maria Papadopouli (maria@csd.uoc.gr)
* Equal contribution

Abstract

We investigate how internal representations emerge across hierarchical processing systems by introducing a neuroscience-inspired framework for analyzing deep spiking neural networks (SNN) through the lens of functional connectivity. Drawing on concepts from systems neuroscience and information theory, we form the first-order functionally connected (1FC) group of a neuron based on its statistically significant pairwise correlations with neurons from the previous layer of a trained SNN architecture. We then track its response properties during inference under various conditions. Our analysis shows that several principles of functional connectivity previously observed in biological cortex are preserved in spiking ResNet architectures. These 1FC ensembles display interesting properties: their aggregate cofiring reliably predicts downstream neuronal responses through a robust, ReLU-like input–output relationship, whose gain scales systematically with ensemble size. Reliable encoding of the presented class emerges only during high 1FC cofiring events, which themselves occur infrequently, indicating that informative representations are concentrated in rare but highly coordinated activity patterns. Under uniform random noise or adversarial perturbations, these response profiles are disrupted, particularly in early and intermediate layers. This enables a targeted high-resolution interrogation at specific nodes and pathways. We showed that the functional connectivity structure is shaped by learning and this structure breaks under weight permutation. These establish 1FC ensembles as a functionally meaningful substrate for input encoding and information transfer, with potential implications in designing targeted fine-grained diagnostics on the information flow.

1 Introduction

Neural networks are not merely fitting data or minimizing training error; they are learning structured internal representations, exhibiting consistent patterns across tasks and scales [1–5]. However understanding how representations emerge across hierarchical processing systems remains a central challenge in computational neuroscience and machine learning. Inspired by systems neuroscience, we study this question in deep spiking neural architectures, with an emphasis on the functionally emergent connectivity. In neuroscience, both spontaneous activity—observed in the absence of strong sensory input—and stimulus-evoked responses exhibit structured, non-random patterns, indicating the

Preprint.

presence of organized collective dynamics. Neuronal ensembles—groups of neurons that *co-activate more frequently than expected by chance*—are increasingly viewed as fundamental units of cortical computation, enabling efficient encoding and transmission of shared information. A substantial body of work has identified structured patterns in both spontaneous and evoked population activity, supporting this ensemble-based view of neural processing [6–19]. In neocortex, cortico-cortical connectivity is weak, necessitating that ensembles of neurons coordinate to form computational units. Functional connectivity analysis offers a principled approach to identifying such ensembles, revealing networks of neurons linked through shared anatomical and statistical dependencies, forming canonical elements of information processing [20–25].

Building on this idea, we have proposed that all neurons of a layer in the visual cortex with statistically significant pairwise correlations with a *target neuron* compose its *first-order functionally connected (1FC)* ensemble in that layer, forming elementary cortical computational units (see attached manuscript in supplement). We then characterized their properties and couplings. This framework yields mechanistic insight into cortical circuit organization, complementary to structural connectivity, helping us also to bridge the study of cortical circuit organization with deep-learning architectures for assessing neuro-computing hypotheses.

Spiking neural networks (SNNs) consist of layers of simplified (computational) neurons, that mimic the hierarchical, integrative properties and sparse firing of biological circuits[26, 27]. Here we apply this framework to study the structure of these representations across the processing hierarchy in SNNs. This framework enables a targeted high-resolution interrogation, placing "magnifying lens" at critical nodes and pathways within the architecture, *during inference*, offering direct insight into the information flow. Unlike conventional interpretability approaches that summarize model behavior through attribution maps/scores [28–30], feature visualizations [31, 32], mechanistic approaches like causal tracing [33], or post hoc saliency analyses [34–36], this framework provides a node-level, pathway-specific view of computation. Rather than asking only which input features or units are important, it identifies *when* and *through which functional routes* information is propagated within the architecture during inference.

Innovation: We apply systems network neuroscience and information theory to reverse engineer the structure of learned *circuit* representations during inference. We do this by studying: **1)** the architecture of 1FC ensembles that emerge from identifying significant pairwise functional connections between neurons in consecutive layers during testing, and **2)** the *response properties* of these ensembles across layers that shape the information flow during inference, and how they change under different input testing conditions. While existing interpretability approaches predominantly analyze individual neuron activations or population-based layer-wise representations, this work takes a distinct approach by examining *functional* connectivity based on *pairwise correlations* of neurons between *consecutive layers*, to characterize the structure of information flow in SNNs.

Contributions: Although, we do not intend to claim a precise correspondence between SNNs and visual cortex, which have substantial differences [37], we find it interesting that several principles of the functional connectivity identified in mouse visual cortex (see attached manuscript in supplement) are conserved in SNNs. Both uniform random noise and adversarial perturbations induce changes in these response properties, with the most significant impact observed in early and intermediate processing layers. This provides an opportunity for targeted, high-resolution probing of specific nodes and pathways, particularly in regimes characterized by strong 1FC co-firing events. As expected, functional connectivity architecture emerges during learning and degrades markedly under weight permutation. Taken together, these observations support the interpretation of 1FC ensembles as a functionally relevant substrate for input encoding and information transmission, enabling targeted fine-grained diagnostics on the information flow.

Limitations: So far we tested this approach in ResNet18 SNNs and spiking-neural units of spiking-layers that extend a ResNet32 feature extractor[38], pre-trained on CIFAR100[39] and ResNet9 SNNs trained on DVS Gesture Dataset. We also have only tested the framework for the Fast Gradient Sign Method (FGSM) adversarial attack[40] and additive noise. It is part of our ongoing effort to test it with other architectures, activation functions, and adversarial conditions and analyze its diagnostic power in efficiently detecting anomalies using test datasets of different size, richness, and noise levels. Going forward, integrating this framework with causal perturbations will be key to clarifying the mechanistic role of specific ensembles in learning, as well as to probing how these mechanisms fail under adversarial conditions.

2 Related Work

Research on internal network interpretability in ANNs and SNNs has developed along several connected lines: probing intermediate representations for class-relevant information, comparing representational geometry across layers and models, attributing predictions to spatial regions or channels, and identifying influential hidden units [41–43]. One approach treats intermediate layers as objects to be probed. Alain and Bengio [44] showed that linear classifier probes can track where task-relevant information becomes linearly separable across depth, turning internal representations into diagnostic objects. This framework summarizes a layer by its ability to decode a target class, as opposed to the internal organization of neurons that support it. Representational comparison methods enable cross-layer and cross-model comparisons. For example, SVCCA [45] enabled comparisons by factoring out irrelevant basis changes, while Kornblith et al. [4] identified limitations of CCA-based approaches in high-dimensional settings and advocated for centered kernel alignment as a more reliable statistic. These methods work as tools for investigations that remain whole-layer summaries; they characterize what a representation contains, not how its constituent neurons are functionally organized.

A parallel branch explains predictions through class-discriminative spatial attribution. CAM [34] demonstrated that class-specific weights can be projected onto convolutional feature maps to identify decision-relevant image regions. Grad-CAM [35] generalized this using gradients into any chosen convolutional layer, removing architectural constraints. Ablation-CAM [36] proposed a gradient-free alternative based on feature-map ablation, reducing sensitivity to gradient pathologies. These methods make class importance spatially visible and are central in visualization of unit receptive fields, but they rely on predictions of models to ascertain feature representation of individual nodes or units. Previous work on information flow within the network concerns tracking the activations of the nodes. For example, Wang et al. [46] identifies critical data paths which are learnt by gating over channels to recover the semantically selective pathway most responsible for a prediction. This frames explanation in terms of internal pathways rather than input saliency. At the unit level, conductance [29] tracks how attribution flows through individual neurons, outperforming raw activation measures. Bau et al. [30] showed that ablating small subsets of class-aligned units can strongly impair class-specific performance suggesting that relevant structure is often concentrated in ensembles rather than isolated neurons. For SNNs, Youngeun and Priyadarshini [47] adapted the traditional visual explanation through Spike Activation Maps based on inter-spike intervals, providing a gradient-free, SNN-native attribution baseline, though still focused on individual neuron salience over time.
While existing interpretability approaches primarily focus on individual neuron activations or layer-wise representations, we instead examine functional connectivity through pairwise correlations between neurons in consecutive layers, enabling a structured characterization of information flow in SNNs.

3 Functional Connectivity Methodology

Network Neuroscience Background. We recently found that biological neurons, along with their functionally-connected neighbors form elementary multi-neuronal ensembles that act as putative information-processing "primitives" both within and across granular and supragranular layers in mouse primary visual cortex[23]. Such ensembles, being flexible and multiplexing, may support task-specific processing, adaptation, and learning through dynamic reconfiguration of their interactions. We found that the firing probability of layer 2/3 ("putative output") neurons in mouse primary visual cortex exhibits a ReLU-like nonlinearity, emerging when 13% of L4-1FC "putative inputs" co-fire, yielding sparse yet reliable responses. Moreover, their responses depend on the count (N), not the identity, of co-active L4-1FC partners, with response sensitivity scaling as a power law in N. The corresponding L4-1FC and L2/3-1FC ensembles form information pathways, a view reinforced by their shared representational content. Our approach is inherently local and contrasts with prior methods that use regression or dimensionality reduction to model inter-areal communication, behavior, or stimulus encoding [48–51]. Recent advances have mapped anatomical connectivity with unprecedented resolution [52]. However, structural connectivity alone does not specify how functional ensembles interact across cortical layers. In contrast, our approach focuses on functional connectivity, capturing coordinated, time-varying interactions among specific neurons to reveal ensemble organization at single-cell resolution.

Deep Learning Spiking Model. To address our objective to study how spiking representations emerge across hierarchical processing, we selected the *convolutional ResNet18* that supports both

layer-wise analysis, biologically meaningful inductive biases, and practical considerations (e.g., sufficient depth and computational tractability). Moreover, the hierarchical stages of CNNs and ResNet have been extensively mapped onto the ventral visual stream in prior work[53, 54], further supporting the use of SEW-ResNet as a suitable model system for probing the development of spiking representations. We focused on the main residual branch through successive layers. We analyzed the spiking responses of a convolutional SNN trained on CIFAR100 [39], using the Spike-Element-Wise (SEW) ResNet[55]. The SEW-ResNet employs a Parametric Leaky Integrate-and-Fire (PLIF) neuron with a learnable membrane time constant [56], and is structured as 8 residual blocks each containing two convolutional spiking layers in the order of Convolution→Batch Normalization(BN)→Spiking Neuron (CONV→BN→PLIF), yielding 16 feature-extracting layers in total. Final representations are aggregated via adaptive average pooling across spatial dimensions, before being passed to a linear classification head. The resulting logits are averaged across the $T = 4$ timesteps (chosen as per [55]), and a single softmax and loss computation is performed on this averaged output to produce the final class prediction. As the architecture was originally designed for ImageNet, we adapt the stem convolution from a 7×7, 128-channel configuration to a 3×3, 64-channel kernel to match CIFAR100 inputs. Crucially, the network operates at a mean firing rate of ~ 0.072 spikes per sample (i.e. image) per neuron (Fig. A1) which is consistent with the sparse activity observed in the biological visual cortex and thus provides a biologically plausible regime in which to study sparse functional structure. The SEW-ResNet-18 comprises 11.2 million parameters and achieves 72.18% accuracy when trained and evaluated on CIFAR100 (Sec. A.1).

Image Presentation and Recorded Spike Trains. Each image is presented to the network $T = 4$ times within a single pass, both during training and inference stages; the network produces one classification output per image, with the temporal repetition allowing neurons to accumulate sufficient spiking activity for reliable encoding rather than generating independent predictions at each timestep. All results below are computed on a subset of neurons from each examined layer. Approximately 7,500 - 9000 neurons are sampled per layer, drawn across layers while restricting sampling to a central spatial region for each channel. In the final four convolutional spiking layers (corresponding to the last two blocks), each containing 8,192 neurons, all neurons are included in the analysis. For earlier layers, only neurons located in the central region are selected.

Spike Trains. Passing the full CIFAR100 test set (10,000 images) through the network yields, for each neuron, a binary spike train of length 40,000—-*four spike- or no-spike responses* per image—-with each neuron's activity arranged spatially as a tensor of shape (C_l, H_l, W_l) corresponding to channel, height, and width at layer l. These spike trains will be used in the pairwise correlation analysis for the identification of the 1FC groups. (These pairwise correlations are the equivalent signal correlations.) The CIFAR100 images, with no distortions, will be called *original input*. Throughout this paper, the membership of 1FC remains fixed, as computed under the reference testing set ("original input").

Pairwise Correlations. The pairwise correlations between neurons are calculated using the spike time tiling coefficient (STTC)[57], a symmetric measurement that is bounded in [-1, 1] (Sec. A.2). STTC was selected because it is largely independent of firing rate and performs favorably relative to thirty-three other commonly used correlation measures. STTC is defined with respect to a measurement window of synchronization Δt. Since in SNNs the exact response timing is known, we set $\Delta t = 0$.

First-Order Functional-Connectivity (1FC) Groups. To extend the analysis beyond pairwise correlations, a natural next step is to consider the graph formed by a node (neuron) and its *statistically significantly* correlated (functionally connected) neighbors at *one degree of separation*. We refer to this set as the neuron's *first-order functional connectivity (1FC)* group. The 1FC group thus represents the simplest multi-neuronal ensemble beyond a single neuron and therefore provides a principled starting point for higher-order analysis. We summarize the extent of each neuron's participation in functional interactions via its degree of connectivity (DoC), i.e., the size of its 1FC group. Here we focus on the 1FC group of a neuron at layer L, based on its *positive statistically significant functional connections* with neurons from the *previous* layer $(L - 1)$, in a ResNet architecture trained on CIFAR100. The 1FC groups identified at testing using the original dataset becomes our functional connectivity reference throughout the paper.

