The Emergence of Partial Conjunctions and Feature Binding in Artificial Neural Network Models of the Human Visual Cortex

Supervisor: Dr. Greg Davis

Summary

The binding problem, or how separately coded visual features are integrated into unified object representations, remains a central challenge in cognitive science. While hierarchical, hard-wired conjunction coding has been proposed as a solution, it has long been considered intractable due to the exponential number of required feature conjunctions. A complementary solution arises through partial conjunctions, which selectively encode subsets of features and thereby reduce representational demands while maintaining unambiguous object coding. In this computational study, I investigated the emergence of partial conjunctions in two models of the human visual cortex: a biologically inspired spiking neural network (SNN) and a modified residual neural network (ResNet). By exposing the SNN model to intrinsically structured visual stimuli and updating it through unsupervised learning, I showed that neurons with partial conjunctional tuning emerged spontaneously; via representational similarity analyses, I revealed that these emerged neurons do not explicitly encode feature combinations in visual stimuli. By training the ResNet model on object-like visual stimuli under supervised classification tasks, I revealed that partial conjunctions arose robustly, regardless of the presence of binding demands. These findings demonstrate that partial conjunctional coding naturally arises in hierarchical neural networks, providing a criterion-free, condition-independent basis for representing complex object-based feature combinations. Together, this work highlights partial conjunction coding as a tractable, efficient, and biologically plausible solution to the binding problem in vision.

Table of Contents

1. Introduction

The feature binding problem in vision (FBP) concerns how the human visual system integrates separately encoded visual features, such as shape, colour, motion, etc., into holistic object representations that one consciously perceives (von der Malsburg, 1981). This problem arises from the experimental observations that neurons in the visual cortex are selectively tuned to distinct features at different retinal locations (Livingstone & Hubel, 1987; Nassi & Callaway, 2009; Taylor & Xu, 2022). Consequently, when multiple objects appear simultaneously in a visual scene, features from different objects would concurrently activate a pool of neurons across the cortex, creating a challenge for the visual system in differentiating which feature belongs to what object. This neural ambiguity, however, does not prevent human vision from accurately recognising objects and their feature compositions. Thus, it is assumed that a dedicated neural mechanism must exists to resolve this challenge (see Di Lollo, 2012 for opposing claims).

Fig 1.1 provides a straightforward illustration of the FBP. Consider a simplified human visual cortex that contains six neurons: two selectively tuned to the colours red and green, two to the shapes square and triangle, and two to the locations top and bottom, which replicates the visual system's feature selectivity. While this cortex is capable of encoding single objects (Fig 1.1a and Fig 1.1b), a binding problem arises when two objects are simultaneously presented as the visual input (Fig 1.1c), as the cortex can no longer unambiguously represent the combination of shape, colour, and location for each object in the scene. Note that, for simplicity, the presented objects in this illustration contain spatially overlapping features that can be bound trivially via locations (Treisman, 1988). However, multipart objects with hierarchical feature compositions, as those in naturalistic scenes, cannot be bound simply by spatial cues.

Understanding the generic binding mechanism that these complex objects require remains one of the biggest open questions about the human brain (Feldman, 2013).
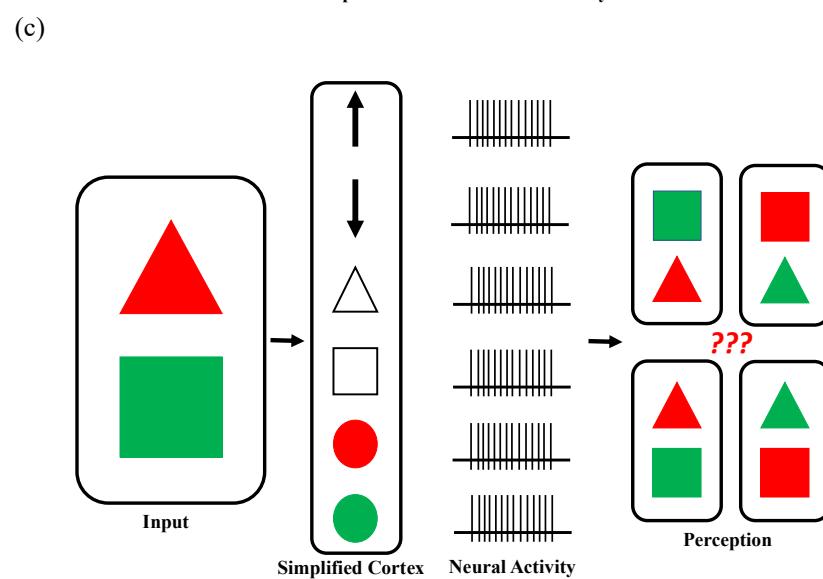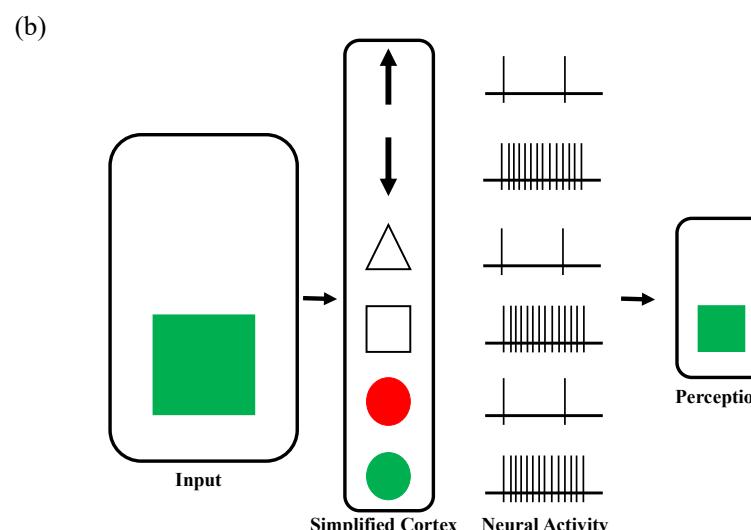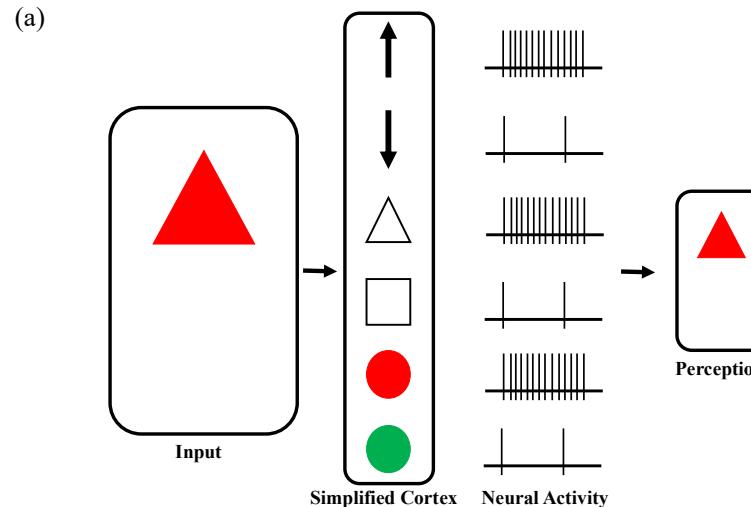
Fig 1.1: A classical illustration of the FBP by Rosenblatt (1962). A cortex with selectively tuned neurons can uniquely encode (a) a single red triangle at the top location, or (b) a single green square at the bottom location, but not (c) the simultaneous presentation of both objects. Icons in the cortex represent neurons with specific feature selectivity, with raster plots conceptually indicate levels of neural activation. Adapted from Velik (2012).

Since the popularisation of the FBP, several binding mechanisms have been proposed, yet each remains flawed or incomplete. The most prominent theories among them include binding by combinatorial coding (Barlow, 1972; Riesenhuber & Poggio, 1999; Roelfsema, 2006; Fig 1.2a), which uses neurons tuned to complete sets of feature conjunctions to directly represent feature combinations, binding by synchrony (von der Malsburg, 1981; Fig 1.2b), which applies synchronised neural oscillations to temporally group neurons encoding the same object, and binding by rate-enhancement (Treisman & Gelade, 1980; Reynolds & Desimone, 1999; Fig 1.2c), which utilises attention-induced neural activations to bind neurons activated by the same object. Despite some theoretical and experimental supports (hardcoded: Riesenhuber & Poggio, 1999; Mel, 1997; synchrony: Singer & Gray, 1995; Fries et al., 2007; rate-enhancement: Roelfsema et al., 1998; Jeurissen et al., 2016), all three solutions are incomplete and required a defined neural mechanism (Roelfsema, 2023).
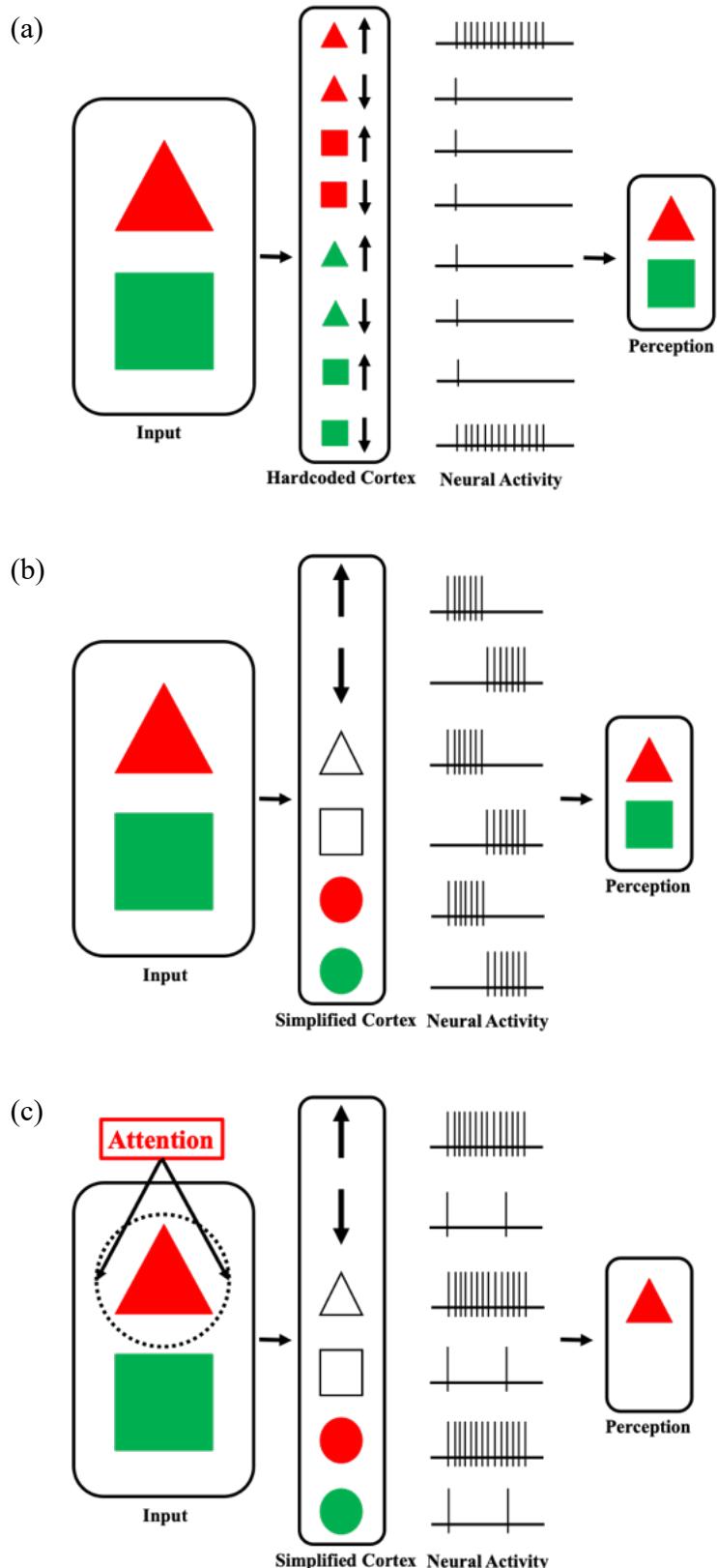
Fig 1.2: Solving the FBP via (a) binding by hardcoded conjunction, (b) binding by synchrony, and (c) binding by rate-enhancement. Icons in the cortex represent neurons with specific feature selectivity, with raster plots conceptually indicate the level of neural activations. Adapted from Velik (2012).

Partial conjunctions are subsets of adjacent elements extracted from a continuous sequence, traditionally used in the field of linguistics for text analysis (Cavnar & Trenkle, 1994). By decomposing a word (e.g., *biology*) into a set of interconnected letter pairs (e.g., {*bi, io, ol, lo, og, gy*}), partial conjunctions allow the unambiguous and efficient representation of letter orders in common words (Mel & Fiser, 2000). Given the intricate similarity between assigning letters to words and binding features to objects, partial conjunctions could provide an alternative solution to the FBP (Sommer, 2013; Fig 1.3). Rather than relying on hardcoded conjunctions or external binding mechanisms, the visual cortex could use partial conjunctional units to encode partial feature combinations in visual inputs, which, collectively, enable the unambiguous representation of each object's feature composition in the scene.
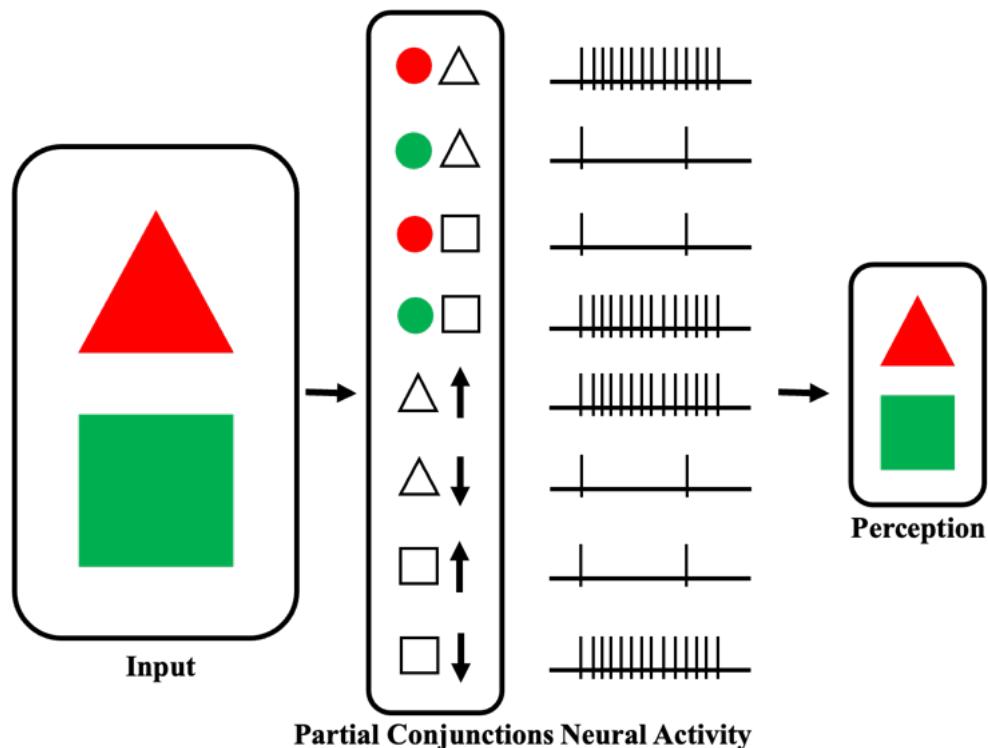

Fig 1.3: Solving the FBP by partial conjunctions. Icons in the cortex represent neurons with specific feature selectivity, with raster plots conceptually indicate levels of neural activation.

Partial conjunctions are not entirely new to visual neuroscientists, as recent studies have explored these units as potential mechanisms for feature binding. Morita et al. (2010) demonstrated that rapid flashes of partial feature pairs could reliably induce illusions of holistic objects in both human observers and a connectionist neural network of partial conjunctions. Furutate et al. (2019) showed that, using a perceptual conditioning task, partial feature representations could accurately explain human performances on learning feature combinations. Schneegans & Bays (2017) further revealed that human recall errors in visual working memory are best predicted by a neural population encoding partial feature-location conjunctions. These studies have shown strong evidence for the involvement of partial conjunctions in feature binding. However, a central question remains: how do partial conjunctions arise in the visual system? Do they emerge spontaneously through the brain's intrinsic plasticity (Fiser & Lengyel, 2022), develop via supervised or reinforcement visual learning processes (Botvinick et al., 2020), or are they innately hardwired (Li et al., 2020)? Addressing this question was the primary focus of the present study.

Beyond traditional experimental approaches, modelling and simulation studies have become an increasingly important tool for investigating the human brain (Fan & Markram, 2019). By avoiding the technical challenges and ethical constraints of in vivo experiments, computational models enable cognitive neuroscientists to explore broader research questions with flexible experimental designs. The present study combined two complementary computational models of the human visual system: spiking neural networks (SNNs) and residual neural networks (ResNets).