Response Function of a Neuron. We computed the response of a neuron as a function of the aggregate co-firing within its input 1FC group. We then examine how aggregate co-firing within an input 1FC group relates to the activity of its downstream "output" neuron. We characterize this response function and examine its behavior under different conditions. The membership of its 1FC

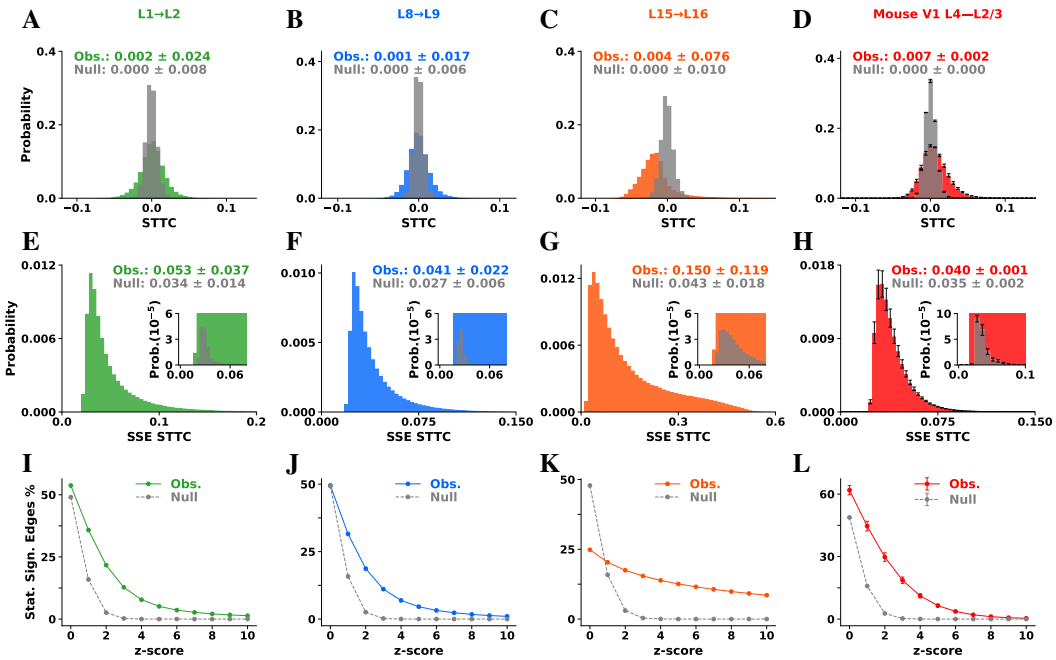

Figure 1: **Inter-Layer Functional Connectivity through Pairwise STTC**. Pairwise STTC was computed between the spiking activity of neurons in consecutive layer pairs during inference over the test set of CIFAR100 dataset ("original" data). Columns correspond to layer pairs L2→L3 (green), L9→L10 (blue), and L16→L17 (orange); The final column shows the same analysis in mouse primary visual cortex (V1), computed across cortical layers L4 to L2/3 and aggregated over 5 mice (red). For each panel, a separate null distribution (grey) was computed by randomly circular shifting the spiketrains 150 times. **(A–D)** Distributions of pairwise STTC values overlaid with null distributions derived from circularly shifted controls. **(E–H)** Distributions of statistically significant positively correlated edges, defined by a z-score threshold of $z > 4$. **(I-L)** Percentage of statistically significant edges as a function of z-score. For the mouse data, histograms reflect the distribution of mean values across mice, with error bars denoting SEM; z-scores for all data were computed using the mean and standard deviation of the circular-shift null distributions. The pairwise correlations are weak but statistically significant.

group remains fixed (as computed under the original test). The spike trains that will be used for estimating the response will be the ones recorded under the corresponding test input. Thus, the response of a neuron may change under different conditions (despite keeping the membership of its 1FC membership as the one identified during original input–no distortion).

Informativeness of a Neuron. To quantify the informativeness of individual neurons, we computed the mutual information (MI) between each neuron's spike train and the true class labels across the test set. MI was calculated between the spike counts (number of spikes out of a maximum possible 4) per image and both the categorical class labels and their one-hot encoded representations, the latter yielding a class-wise MI profile for each neuron. MI is defined as $I(X;Y) = H(X) - H(X|Y)$, where $H(X)$ denotes the Shannon entropy of the neuron's spiking activity and $H(X|Y)$ is the conditional entropy given the class label, such that MI reflects the reduction in uncertainty about a neuron's firing upon observation of the true class.

Distorted Image Presentation. For analysis under *additive noise*, each test image is corrupted with noise (Sec.A.8), prior to inference, and the resulting spiking activity is recorded and analyzed equivalently. For the *adversarial perturbation* condition specifically, analysis is conducted on a representative subset of 30 classes, owing to the computational overhead of per-sample gradient computation required by FGSM (Fig.A11). The spike trains produced under these image presentations will be used for computing the response of the neurons under distorted image presentation, keeping however the membership of their 1FC groups that was computed on the original image set.

Exclusion Criteria. Neurons with a mean firing rate below 0.01 spikes per sample are designated as *silent* and are excluded from all analyses ($\sim 9.5\%$ excluded per layer). Of the sampled neurons per layer, approximately 90.5% meet the activity threshold and are designated as *active* neurons; all subsequent analyses are conducted exclusively on this active population (See Fig. A1).

4 Results

Statistically significant but weak pairwise correlations. We first analyze the functional connectivity of the model on the original test set. We observed significant functional coupling between neurons of each layer of the network as well as across source-target layer pairs (Figs 1). Observed pairwise STTC values extended well-beyond their corresponding null distributions in both positive and negative directions, with a stronger positive tail (Fig. 1, A-C), with the structural edges contributing to the tails (Fig. A7 A-F) at the first layers. At a conservative threshold of z-score>4, approximately 7.8% of neuronal pairs from L1→L2 exhibit significant functional correlations compared to $\sim 7\%$ in L8→L9 and $\sim 13.9\%$ edges from L15→L16 (Fig.1 I-L).

Response of a Neuron as a Function of its input 1FC Cofiring well-fitted by ReLU. We examined how the activity of a downstream "output" neuron at layer L relates to the aggregate co-firing of its input $(L-1)$ 1FC group. Across the population, the firing probability of "output" neurons was well described by rectified linear unit (ReLU)-like input–output relationships as a function of their corresponding input-1FC group co-firing. Notably, some neurons reached firing probabilities approaching unity (Fig. 2). This functional form is of particular interest, as ReLU nonlinearities confer key computational advantages—including sparsity, efficient learning, and representational efficiency—in both biological and artificial neural systems [58, 59]. We therefore focused on neurons whose responses were well fit by a ReLU model ($R^2 \geq 0.8$). The slope of the response functions scaled inversely with the size of the corresponding input-1FC groups: neurons associated with smaller groups exhibited steeper input–output functions, whereas those linked to larger groups showed shallower slopes. We found that when cofiring was rescaled by N^{α}, where N is input 1FC group size, and $\alpha \sim 0.6 - 1$ depending the depth, the response functions collapsed onto a common curve (A3).

Two Regimes in the Response. We also observed the existence of weak and strong response regimes: Figures 2 M shows a neuron whose response (green) is plotted alongside the cumulative probability distribution of cofiring events within its input-1FC group. The firing probability clearly begins to increase only for events located deep in the tail of the cofiring event size distribution. In general, neurons remain inactive for most cofiring events in their 1FC groups, with their firing probability *rising sharply* only once the event size exceeds approximately 8% of the largest and *rarest* events. This threshold corresponds to at least 5-16% of 1FC neurons firing *synchronously* (Fig. 2 Q-S inset).

Representational Significance. To assess the representational significance of the first-order functional connectivity modules in encoding the image being presented, we plot the similarity between the class of the image being presented at a frame with the preferred class of the "index" neuron as a function of the number of its input 1FC cofiring events (y-axis). The larger the number of cofiring, the higher the similarity. That is, at the strong response regime, the neurons of the deeper layer encode more reliably the image being presented. As we ascend the layers, neurons exhibit higher MI for a class, consistent with the improved class discrimination trend in CNNs[60]. Especially in last layer (L16), the higher the degree of connectivity, the larger the MI. A cluster of hub neurons with high degree of connectivity and MI is prominent (Fig. A6). To further examine the relationship between the index neuron firing at a given layer (say layer L) and the cofiring events of its layer $L-1$ 1FC group, we applied several classifiers to predict on a frame by frame basis, whether the neuron fires from the cofiring events that occurred in the preceding layer's 1FC group. Preserving the identities of individual neurons within input–1FC groups does not provide significant additional predictive value beyond the aggregate co-firing signal when modeling the activity of the corresponding output neuron (see A.6, Fig. A9). This suggest that ensemble co-firing magnitude may play a more prominent role in shaping downstream responses than the specific identities of the contributing neurons.

Temporal Robustness of the Response. To verify that the 1FC groups are not an artifact of the specific images used to estimate functional connectivity, we assess the temporal robustness of response functions across held-out input samples. The consistency of response function slopes across batches and across the full range of input images confirm that the 1FC groups reflect stable functional structure rather than sample-specific co-activation (Figs. A4, A5, Sec. A.3).

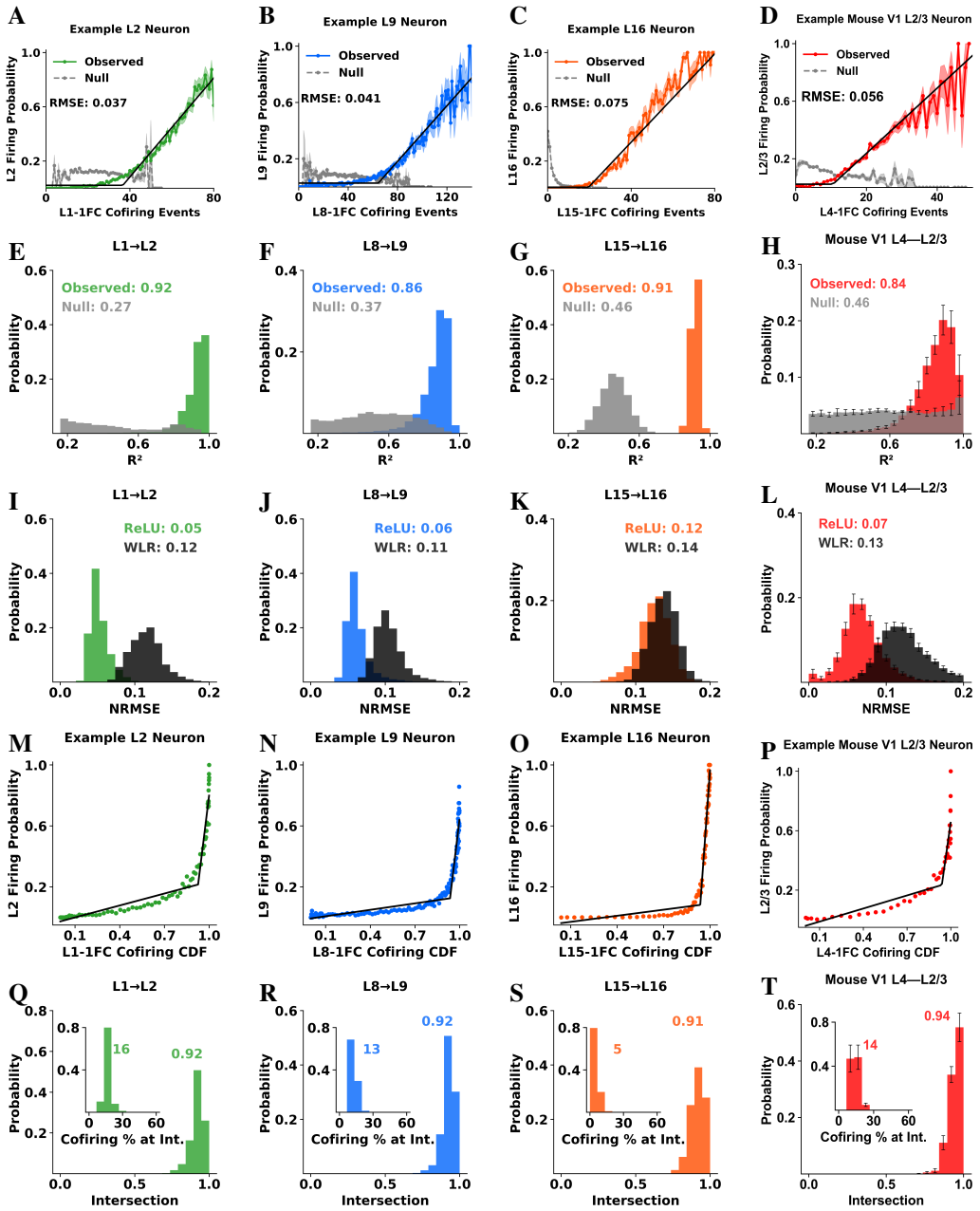

Figure 2: **Response of a Neuron as a Function of the Number of Cofiring of its Input 1FC group.** Columns 1–3 show results from the SNN (for layer pairs L1→L2, L8→L9, L15→L16), and column 4 shows corresponding analyses from mouse primary visual cortex (V1). **(A–D)** Example response functions of single output neurons, showing firing probability as a function of the number of cofiring events in their respective input-layer 1FC groups. **(E–H)** R^2 scores of the ReLU fits (black lines in A–D). **(I–L)** Normalized root mean square error (NRMSE) comparing linear (black) and ReLU (red) fits. **(M–P)** Firing probability as a function of the cumulative distribution (CDF) of group cofirings, highlighting sensitivity to rare, high-cofiring events. **(Q–T)** Cofiring percentages at the intersection points defining the transition to the fast-rising response regime, with insets showing the corresponding distributions of input-layer 1FC cofiring percentages. For the mouse data (column 4), error bars in **(H)**, **(L)**, **(P)**, and **(T)** indicate SEM across $n = 5$ mice. On average, the transition to the fast-rising regime occurs at CDF ≈ 0.92, corresponding to $\sim 33\%$ of the input group firing together. Analyses include only neurons with firing probability ≥ 0.6 and ReLU fit $R^2 \geq 0.8$; neurons with response-region slope < 1 were excluded ($\sim 58 \pm 13\%$ across mice). The slope was estimated as the change in response probability between two cofiring levels divided by the change in their corresponding ECDF values. The slope histograms are presented in Fig.4A.

Properties of a Neuron. We then characterize a neuron based on the main properties of its response on a certain test set, namely the *slope* of the ReLU fit, the 1FC-cofiring % at the intersection of the two response regimes, the response rise intersection (CDF of the 1FC-cofiring), and the semantic rank similarity. For each neuron, we also compute the membership and size of its 1FC group (i.e., degree of connectivity) and its informativeness with respect to the original test set.

Robustness of Neuronal Response Functions Under Additive Uniform Noise. Here we analyze the response of the neurons on the test set with random noise (Sec. A.8). For each neuron, we consider its 1FC group as formed during the original test set. All neurons exhibited lower firing rates under noise input (Fig. A12), L15-1FC groups of L16 neurons exhibited infrequent co-firing events (Fig. 5C) and had large deviations in their response slopes (Fig. 4C).

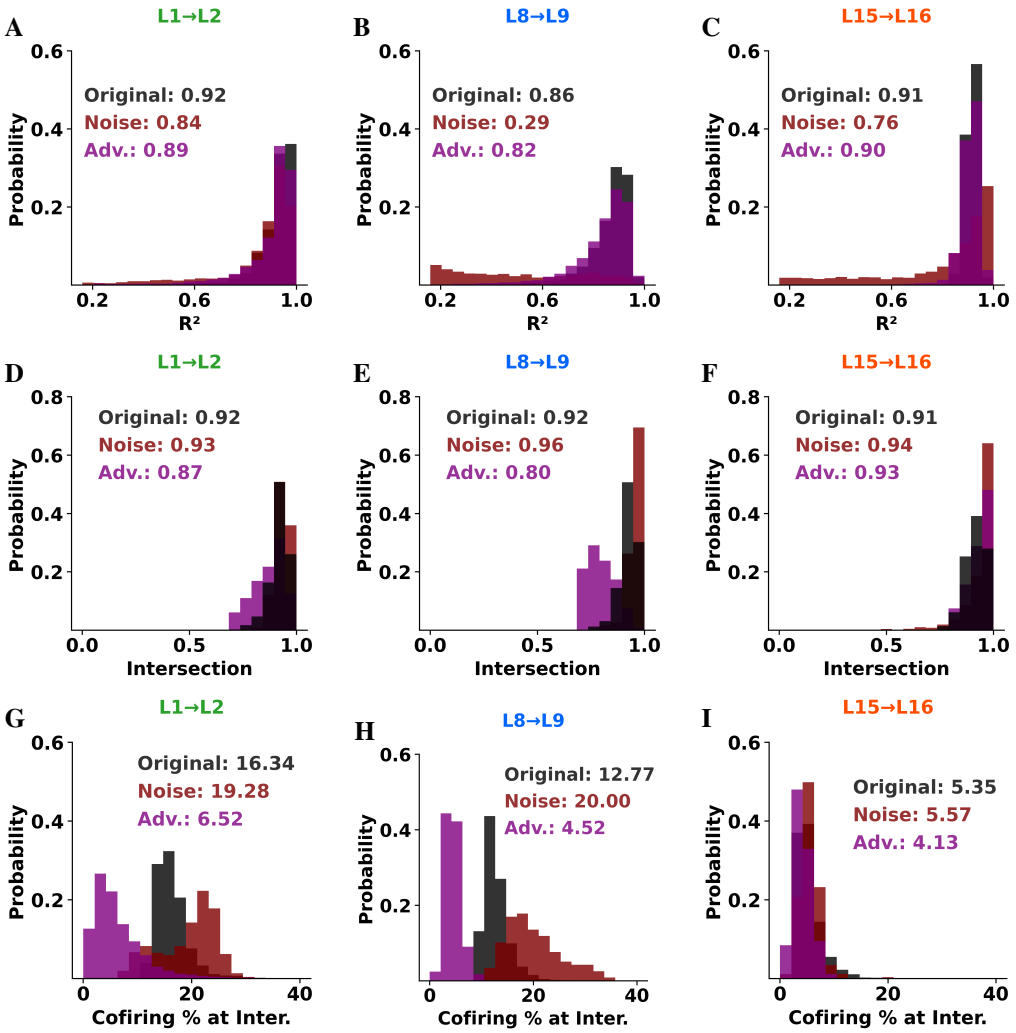

Figure 3: **Response characteristics as they change under different conditions**. **(A-C)** Histograms of R^2 values resulting from ReLU fits of neuronal responses as a function of their respective input-layer 1FC group cofirings, for layer pairs L1→L2, L8→L9, L15→L16 for *original* input (black), *random noise* (maroon), and *adversarial* input (purple). **(D-F)** Histograms of the intersection points corresponding to the transition to the fast-rising response regime for the same three layer pairs. **(G-I)** Distributions of input-layer 1FC group cofiring percentages at these intersection points. On average, under the original input, this transition occurs at CDF ≈ 0.92, corresponding to $\sim 16\%$ of the 1FC cofiring. Under noise and adversarial conditions, these histograms of the first and intermediate layers change significantly. Neurons with $R^2 < 0.8$ or firing rate < 0.01 Hz were excluded.

Robustness of the Response under Adversarial Noise. To probe the sensitivity of individual neuronal response functions under targeted input perturbation (Sec. A.8), we applied the Fast Gradient Sign Method (FGSM) [40] adversarial attack to the test set. FGSM generates adversarial examples by perturbing the input in the direction of the gradient of the loss with respect to the input. Under adversarial input, individual neuronal response functions preserved qualitatively similar trends to those observed under clean input, suggesting a degree of structural robustness in the population-level response organization. To quantify this at the single-neuron level, we computed the difference in the slopes of the response functions between adversarial and original inputs, and examined this quantity as a function of each neuron's mutual information with respect to the output class (Fig. A13). Neurons exhibiting high MI showed minimal deviation in response function slope between original and adversarial conditions across all layers, suggesting that informationally relevant neurons maintain stable tuning under perturbation. In contrast, neurons with low MI displayed substantially greater variability in slope differences, suggesting that weakly informative neurons are more susceptible to adversarial perturbation, with this susceptibility attenuating towards deeper layers. In the presence of adversarial perturbations, L16 neurons with high MI remain relatively informative (see Figs. 5B, A10 B). Notably these L16 high-MI neurons also exhibit a high degree of connectivity (Fig. A6), which allows them to integrate input from multiple upstream nodes. This dense connectivity may confer a degree of robustness, enabling these neurons to partially mitigate the impact of adversarial perturbations through aggregation and implicit noise averaging. Nevertheless, overall model performance (Fig. A11A) is still degraded due to the propagation of noisy information from nodes with smaller degree of connectivity to the output layer. The adversarial input yields images with imperceptible degradation, whereas images with additive random noise input introduce substantial visual corruption Fig. A11B, noticable in the response function (e.g., Fig.5B, A14, A16).

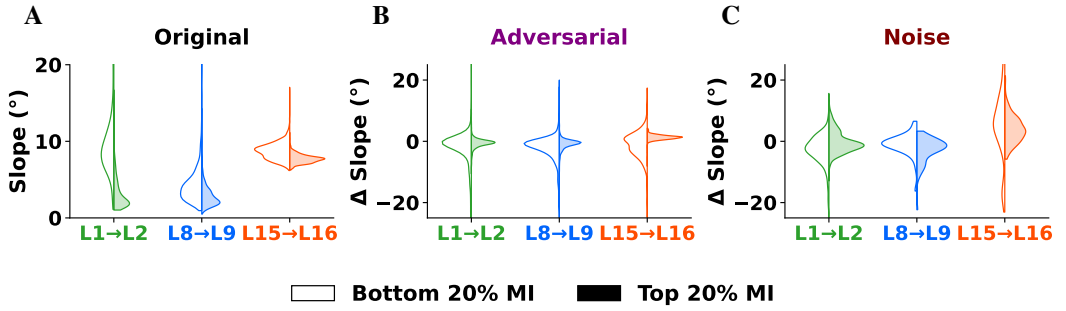

Figure 4: **Slopes of response functions under different input conditions (A)** Violin plots of neuronal response slopes for layer pairs L1→L2, L8→L9, and L15→L16. For each violin, the left half corresponds to neurons with MI below the 20th percentile, and the right half to neurons with MI above the 80th percentile. **(B)** Same as **(A)**, but showing the *difference* in slopes between *adversarial* and original test, respectively. **(C)** Same as **(B)** but on the *difference* in slopes between *uniform noise* and original test.(Fig. A13)

Diagnostics: Detection of Adversarial Frames We developed a diagnostic framework to detect adversarial perturbations using cofiring patterns of 1FC groups in ResNet output layers. 1FC groups were identified across layers L1–L16 from 10,000 training examples of CIFAR100. Next, spike trains (T=4, length 40,000) were generated for clean and adversarial inputs from the full CIFAR-100 test set. Finally, a logistic regression classifier was trained on the resulting cofiring statistics to distinguish adversarial from clean inputs.

5 Discussion, Conclusions, and Future Work

Our results support the view that first-order functionally connected (1FC) ensembles act as computational modules that mediate information processing across hierarchical layers. In a convolutional architecture, neurons receiving common input through the same convolutional kernel may exhibit correlated activity by virtue of its input and structural organization (Fig. A2). To contrast direct vs. indirect connections, we partition each neuron's 1FC group into **i)** the *structural* subset (*STR*), comprising source-layer neurons whose activations are routed to the reference neuron via the convolutional kernel, representing (*direct* connections); and **ii)** the subset (*nSTR*) of *all remaining* 1FC members, whose co-activation with the reference neuron cannot be attributed to direct information

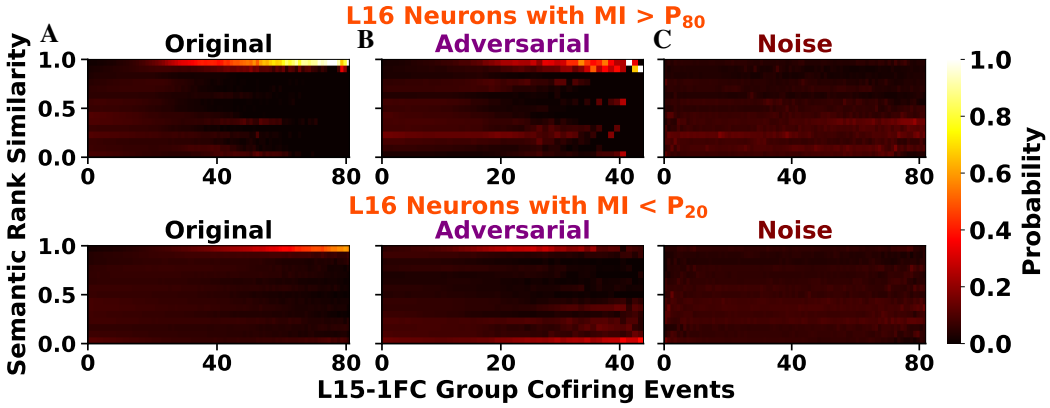

Figure 5: **Information encoding of neurons as a function of their 1FC cofiring. (A-Top)** Heatmap of semantic rank similarity for L16 neurons in the top 20th percentile of class-wise mutual information (MI), shown as a function of L15-1FC group co-firing events under original input. Each bin indicates the probability that co-firing frames match the neuron's preferred class. **(A)** The larger the L15-1FC cofiring, the higher the semantic similarity of L16-neurons. **(B-C Top)** Same as (A) under adversarial and random noise input, respectively, with L15-1FC groups fixed (formed under original input). High-MI neurons remain relatively informative under adversarial perturbations. Their higher degree of connectivity (Fig. A6) may support robustness via input aggregation. **(A-C Bottom)** Same analysis for neurons in the lowest 20th percentile of MI. For each co-firing bin, probabilities are normalized to sum to 1 across semantic rank similarity (Fig. A10).

flow and must therefore reflect feature-level correlations learned during training (in part via indirect structural connections). For each neuron, we estimate all its statistically significant functional connections with all neurons in the previous layer (Fig. A7). As expected, the first layer (L1) is driven by the spatial locality of the input, which shapes the structural connections, resulting to higher STTC weights and larger normalized degree of connectivity. Interestingly, at the last layer (L15→L16), the pairwise correlations exhibit similar structure (Figs. A7 C,F, I). Note that structural connectivity alone does not capture the complete view of the coordinated cofiring at a *wider level* across the hierarchical processing pipeline, which the proposed functional connectivity analysis reflects (Fig. A8).

The 1FC ensembles display interesting properties: their aggregate cofiring reliably predicts downstream neuronal responses through a robust, ReLU-like input–output relationship, whose gain scales systematically with ensemble size. Notably, neuronal responses are characterized by two regimes—a largely inactive or weak-response regime and a strong-response regime. Reliable encoding of the presented class emerges only during high 1FC cofiring events, which themselves occur infrequently, indicating that informative representations are concentrated in rare but highly coordinated activity patterns. Under uniform random noise or adversarial perturbations, keeping fixed the 1FC groups identified during the original testing, the response regimes of these neurons are disrupted, particularly in early and intermediate layers, with downstream degradation in class-specific encoding at later stages (Figs. A14,A16,A15).

We also examined the relationship between what has been learned and functional connectivity by permuting the weights across edges within each layer and reapplying the functional connectivity analysis (See Fig. A17). While the overall distribution of STTC weights and the proportion of statistically significant connections remain largely unchanged in early and intermediate layers, the characteristic response properties are lost. Moreover, deeper layers fail to form coherent 1FC groups altogether, underscoring the importance of functional connectivity for the emergence of meaningful ensembles. This establishes 1FC ensembles as a functionally meaningful substrate for input encoding and information transfer. The response properties of *targeted* neurons, selected based on their mutual information with their preferred class and 1FC size, at the onset of a strong-response regime, can be used to provide alerts about the information flow. This framework thus, establishes a conceptual foundation for conducting diagnostics on the information transfer.

Acknowledgments and Disclosure of Funding

This work has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 101007926 as well as from the Hellenic Foundation Research Institute (HFRI) with the neuron-AD project number 04058 and neuronXnet project number 2285 (PI: Maria Papadopouli). It has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program. Finally, this research was also supported by R01 NS113890, and R21 NS127299 (PI: Stelios Smirnakis). AWS resources were provided by the National Infrastructures for Research and Technology GRNET and funded by the EU Recovery and Resiliency Facility.

References

[1] Jamie Simon, Daniel Kunin, Alexander Atanasov, Enric Boix-Adserà, Blake Bordelon, Jeremy Cohen, Nikhil Ghosh, Florentin Guth, Arthur Jacot, Mason Kamb, et al. There will be a scientific theory of deep learning. *arXiv preprint arXiv:2604.21691*, 2026.

[2] Minhyuk Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. In *International Conference on Machine Learning (ICML)*, 2024.

[3] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[4] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, 2019.

[5] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8): 1798–1828, 2013.

[6] Nó R de Lorente. Cerebral cortex: architecture, intracortical connections, motor projections. *Physiology of the nervous system*, pages 288–330, 1949.