SNNs provide a biologically grounded framework for in silico brain simulations (Maass, 1997). By incorporating biophysically realistic neurons that represent information via discrete spikes

(Hodgkin & Huxley, 1952; Izhikevich, 2003; Burkitt, 2006), SNNs replicate key aspects of neuronal computation in the human brain (Yamazaki et al., 2022). They have been successfully applied to studies of the FBP (Miconi & VanRullen, 2010; Martin & von der Heydt, 2015; Isbister et al., 2018), where their ability to replicate temporal spiking and Hebbian learning offers unique insights into cortical mechanisms. SNNs thus provide an ideal model for testing whether partial conjunctions can emerge spontaneously through an unsupervised, biologically plausible learning process.

ResNets provide a complementary framework for studying task-driven visual cognition (He et al., 2016). With hierarchically organized convolutional layers and residual connections, ResNets support robust performance on various visual recognition tasks (He et al., 2016), often reaching human-level benchmarks (Geirhos et al., 2018; Celeghin et al., 2023). Importantly, the visual representations learned by ResNets closely match patterns of cortical activity in the human visual system (Wen et al., 2018; Rajalingham et al., 2018). This makes ResNets useful computational models for studying how explicit task demands influence the emergence of partial conjunctions in human vision.

In this study, I investigated the emergence of partial conjunctions in a SNN and a ResNet model of the human visual cortex. In the SNN model, I passively exposed the model to a set of object-like visual stimuli and updated it through unsupervised synaptic learning rules. By analysing the resulting neural tuning properties, I revealed the spontaneous emergence of partial conjunctions in the model. Furthermore, by decoding the recorded neural responses via representational similarity analysis, I revealed that these emerged neurons do not explicitly encode feature combinations in visual stimuli. In the ResNet model, I employed supervised learning to train the model on two distinct classification tasks: a perceptual task requiring no
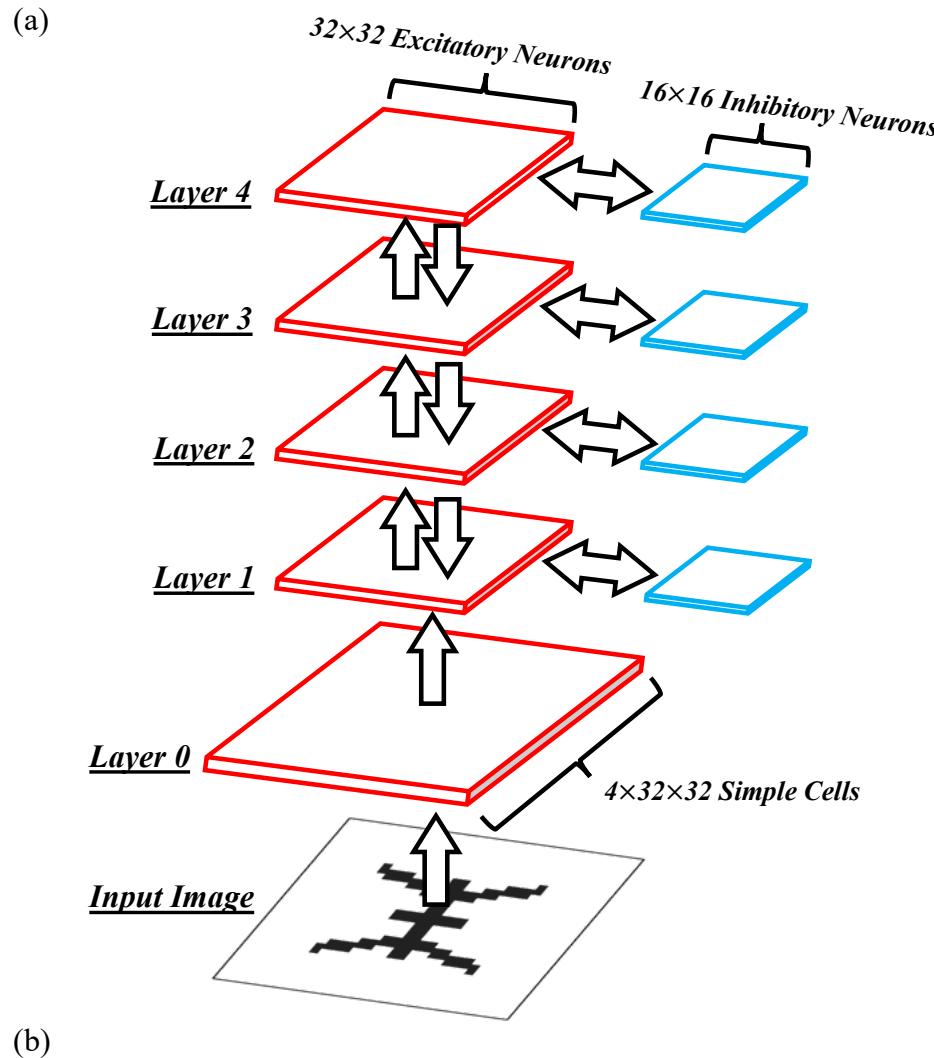
feature binding and a binding task with implicit binding demands. By comparing the emerged tuning properties under different training objectives, I demonstrated the stable formation of partial conjunctions in both classification tasks, regardless the task-imposed binding demands. Together, these results highlight the natural emergence of partial conjunctions in hierarchical neural networks, which provide a criterion-free, condition-independent neural architecture for solving the FBP.

2. Spiking Neural Network Model: Method

2.1 Model Architecture

Inspired by Eguchi et al. (2018), I constructed a spiking neural network model of the human visual cortex that implemented hierarchical network architectures (Serre, 2012), leaky integrate-and-fire neurons (Burkitt, 2006), and spike-timing dependent plasticity (Perrinet et al., 2001) to simulate the structure, dynamics, and flexibility of a human visual system.

Network architecture. The model consisted of five two-dimensional cortical layers arranged in a bottom-up hierarchy, which approximated the human ventral visual pathway for object recognition (Kravitz et al., 2013; Fig 2.1). Layer 0 contained hardwired excitatory cells that preprocessed raw visual images. Layer 1 to 4 contained subpopulations of excitatory and inhibitory neurons that progressively propagated and transformed the preprocessed visual stimuli. In total, the model consisted of 4096 simple cells (4 for 32×32 retinal locations), 4096 excitatory neurons (32×32 for 4 layers), and 1024 inhibitory neurons (16×16 for 4 layers).

(a)

32×32 Excitatory Neurons

16×16 Inhibitory Neurons

Layer 4

Layer 3

Layer 2

Layer 1

Layer 0

4×32×32 Simple Cells

Input Image

(b)
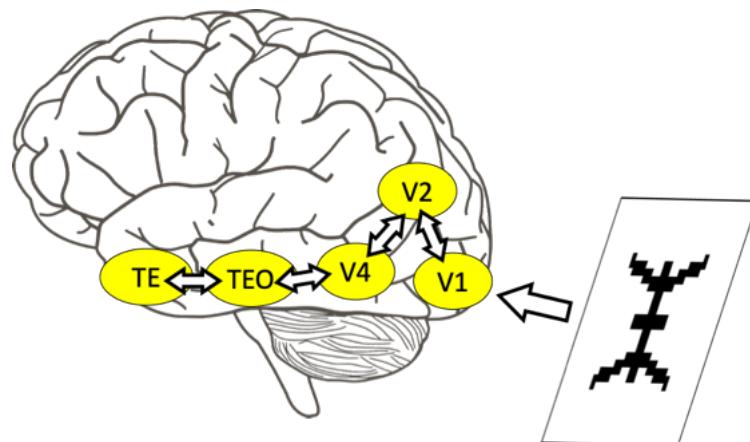
Fig 2.1: (a) Overall architecture of the SNN model and (b) its similarity with the human ventral visual pathway. Red and blue blocks indicate excitatory and inhibitory neural subpopulation, respectively. Arrows indicate permitted directions of information flow among neural subpopulations in the model. Layer 0 to 5 in the model respectively corresponded to V1, V2, V4, posterior inferior temporal cortex (TEO), and anterior inferior temporal cortex (TE) in the human visual system.

Synaptic connection. Neurons in the SNN model communicated through conductance-based synapses. Specifically, excitatory neurons formed feedforward and feedback connections with excitatory neurons in successive layers, as well as lateral connections with both excitatory and inhibitory neurons in the same layer. Inhibitory neurons formed only local lateral synaptic connections with excitatory neurons in the same layer. These connections were spatially confined, where neurons formed synapses primarily within a defined projection radius (Fig 2.2). Each presynaptic neuron selected its postsynaptic targets by sampling from a bivariate Gaussian distribution, with its presynaptic spatial coordinates as the mean and its projection radius as the covariance. The projection radius was gradually increased across cortical layers to replicate the expanding neural receptive fields observed along the ventral visual pathway (Kravitz et al., 2013).
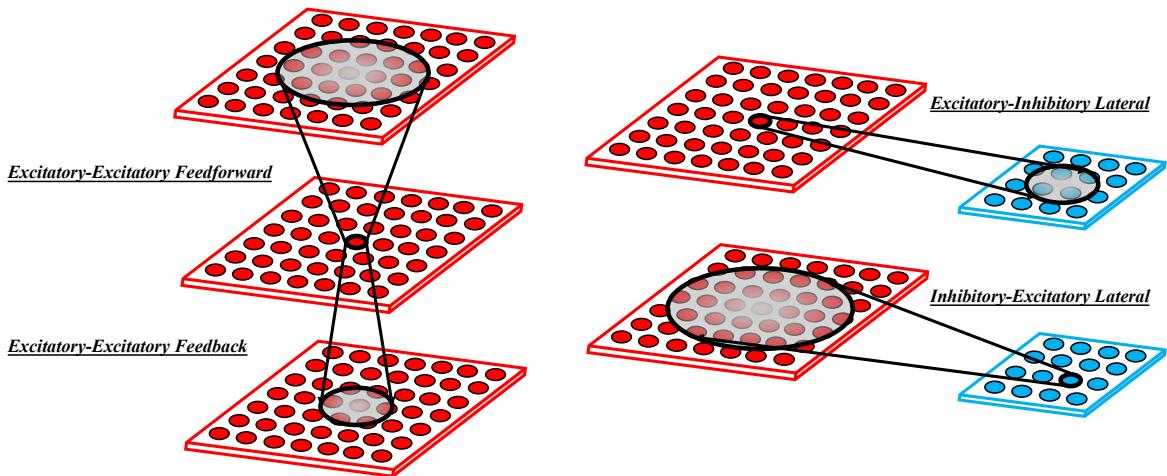
Fig 2.2: Synaptic connections among neurons in the SNN model. Red and blue circles indicate excitatory and inhibitory neurons, respectively, with presynaptic cells mark with dark edges and their postsynaptic projection radii mark with grey shaded areas.

Simple Cells. Cells in Layer 0 were modelled as excitatory simple cells (Hubel & Wiesel, 1962), which fire strongly to oriented edges in visual stimuli at specific retinal locations. When a light

impulse occurred at a retinal coordinate (x, y), the firing rate of a simple cell, f, followed a modified Gabor filter (Kruizinga & Petkov, 1999):

$$f = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \cos\left(2\pi \frac{x'}{\bar{\lambda}} + \psi\right) \tag{1}$$

with

$$\begin{cases} x' = x \cdot \cos\theta + y \cdot \sin\theta \\ y' = y \cdot \cos\theta - x \cdot \sin\theta \\ \sigma = \sqrt{\frac{\ln 2}{2}} \cdot \frac{\bar{\lambda}(2^b + 1)}{\pi(2^b - 1)} \end{cases} \tag{2}$$

where θ is the cell's preferred edge orientation, σ and γ set the geometric properties of the cell's receptive field, $\bar{\lambda}$ and b control the cell's spatial frequency, and ψ determines the phase in the cell's response. For each of the 32×32 retinal locations, a group of simple cells with four different preferred orientation was deployed ($\theta = 0°$, 45°, 90°, and 135°, respectively).

Simple cells emitted all-or-none spike trains with equally spaced spikes as output, with interspike interval, T, determined by:

$$T = \frac{f_{max}}{f} \omega \tag{3}$$

where f_{max} is the maximum f among the entire simple cell population, and ω represents the upper bound of a simple cell's firing intensity.

Neuron Equations. Excitatory and inhibitory neurons in layer 1 to 4 were modelled as leaky integrate-and-fire neurons, which replicated the biophysical properties of pyramidal cells in mammalian neocortex (Jolivet et al., 2006). At time t, the membrane potential of a neuron i, denoted as $V_i(t)$, was updated via:

$$\tau_m^\gamma \frac{dV_i(t)}{dt} = V_0^\gamma - V_i(t) + R^\gamma I_i(t) \tag{4}$$

where τ_m^γ represents the membrane time constant, V_0^γ is the resting potential of the neuron, $I_i(t)$ is the total synaptic current received by neuron i at time t, R^γ is the membrane resistance, and γ index indicates the neuron type. This formulation ensured that $V_i(t)$ passively decayed toward the resting state V_0^γ while being actively driven by synaptic inputs.

Neurons emitted all-or-none spikes. At time t, the binary spike state of neuron i, $\delta_i(t)$, was computed via:

$$\delta_i(t) = \begin{cases} 1, & \text{if } V_i(t) > \Theta^\gamma \text{ and } \Delta\delta_i(t) > \tau_R \\ 0, & \text{elsewise} \end{cases} \tag{5}$$

where Θ^γ is the threshold potential, and $\Delta\delta_i$ is the time elapsed since the neuron's last spiking event. After a spike was emitted, the neuron entered an absolute refractory period that lasted for a duration of τ_R, during which $V_i(t)$ was fixed at the neuron's hyperpolarised potential, V_H^γ, and no spiking was allowed.

The current term $I_i(t)$ in Equation 4 was defined as the sum of all excitatory and inhibitory synaptic currents injected into the neuron i:

$$I_i(t) = \sum_j \left(\hat{V}^\gamma - V_i(t) \right) g_{ij}(t) \tag{6}$$

where $\hat{V}^\gamma$ is the synaptic reversal potential, and $g_{ij}(t)$ is the synaptic conductance from presynaptic neuron j to postsynaptic neuron i, which was updated via:

$$\frac{dg_{ij}(t)}{dt} = -\frac{g_{ij}(t)}{\tau_r} + \lambda \cdot \Delta g_{ij}(t) \delta_j(t) \tag{7}$$

where τ_r is the synaptic time constant, λ is the biological scaling factor, $\Delta g_{ij}(t)$ is the synaptic efficacy, and $\delta_j(t)$ represents the presynaptic spiking event as defined in Equation 5.

Unsupervised learning. Spike-timing dependent plasticity was applied to modify excitatory-excitatory synaptic efficacies based on the temporal correlation between presynaptic and postsynaptic spiking. Following Perrinet et al. (2001), this rule was implemented using the updating equation:

$$\frac{d\Delta g_{ij}(t)}{dt} = \rho \left[\left(1 - \Delta g_{ij}(t) \right) \cdot C_{ij}(t) \delta_i(t) - \Delta g_{ij}(t) \cdot D_i(t) \delta_j(t) \right] \tag{8}$$

where ρ is the learning rate, $C_{ij}(t)$ quantifies presynaptic activity, and $D_i(t)$ quantifies postsynaptic activity. The value of $\Delta g_{ij}(t)$ was clamped between [0, 1] to ensure numerical stability.

$C_{ij}(t)$ and $D_i(t)$ in Equation 8 were updated by:

$$\frac{dC_{ij}(t)}{dt} = -\frac{C_{ij}(t)}{\tau_C} + \alpha_C \left(1 - C_{ij}(t)\right) \cdot \delta_j(t) \tag{9}$$

and

$$\frac{dD_i(t)}{dt} = -\frac{D_i(t)}{\tau_D} + \alpha_D \left(1 - D_i(t)\right) \cdot \delta_i(t) \tag{10}$$

where τ_C and τ_D are the presynaptic and postsynaptic time constants, and α_C and α_D are proportionality constants that describe the presynaptic and postsynaptic learning resources. Conceptually, $C_{ij}(t)$ represented the concentration of glutamate neurotransmitter residue in the synaptic cleft, and $D_i(t)$ represented the number of unblocked glutamate receptors on the postsynaptic membrane. Both variables decayed passively and were activated asymptotically by presynaptic or postsynaptic spiking.

Equation 8 to 10 ensured that changes in synaptic efficacy were governed by spike timing. When a presynaptic neuron fired shortly before a postsynaptic neuron, the presynaptic term involving $C_{ij}(t)$ in Equation 8 dominated, which increased $\Delta g_{ij}(t)$. Conversely, when a postsynaptic neuron fired before a presynaptic neuron, the postsynaptic term involving $D_i(t)$ dominated, which decreased $\Delta g_{ij}(t)$. The magnitude of change in $\Delta g_{ij}(t)$ was inversely proportional to the interval between presynaptic and postsynaptic spikes. Hence, a shorter spiking interval, indicating a stronger neuronal correlation, produced a greater learning effect.