[7] Markus Meister, Rachel O. L. Wong, Denis A. Baylor, and Carla J. Shatz. Synchronous bursts of action potentials in ganglion cells of the developing mammalian retina. *Science*, 252(5008):939–943, May 1991. ISSN 1095-9203. doi: 10.1126/science.2035024. URL `http://dx.doi.org/10.1126/science.2035024`.

[8] M. Abeles, H. Bergman, E. Margalit, and E. Vaadia. Spatiotemporal firing patterns in the frontal cortex of behaving monkeys. *Journal of Neurophysiology*, 70(4):1629–1638, October 1993. doi: 10.1152/jn.1993.70.4.1629. URL `https://doi.org/10.1152/jn.1993.70.4.1629`.

[9] Amiram Grinvald, Amos Arieli, Misha Tsodyks, and Tal Kenet. Neuronal assemblies: Single cortical neurons are obedient members of a huge orchestra. *Biopolymers*, 68(3):422–436, February 2003. doi: 10.1002/bip.10273. URL `https://doi.org/10.1002/bip.10273`.

[10] Donald O Hebb. The first stage of perception: growth of the assembly. *The Organization of Behavior*, 4(60):78–60, 1949.

[11] Dario L Ringach. Spontaneous and driven cortical activity: implications for computation. *Curr Opin Neurobiol*, 19(4):439–444, August 2009.

[12] Luis Carrillo-Reid, Weijian Yang, Yuki Bando, Darcy S. Peterka, and Rafael Yuste. Imprinting and recalling cortical ensembles. *Science*, 353(6300):691–694, August 2016. doi: 10.1126/science.aaf7560. URL `https://doi.org/10.1126/science.aaf7560`.

[13] Stefano Panzeri, Monica Moroni, Houman Safaai, and Christopher D Harvey. The structures and functions of correlations in neural population codes. *Nature Reviews Neuroscience*, 23(9), 2022.

[14] Bruno B Averbeck, Peter E Latham, and Alexandre Pouget. Neural correlations, population coding and computation. *Nature reviews neuroscience*, 7(5), 2006.

[15] R. Yuste. From the neuron doctrine to neural networks. *Nature Reviews Neuroscience*, 2015. URL `https://doi.org/10.1038/nrn3962`.

[16] Rosa Cossart, Dmitriy Aronov, and Rafael Yuste. Attractor dynamics of network up states in the neocortex. *Nature*, 423(6937):283–288, May 2003. ISSN 1476-4687. doi: 10.1038/nature01614. URL `http://dx.doi.org/10.1038/nature01614`.

[17] Jae-eun Kang Miller, Inbal Ayzenshtat, Luis Carrillo-Reid, and Rafael Yuste. Visual stimuli recruit intrinsically generated cortical ensembles. *Proceedings of the National Academy of Sciences of the United States of America*, 111(38):E4053—61, September 2014. ISSN 0027-8424. doi: 10.1073/pnas.1406077111. URL `https://europepmc.org/articles/PMC4183303`.

[18] Eleonora Russo and Daniel Durstewitz. Cell assemblies at multiple time scales with arbitrary lag constellations. *Elife*, 6:e19428, 2017.

[19] Rafael Yuste, Rosa Cossart, and Emre Yaksi. Neuronal ensembles: Building blocks of neural circuits. *Neuron*, 2024. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2023.12.008. URL `https://www.sciencedirect.com/science/article/pii/S0896627323009674`.

[20] Tal Kenet, Dmitri Bibitchkov, Misha Tsodyks, Amiram Grinvald, and Amos Arieli. Spontaneously emerging cortical representations of visual attributes. *Nature*, 425(6961):954–956, Oct 2003. ISSN 1476-4687. doi: 10.1038/nature02078. URL `https://doi.org/10.1038/nature02078`.

[21] Artur Luczak, Peter Barthó, and Kenneth D. Harris. Spontaneous events outline the realm of possible sensory responses in neocortical populations. *Neuron*, 62(3):413–425, May 2009. doi: 10.1016/j.neuron.2009.03.014. URL `https://doi.org/10.1016/j.neuron.2009.03.014`.

[22] Ganna Palagina, Jochen F. Meyer, and Stelios M. Smirnakis. Inhibitory units: An organizing nidus for feature-selective sub-networks in area v1. *bioRxiv*, 2018. doi: 10.1101/282988. URL `https://www.biorxiv.org/content/early/2018/03/15/282988`.

[23] Maria Papadopouli, Ioannis Smyrnakis, Emmanouil Koniotakis, Mario-Alexios Savaglio, Christina Brozi, Eleftheria Psilou, Ganna Palagina, and Stelios Manolis Smirnakis. Brain orchestra under spontaneous conditions: Identifying communication modules from the functional architecture of area v1. *bioRxiv*, 2024. doi: 10.1101/2024.02.29.582364. URL `https://www.biorxiv.org/content/early/2024/04/25/2024.02.29.582364`.

[24] Mario Alexios Savaglio, Christina Brozi, Eleftheria Psilou, Chryssalenia Koumpouzi, Marianna Papadostefanaki, Christos Vasilakos, Spyros Nessis, Emmanouel Smyrnakis, Vasileios Akoutas, Georgios A. Keliris, Ganna Palagina, Stelios M. Smirnakis, and Maria Papadopouli. Direction of motion decoding in mouse v1: Neuron predictive power relates to functional connectivity organization. In *International Joint Conference on Neural Networks (IJCNN)*, 2025.

[25] Mario Alexios Savaglio, Christina Brozi, Stelios M. Smirnakis, and Maria Papadopouli. On temporal robustness & brain-state stability of functional connectivity in mouse primary visual area v1 compared to higher visual area al. In *IEEE 25th International Conference on Bioinformatics and Bioengineering (BIBE) Best Paper Award*, 2025.

[26] Mackenzie Weygandt Mathis, Adriana Perez Rotondo, Edward F Chang, Andreas S Tolias, and Alexander Mathis. Decoding the brain: From neural representations to mechanistic models. *Cell*, 187(21):5814–5832, 2024.

[27] Amirhossein Tavanaei, Masoud Ghodrati, Saeed Reza Kheradpisheh, Timothée Masquelier, and Anthony Maida. Deep learning in spiking neural networks. *Neural networks*, 111:47–63, 2019.

[28] Reduan Achtibat, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. From attribution maps to human-understandable explanations through concept relevance propagation. *Nature machine intelligence*, 5(9):1006–1019, 2023.

[29] Kedar Dhamdhere, Mukund Sundararajan, and Qiqi Yan. How important is a neuron. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SylKoo0cKm.

[30] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 117(48):30071–30078, 2020. doi: 10.1073/pnas.1907375117. URL https://www.pnas.org/doi/abs/10.1073/pnas.1907375117.

[31] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 2018. doi: 10.23915/distill.00010. https://distill.pub/2018/building-blocks.

[32] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001. https://distill.pub/2020/circuits/zoom-in.

[33] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 36, 2022. arXiv:2202.05262.

[34] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. *CoRR*, abs/1512.04150, 2015. URL http://arxiv.org/abs/1512.04150.

[35] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.*, 128(2):336–359, February 2020.

[36] Saurabh Desai and Harish G Ramaswamy. Ablation-CAM: Visual explanations for deep convolutional network via gradient-free localization. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, March 2020.

[37] Jeffrey S. Bowers, Gaurav Malhotra, Federico Adolfi, Marin Dujmović, Milton L. Montero, Valerio Biscione, Guillermo Puebla, John H. Hummel, and Rachel F. Heaton. On the importance of severely testing deep learning models of cognition. *Cognitive Systems Research*, 82:101158, 2023. doi: 10.1016/j.cogsys.2023.101158.

[38] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[39] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://api.semanticscholar.org/CorpusID:18268744.

[40] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2014. URL https://api.semanticscholar.org/CorpusID:6706414.

[41] Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Comput. Surv.*, 57(9):1–52, September 2025.

[42] Nicholas Goldowsky-Dill, Christopher Macleod, Lucas Jun Koba Sato, and Aryaman Arora. Localizing model behavior with path patching. *ArXiv*, abs/2304.05969, 2023. URL https://api.semanticscholar.org/CorpusID:258079237.

[43] Shriyank Somvanshi, Md Monzurul Islam, Amir Rafe, Anannya Ghosh Tusti, Arka Chakraborty, Anika Baitullah, Tausif Islam Chowdhury, Nawaf Alnawmasi, Anandi Dutta, and Subasish Das. Bridging the black box: A survey on mechanistic interpretability in ai. *ACM Comput. Surv.*, 58 (8), February 2026. ISSN 0360-0300. doi: 10.1145/3787104. URL https://doi.org/10.1145/3787104.

[44] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *ArXiv*, abs/1610.01644, 2016. URL https://api.semanticscholar.org/CorpusID:9794990.

[45] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6076–6085. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7188-svcca-singular-vector-canonical-correlation-analysis-for-deep-learning-dynamics-and-int pdf.

[46] Yulong Wang, Hang Su, Bo Zhang, and Xiaolin Hu. Interpret neural networks by identifying critical data routing paths. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8906–8914. IEEE, June 2018.

[47] Kim Youngeun and Panda Priyadarshini. Visual explanations from spiking neural networks using inter-spike intervals. *Sci. Rep.*, 11(1):19037, September 2021.

[48] Adam Kohn, Ruben Coen-Cagli, Ingmar Kanitscheider, and Alexandre Pouget. Correlations and neuronal population information. *Annual review of neuroscience*, 39, 2016.

[49] Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Charu Bai Reddy, Matteo Carandini, and Kenneth D. Harris. Spontaneous behaviors drive multidimensional, brainwide activity. *Science*, 364(6437), April 2019. doi: 10.1126/science.aav7893. URL https://doi.org/10.1126/science.aav7893.

[50] João D Semedo, Amin Zandvakili, Christian K Machens, Byron M Yu, and Adam Kohn. Cortical areas interact through a communication subspace. *Neuron*, 102(1):249–259.e4, February 2019.

[51] Jermyn Z See, Craig A Atencio, Vikaas S Sohal, and Christoph E Schreiner. Coordinated neuronal ensembles in primary auditory cortical columns. *eLife*, 7:e35587, jun 2018. ISSN 2050-084X. doi: 10.7554/eLife.35587. URL https://doi.org/10.7554/eLife.35587.

[52] J. Alexander Bae, Mahaly Baptiste, Maya R. Baptiste, Caitlyn A. Bishop, Agnes L. Bodor, Derrick Brittain, Victoria Brooks, JoAnn Buchanan, Daniel J. Bumbarger, Manuel A. Castro, Brendan Celii, Erick Cobos, Forrest Collman, Nuno Maçarico da Costa, Bethanny Danskin, Sven Dorkenwald, Leila Elabbady, Paul G. Fahey, Tim Fliss, Emmanouil Froudarakis, Jay Gager, Clare Gamlin, William Gray-Roncal, Akhilesh Halageri, James Hebditch, Zhen Jia, Emily Joyce, Justin Ellis-Joyce, Chris Jordan, Daniel Kapner, Nico Kemnitz, Sam Kinn, Lindsey M. Kitchell, Selden Koolman, Kai Kuehner, Kisuk Lee, Kai Li, Ran Lu, Thomas Macrina, Gayathri Mahalingam, Jordan Matelsky, Sarah McReynolds, Elanine Miranda, Eric Mitchell, Shanka Subhra Mondal, Merlin Moore, Shang Mu, Taliah Muhammad, Barak Nehoran, Erika Neace, Oluwaseun Ogedengbe, Christos Papadopoulos, Stelios Papadopoulos, Saumil Patel, Guadalupe Jovita Yasmin Perez Vega, Xaq Pitkow, Sergiy Popovych, Anthony Ramos, R. Clay Reid, Jacob Reimer, Patricia K. Rivlin, Victoria Rose, Zachary M. Sauter, Casey M. Schneider-Mizell, H. Sebastian Seung, Ben Silverman, William Silversmith, Amy Sterling, Fabian H. Sinz, Cameron L. Smith, Rachael Swanstrom, Shelby Suckow, Marc Takeno, Zheng H. Tan, Andreas S. Tolias, Russel Torres, Nicholas L. Turner, Edgar Y. Walker, Tianyu Wang, Adrian Wanner, Brock A. Wester, Grace Williams, Sarah Williams, Kyle Willie, Ryan Willie, William Wong, Jingpeng Wu, Chris Xu, Runzhe Yang, Dimitri Yatsenko, Fei Ye, Wenjing Yin, Rob Young, Szi-chieh Yu, Daniel Xenes, Chi Zhang, and The MICrONS Consortium. Functional connectomics spanning multiple areas of mouse visual cortex. *Nature*, 640(8058): 435–447, Apr 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-08790-w. URL https://doi.org/10.1038/s41586-025-08790-w.

[53] Daniel L K Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.*, 19(3):356–365, March 2016.

[54] Haiguang Wen, Junxing Shi, Wei Chen, and Zhongming Liu. Deep residual network predicts cortical representation and organization of visual features for rapid categorization. *Sci. Rep.*, 8 (1):3752, February 2018.

[55] Wei Fang, Zhaofei Yu, Yanqi Chen, Tiejun Huang, Timothée Masquelier, and Yonghong Tian. Deep residual learning in spiking neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 21056–21069. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/afe434653a898da20044041262b3ac74-Paper.pdf.

[56] Wei Fang, Zhaofei Yu, Yanqi Chen, Timothee Masquelier, Tiejun Huang, and Yonghong Tian. Incorporating learnable membrane time constant to enhance learning of spiking neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2661–2671, October 2021.

[57] Catherine S. Cutts and Stephen J. Eglen. Detecting pairwise correlations in spike trains: An objective comparison of methods and application to the study of retinal waves. *Journal of Neuroscience*, 34(43):14288–14303, 2014. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.2767-14.2014. URL https://www.jneurosci.org/content/34/43/14288.

[58] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[59] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 315–323, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL https://proceedings.mlr.press/v15/glorot11a.html.

[60] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer International Publishing.

[61] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. URL https://arxiv.org/abs/2103.00020.

[62] Aditi Aravind. Sttcpy: Efficient spike time tiling coefficient calculations for large-scale spiking datasets, 2024. URL https://github.com/aditiaravind/STTCPy.

A Technical appendices and supplementary material

A.1 SNN Architecture and Training Details

From a practical standpoint, hierarchical representation learning necessitates sufficient network depth—something shallow or linear models cannot provide without a significant loss in classification performance. ResNet-18 offers a well-characterized hierarchy that is deep enough to enable systematic layer-by-layer analysis, while remaining computationally tractable. From a theoretical perspective, convolutional architectures embody an inductive bias—fixed local connectivity combined with learned weight sharing—that serves as a structural analogue to the organization of the visual cortex. In biological systems, feedforward anatomical wiring gives rise to functionally distinct cell types. This parallel between imposed structural connectivity and emergent functional organization is central to our analysis. Moreover, the hierarchical stages of CNNs and ResNet have been extensively mapped onto the ventral visual stream in prior work[53, 54], further supporting the use of SEW-ResNet as a suitable model system for probing the development of spiking representations. Our analysis focuses on the residual stream, tracing the direct flow of information through successive layers.

Training Details : The SEW-ResNet-18 was trained on CIFAR-100, which comprises 60,000 32×32 color images across 100 classes, partitioned into 50,000 training and 10,000 test images. The stem convolution was adapted from a 7×7, 128-channel configuration to a 3×3, 64-channel kernel to suit CIFAR-100 input dimensions. The feature of SEW ResNet is that the dropout layers also have the same structure, where if the stride of a layer is more than 1, the residual connections have the block structure (CONV→BN→PLIF) to downsample the input. Training was conducted for 320 epochs with a batch size of 256 images, using the SGD optimizer with a learning rate of 5×10^{-4}. Gradients were propagated through the non-differentiable spiking nonlinearity via surrogate gradient backpropagation, using the sigmoid surrogate function. The model checkpoint from epoch 314 was selected as the final model on the basis of best test accuracy, achieving 72.18% on the CIFAR-100 test set. Each image was presented across $T = 4$ timesteps during both training and inference. The following data augmentation strategies were applied during training: random cropping with padding of 4 pixels, random horizontal flipping, and Gaussian blur (kernel size 3×3, $\sigma \in [0.05, 0.15]$,) applied with a probability of 0.5.

Neuron Sampling : For each convolutional spiking layer, spike trains were recorded during inference over the full CIFAR-100 test set. Rather than recording all neurons, we sample a spatially contiguous column of approximately 8,000 neurons per layer to balance computational tractability with representational coverage. Specifically, for a layer with shape (C, H, W), a square spatial patch of side length $s \approx \lfloor \sqrt{8000/C}$ is extracted from the center of each feature map, where s is clipped to not exceed H or W. This yields a set of $C \times s \times s$ neurons, uniformly covering all channels while remaining spatially localized to the central region of each feature map. In the final four convolutional spiking layers, where the spatial dimensions are sufficiently small, all 8192 neurons are included. The recorded spike trains from this sampled population form the basis for all functional connectivity, mutual information, and robustness analyses reported in this paper.

Exclusion Criteria. Neurons with a mean firing rate below 0.01 spikes per sample are designated as *silent* and excluded from all analyses ($\sim 9.5\%$ excluded per layer). This firing-rate threshold was chosen to ensure sufficient spiking activity for reliable estimation of functional connectivity, response functions, and information-theoretic measures. Of the sampled neurons per layer, approximately 90.5% meet this activity criterion and are designated as *active* neurons; all subsequent analyses are conducted exclusively on this active population. For analyses involving response-function fitting, additional R^2 thresholds were applied where indicated in the corresponding figures. These thresholds were selected to ensure reliable characterization of neuronal response regimes and were chosen to remain consistent with prior analyses performed in mouse visual cortex datasets.

A.2 STTC

We use the spike time tiling coefficient (STTC), introduced by Cutts and Eglen [57], which performs favorably against 33 other commonly-used measures and is less sensitive to firing rate. The STTC of neuron A relative to B is defined as:

$$\mathrm{STTC}(\Delta t) = \frac{1}{2}\left(\frac{P_A - T_B}{1 - P_A T_B} + \frac{P_B - T_A}{1 - P_B T_A}\right) \tag{1}$$

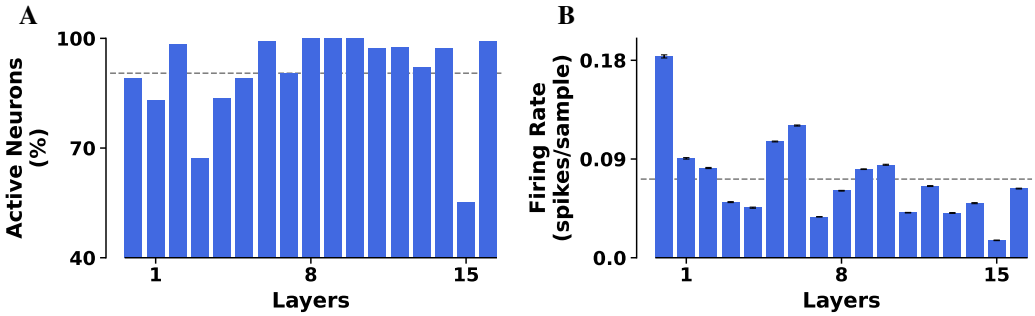

Figure A1: **Neuronal activity per layer. (A)** Percentage of active neurons (neurons having firing rate > 0.01 spikes per sample) among all sampled neurons across layers. The grey dashed line indicates the mean activity level across layers. **(B)** Firing rate of active neurons, defined as the proportion of spikes per input sample for each layer. The grey dashed line indicates the mean firing rate across layers (~ 0.072 spikes per sample). Approximately 9.5% of neurons (out of ~ 8300 per layer) are excluded due to inactivity. We observe that neuronal responses are inherently sparse, despite no explicit sparsity constraint during training.

where P_A is the proportion of firing events of neuron A found within an interval ($\pm\Delta t$) around each firing event of neuron B, T_B is the proportion of the total observed spikes that falls within an interval $\pm\Delta t$ around each firing event of neuron B, and likewise for T_A and P_B. STTC is a symmetric measure and is normalized to lie within $[-1, 1]$. STTC is defined with respect to a measurement window of synchronization Δt. Since we use an artificial SNN where we know the exact response timing, we set Δt= 0.

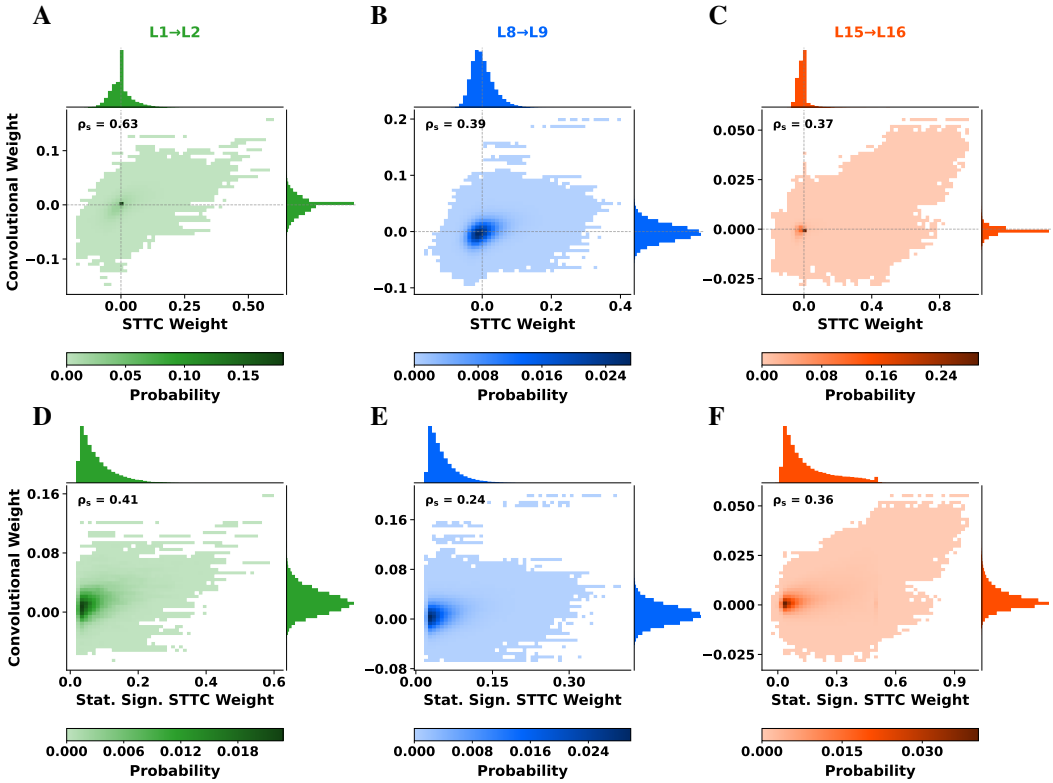

Figure A2: **Structural Connections: Convolutional kernel weight vs. inter-layer STTC weight**. **(A)** Two-dimensional histogram of convolutional weights versus corresponding inter-layer STTC weights for Layer 1→2; shading (light to dark) indicates probability of occurrence. Marginal histograms (top and right) show the distributions of STTC and convolutional weights, respectively. **(B)**, **(C)** Same as **(A)** for Layer pairs 8→9 and 15→16. **(D-F)** Same as **(A-C)**, respectively, restricted to positive statistically significant edges ($z > 4$). Inset: Spearman's rank correlation between STTC weight and convolutional weight.Across all examined layer pairs, the joint distributions reveal a consistent alignment between STTC weights and the corresponding convolutional kernel weights; however, this correspondence is modestly attenuated when restricted to statistically significant positive edges **(D-F)**, suggesting that strong functional correlations are not exclusively confined to connections with large learned weights.

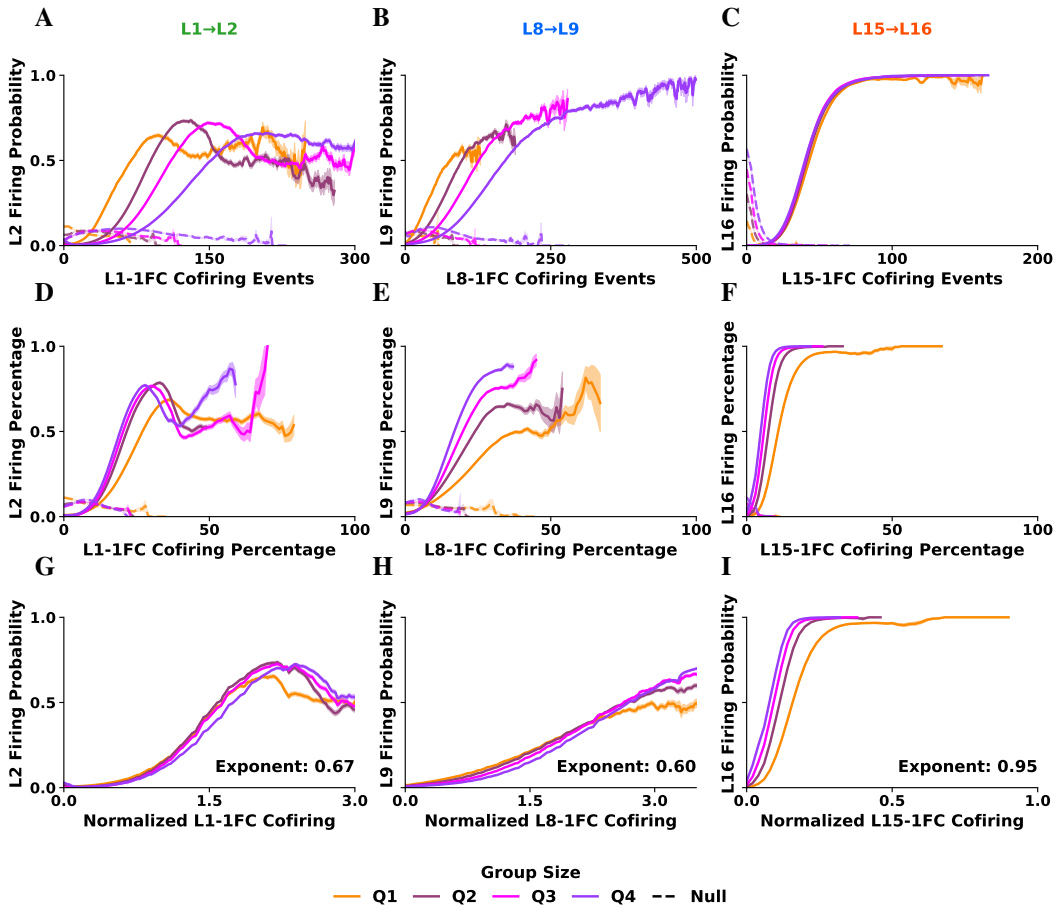

Figure A3: **Response functions across group size conditions. (A–C)** Aggregate firing probability of target-layer neurons as a function of input 1FC co-firing events, stratified by group size quartiles (Q1–Q4), for layer pairs L2→L3 **(A)**, L8→L9 **(B)**, and L15→L16 **(C)**. Functional groups (1FC) are defined once using the full test dataset and held fixed; the null condition (dashed gray) shows the mean ± SEM of size-matched control groups without significant functional correlations. **(D–F)** Same as **(A–C)**, with responses expressed as a function of the percentage of co-firing events (normalized by total group size). **(G–I)** Same as **(A–C)**, but co-firing is normalized by a normalization factor - N^{α}, with N as the group size and $\alpha = 0.67$, 0.6, and 0.95 for L2→L3, L8→L9, and L15→L16, respectively.

A.3 Temporal Robustness of Responses

To verify that the 1FC groups are not an artifact of the specific images used to estimate functional connectivity, we assess the temporal robustness of neuronal firing response functions across held-out input samples. For each neuron, we compute firing response functions over four independent batches of test images and compare the resulting slopes, retaining the same 1FC group assignments throughout. The distribution of slope differences across all four batches remains tightly concentrated around zero, with a negligible proportion of outliers (Fig. A5). The consistency of response function slopes across batches and across the full range of input images confirm that the 1FC groups reflect stable functional structure rather than sample-specific co-activation.