All neuronal and synaptic parameters were summarised in Appendix I, with values adopted from available neurophysiological measurements of the mammalian visual cortex (Perrinet et al., 2001; Troyer et al., 1998; Eguchi et al., 2018).

2.2 Visual Stimuli

I designed a set of intrinsically structured, object-like visual stimuli to investigate feature binding in the SNN model. Each stimulus consisted of a central body and several bilaterally symmetrical limbs (Fig 2.3a), with its shape defined by three orthogonal feature dimensions (Fig 2.3b):

- *Thorax length*, the distance between the branching points of front and hind limbs

- *Arm length*, the extension distances of middle limbs

- *Leg angle*, the branching angle of the front and hind limbs
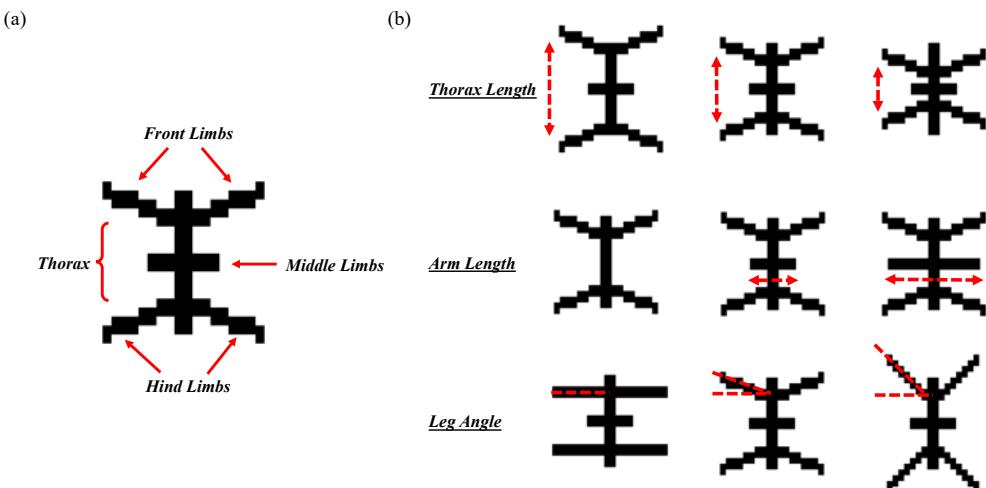
**Fig 2.3: (a) Overall shape of the designed object-like visual stimulus and (b) its three underlying feature dimensions.**

This unique shape configuration encouraged the model to learn the abstract geometric relationships among different parts of an object without localised spatial cues, a core challenge of the binding problem. I generated each stimulus as a 32×32-pixel binary image with a 2-pixel

stroke width. To avoid undesired pixel overlaps, three representative values along each feature dimension were selected to produce a set of 27 object-like visual stimuli for subsequent training and testing (Fig 2.4). All stimuli were generated using NumPy v2.2.0 in Python 3.12.

Fig 2.4: 27-object stimulus set generated for model training and testing. The selected values from each dimension: [*16, 11, 5 pixels*] for thorax length, [*0, 8, 16 pixels*] for arm length, and [*0°, 22.5°, 45°*] for leg angle.

2.3 Model Execution

I built the SNN model using the Numba CUDA v0.61.0 (Lam et al., 2015) in Python 3.12 and executed it on GPU-accelerated (NVIDIA A100-SXM) clusters provided by the Cambridge Center for Data Driven Discovery. Before training, excitatory-excitatory synaptic efficacies were randomised by sampling uniformly from [0, 1], and all inhibitory synaptic efficacies were fixed at 1.

I trained the model on a total of 5,400 stimuli, each randomly selected, with replacement, from the designed object-like visual stimuli. Each stimulus was displayed at a randomised retinal location for 2,000 ms, during which the model was updated according to the aforementioned equations, with differential equations solved discretely using the forward Euler method with a timestep of $\delta t = 0.1$ ms. After each image presentation, all neuronal and synaptic activities were reset to their resting states, except the updated synaptic efficacies.

I evaluated neurons in the model over eleven testing sessions: one before training, one after training, and nine intermediate sessions during training (once per 540 stimuli trained). Each visual stimulus in the stimulus set was displayed at the centre of retina for 1,000 ms with synaptic plasticity disabled, during which the firing rates of layer 4 excitatory neurons, the model's output units, were recorded for subsequent analyses.

2.4 Neural Tuning Analysis

I applied information theory to analyse neural tuning properties. All tuning analyses were conducted using NumPy v2.2.0 and SciPy v1.15 in Python 3.12.

Following Luczak (2024), I quantified a neuron's selectivity for a given feature dimension g (e.g., *leg angle*) via the information entropy, H_g, carried by its firing responses:

$$H_g = -\sum_{n \in g} P(r|n) \cdot \log_2 P(r|n) \tag{11}$$

where n denotes a specific feature value along g (e.g., *45°*), and $P(r|n)$ is the conditional probability of firing (i.e., the neuron's normalised average firing rate) given that feature n is displayed. A neuron carries more information about g should exhibit a more unevenly distributed firing rate along g, thus yielding a lower H_g.

The neuron's normalised feature selectivity, λ_g, was then computed via:

$$\lambda_g = 1 - \frac{\lambda_g}{\log_2 \|g\|} \tag{12}$$

22

where $\|g\| = 3$ is the number of distinct feature values along g. Since $\log_2 \|g\|$ is the theoretical upper bound of H_g, λ_g ranges from 0 (no selectivity) to 1 (maximum selectivity).

To quantify the number of feature dimensions to which a neuron is selective for, I defined a neuron's *effective dimensionality*, D_{eff}, as the scale-independent participation ratio (Recanatesi et al., 2022) of its λ_g across all three orthogonal feature dimensions ($g \in \{1, 2, 3\}$):

$$D_{\text{eff}} = \frac{(\sum_{g=1}^3 \lambda_g)^2}{\sum_{g=1}^3 (\lambda_g)^2} \tag{13}$$

To measure a neuron's absolute sensitivity to its preferred feature values, I defined a neuron's *tuning strength*, ε, as its largest normalised feature selectivity across the three feature dimensions in the stimulus set:

$$\varepsilon = \max_g(\lambda_g) \tag{14}$$

By definition, D_{eff} ranges from 1 (selective for a single feature dimension) to 3 (selective for all three feature dimensions) and ε ranges from 0 (zero sensitivity to visual information) to 1 (maximum sensitivity to visual information). Thus, the expected tuning profiles of idealised neurons were:

- *Partial conjunctional neurons*: $D_{\text{eff}} = 2$ and $\varepsilon = 1$

- *Feature-selective neurons*: $D_{\text{eff}} = 1$ and $\varepsilon = 1$

- *Object-selective neurons*: $D_{\text{eff}} = 3$ and $\varepsilon = 1$

- *Untuned neurons*: $\varepsilon = 0$

To categorise neural tuning emerged in the SNN model, I adopted the following classification boundaries:

- *Partial conjunctional neurons*: $1.5 < D_{\text{eff}} \leq 2.5$ and $\varepsilon > 0.2$

- *Non-partial conjunctional neurons*: $(D_{\text{eff}} \leq 1.5$ or $D_{\text{eff}} > 2.5)$ and $\varepsilon > 0.2$

- *Sharply tuned neurons*: $\varepsilon > 0.2$

- *Untuned neurons*: $\varepsilon \leq 0.2$

2.5 Representational Similarity Analysis

I applied representational similarity analysis (Kriegeskorte et al., 2008) to evaluate whether a neural population explicitly encodes feature combinations in visual stimuli. I first computed the pairwise representational dissimilarities among visual stimuli and among their elicited neural responses. For a pair of stimuli $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$, its stimulus dissimilarity, S_{ij}, was defined as:

$$S_{ij} = \frac{\text{H}_d(\boldsymbol{f}_i, \boldsymbol{f}_j)}{3} \qquad (15)$$

where H_d, the Hamming distance, counts the number of unshared feature values, which ranges from 1 (one feature unshared) to 3 (all features unshared). Similarly, for the corresponding neural responses $\boldsymbol{r}_i$ and $\boldsymbol{r}_j$, its neural dissimilarity, R_{ij}, was defined as:

$$R_{ij} = \frac{1 - \text{Corr}(\boldsymbol{r}_i, \boldsymbol{r}_j)}{2} \qquad (16)$$

where $\text{Corr}(\boldsymbol{r}_i, \boldsymbol{r}_j)$ indicates Pearson correlation coefficient between $\boldsymbol{r}_i$ and $\boldsymbol{r}_j$. All S_{ij} and R_{ij} ranged from 0 to 1.

I then assembled all S_{ij} and R_{ij} values into representational dissimilarity matrices (RDMs), which were 27×27 symmetrical matrices with each row and column indexing a specific visual stimulus or its elicited neural response. The embedding spaces of the visual stimuli or the neural responses were visualised by projecting their RDMs onto two-dimensional maps using the SMACOF algorithm (Kruskal, 1964) with 10 different initialisations and 300 maximum iterations. The similarity score between the stimulus set and the targeted neural population were quantified by Spearman's rank-order correlation across the off-diagonal entries of their RDMs.

All representational similarity analyses were conducted using NumPy v2.2.0 in Python 3.12, with the SMACOF algorithm applied using scikit-learn v1.16 in Python 3.12.

2.6 Statistical Analysis

I tested hypotheses regarding pairwise group-level differences in mean using one-way repeated measures ANOVA with Bonferroni's *post hoc* comparisons. Assumptions for ANOVA were first verified [1], and if violated, nonparametric Friedman's test with Conover's *post hoc* comparisons was conducted instead. Hypotheses regarding correlations were tested using Spearman rank-based correlation analysis. A significant level of $\alpha = 0.05$ was applied to all statistical tests. All statistical analyses were conducted in JASP v0.18.3 (JASP Team, 2019).

---

[1] Assumption of normality tested by Shapiro-Wilk test, homogeneity of variance tested by Levene's test, and sphericity tested by Mauchly's test.

3. Spiking Neural Network Model: Results

3.1 Emergence of Partial Conjunctional Tuning

Do partial conjunctions emerge spontaneously through unsupervised synaptic learning processes? I trained the SNN model by exposing it to 5,400 object-like visual stimuli. Over the course of training, synaptic efficacies of excitatory-excitatory connections rapidly evolved from a uniform distribution ($M = 0.501$, $SD = 0.289$) to a symmetrical peak ($M = 0.503$, $SD = 0.210$), with both its mean and standard deviation roughly converged to stable values, albeit with fluctuations, indicating a reasonably well-behaved learning process (Fig 3.1).
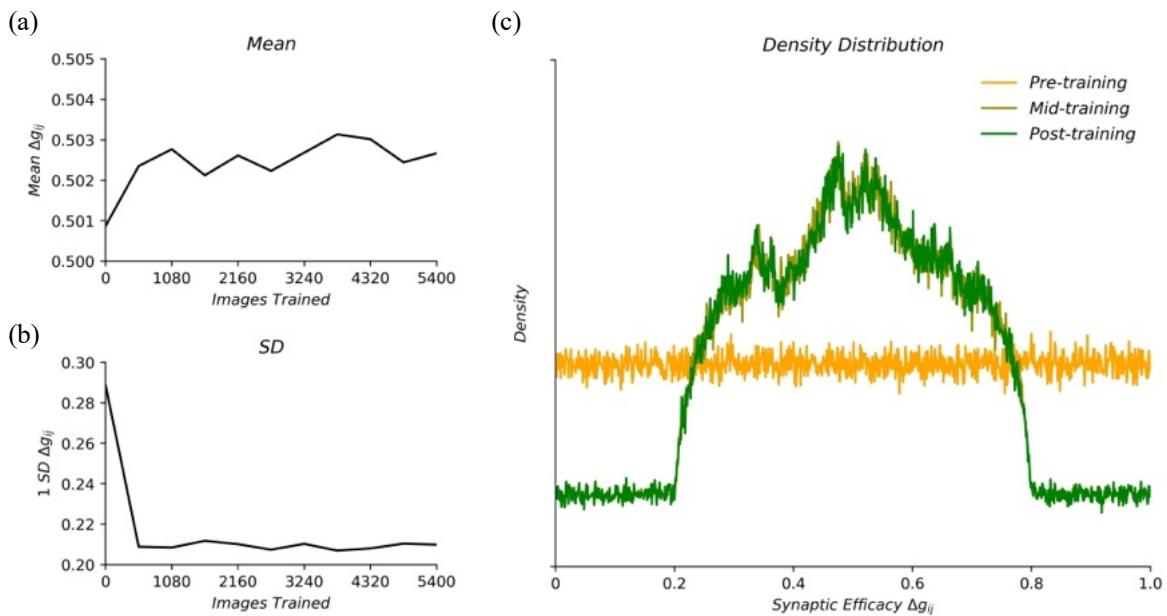

Fig 3.1: Variations in excitatory-excitatory synaptic efficacies Δg_{ij} across training, including their (a) mean, (b) standard deviation, and (c) distribution. *Pre-training* = 0 images trained. *Mid-training* = 2700 images trained. *Post-training* = 5,400 images trained.

To examine whether partial conjunctions emerged through the model's intrinsic plasticity, I analysed neurons' effective dimensionality D_{eff} and tuning strength ε before, midway, and after training. As training progressed, the distribution of D_{eff} shifted from a right-skewed curve to a concentrated peak around $D_{\text{eff}} \approx 2.3$, approaching the expected $D_{\text{eff}} = 2.0$ for ideal partial

conjunctions (Fig 3.2). Consistent with this observation, the population mean of D_{eff} varied significantly across training, despite with a small effect size (Friedman test: $\chi^2(2) = 60.8$, $p < .001$, Kendall's $W = 0.03$), which decreased non-significantly from pre- to mid-training and then further decreased significantly by post-training (Table 1.1a). The distribution of ε, in contrast, remained heavily skew towards $\varepsilon = 0$ throughout training, far from the expected $\varepsilon = 1$ for sharply tuned neurons (Fig 3.2b). Although significant variations in the population mean of ε were detected (Friedman test: $\chi^2(2) = 228.8$, $p < .001$, Kendall's $W = 0.11$), it first decreased from pre- to mid-training and then increased by post-training, resulting in a nonsignificant net change across training (Table 1.1b). This consistently low ε was likely due to the dense lateral inhibitory connections within each cortical layer, which prevents the bulk formation of sharply tuned neurons. Together, these findings reveal the emergence of population-wide, low-strength, partial conjunctional tuning in the SNN model.

(a)

(b)
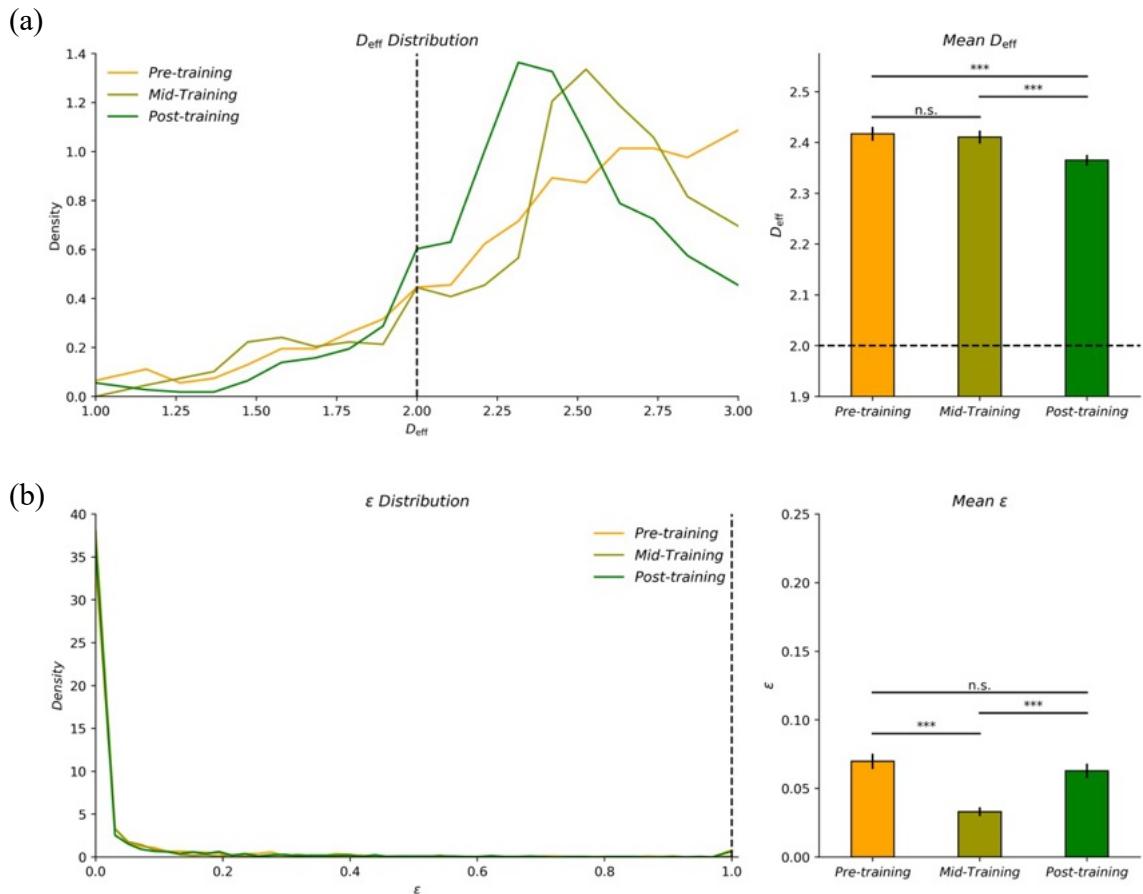
Fig 3.2: Distributions and population mean of (a) effective dimensionality D_{eff} and (b) tuning strength ε for neurons before, midway, and after training. Distributions estimated using 20 evenly spaced histogram-bins. *Pre-training*: 0 images trained. *Mid-training*: 2700 images trained. *Post-training*: 5,400 images trained. Dotted lines indicate the expected D_{eff} and ε for an ideal partial conjunctional neural population. Error bars indicate $\pm\ 1\ SEM$. *** represents $p < .001$.