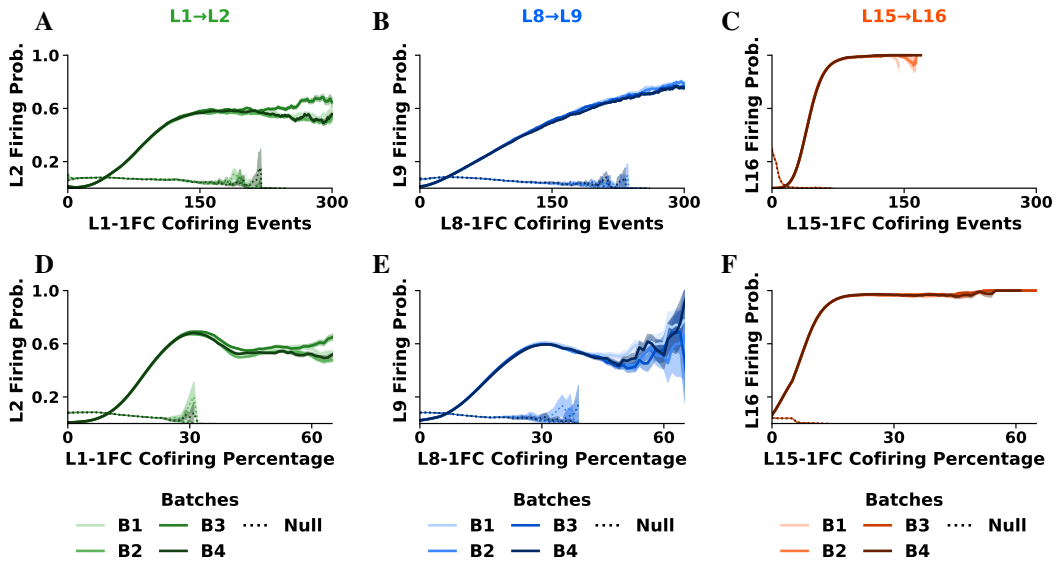

Figure A4: **Temporal robustness of spiking responses.** **(A–C)** Aggregate firing probability of target-layer neurons as a function of input 1FC co-firing events, evaluated for layer pairs L2→L3 **(A)**, L8→L9 **(B)**, and L15→L16 **(C)**. Functional groups (1FC) are defined once using the full test dataset and held fixed; firing responses are then computed separately across four input batches (light to dark shading). Shaded regions denote SEM. The null condition (dotted gray) shows the mean ± SEM of size-matched control groups drawn from neurons without statistically significant functional correlations. **(D–F)** Same as **(A–C)**, shown as a function of the percentage of co-firing events. Consistent responses across batches indicate that firing probabilities are not biased by input composition or batch order.

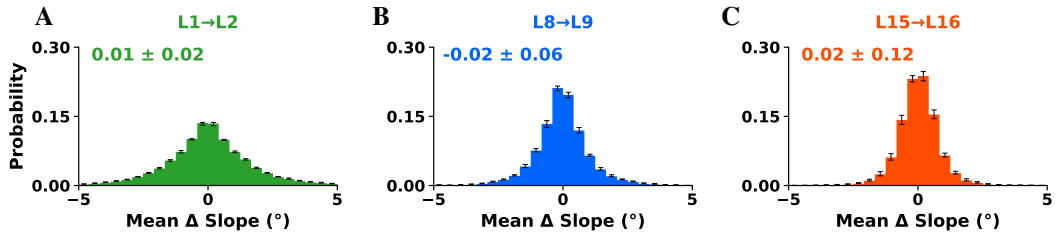

Figure A5: **Stability of response function slopes across batches.** For layer pairs L1→L2 **(A)**, L8→L9 **(B)**, and L15→L16 **(C)**, we quantify variability in neuron-wise response function slopes across input batches. 1FC input functional groups are defined on the full dataset and held fixed. For each neuron, slopes are computed independently for each of four batches (B1,...,B4). Pairwise differences in slopes are then evaluated across all 6 batch combinations (e.g., B2−B1, B3−B1). For each batch pair, we obtain a distribution of slope differences across neurons; panels show the mean ± SEM of these distributions aggregated over all batch pairs. Inset text shows the mean (± SD) of the distributions across batch pairs. Small mean differences indicate that response slopes are stable across batches. Neurons whose response-curve fits yielded $R^2 \leq 0.8$ were excluded from the analysis.(2424, 5946 and 11 neurons were excluded from the analysis for L1→L2, L8→L9, L15→L16 respectively.)

A.4 Mutual Information

A *one-hot encoding* is a mapping from a categorical set to binary vectors. Let $\mathcal{C} = \{c_1, \ldots, c_K\}$ be a set of K categories (here classes of input images). The one-hot encoding is the function

$$\phi : \mathcal{C} \to \{0, 1\}^K$$

defined by

$$\phi(c_i) = e_i,$$

where $e_i \in \mathbb{R}^K$ is the i-th standard basis vector, i.e.,

$$(e_i)_j = \begin{cases} 1 & \text{if } j = i, \\ 0 & \text{otherwise.} \end{cases}$$

To quantify the informativeness of individual neurons, we computed the mutual information (MI) between each neuron's spike train and the true class labels across the test set. MI was calculated between the binary spiking activity of each neuron (spike or no spike) and both the categorical class labels and their one-hot encoded representations, the latter yielding a class-wise MI profile for each neuron. MI is defined as

$$I(X;Y) = H(X) - H(X|Y), \tag{2}$$

where $H(X)$ denotes the Shannon entropy of the neuron's spiking activity and $H(X|Y)$ is the conditional entropy given the class label, such that MI reflects the reduction in uncertainty about a neuron's firing upon observation of the true class.

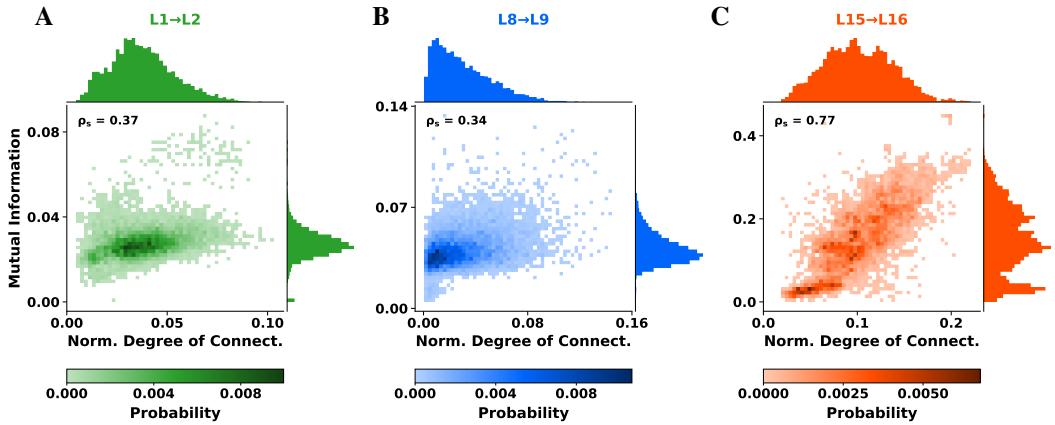

Figure A6: **Relationship between normalized degree of connectivity and mutual information**. **(A)** Two-dimensional histogram of mutual information versus corresponding normalized degree of connectivity for Layer 1→2; shading (light to dark) indicates probability of occurrence. Marginal histograms (top and right) show the distributions of normalized degree of connectivity and mutual information, respectively. **(B)**, **(C)** Same as **(A)** for Layer pairs 8→9 and 15→16. Especially in L16, the higher the degree of connectivity, the larger the MI. A cluster of hub neurons with high degree of connectivity and MI is prominent. Spearman correlations are reported in each plot. Positive correlation between MI and DoC is most pronounced in the final layer pair **(C)**, suggesting that deeper layers progressively concentrate class-discriminative information within a subset of highly interconnected neurons.

A.5 Differences in functional edge types: structural and non-structural edges.

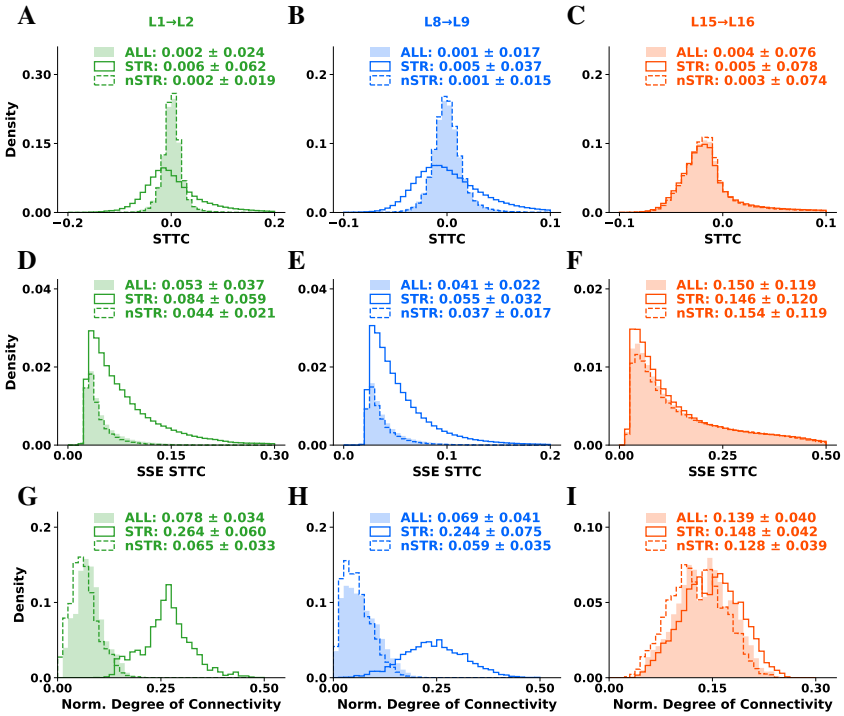

Figure A7: **Inter-Layer First-Order Functionally Connected (1FC) Groups. (A-C)** show pairwise STTC weights (ALL; filled), structural edges (STR; solid line), and non-structural edges (nSTR; dashed line) for layer-pairs L2→L3, L9→L10, and L16→L17 respectively. **(D-F)** show the STTC weights of statistically significant edges (SSEs) having z-score > 4 for the three connection types. **(G-I)** shows the distribution of the normalized degree of connectivity (DoC) of each output neuron. For ALL edges, degree is expressed as a fraction of the total number of neurons in the source layer. For structural edges, normalization is with respect to the total number of possible structural connections defined by the convolutional kernel ($c_{\text{in}} \times k \times k$). For non-structural edges, normalization is with respect to the remaining possible connections (ALL $-$ STR). Early and middle layer pairs show higher structural connectivity, whereas in the final layer pair structural and non-structural contributions become comparable, indicating increased purely functional connectivity in deeper layers.

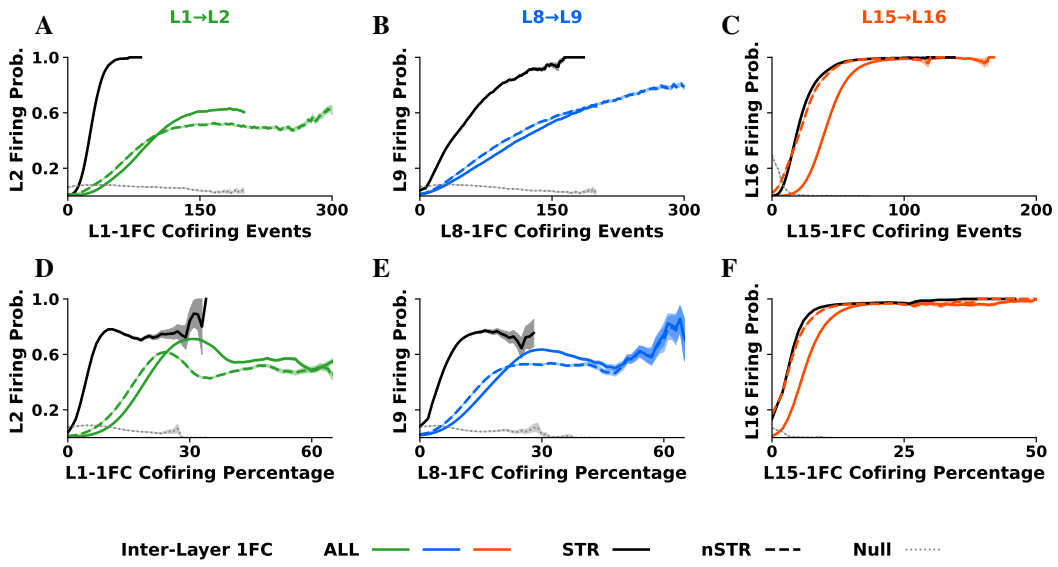

Figure A8: **Aggregate neuronal responses as a function of each neuron's 1FC neighbors from the previous layer, under three cases: all its 1FC neighbors (ALL), only those with structural connections (STR), and the remaining ones (nSTR). (A)** For each L2 neuron, we estimated its firing probability as a function of co-firing events among its all its L1-FC neighbors ALL; solid green), only its STR L1-1FC neighbors (solid black), and only its nSTR (dashed green). Results are aggregated across all L2 neurons. Shaded regions indicate the SEM. The null condition (dotted gray) represents the mean ± SEM of a control group matched in size to the 1FC set, sampled from neurons without statistically significant functional correlations. **(D)** Same as **(A)**, but plotted as a function of the percentage of co-firing events. **(B–C, E–F)** Corresponding analyses for layer pairs L8→L9 layer connectivity and L15→L16 layer connectivity, respectively. All neurons were included.

A.6 Impact of Identity of Neurons in 1FC

Preserving the identities of individual neurons within input–1FC groups does not provide significant additional predictive value beyond the aggregate co-firing signal when modeling the activity of the corresponding output neuron. In particular, a linear-SVM kernel (Fig. A9 A-D) classifier that explicitly incorporated neuron identity exhibit a relatively small improved performance to models relying solely on total co-firing (consistent results also with Random Forest, Naive Bayes, and Logistic Regression; Fig. A9 E-P). These findings suggest that ensemble-level co-firing magnitude may play a more prominent role in shaping downstream responses than the specific identities of the contributing neurons.

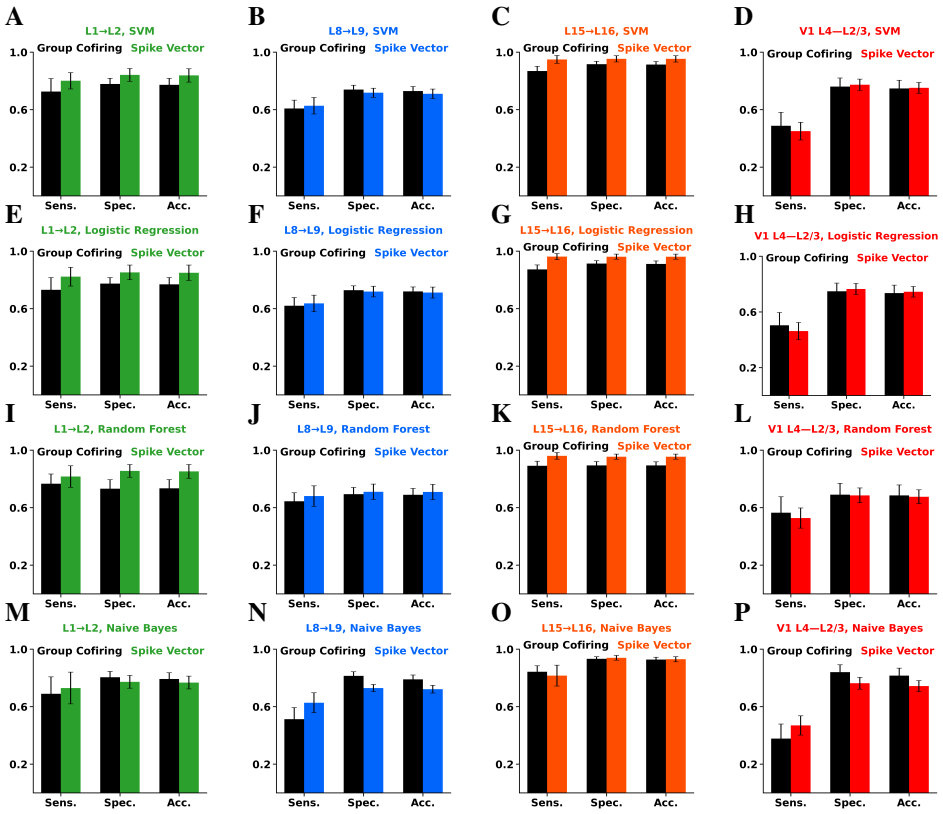

Figure A9: **Predicting neural response using its -1FC group**. The knowledge of the identity of the 1FC neighbors of an index neuron, in comparison to the aggregate 1FC cofiring, has a relatively small impact on the classification accuracy in determining the firing of the index neuron. We employ Support Vector Machines **(A-D)**, Logistic Regression **(E-H)**, Random Forests **(I-L)**, and Naive Bayes Classifers **(M-P)** to predict the inferred firing of the index neuron during the original test dataset using as input: the number of cofiring of its 1FC groups (black), or the individual spike trains of the 1FC group of the index neuron (color; green, blue, orange for layer-pairs L1→L2, L8→L9 and L15→L16 respectively). Neurons with 1FC group sizes < 15 were excluded from this analysis. Note that the input of the individual spike trains (color) includes the identity of the neurons that fire. We compared their performance in terms of sensitivity, specificity, and accuracy.

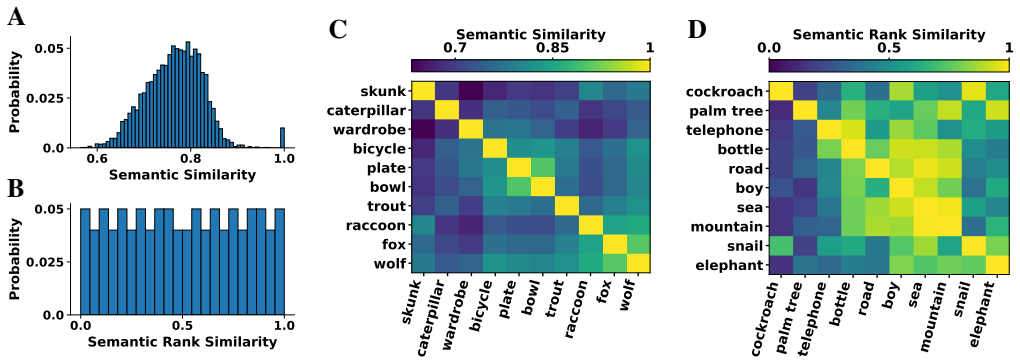

Figure A10: **Semantic Similarity of CIFAR100 classes (A, B)** Distributions of CLIP-based semantic similarity and semantic rank similarity across CIFAR-100 classes. **(C, D)** Heatmaps of pairwise semantic similarity and semantic rank similarity for a subset of 10 randomly selected classes. While both representations encode similar relational structure, rank similarity reduces distributional imbalance by emphasizing relative ordering over absolute cosine distances, resulting in a more uniform comparison across classes.

A.7 Semantic Similarity

We examined whether neurons preferentially encode classes that are semantically related in human-interpretable terms. To quantify semantic relationships among CIFAR-100 labels, we compute embeddings for each class using CLIP, acontrastive language-image pre-trained model [61], and define semantic similarity via cosine similarity between embedding vectors. Formally, semantic distance is given by cosine distance, and semantic similarity is defined as $1-$ cosine distance.

Because CIFAR-100 class definitions are not uniformly distributed in semantic space, raw cosine similarities exhibit a biased and uneven distribution. To mitigate this, we introduce a semantic rank similarity metric. For each class, all other classes are ranked according to their semantic distance, and ranks are normalized by the total number of classes. This yields a uniform representation of relative semantic proximity, where each class has maximal self-similarity and progressively lower similarity assigned according to rank order rather than absolute distance. (See Fig A10 for distributions of semantic similarity and semantic rank similarity.)

A.8 Additive Noise conditions

We analyze the functional connectivity under two types of additive noise, namely uniform random noise and adversarial noise via targeted input perturbation, based on the Fast Gradient Sign Method (FGSM)[40].

Uniform Random Noise. We probed the robustness of the neuronal responses under uniform random noise added to the image. (See Fig. A11)

$$X' = 0.1 \times X + 0.9 \times \mathcal{U}\left[0, 1\right). \tag{3}$$

Under uniform additive noise, firing rates were substantially lower than under original input conditions (Fig. A12). Information encoding was also altered: although L15-1FC groups of L16 neurons could still exhibit high co-firing events, these events occurred less frequently and were largely associated with images unrelated to the neuron's preferred class (Fig. A16 A). Consequently, the spiking activity of L16 neurons exhibited reduced entropy and lower class-specific information content (Fig. A16 B, C).

Adversarial noise under targeted input perturbation. We applied the Fast Gradient Sign Method (FGSM) Goodfellow et al. [40] adversarial attack to the test set. FGSM generates adversarial examples by perturbing the input in the direction of the gradient of the loss with respect to the input, scaled by a perturbation magnitude ϵ:

$$X' = X + \epsilon \times \text{sign}(\nabla_X L(X, y)). \tag{4}$$

where, x is the input image, y is the network's prediction for x and $(\nabla_x L(\theta, x, y))$ is the gradient of the loss function with respect to x. Since spiking neural networks are non-differentiable by nature,

gradient computation is non-trivial; however, the use of surrogate gradient backpropagation during training, specifically the sigmoid surrogate function, afforded a straightforward extension of FGSM to our SNN, as the surrogate gradients with respect to the input were directly available. Adversarial examples were generated at $\epsilon = 4$, and neuronal firing probabilities were measured as a function of the 1FC defined using the original test images.

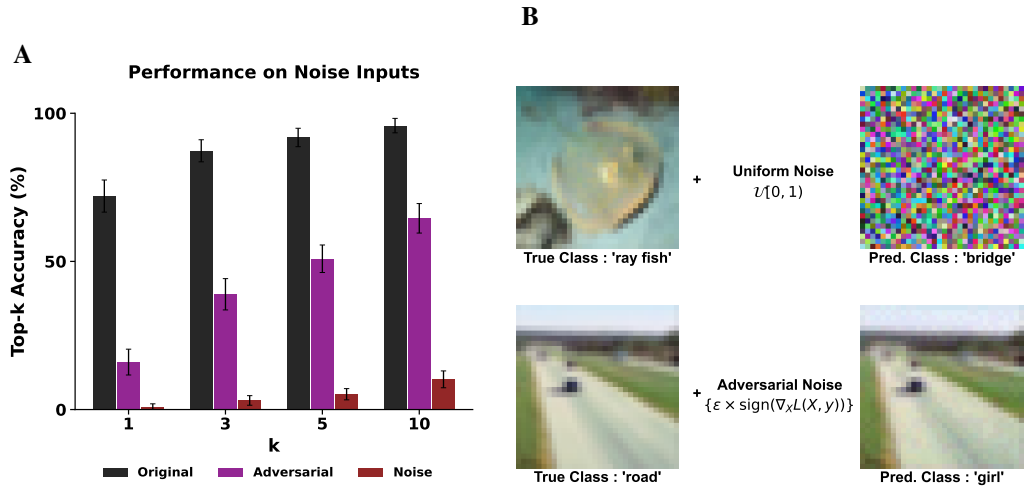

Figure A11: **Performance of the network under additive noise and adversarial perturbations.** **(A)** Bars show top-k accuracy: the proportion of samples for which the ground-truth label is contained within the model's k highest-scoring predictions—for $k = 1$, 3, 5, and 10, evaluated on original, noise-corrupted, and adversarial inputs. Error bars denote standard deviation across batches of input. **(B)** Example test images ("ray fish" and "road") with added perturbations. Top row: uniform random noise. Bottom row: adversarial noise. The perturbed images are misclassified as "bridge" and "girl," respectively.

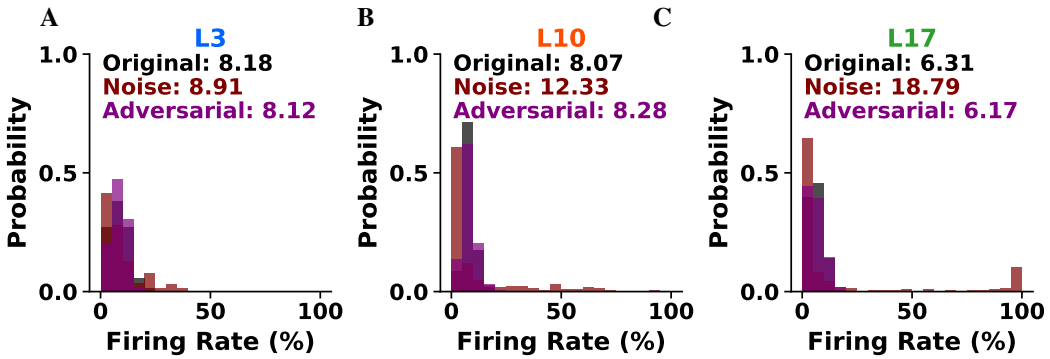

Figure A12: **Firing rates under original test, noisy and adversarial input.** **(A)**. The histogram of the event rate of each neuron in L2 under the noisy test input (maroon), adversarial noise (purple) and original test input (black). **(B-C)**. Same, for neurons in L9 and L16, respectively. Each panel shows active neurons with firing rate > 0.01 spikes/sample. Inset: Mean event rate, for each layer.

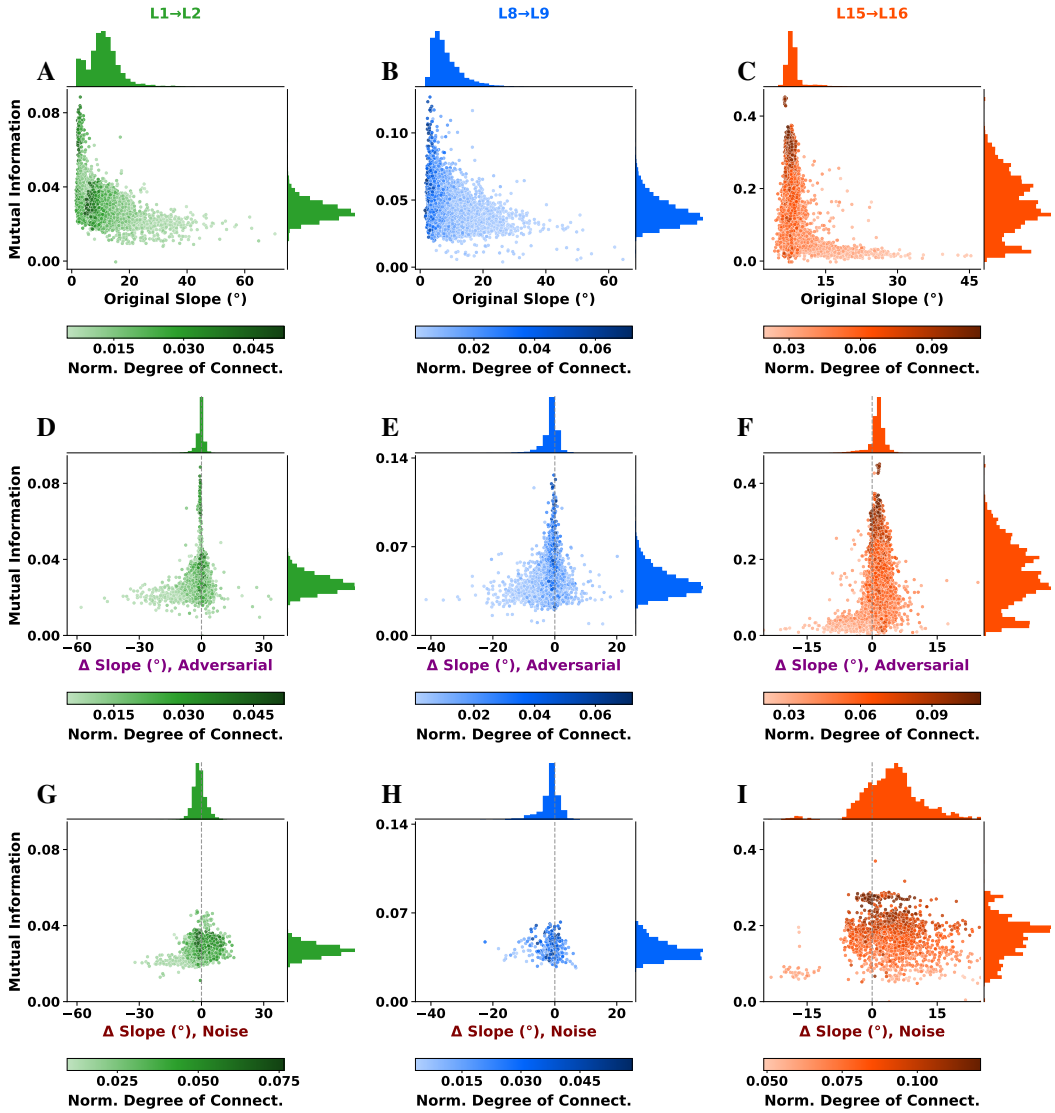

Figure A13: **Differences in the slope of a neuron's response under additive noise input relative to its response under original input.** **(A–C)** Scatter plots of mutual information (MI) versus the slope of neuronal response functions for layer pairs L1→L2, L8→L9, and L15→L16, respectively, where slope is computed as a function of input-layer 1FC group cofiring events under original input. **(D-F)** MI versus the change in slope under adversarial noise for the same layer pairs. **(G-I)** MI versus the change in slope under uniform additive noise. Neurons with lower MI and connectivity exhibit greater sensitivity to noise, whereas higher-MI neurons show more robust (less sensitive) responses. Neurons with response-curve fits yielding $R^2 \leq 0.8$ were excluded from the analysis.