Table 1.1a: Comparisons of D_{eff} before, midway, and after the training.

Condition	Summary		Pairwise vs. Pre-training		Pairwise vs. Mid-training	
	M	*SD*	*t*	*p_bonf*	*t*	*p_bonf*
Pre-training	2.43	0.44	/	/	/	/
Mid-training	2.41	0.41	1.91	0.17	/	/
Post-training	2.36	0.35	7.50	<u>*<.001*</u>	5.59	<u>*<.001*</u>

Pairwise results from Conover's *post hoc* comparisons. *p_bonf*: *p*-value after Bonferroni correction.

**Table 1.1b: Comparisons of ε before, midway, and after the training.**
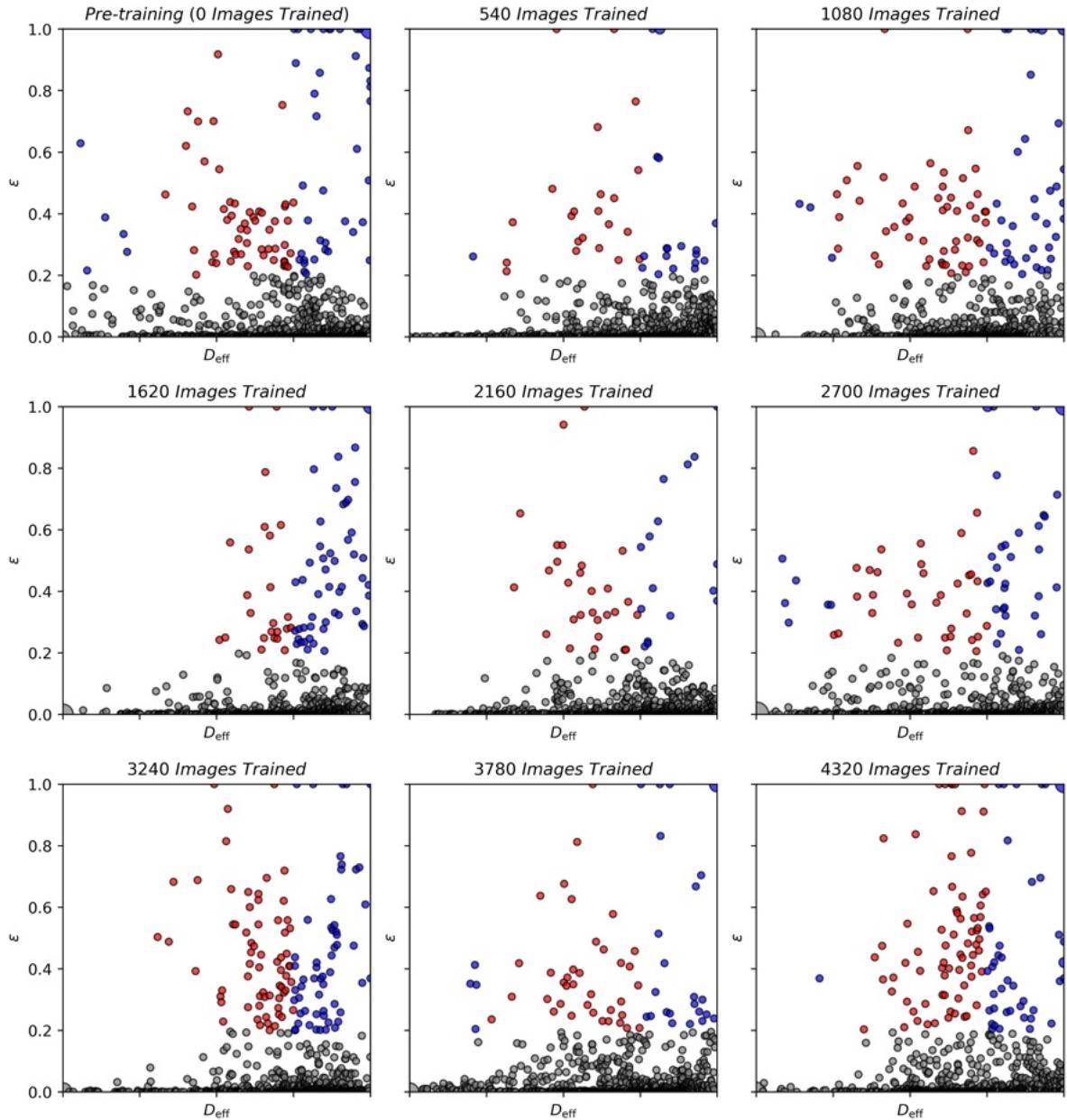
Condition	Summary		Pairwise vs. Pre-training		Pairwise vs. Mid-training	
	M	*SD*	*t*	*p_bonf*	*t*	*p_bonf*
Pre-training	0.07	0.18	/	/	/	/
Mid-training	0.03	0.10	13.1	<u>*<.001*</u>	/	/
Post-training	0.06	0.17	0.01	1.00	13.1	<u>*<.001*</u>

Pairwise results from Conover's *post hoc* comparisons. *p_bonf*: *p*-value after Bonferroni correction.

3.2 Formation of Shapely Tuned Partial Conjunctions

To further investigate the emergence of sharply tuned partial conjunctions in the model, I assessed each neuron's tuning profile during training across eleven evenly spaced testing sessions. As shown in Fig 3.3, most neurons remained untuned throughout training, but a subset of neurons acquired high ε values and gradually clustered between $D_{\text{eff}} \approx 2.0$ and 3.0. I then classified these sharply tuned neurons as partial or non-partial conjunctions based on their D_{eff} values and tracked their population sizes as training progressed (Table 1.2). Both populations initially declined from $n \approx 50$ to $n \approx 20$, likely due to an early synaptic pruning event. Subsequently, the number of partial conjunctions rebounded and peaked at $n = 71$, while the

number of non-partial conjunctions oscillated between $n = 16$ and $n = 61$ (Fig 3.4a). As a result, while the total number of sharply tuned neurons in the model remained roughly unchanged at $n \approx 100$ throughout training, the proportion of partial conjunctions among them increased from 51% to 58% across training, despite with large fluctuations (Fig 3.4b). Together, these findings demonstrate the gradual, yet unstable, emergence of sharply tuned partial conjunctions in the SNN model.
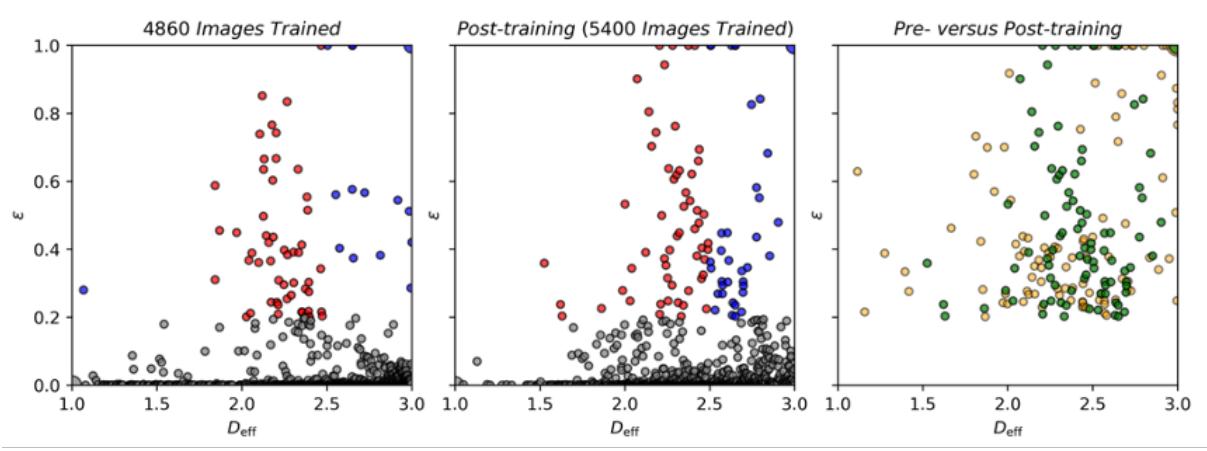
Fig 3.3: Neurons' tuning profiles during training measured across eleven testing sessions. Red, blue, and grey datapoints respectively indicate neurons classified as sharply tuned partial conjunctions, sharply tuned non-partial conjunctions, and untuned neurons. In the bottom right subplot, orange and green datapoints respectively indicate all sharply tuned neurons pre- and post-training. Marker sizes reflect the number of overlapping neurons that shared the same tuning profile.

Table 1.2: Number and relative proportion of sharply tuned partial conjunctions, sharply tuned non-partial conjunctions, and untuned neurons during training.

Images Trained	Partial Conjunctions		Non-partial Conjunctions		Untuned Neurons	
	n	*%*	*n*	*%*	*n*	*%*
0	53	5.12	51	4.98	920	89.8
540	22	2.15	20	1.95	982	95.9
1080	52	5.78	43	4.20	929	90.7
1620	24	2.34	55	5.37	945	92.3
2160	28	2.73	16	1.56	980	95.7
2700	34	3.32	43	4.20	947	92.5
3240	65	6.35	56	5.47	903	88.2
3780	37	3.61	31	3.03	956	93.4
4320	71	6.93	48	4.69	905	88.4
4860	52	5.01	18	1.76	954	93.2
5400	56	5.47	40	3.91	928	90.3

*n*: Raw counts. *%*: Proportion among all neurons ($n = 1024$) across the cortical layer.
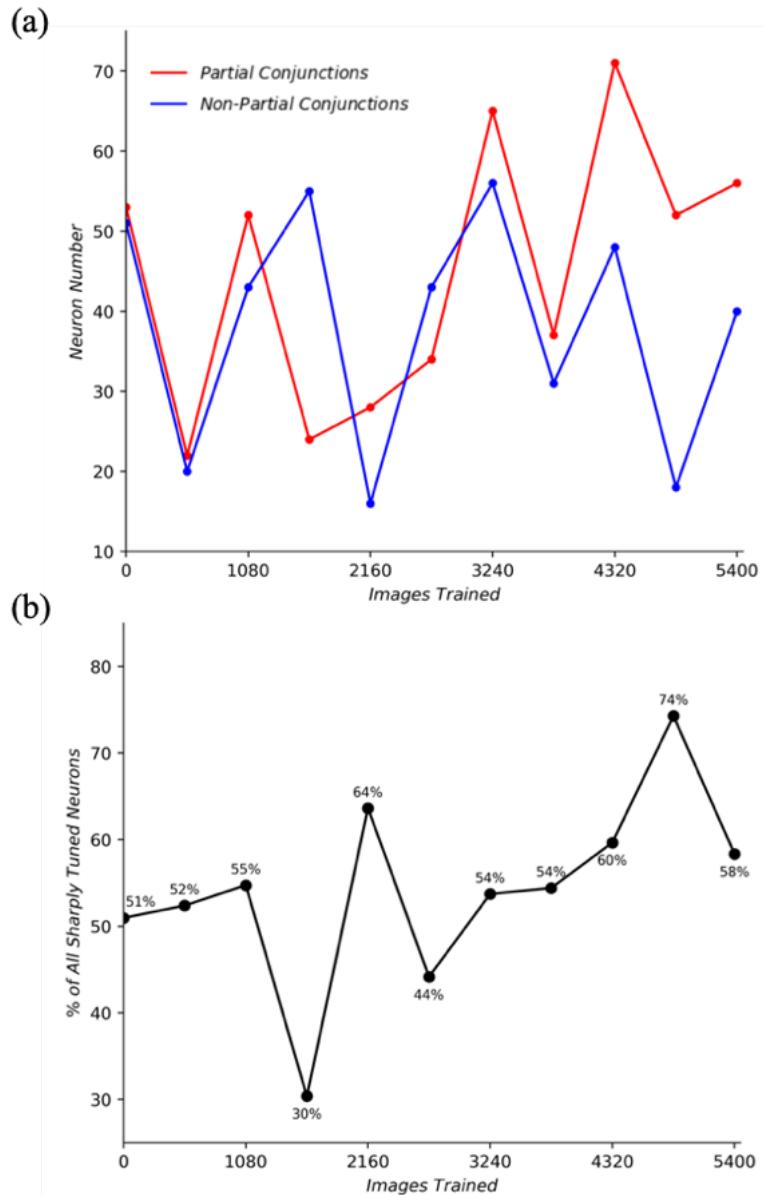
Fig 3.4: (a) Absolute size and (b) relative proportion of the partial conjunction population among all sharply tuned neurons as training progressed. Red and blue curves indicate the number of partial and non-partial conjunctions, respectively.

3.3 Feature Representations among Partial Conjunctions

Can the emerged partial conjunctions explicitly encode feature combinations in the object-like visual stimuli? To address this question, I applied representational similarity analysis to assess whether partial conjunctional neural responses directly reflect the stimulus set's underlying feature space.

To qualitatively compare the representational similarities between visual stimuli and the recorded neural responses, I computed the pairwise stimulus and partial conjunctional neural dissimilarities, constructed their respective RDMs, and projected them onto two-dimensional embedding spaces via multidimensional scaling. While the stimulus representations formed evenly distributed subgroups, reflecting the continuous variation in their underlying feature combinations (Fig 3.5a), the corresponding partial conjunctional representations formed tightly spaced clusters with no clear grouping pattern (Fig 3.5b). This discrepancy indicates a substantial mismatch between the stimulus set and partial conjunctional neural responses in representational geometries.
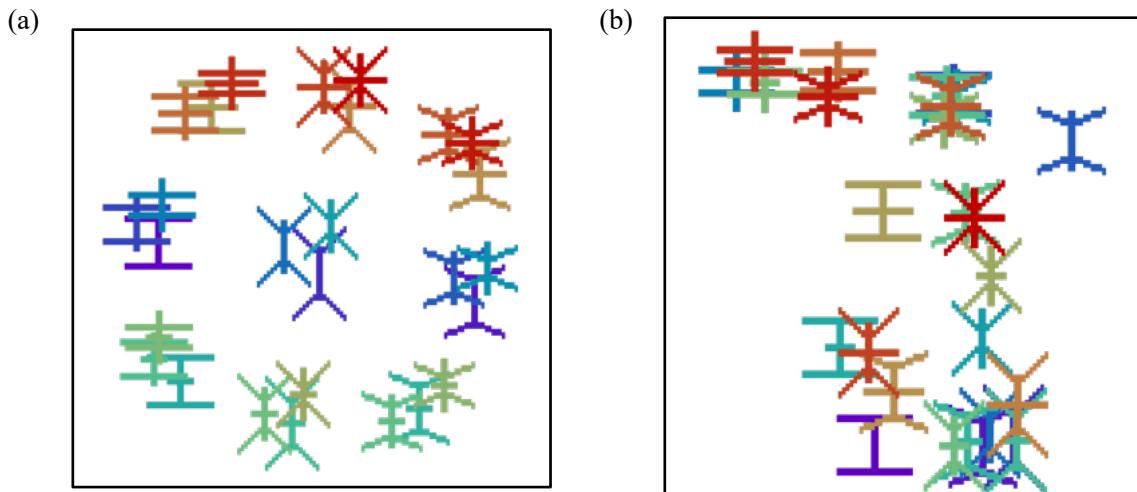

Fig 3.5: Two-dimensional embedding spaces of (a) visual stimuli and (b) their elicited partial conjunctional neural responses. Falsely coloured items indicate stimulus representations in (a) and neural representations in (b), with distances between items reflect pairwise dissimilarities.

To further assess the representational similarity between the stimulus set and partial conjunctional neural responses, I computed the rank-based similarity score between the stimulus RDM (Fig 3.6a) and the partial conjunctional neural RDM (Fig 3.6b). For comparison, I also included RDMs for non-partial conjunctions, untuned neurons, and the combined whole

cortex in the analysis. All neural RDMs exhibited statistically significant yet modest correlations with the stimulus RDM (Spearman's rank-order correlation: $\rho = 0.16$ to 0.29, $p < .003$; Fig 3.6c and Table 1.3), indicating that each neural population in the cortex shares, to some extent, representational similarity with the stimulus set. Notably, the partial conjunctional neural RDM exhibited the highest similarity score with the stimulus RDM, while its counterpart, the non-partial conjunctional neural RDM, exhibited the lowest similarity score, suggesting that partial conjunctional neural responses align most closely with the stimulus set's underlying feature space. However, even this highest score is still low in absolute magnitude and only marginally exceeds that of the other neural populations. Combined with the qualitative mismatch in the embedding space, all results indicate that the emerged partial conjunctions do not explicitly encode feature combinations in visual stimuli.
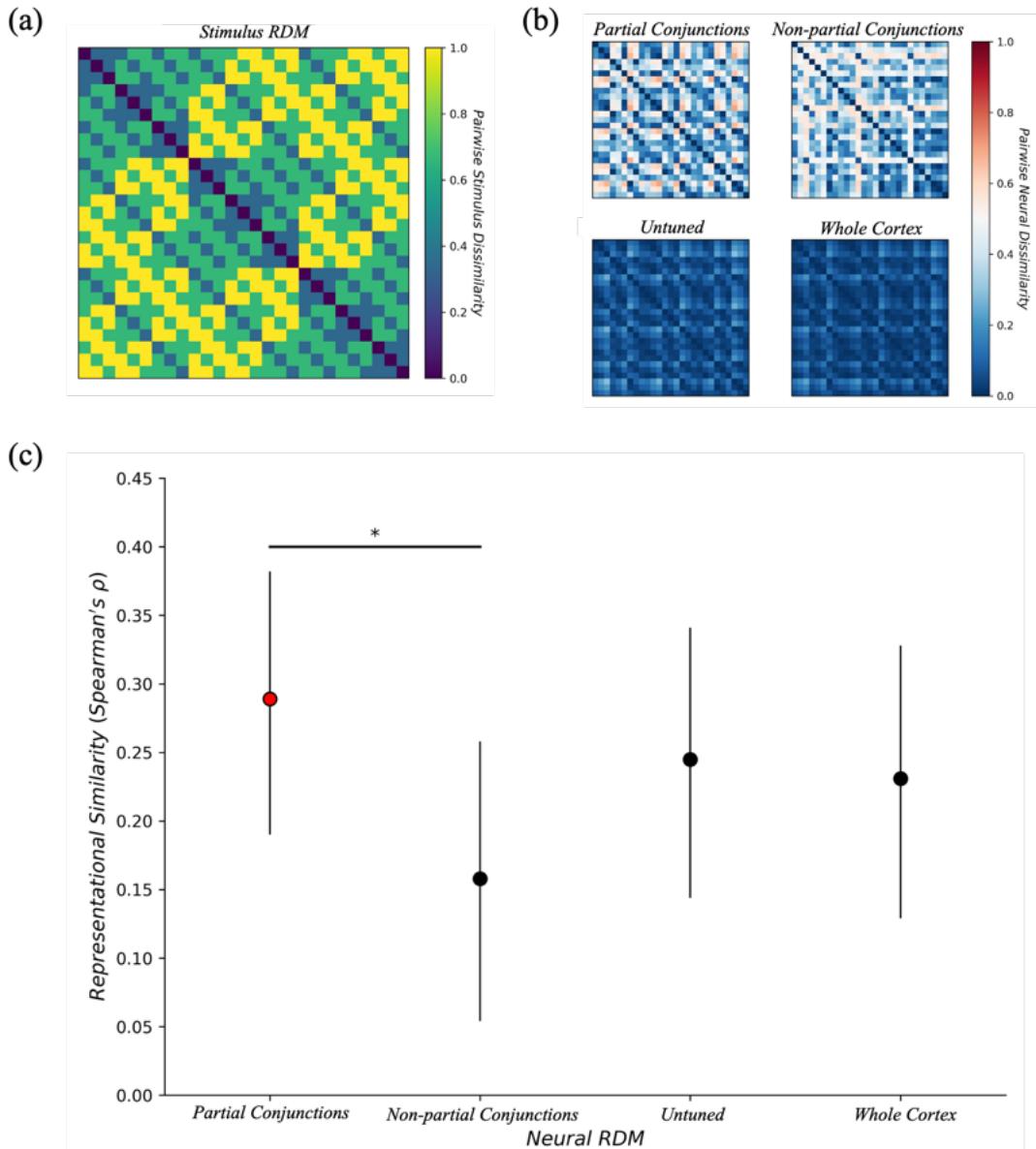
Fig 3.6: (a) Stimulus and (b) neural RDMs constructed, with (c) the similarity score (Spearman correlation coefficient) between the stimulus RDM and each neural RDM. Error bars indicate 95% confidence intervals. * represents pairs with non-overlapping confidence intervals.

Table 1.3: Representational similarity score between the stimulus RDM and each neural RDM.

Neural RDM	Spearman's ρ	95% Confidence Interval	p
Partial Conjunctions	0.29	[0.19, 0.38]	<.001
Non-partial Conjunctions	0.16	[0.05, 0.26]	<.003
Untuned	0.25	[0.14, 0.34]	<.001
Whole Cortex	0.23	[0.13, 0.33]	<.001

4. Residual Neural Network Model: Method

The SNN model revealed preliminary evidence for the spontaneous emergence of partial conjunctions, yet their population sizes and tuning properties were highly unstable, and their representations did not explicitly align with the visual stimuli. Several factors may account for this instability:

- Lack of a global objective to constrain learning. Training large-scale SNNs via local spike-timing dependent plasticity is known to be numerically unstable and often leads to model divergence (van Hemmen, 1997; Pfeiffer & Pfei, 2018).

- Unnaturalistic visual stimuli. The object-like visual stimuli used for training were highly abstract and did not provide clearly separable feature dimensions, which limited the robustness of feature-based learning.

- Limitations of tuning analysis. The information-theoretic approach of tuning analysis collapsed neural firing responses onto each individual feature dimensions, which ignored higher-order interactions among dimensions and made the analysis hypersensitive to small variations in firing rates.

To address these limitations, I implemented an alternative ResNet model with a new stimulus set, an updated training protocol, and a revised tuning analysis to further investigate the emergence of partial conjunctions under supervised learning.

4.1 Model Architecture

Following He et al. (2016), I implemented a modified 18-layer ResNet to simulate the human visual system. The model consisted of an input convolutional layer, a max-pooling layer, four

layers of residual blocks with progressively increasing channel depth, a global average pooling layer, and a fully connected classification head (see Fig 4.1 for the detailed architecture). To adapt the model for grayscale image inputs and the binary classification tasks applied in this study, I modified the original architecture by configurating the first convolutional layer to accept grayscale inputs, reducing its kernel size to 5×5, and replacing the final fully connected layer with a two-unit classifier. These adjustments ensured that the model could perform hierarchical feature extraction under specific task constraints, with earlier layers capturing basic visual features and later layers encoding complex feature combinations.
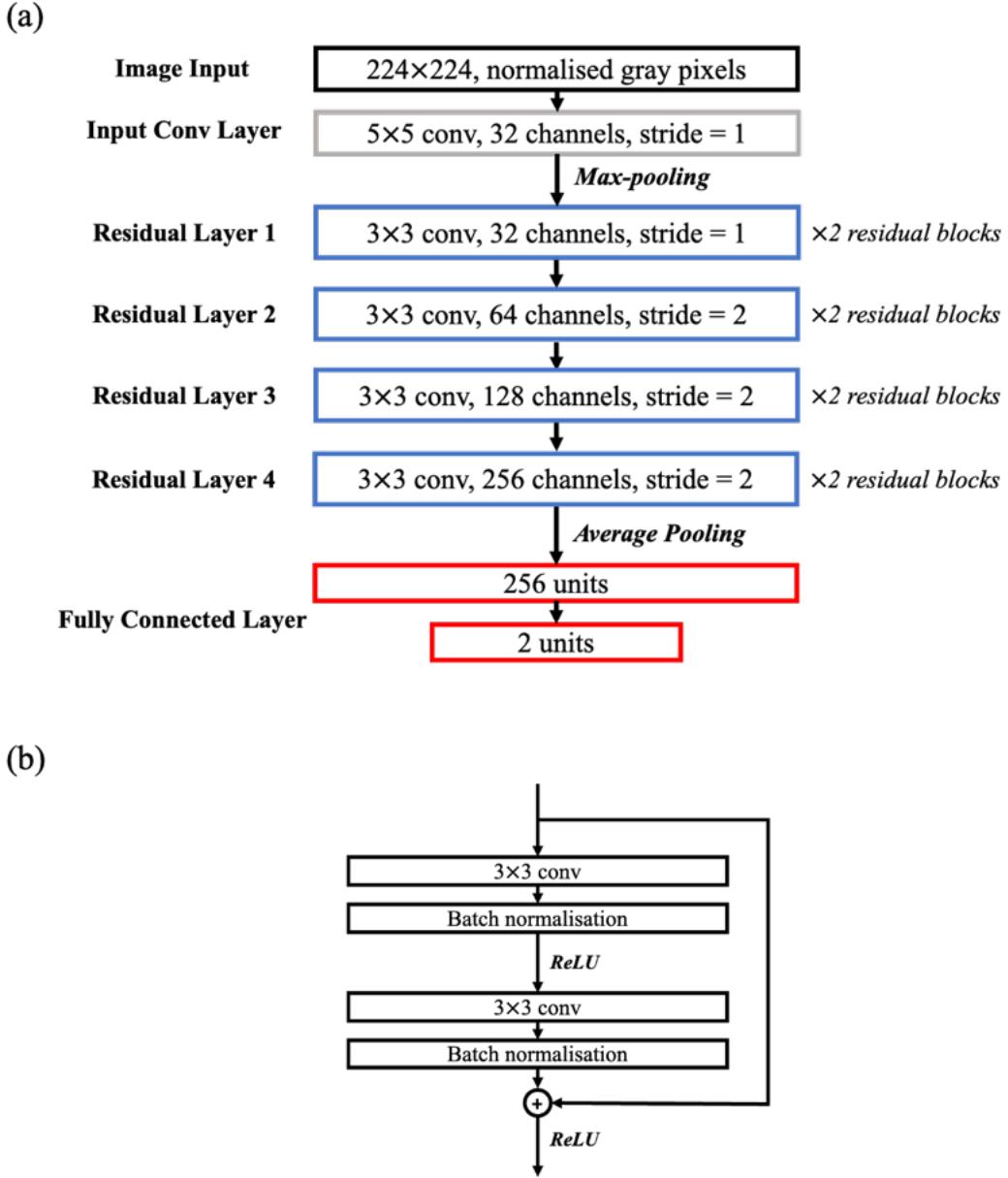
Fig 4.1: (a) Overall architecture of the modified ResNet model used in this study, and (b) the detailed structure of a residual block. Relative to the original ResNet proposed by He et al. (2016), this model has several adaptations: a 5×5 instead of a 7×7 input layer convolutional filter, a single grayscale channel instead of three RGB input channels, and a reduced fully connected classification head with only two output units.

4.2 Visual Stimuli

I designed a set of naturalistic insect-like stimuli to investigate feature binding in structurally complex multi-part objects. Similar to those used for the SNN, each stimulus was defined by

three orthogonal feature dimensions: the antenna, the head, and the limb. Each part could take two possible forms, which resulted in binary variation along each feature dimension (Fig 4.2). To control low-level image properties and pixel counts across stimuli, each binary feature variant was created as the rotated conjugate of its counterpart. By permutating along the three possible feature dimensions, I generated a complete set of 8 unique stimulus for model training, each rendered as a 224×224-pixel binary image using NumPy v2.2.0 in Python 3.12.
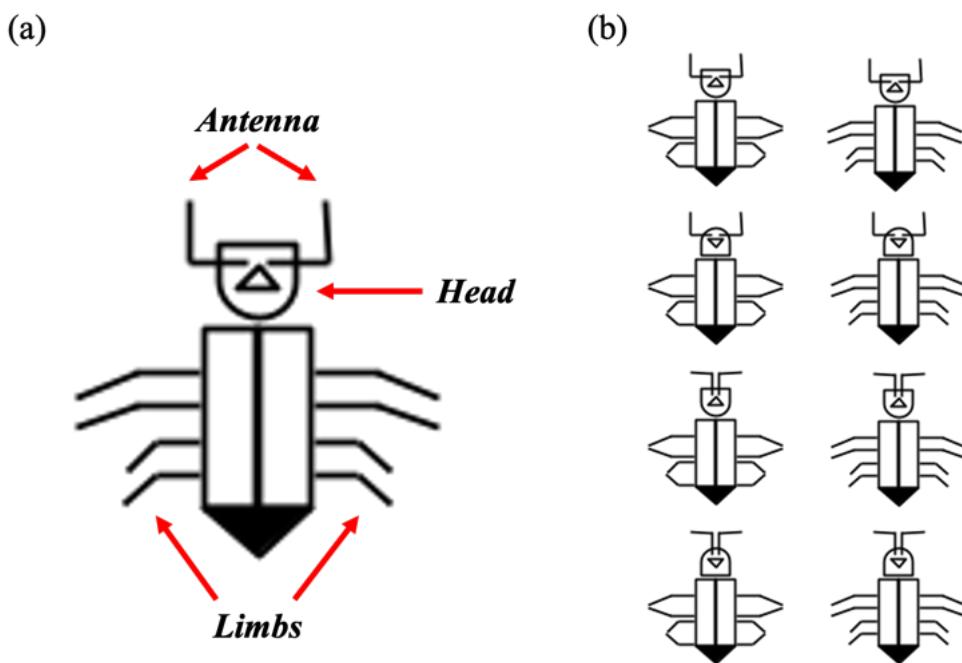

Fig 4.2: (a) Structure of the designed naturalistic insect stimulus, with three orthogonal feature dimensions (head, body, tail). (b) The full naturalistic insect stimulus set generated for model training and tuning analysis.

4.3 Model Training

Classification task. I trained the ResNet model on two image classification tasks derived from the insect stimulus set. In the *perceptual task* (Fig 4.3a), training labels were defined solely by a single feature dimension, which served as a baseline measure of classification performance without binding demands. In the *binding task* (Fig 4.3b), training labels were defined by

specific combinations of each insect's antenna, head, and limb dimensions, which introduced

an implicit binding requirement that forced the model to learn and integrate different parts.
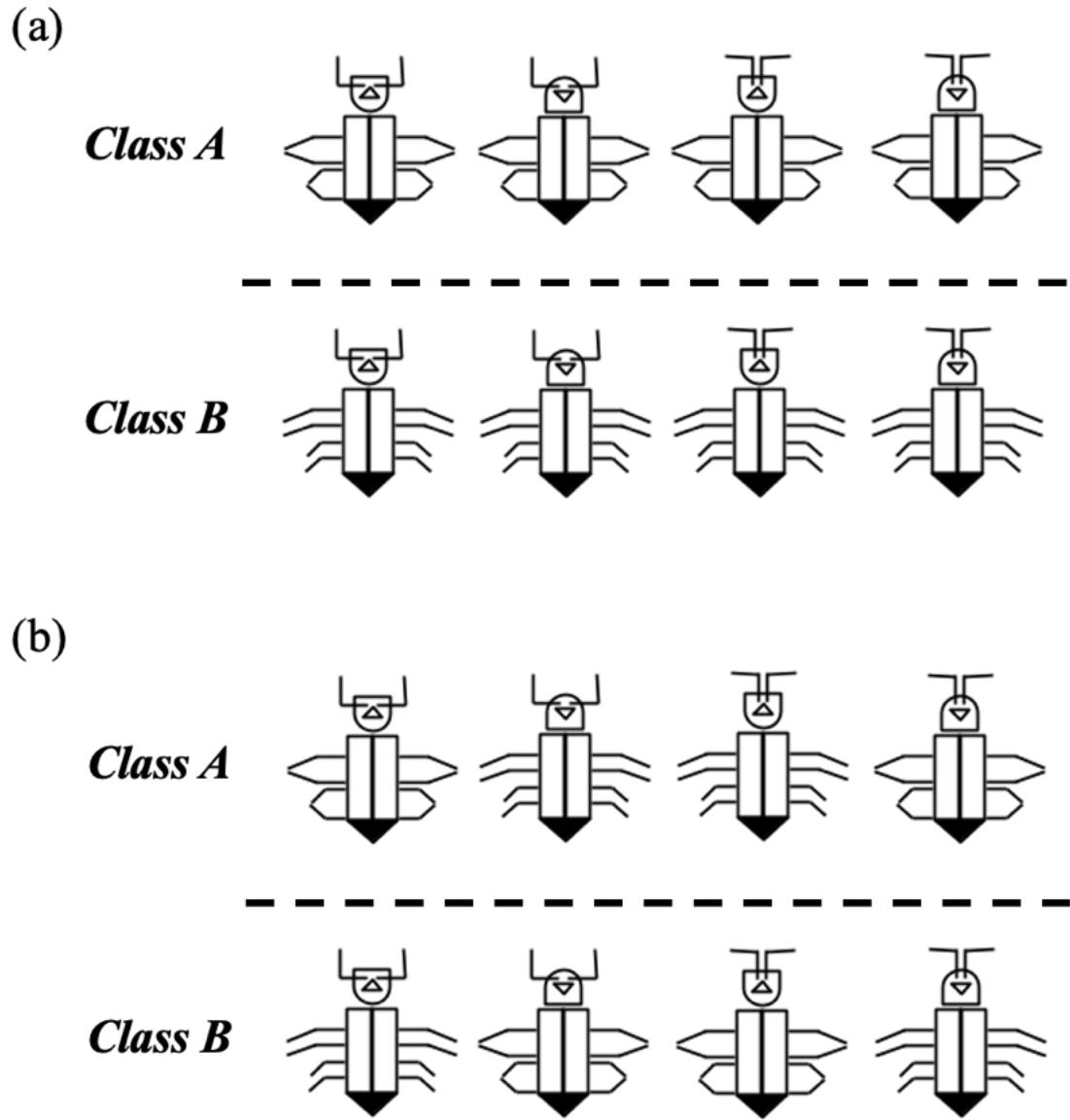

Fig 4.3: Classification rules for (a) the perceptual task and (b) the binding task. For the perceptual task, labels were assigned based on one specific feature dimension. In the above example, insects with the same limb configuration were grouped into the same class. For the binding task, labels were assigned using an alternative labelling scheme, where insects differing along an even number of feature dimensions were grouped into the same class. This design ensured that the classification task cannot be completed based on any single feature dimension or trivial local image properties.