A.9 Limitations

So far we tested this approach in ResNet18 SNNs and spiking-neural units of spiking-layers that extend a ResNet32 feature extractor[38], pre-trained on CIFAR100 and ResNet9 SNNs trained on DVS Gesture Dataset. We also have only tested the framework for the Fast Gradient Sign Method (FGSM) adversarial attack[40] and additive noise. It is part of our ongoing effort to test it with other architectures, activation functions, and hyperparameters. The impact of T=4 (in the image presentation of the SNN model) on STTC will be further examined.

Assessing the accuracy of the diagnostics for test datasets of different size, distribution of class images, diversity/richness of the images, types of distortion (e.g., noise levels, blurriness, adversarial

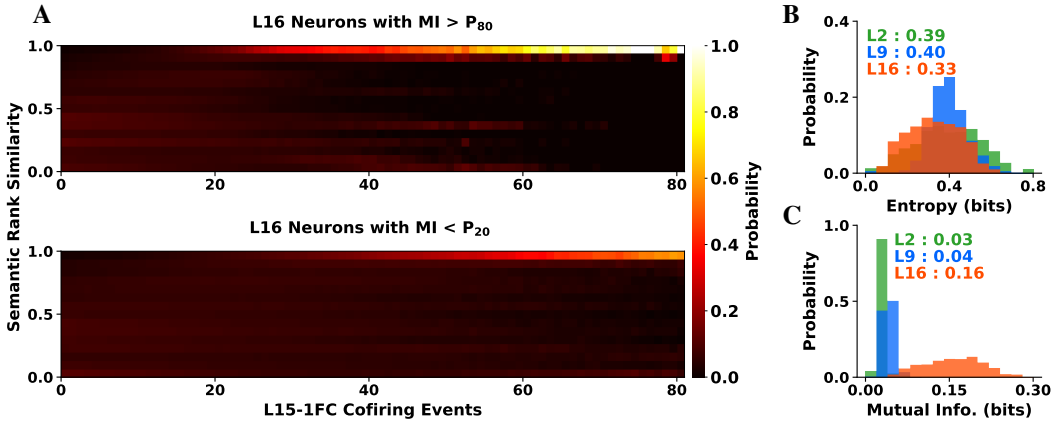

Figure A14: **Information encoding within output neurons and their 1FC groups (A)** Top: heatmap of semantic rank similarity for L16 neurons whose class-wise mutual information lies in the top 20th percentile, shown as a function of L15-1FC group co-firing events. Each bin reflects the probability that the class of frames with N co-firing events matches the L16 neuron's preferred class (defined as the class with maximum class-wise mutual information). Bottom: same analysis for neurons in the lowest 20th percentile. For each co-firing bin, probabilities are normalized to sum to 1 across semantic rank similarity. **(B)** Entropy of L2, L9, and L16 spike trains. **(C)** Total mutual information of L2, L9, and L16 neurons with the output, quantifying the ability of their spike trains to distinguish among output classes.

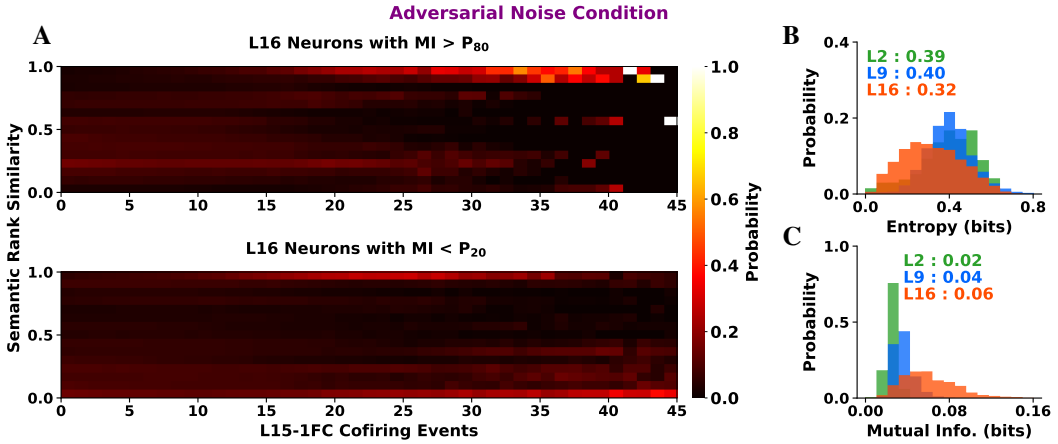

Figure A15: **Information encoding within output neurons and their 1FC groups under Adversarial Noise conditions (A)** Top: heatmap of semantic rank similarity for L16 neurons whose class-wise mutual information lies in the top 20th percentile, shown as a function of L15-1FC group co-firing events under adversarial noise conditions. Each bin reflects the probability that the class of frames with N co-firing events matches the L16 neuron's preferred class (defined as the class with maximum class-wise mutual information). L15-1FC group identities and L16 neuron's preferred classes are kept fixed from the original input without adversarial noise. Bottom: same analysis for neurons in the lowest 20th percentile. For each co-firing bin, probabilities are normalized to sum to 1 across semantic rank similarity. **(B)** Entropy of L2, L9, and L16 spike trains under adversarial noise conditions. **(C)** Total mutual information of L2, L9, and L16 neurons with the output under adversarial noise conditions, quantifying the ability of their spike trains to distinguish among output classes.

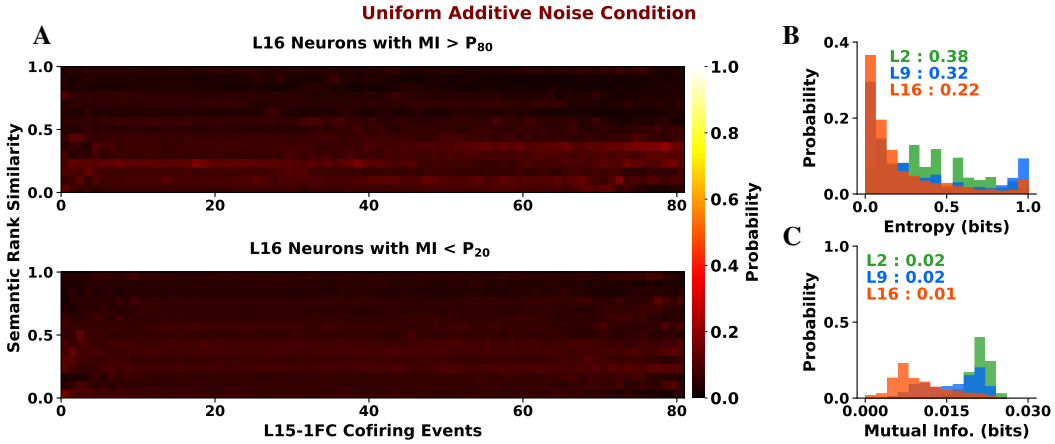

Figure A16: **Information encoding within output neurons and their 1FC groups under Uniform Additive Noise conditions (A)** Top: heatmap of semantic rank similarity for L16 neurons whose class-wise mutual information lies in the top 20th percentile, shown as a function of L15-1FC group co-firing events under uniform additive noise conditions. Each bin reflects the probability that the class of frames with N co-firing events matches the L16 neuron's preferred class (defined as the class with maximum class-wise mutual information). L15-1FC group identities and L16 neuron's preferred classes are kept fixed from the original input without additive noise. Bottom: same analysis for neurons in the lowest 20th percentile. For each co-firing bin, probabilities are normalized to sum to 1 across semantic rank similarity. **(B)** Entropy of L2, L9, and L16 spike trains under uniform additive noise conditions. **(C)** Total mutual information of L2, L9, and L16 neurons with the output under uniform additive noise conditions, quantifying the ability of their spike trains to distinguish among output classes.

attacks) is part of our ongoing research. We plan to explore the required size of test set (batch size) and size of the population of neurons to be sampled for detecting reliably when abnormalities in the input or in the behaviour of the architecture occur.

Going forward, integrating this framework with causal perturbations will be key to clarifying the mechanistic role of specific ensembles in learning, as well as to probing how these mechanisms fail under adversarial conditions.

A.10 Computational Efficiency

One of the most computationally intensive steps is estimating the statistical significance of pairwise STTC correlations for all neuron pairs (considered in the analysis), as it requires generating 150 circularly shifted null spike trains to compute the control STTC weight.

Compute Resources. Model training was performed on local GPU servers and the CINECA Leonardo Booster HPC facility (NVIDIA A100 GPUs). Training the SEW-ResNet-18 for 320 epochs on CIFAR-100 required approximately 1.5 GPU hours on 1 GPU. Post-training analyses were considerably more computationally demanding. Pairwise STTC computation is memory-intensive despite algorithmic optimization via vectorization [62]. It achieves near-linear time complexity $\mathcal{O}(N)$, where N is the length of the spike trains being compared; statistical validation further requires 150 circular shift controls per neuron pair. Computing STTC and all controls across all examined layers for a single model required approximately 1 GPU hour on the Leonardo Booster. Mutual information estimation and piecewise linear response function fitting, both performed at the single-neuron level across the sampled population of $\sim$8,000 neurons per layer, were executed on local servers and required approximately 2–3 hours per layer. All inference passes for spike train recording, including clean, noisy, and adversarial conditions, were performed on 1 GPU with a total inference time of approximately <10 minutes for the full test set.

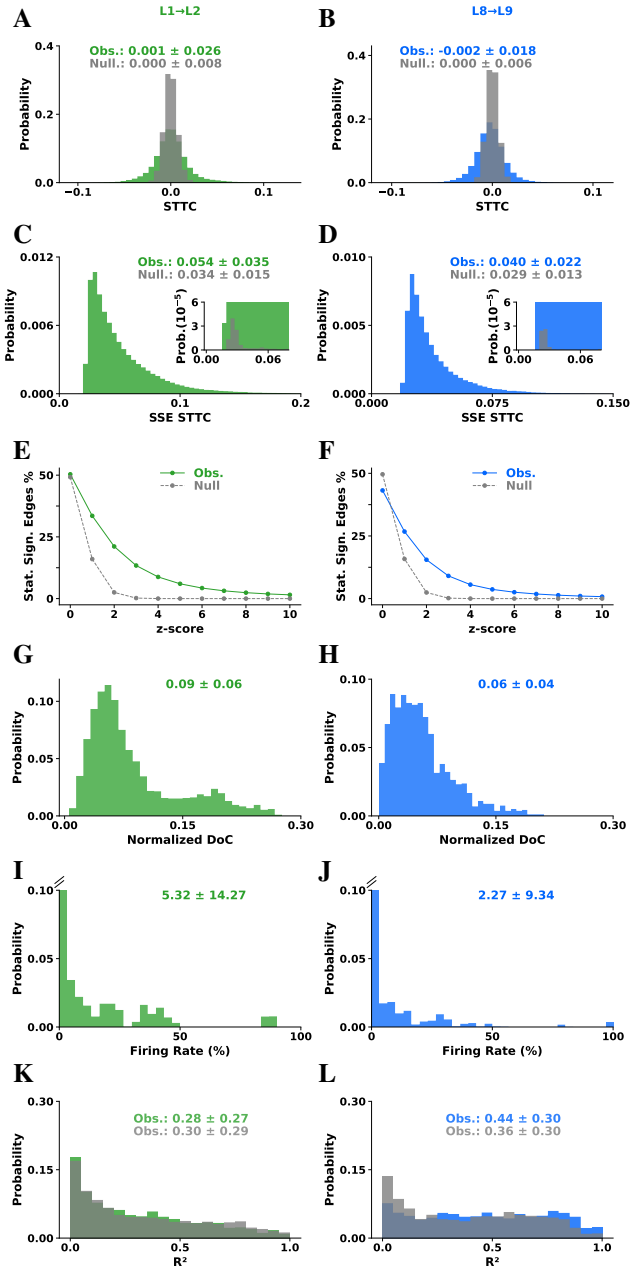

Figure A17: **Inter-layer functional connectivity under shuffled-weight conditions.** Pairwise STTC was computed between neurons in consecutive layers during inference on CIFAR100 using the trained network with shuffled weights. Columns correspond to layer pairs L2→L3 (green) and L9→L10 (blue). A null distribution (grey) was generated via 150 circular shifts of spike trains. **(A, B)** STTC distributions with corresponding nulls. **(C, D)** Distributions of statistically significant positive edges ($z > 4$). **(E, F)** Fraction of significant edges as a function of z-score. **(G, H)** Degree of connectivity distributions (normalized by total possible edges). **(I, J)** Firing rate distributions of active neurons (input and output combined). **(K, L)** Distributions of R^2 values from response-function fits. Neurons were considered active if firing rate > 0.01 spikes/sample (yielding $< 40\%$ and $< 10\%$ active neurons across layers). z-scores were computed relative to the circular-shift null distributions.

A.11 Broader Implications

A potential implication of this work involves the design of diagnostic frameworks for spiking neural networks (SNNs), for assessing the sequence of input images for testing/inference.

In general, by emulating fundamental features of biological computation—such as sparse and event-driven activity—SNN models offer a compelling paradigm for efficient information processing, which neuromorphic hardware systems exploit to deliver substantial gains in energy efficiency. Moreover, the shared features of functional connectivity between SNNs and the mouse visual cortex may enable the use of SNNs as testbeds for evaluating neurocomputational hypotheses about cortical connectivity under diverse conditions.