Model Training. I built the ResNet model using PyTorch v2.7.1 (Paszke et al., 2019) in Python 3.12 and trained it on an Apple Silicon M2 Pro via the MPS backend (Apple Inc., 2022). Before training, the model was initialised with random weights and configured to process grayscale inputs normalised to the range [–1, 1].

Each training run was conducted for 50 epochs in total. In every epoch, the full stimulus set was presented repeatedly for 5 iterations in a randomised order, with each insect presented randomly jittered across the 224×224-pixel image field. The model was optimised using stochastic gradient descent with a Nesterov momentum of 0.9, a learning rate of 1×10^{-5}, a batch size of 8, and cross-entropy loss as the objective function. After each training epoch, the model's performance was validated on the original stimulus set without spatial jittering. Loss and classification accuracy on the training and validation sets were tracked across epochs to monitor learning progress. Final classification performance was quantified using receiver-operating-characteristic (ROC) analysis, summarized by the area under the ROC curve (AUC; Bradley, 1997).

For both classification tasks, the full training, validation, and analysis run was repeated 10 times with randomised seeds to obtain robust performance estimation and tuning analysis.

4.4 Tuning Analysis

I applied variance partitioning to analyse the tuning properties of units in the ResNet model. All analyses were conducted using scikit-learn v1.7.1 in Python 3.12.

Unit activity recording. After each training and validation run, I presented the original stimulus set, without spatial jittering, to the model. From the input convolutional layer and the residual

layer 1 to 4, I recorded each convolutional channels' activation map and reduced it to a 1×1 representation by average pooling, such that every convolutional channel corresponded to a single unit in the subsequent analyses. These units' responses were interpreted as approximations of neural firing rates. For each unit, I obtained 8 activation values, one for each object in the stimulus set.

Variance partitioning. I used ANOVA-based variance partitioning with Type-III sums of squares (McFarquhar, 2016) to decompose each unit's response variance into *feature-selective*, *partial conjunctional*, and *object-selective* tuning components. The relative magnitudes of these components provide a quantitative characterisation of each unit's tuning properties.

For each unit, I modelled variance in its responses to the visual stimuli using a factorial design matrix $X \in \mathbb{R}^{8 \times 8}$ of the form:

$$X = [\mathbf{1}, \quad H, \quad B, \quad T, \quad H \cdot B, \quad B \cdot T, \quad H \cdot T, \quad H \cdot B \cdot T] \qquad (17)$$

where H, B, and T denote the head, body, and tail dimensions, respectively, coded using -1 or 1 to represent circle or cross. Column 1 encodes the intercept or the baseline activation. Columns 2 to 4 encode the main effect, which correspond to feature-selective tuning. Column 5 to 7 encode the 2-way interaction, which correspond to partial conjunctional tuning. Column 8 encodes the 3-way interaction, which correspond to object-selective tuning.

For a given tuning component g (feature-selective, partial conjunctional, or object-selective tuning) with submatrix X_g, the associated orthogonal projection matrix, P_g, is defined as:

$$P_g = X_g \left(X_g{}^T X_g \right)^{-1} X_g{}^T \tag{18}$$

The residual response matrix, R_g, can then be calculated as:

$$R_g = \left(I - \bar{P}_g \right) \cdot Y \tag{19}$$

where $\bar{P}_g$ is the projection matrix onto all tuning components other than g. Here, the response matrix $Y \in \mathbb{R}^{8 \times m}$ is the normalised activity matrix of m units across 8 insects in the stimulus set:

$$Y_{ij} = \frac{\bar{Y}_{ij} - \mu_j}{\sigma_j} \tag{20}$$

where $\bar{Y}_{ij}$ is the recorded raw activation of unit j under stimulus i, μ_j is the mean activation of unit j across the stimulus set, and σ_j is the corresponding standard deviation.

From Equation 18 and 19, the unique sum of squares for a specific tuning component g, SS_g, is therefore:

$$SS_g = \left\| P_g \cdot R_g \right\|_F^2 \tag{21}$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

To quantify the proportion of variance in a unit's activities explained by a given tuning component g, I defined a unit's *tuning selectivity* for g, λ_g, as:

$$\lambda_g = \frac{SS_g}{\sum_h SS_h} \tag{22}$$

which directly assess the relative contribution of feature-selective, partial conjunctional, and object-selective tuning to a unit's responses.

This sum of squares approach ensured that the relative strengths of feature-selective, partial conjunctional, and object-selective tuning was estimated symmetrically and uniquely, thereby providing an unbiased characterisation of units' tuning properties.

4.5 Statistical Analysis
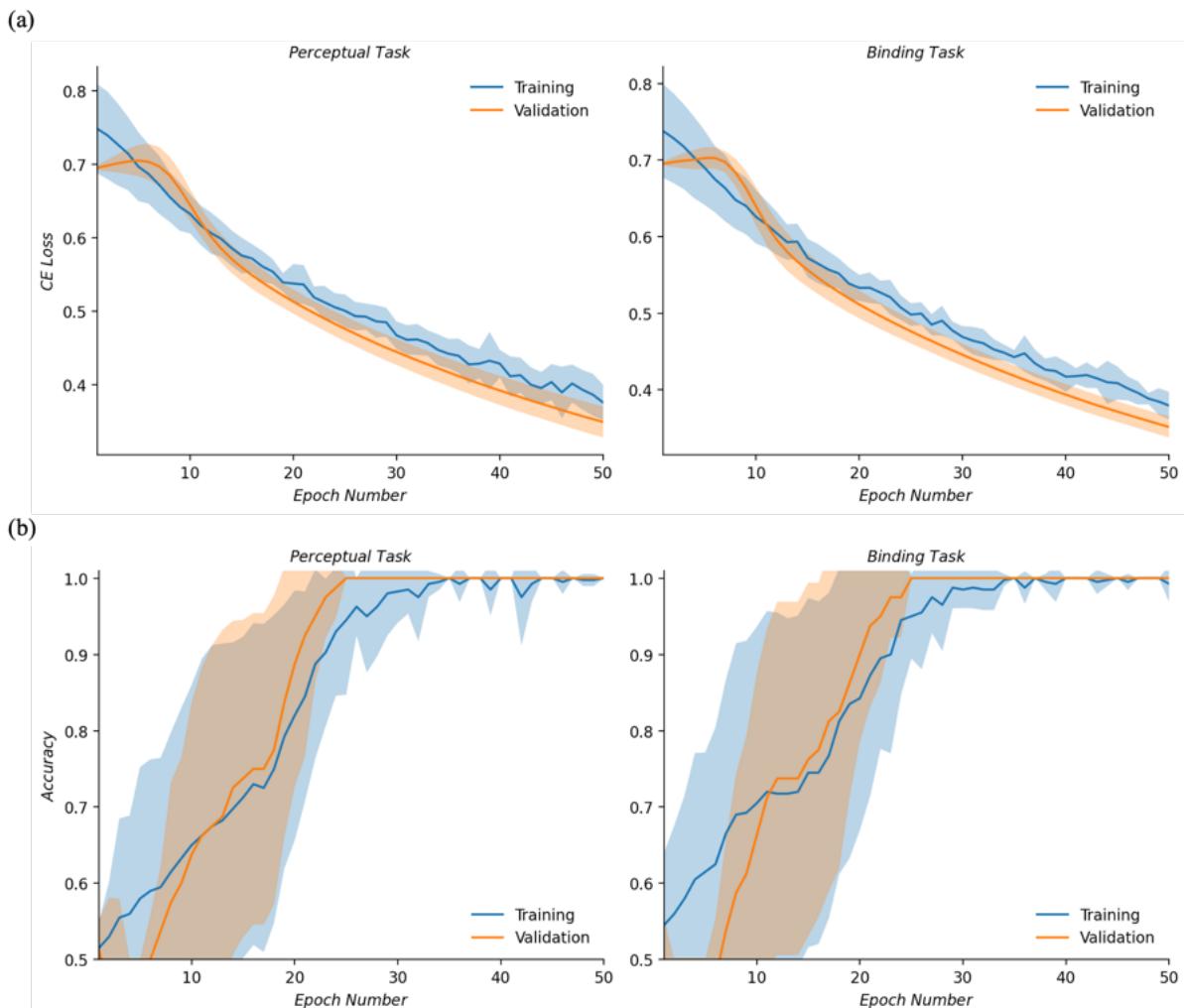
I tested hypotheses regarding pairwise group-level differences in mean using independent samples t-tests. Assumptions for t-tests were first verified[2], and if violated, nonparametric Mann-Whitney U test was conducted instead. A significant level of $\alpha = 0.05$ was applied to all statistical tests. All statistical analyses were conducted in JASP v0.18.3 (JASP Team, 2019).

---

[2] Assumption of normality tested by Shapiro-Wilk test and homogeneity of variance tested by Levene's test.

5. Residual Neural Network Model: Result

5.1 Learning Progress and Classification Performance

Can supervised learning yield a stable learning outcome? I trained the ResNet model on the new stimulus set under two classification objectives: the perceptual task and the binding task. For both tasks, cross-entropy loss and classification accuracy during training and validation converged to stable values across training runs, with ROC analysis demonstrated near-perfect classification performances (perceptual task AUC: $M = 1.00, SD = 0.00$; binding task AUC: $M = 1.00, SD = 0.00$), indicating a robust and well-behaved learning process under both training objectives (Fig 5.1; Table 2.1).

(c)
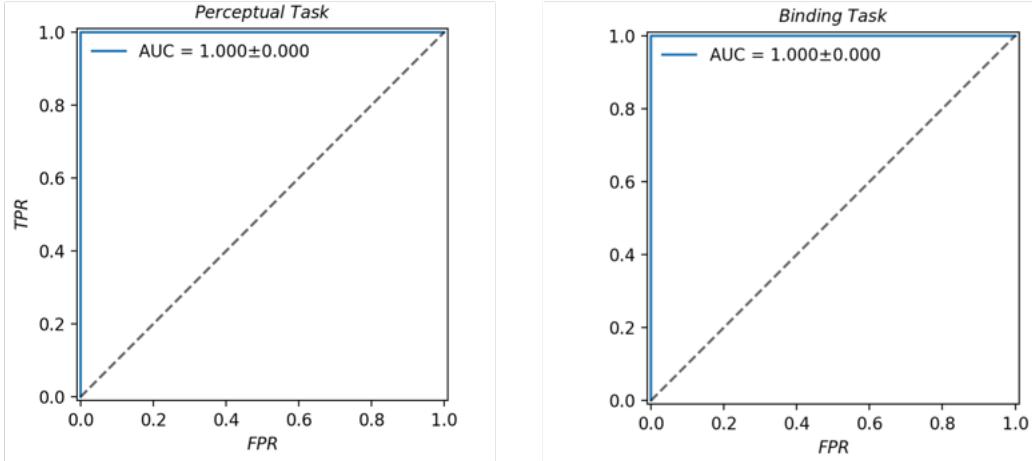
Fig 5.1: Model's learning progress and classification performance in the perceptual and binding task, including (a) cross-entropy loss, (b) classification accuracy, and (c) ROC curves. Blue and orange curves indicate training and validation runs, respectively. Solid lines show mean values across 10 randomised runs and shaded areas indicate $\pm\ 1\ SD$.

Table 2.1: Model's cross-entropy loss and classification accuracy.

Metric	Perceptual Task		Binding Task	
	M	*SD*	*M*	*SD*
Training Loss	0.38	0.02	0.38	0.02
Training Acc.	1.00	0.00	0.99	0.02
Validation Loss	0.35	0.02	0.35	0.01
Validation Acc.	1.00	0.00	1.00	0.00

Values obtained during the final (50th) training epoch, averaged across 10 randomised runs.

5.2 Emergence of Partial Conjunctional Tuning

Do partial conjunctions emerge spontaneously through supervised learning under different task constrains? After training the ResNet model, I extracted unit activations from the input convolutional layer and residual layers 1 to 4 and applied variance partitioning as described in Section 4.4. For each unit, I estimated its feature-selective, partial conjunctional, and object-selective tuning selectivity (λ_g; Equation 22) in both the perceptual and binding tasks, which quantifies the contribution of different tuning components to a unit's activity.

Tuning selectivity of units shifted systematically across successive layers of the ResNet model. Regardless the training objectives, in the input convolutional layer, tuning was exclusively dominated by feature-selective components. With increasing network depth, feature-selective tuning decreased, whereas both partial conjunctional and object-selective tuning increased (Fig 5.2). This progressive shift toward more complex tuning profiles across the model likely reflects the combinatorial convergence of inputs arising from the hierarchical architecture of the model.
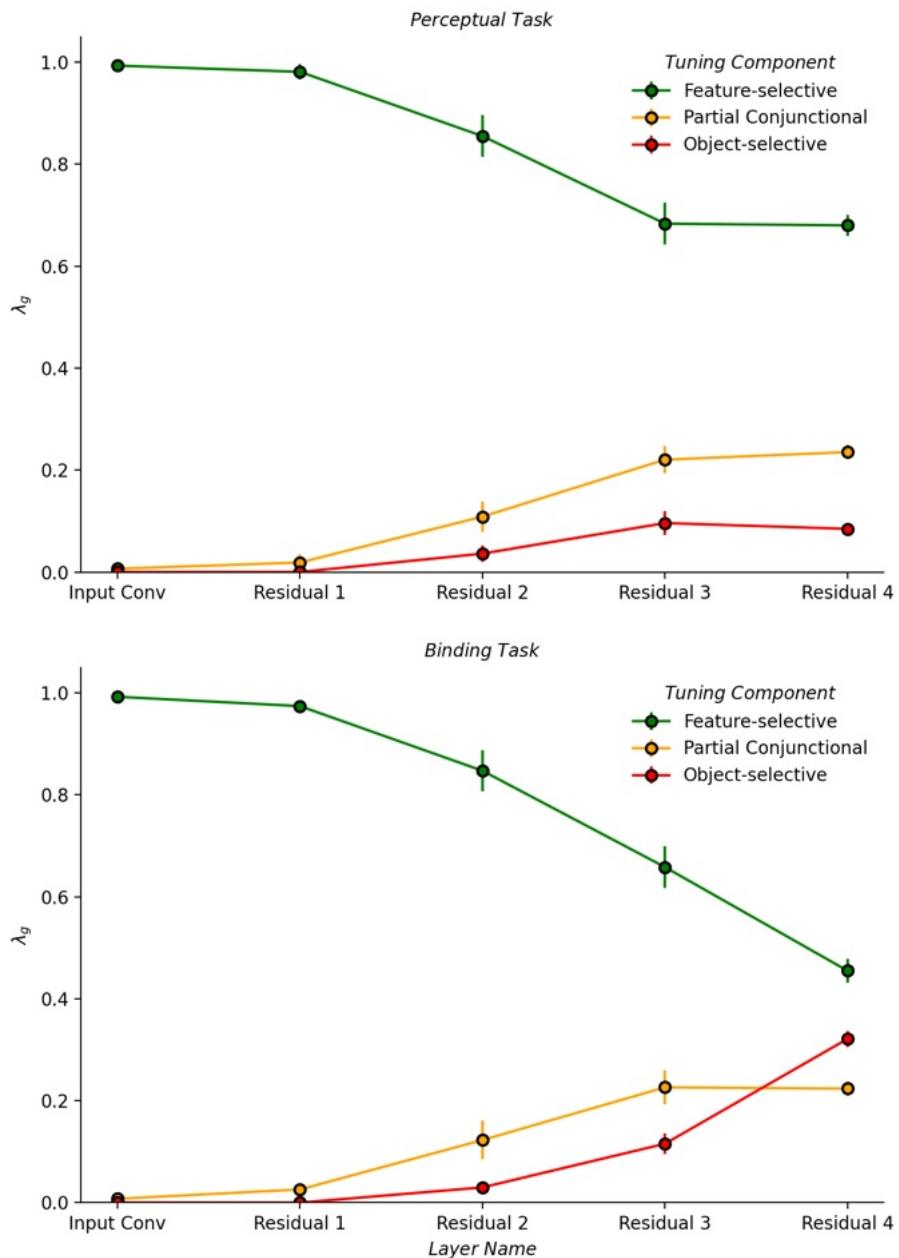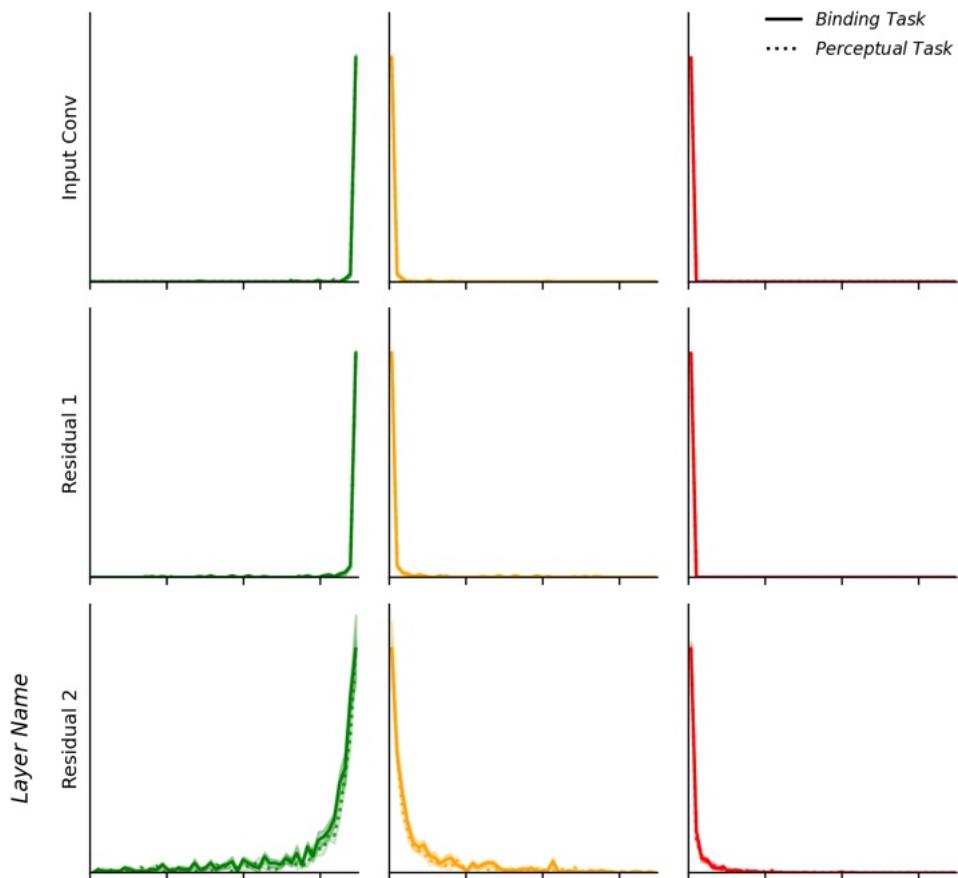
Fig 5.2: Emerged layer-averaged tuning selectivity across different layers of the ResNet model. Green, orange, and red curves indicate feature-selective, partial conjunctional, and object-selective tuning selectivity, respectively. Data points and error bars indicate mean ± 1 SD across 10 randomised runs.

To assess how task-imposed binding demands shaped tuning across layers, I compared the emerged tuning selectivity of units between the perceptual and binding tasks. In early and middle layers of the model, including the input convolutional layer and residual layers 1 to 3, tuning properties of units were largely unaffected by the training objectives: for all three tuning components, the distributions of unit-wise raw tuning strength (SS_g; Equation 21) overlapped closely between the perceptual and binding task (Fig 5.3), and the layer-averaged tuning selectivity did not differ significantly across tasks (Mann-Whitney U test: $33 < W < 78$, $p > 0.04$; Fig 5.4; Table 2.2).
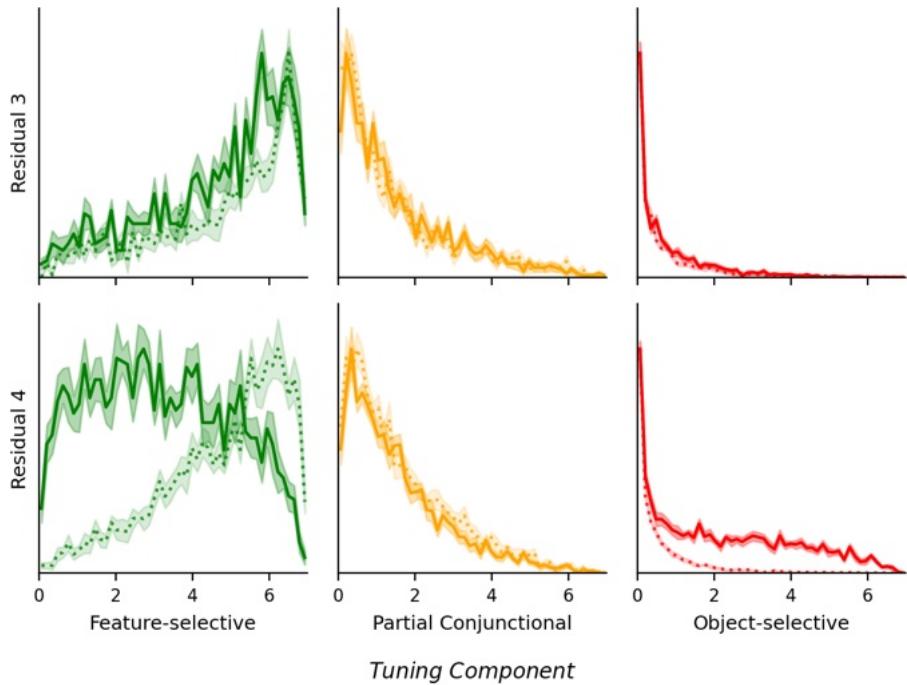
Fig 5.3: Emerged distributions of unit-wise tuning strength across layers of the ResNet model. Light and dark curves indicate the perceptual and binding tasks, respectively. Green, orange, and red curves denote feature-selective, partial conjunctional, and object-selective raw tuning strength (SS_g; Equation 21). The x-axis indicates SS_g values. The y-axis shows the relative density distribution of units estimated from a 50-bin histogram. Lines and shaded areas indicate mean $\pm 1\ SEM$ across 10 randomised runs.

Residual layer 4 showed the largest task-imposed differences in emerged tuning properties. Compared with the perceptual task, training on the binding task produced fewer feature-selective and more object-selective units in this layer, as evident by the leftward shift in unit-wise feature-selective tuning strength distribution and the elevated tail in object-selective tuning strength distribution (Fig 5.3). Consistent with this pattern, layer-averaged tuning selectivity in the residual layer 4 was significantly lower for the feature-selective tuning component (Mann-Whitney U test: $W = 0$, $p < 0.001$) and higher for the object-selective tuning component in the binding than in the perceptual task (Mann-Whitney U test: $W = 100$, $p < 0.001$; Fig 5.4; Table 2.2a; Table 2.2c).
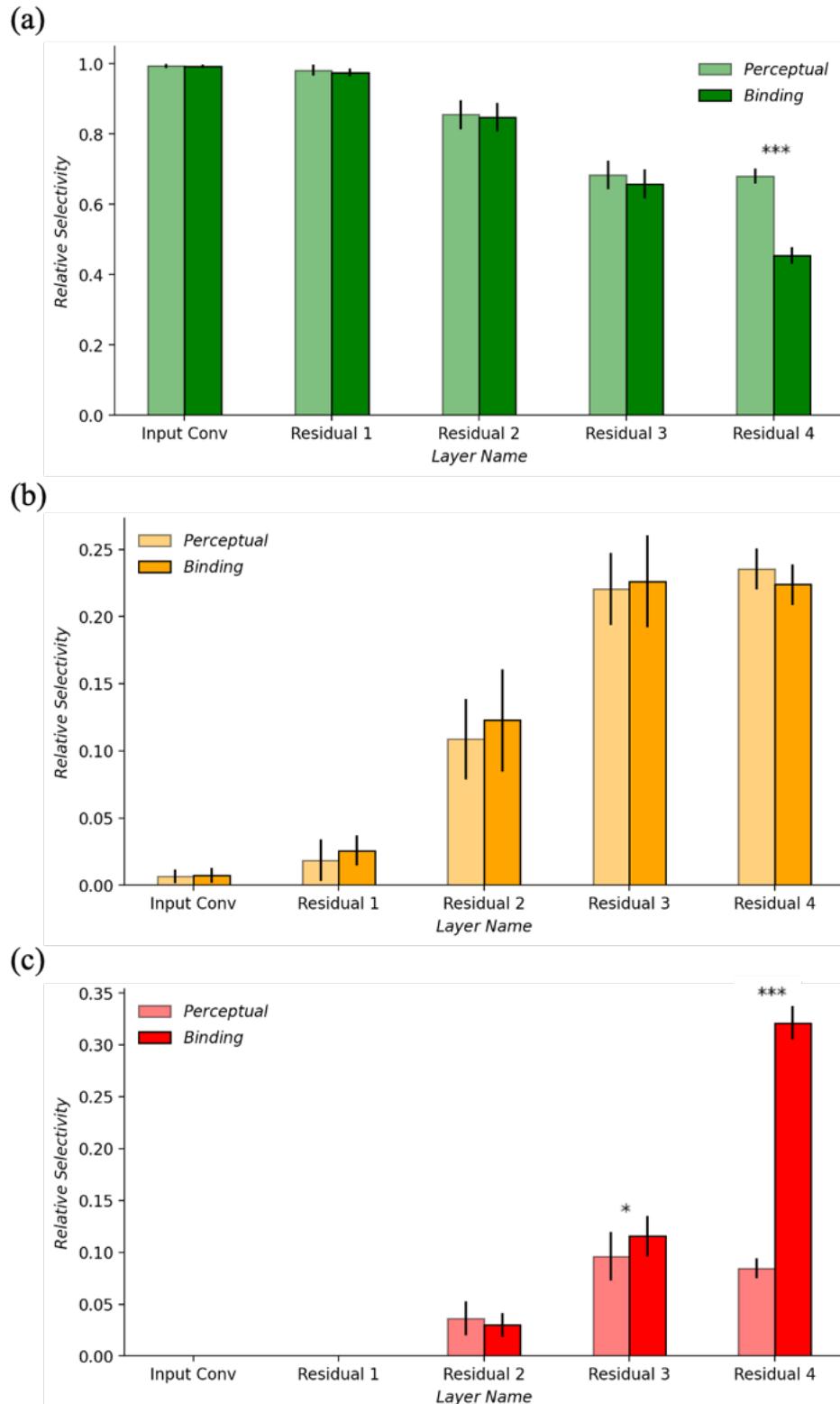
**Fig 5.4: Emerged layer-averaged (a) feature-selective, (b) partial conjunctional, and (c) object-selective tuning selectivity across layers of the ResNet model**. Light and dark bars indicate the perceptual and binding tasks, respectively. Green, orange, and red curves denote feature-selective, partial conjunctional, and object-selective tuning selectivity (λ_g; Equation 22). Height and error bars

indicate mean ± 1 *SEM* across 10 randomised runs. * and *** represents $p < .05$ and $p < .001$, respectively.

Table 2.2a: Comparisons of layer-averaged feature-selective tuning selectivity between the perceptual and binding task.

Layer Name	Tuning Selectivity λ_g				Perceptual vs. Binding task	
	Perceptual Task		Binding Task		W	p
	M	SD	M	SD		
Input Conv.	0.99	0.01	0.99	0.01	48	0.58
Residual 1	0.98	0.02	0.97	0.01	33	0.22
Residual 2	0.86	0.04	0.85	0.04	45	0.74
Residual 3	0.68	0.04	0.66	0.04	36	0.32
Residual 4	0.68	0.02	0.46	0.02	0	*<.001*

Table 2.2b: Comparisons of layer-averaged partial conjunctional tuning selectivity between the perceptual and binding task.

Layer Name	Tuning Selectivity λ_g				Perceptual vs. Binding task	
	Perceptual Task		Binding Task		W	p
	M	SD	M	SD		
Input Conv.	0.01	0.01	0.01	0.01	52	0.58
Residual 1	0.02	0.02	0.03	0.01	67	0.22
Residual 2	0.11	0.03	0.12	0.04	61	0.44
Residual 3	0.22	0.03	0.23	0.03	56	0.68
Residual 4	0.24	0.02	0.22	0.02	25	0.06

Table 2.2c: Comparisons of layer-averaged object-selective tuning selectivity between the perceptual and binding task.

Layer Name	Tuning Selectivity λ_g				Perceptual vs. Binding task	
	Perceptual Task		Binding Task		W	p
	M	SD	M	SD		
Input Conv.	0.00	0.00	0.00	0.00	50	1.00
Residual 1	0.00	0.00	0.00	0.00	55	0.74
Residual 2	0.04	0.02	0.03	0.01	42	0.58
Residual 3	0.10	0.02	0.12	0.02	78	*0.04*
Residual 4	0.09	0.02	0.32	0.01	100	*<.001*

By contrast, partial conjunctional tuning showed no significant task-imposed differences across layers, either in unit-wise tuning-strength distributions (Fig 5.3) or in layer-averaged tuning selectivity (Mann-Whitney U test: $W = 25$, $p = 0.06$; Fig 5.4; Table 2.2b). Partial conjunctions thus appeared robust across training objectives, emerging regardless of binding demands, indicating that supervised learning in the ResNet model naturally and task-independently gives rise to partial conjunctional tuning.

6. Discussion

In this study, I investigated the emergence of partial conjunctions in artificial neural network models of the human visual cortex. By presenting the SNN model with object-like visual stimuli and updating it via unsupervised learning, I showed the gradual emergence of partial conjunctions in the model, specifically, both low-amplitude, partial conjunctional tuning at population level and unstable, sharply tuned partial conjunctions at neuronal levels. By decoding their firing patterns, I further demonstrated that the emerged partial conjunctions do not explicitly encode feature combinations in visual stimuli. By training the ResNet model under different classification tasks and comparing the emerged tuning properties, I demonstrated the stable formation of partial conjunctions regardless the task-imposed binding demands. Together, these results highlight the natural emergence of partial conjunctions in hierarchical neural network models of the human visual system, which provide a context-independent neural architecture for solving the FBP.

A key characteristic of the partial conjunctions that emerged in the SNN model is their instability: even after training on 5,400 images, their tuning properties still had not converged. As I argued previously, this instability most likely reflects the lack of a global objective that constrains learning. An alternative interpretation, however, is that these units function as the subplate neurons in a developing visual cortex (Kanold, 2004), which exert transient influences on the maturation of other neurons and synapses. On this view, units with partial conjunctional tuning in the SNN model are not stable functional representations, but transient by-products of spike-timing dependent plasticity. This idea explains the poor representational alignment and is partially consistent with the task-independent properties of partial conjunctional units in the ResNet model, which suggest that such conjunctions may act as intermediate, context-independent scaffolds rather than final, task-specific representations.

The present findings partially converge with previous works that support the roles of partial conjunctions in feature binding (Morita et al., 2010; Furutate et al., 2019; Schneegans & Bays, 2017). Specifically, I demonstrated that partial conjunctions emerge spontaneously through either unsupervised stimulus exposures or supervised training tasks in hierarchical neural networks. This result thus provides a more neurobiologically plausible basis for models that assume arbitrarily constructed partial conjunctions for binding (Morita et al., 2010; Schneegans & Bays, 2017), by demonstrating that such units can form naturally without the need of innate hardwiring mechanisms.

An interesting question that remains unanswered is the functional role of these emerged partial conjunctions in feature binding. Two promising experimental approaches to this question are *lesion studies*, in which partial conjunctions are selectively removed from a model to quantify their impact on network dynamics and binding-related task performance, and *transfer-learning studies*, in which partial conjunctions are selectively transferred and reused across different tasks to test their adaptability to binding-demands. Although not elaborated in this report, preliminary lesion and transfer-learning studies suggest that these emerged partial conjunctions, particularly in the ResNet model, do contribute to feature binding.

A major limitation of this study is its neglect of active neural mechanisms in binding. Indeed, a full solution to the feature binding problem likely requires both a passive representational architecture, which specifies what to bind, and an active neural mechanism, which determines how to bind. Future work should therefore explore how the emerged partial conjunctions interact with active binding mechanisms, such as binding by synchrony (von der Malsburg,

1981) and binding by rate enhancement (Treisman & Gelade, 1980; Reynolds & Desimone,

1999; Roelfsema, 2023) to truly assess their capacity for feature binding.

Reference

Barlow, H. B. (1972). Single Units and Sensation: A Neuron Doctrine for Perceptual Psychology? *Perception*, *1*(4), 371–394. https://doi.org/10.1068/p010371

Botvinick, M., Wang, J. X., Dabney, W., Miller, K. J., & Kurth-Nelson, Z. (2020). Deep reinforcement learning and its neuroscientific implications. *Neuron, 107*(4), 603-616. https://doi.org/10.1016/j.neuron.2020.06.014

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition, 30*(7), 1145-1159.

Burkitt, A. N. (2006). A Review of the Integrate-and-fire Neuron Model: I. Homogeneous Synaptic Input. *Biological Cybernetics, 95*(1), 1–19. https://doi.org/10.1007/s00422-006-0068-6

Cavnar, W. B., & Trenkle, J. M. (1994, April). N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval* (Vol. 161175, p. 14).

Celeghin, A., Borriero, A., Orsenigo, D., Diano, M., Méndez Guerrero, C. A., Perotti, A., Petri, G., & Tamietto, M. (2023). Convolutional Neural Networks for Vision Neuroscience: Significance, developments, and outstanding issues. *Frontiers in Computational Neuroscience, 17*. https://doi.org/10.3389/fncom.2023.1153572

Di Lollo, V. (2012). The feature-binding problem is an ill-posed problem. *Trends in Cognitive Sciences*, *16*(6), 317–321. https://doi.org/10.1016/j.tics.2012.04.007

Eguchi, A., Isbister, J. B., Ahmad, N., & Stringer, S. (2018). The emergence of polychronization and feature binding in a spiking neural network model of the primate ventral visual system. *Psychological Review*, *125*(4), 545–571. https://doi.org/10.1037/rev0000103

Fan, X., & Markram, H. (2019). A Brief History of Simulation Neuroscience. *Frontiers in Neuroinformatics*, *13*. https://doi.org/10.3389/fninf.2019.00032

Feldman, J. (2013). The neural binding problem(s). *Cognitive Neurodynamics*, *7*(1), 1–11. https://doi.org/10.1007/s11571-012-9219-8

Fiser, J., & Lengyel, G. (2022). Statistical Learning in Vision. *Annual Review of Vision Science*, *8*(Volume 8, 2022), 265–290. https://doi.org/10.1146/annurev-vision-100720-103343

Fries, P., Nikolić, D., & Singer, W. (2007). The gamma cycle. *Trends in Neurosciences*, *30*(7), 309–316. https://doi.org/10.1016/j.tins.2007.05.005

Furutate, M., Fujii, Y., Morita, H., & Morita, M. (2019). Visual Feature Integration of Three Attributes in Stimulus-Response Mapping Is Distinct From That of Two. *Frontiers in Neuroscience*, *13*. https://doi.org/10.3389/fnins.2019.00035

Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. *Advances in neural information processing systems, 31*. https://arxiv.org/abs/1808.08750

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778). https://arxiv.org/abs/1512.03385

Hodgkin, A. L., & Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, *117*(4), 500–544. https://doi.org/10.1113/jphysiol.1952.sp004764

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*(1), 106–154. https://doi.org/10.1113/jphysiol.1962.sp006837

Isbister, J. B., Eguchi, A., Ahmad, N., Galeazzi, J. M., Buckley, M. J., & Stringer, S. (2018). A new approach to solving the feature-binding problem in primate vision. *Interface Focus*, *8*(4), 20180021. https://doi.org/10.1098/rsfs.2018.0021

Izhikevich, E. M. (2003). Simple model of spiking neurons. *IEEE Transactions on Neural Networks*, *14*(6), 1569–1572. https://doi.org/10.1109/TNN.2003.820440

J L van Hemmen. (1997). Hebbian learning, its correlation catastrophe, and unlearning. *Network: Computation in Neural Systems*, *8*(3), V1. https://doi.org/10.1088/0954-898X/8/3/001

Jeurissen, D., Self, M. W., & Roelfsema, P. R. (2016). Serial grouping of 2D-image regions with object-based attention in humans. *eLife*, *5*, e14320. https://doi.org/10.7554/eLife.14320

Jolivet, R., Rauch, A., Lüscher, H.-R., & Gerstner, W. (2006). Predicting spike timing of neocortical pyramidal neurons by simple threshold models. *Journal of Computational Neuroscience*, *21*(1), 35–49. https://doi.org/10.1007/s10827-006-7074-5

Kanold, P. O. (2004). Transient microcircuits formed by subplate neurons and their role in functional development of thalamocortical connections. *Neuroreport*, *15*(14), 2149–2153. https://doi.org/10.1097/00001756-200410050-00001

Kravitz, D. J., Saleem, K. S., Baker, C. I., Ungerleider, L. G., & Mishkin, M. (2013). The ventral visual pathway: An expanded neural framework for the processing of object quality. *Trends in Cognitive Sciences*, *17*(1), 26–49. https://doi.org/10.1016/j.tics.2012.10.011

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2,* 249. https://doi.org/10.3389/neuro.06.004.2008

Kruizinga, P., & Petkov, N. (1999). Nonlinear operator for oriented texture. *IEEE Transactions on Image Processing*, *8*(10), 1395–1407. https://doi.org/10.1109/83.791965

Kruskal, J. B. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, *29*(2), 115–129. https://doi.org/10.1007/BF02289694

Lam, S. K., Pitrou, A., & Seibert, S. (2015). Numba: A LLVM-based Python JIT compiler. *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, 1–6. https://doi.org/10.1145/2833157.2833162

Li, J., Osher, D. E., Hansen, H. A., & Saygin, Z. M. (2020). Innate connectivity patterns drive the development of the visual word form area. *Scientific reports, 10*(1), 18039.https://doi.org/10.1038/s41598-020-75015-7

Livingstone, M. S., & Hubel, D. H. (1987). Psychophysical evidence for separate channels for the perception of form, color, movement, and depth. *Journal of Neuroscience*, *7*(11), 3416–3468. https://doi.org/10.1523/JNEUROSCI.07-11-03416.1987

Luczak, A. (2024). Entropy of Neuronal Spike Patterns. *Entropy*, *26*(11), Article 11. https://doi.org/10.3390/e26110967

Maass, W. (1997). Networks of spiking neurons: The third generation of neural network models. *Neural Networks*, *10*(9), 1659–1671. https://doi.org/10.1016/S0893-6080(97)00011-7

Martin, A. B., & von der Heydt, R. (2015). Spike Synchrony Reveals Emergence of Proto-Objects in Visual Cortex. *Journal of Neuroscience*, *35*(17), 6860–6870. https://doi.org/10.1523/JNEUROSCI.3590-14.2015

McFarquhar, M. (2016). Testable hypotheses for unbalanced neuroimaging data. *Frontiers in neuroscience, 10,* 270. https://doi.org/10.3389/fnins.2016.00270

Mel, B. W. (1997). SEEMORE: Combining Color, Shape, and Texture Histogramming in a Neurally Inspired Approach to Visual Object Recognition. *Neural Computation*, *9*(4), 777–804. https://doi.org/10.1162/neco.1997.9.4.777

Mel, B. W., & Fiser, J. (2000). Minimizing binding errors using learned conjunctive features. *Neural Computation*, *12*(4), 731–762. https://doi.org/10.1162/089976600300015574

Miconi, T., & VanRullen, R. (2010). The Gamma Slideshow: Object-Based Perceptual Cycles in a Model of the Visual Cortex. *Frontiers in Human Neuroscience*, *4,* 205. https://www.frontiersin.org/articles/10.3389/fnhum.2010.00205

Morita, M., Morokami, S., & Morita, H. (2010). Attribute Pair-Based Visual Recognition and Memory. *PLOS ONE*, *5*(3), e9571. https://doi.org/10.1371/journal.pone.0009571

Nassi, J. J., & Callaway, E. M. (2009). Parallel Processing Strategies of the Primate Visual System. *Nature Reviews. Neuroscience*, *10*(5), 360–372. https://doi.org/10.1038/nrn2619

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems, 32.* https://doi.org/10.48550/arXiv.1912.01703

Perrinet, L., Delorme, A., Samuelides, M., & Thorpe, S. J. (2001). Networks of integrate-and-fire neuron using rank order coding A: How to implement spike time dependent Hebbian plasticity. *Neurocomputing*, *38–40*, 817–822. https://doi.org/10.1016/S0925-2312(01)00460-X

Pfeiffer, M., & Pfeil, T. (2018). Deep learning with spiking neurons: Opportunities and challenges. *Frontiers in neuroscience, 12*, 409662. https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2018.00774/full

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience, 38*(33), 7255-7269. https://www.jneurosci.org/content/38/33/7255

Recanatesi, S., Bradde, S., Balasubramanian, V., Steinmetz, N. A., & Shea-Brown, E. (2022). A scale-dependent measure of system dimensionality. *Patterns*, *3*(8), 100555. https://doi.org/10.1016/j.patter.2022.100555

Reynolds, J. H., & Desimone, R. (1999). The Role of Neural Mechanisms of Attention in Solving the Binding Problem. *Neuron*, *24*(1), 19–29. https://doi.org/10.1016/S0896-6273(00)80819-3

Riesenhuber, M., & Poggio, T. (1999). Are Cortical Models Really Bound by the "Binding Problem"? *Neuron*, *24*(1), 87–93. https://doi.org/10.1016/S0896-6273(00)80824-7

Roelfsema, P. R. (2006). Cortical algorithms for perceptual grouping. Annual Review of
Neuroscience, 29(Volume 29, 2006), 203–227.
https://doi.org/10.1146/annurev.neuro.29.051605.112939

Roelfsema, P. R. (2023). Solving the binding problem: Assemblies form when neurons
enhance their firing rate—they don't need to oscillate or synchronize. *Neuron*, *111*(7),
1003–1019. https://doi.org/10.1016/j.neuron.2023.03.016

Roelfsema, P. R., Lamme, V. A. F., & Spekreijse, H. (1998). Object-based attention in the
primary visual cortex of the macaque monkey. *Nature*, *395*(6700), 376–381.
https://doi.org/10.1038/26475

Rosenblatt, F. (1962). *Principles of Neurodynamics: Perceptrons and the Theory of Brain
Mechanisms*. Spartan Books.

Schneegans, S., & Bays, P. M. (2017). Neural Architecture for Feature Binding in Visual
Working Memory. *Journal of Neuroscience*, *37*(14), 3913–3925.
https://doi.org/10.1523/JNEUROSCI.3493-16.2017

Serre, T. (2013). Hierarchical Models of the Visual System. In D. Jaeger & R. Jung (Eds.),
Encyclopedia of Computational Neuroscience (pp. 1–12). Springer.
https://doi.org/10.1007/978-1-4614-7320-6_345-1

Singer, W., & Gray, C. M. (1995). Visual Feature Integration and the Temporal Correlation
Hypothesis. *Annual Review of Neuroscience*, *18*(1), 555–586.
https://doi.org/10.1146/annurev.ne.18.030195.003011

Sommer, F. T. (2013). Neural oscillations and synchrony as a mechanism for coding,
communication and computation in the visual system.

Taylor, J., & Xu, Y. (2022). Representation of color, form, and their conjunction across the
human ventral visual pathway. *NeuroImage*, *251*, 118941.
https://doi.org/10.1016/j.neuroimage.2022.118941

Treisman, A. (1988). Features and Objects: The Fourteenth Bartlett Memorial Lecture. *The
Quarterly Journal of Experimental Psychology Section A*, *40*(2), 201–237.
https://doi.org/10.1080/02724988843000104

Treisman, A. (1996). The binding problem. *Current Opinion in Neurobiology*, *6*(2), 171–178.
https://doi.org/10.1016/S0959-4388(96)80070-5

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive
Psychology*, *12*(1), 97–136. https://doi.org/10.1016/0010-0285(80)90005-5

Troyer, T. W., Krukowski, A. E., Priebe, N. J., & Miller, K. D. (1998). Contrast-Invariant
Orientation Tuning in Cat Visual Cortex: Thalamocortical Input Tuning and

Correlation-Based Intracortical Connectivity. *Journal of Neuroscience*, *18*(15), 5908–
5927. https://doi.org/10.1523/JNEUROSCI.18-15-05908.1998

Velik, R. (2012). From simple receptors to complex multimodal percepts: A first global
picture on the mechanisms involved in perceptual binding. *Frontiers in Psychology*, *3*.
https://doi.org/10.3389/fpsyg.2012.00259

von der Malsburg, C. (1994). The Correlation Theory of Brain Function. In E. Domany, J. L.
Van Hemmen, & K. Schulten (Eds.), *Models of Neural Networks* (pp. 95–119).
Springer New York. https://doi.org/10.1007/978-1-4612-4320-5_2

Wen, H., Shi, J., Chen, W., & Liu, Z. (2018). Deep residual network predicts cortical
representation and organization of visual features for rapid categorization. *Scientific
reports, 8*(1), 3752. https://www.nature.com/articles/s41598-018-22160-9

Wolfe, J. M. (2012). The binding problem lives on: Comment on Di Lollo. *Trends in
Cognitive Sciences*, *16*(6), 307–308. https://doi.org/10.1016/j.tics.2012.04.013

Yamazaki, K., Vo-Ho, V.-K., Bulsara, D., & Le, N. (2022). Spiking Neural Networks and
Their Applications: A Review. *Brain Sciences*, *12*(7), Article 7.
https://doi.org/10.3390/brainsci12070863

Appendix I: Key Parameters in the SNN model

Model Architectures

Parameter Name	Layer			
	1	*2*	*3*	*4*
Number of *ex* neurons within each layer	32×32	32×32	32×32	32×32
Number of *inh* neurons within each layer	16×16	16×16	16×16	16×16
Number of feedforward connections per *ex* neuron	30	100	100	/
Number of feedback connections per *ex* neuron	{0, 10}	{0, 10}	{0, 10}	{0, 10}
Number of lateral *ex* to *ex* connection per *ex* neuron	{0, 10}	{0, 10}	{0, 10}	{0, 10}
Number of lateral *ex* to *inh* connection per *ex* neuron	30	30	30	30
Number of lateral *inh* to *ex* connection per *inh* neuron	30	30	30	30
Projection radius of feedforward connections	0.5	4.0	6.0	8.0
Projection radius of feedback connections	4.0	4.0	4.0	/
Projection radius of lateral *ex* to *ex* connections	2.0	2.0	2.0	2.0
Projection radius of lateral *ex* to *inh* connection	0.5	0.5	0.5	0.5
Projection radius of lateral *inh* to *ex* connections	4.0	4.0	4.0	4.0

Parameters values adopted from Eguchi et al., (2018); *ex*: excitatory neuron; *inh*: inhibitory neuron; {0, 10}: randomly drawn from [0, 10].

Neural and Synaptic Parameters

Parameter Name		Adopted Value	Source
Simple Cells *(Eq 1 - 3)*	θ	$0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}$	**c**
	γ	0.5	**c**
	$\bar{\lambda}$	2	**c**
	b	1.5	**c**
	ψ	π	**c**
	ω	100 Hz	**c**
Neurons *(Eq 4 - 7)*	τ_m^γ	*ex*: 20 ms *inh*: 12 ms	**a**
	V_0^γ	*ex*: -74 mV *inh*: -82 mV	**a**
	R^γ	*ex*: 40 MΩ *inh*: 55 MΩ	**a**
	Θ^γ	*ex*: -53 mV *inh*: -53 mV	**a**
	τ_R	20 ms	**a**
	V_H^γ	*ex*: -57 mV *inh*: -58 mV	**a**
	τ_r	*ex* to *ex* connection: 2 ms *ex* to *inh* connection: 40 ms *inh* to *ex* connection: 80 ms	**a**
	λ	*layer 1*: 0.4 nS *layer 2* to *layer 4*: 1.6 nS	**c**
Learning *(Eq 8 -10)*	ρ	50.0	*
	τ_C	5 ms	**b**
	τ_D	5 ms	**b**
	α_C	0.5	**b**
	α_D	0.5	**b**

Sources of values adopted: **a** from Troyer et al., 1998; **b** from Perrinet et al., 2001; **c** from Eguchi et al., 2018; * are tuned parameter for simulations; *ex*: excitatory neuron; *inh*: inhibitory neuron.