# Trajectory-Based Neural Darwinism in CNNs: Variation, Competition, and Selective Retention

**Anonymous authors**
**Paper under double-blind review**

## Abstract

Understanding how neural networks develop and stabilize their internal representations remains a central challenge in deep learning. Inspired by Edelman's theory of Neural Darwinism, we investigate whether competitive dynamics analogous to neuronal group selection emerge in artificial neural networks during training. Through detailed trajectory analyses of neuron activations, weights, and cumulative representational change across convolutional neural networks (CNNs) including three-layer MLP-Net, ResNet-18, VGG-16, and ResNet-50, we uncover consistent patterns of variation, competition, and selective retention. Ablation studies reveal that networks tolerate removal of large fractions of neurons without accuracy degradation, indicating high redundancy; however, beyond a critical threshold, performance collapses as the core subset of task-critical neurons is disrupted. Across multiple datasets and architectures, neuron trajectory dynamics show that survived neurons sustain longer, more coherent representational paths, stronger weight norms, and higher activations, while eliminated neurons stagnate or fade toward representational silence. Overall, our findings are consistent with a Darwinian view of representation learning: CNNs exhibit robustness through redundancy at early stages, followed by selective consolidation of highly specialized neurons in deeper layers.

## 1 Introduction

The success of deep learning is often attributed to its capacity for hierarchical feature learning Chizat & Netrapalli (2024); Banerjee (2025), yet the internal principles that govern representational stability and neuron specialization remain incompletely understood. Existing analyses have largely focused on optimization dynamics or information-theoretic perspectives Butakov (2024), while comparatively little attention has been paid to potential competitive mechanisms operating at the level of individual neurons. In neuroscience, Edelman's theory of Neural Darwinism posits that neuronal populations undergo variation, competition, and selective retention, yielding stable yet adaptable functional circuits.

Motivated by this perspective, we seek to determine whether analogous Darwinian dynamics emerge in convolutional neural networks (CNNs). Specifically, we investigate whether subsets of neurons demonstrate differential survival and elimination, how these processes evolve across depth, and whether they ultimately shape the robustness and specialization of internal representations. To this end, we develop a multi-perspective framework combining dynamic trajectory analysis (PCA-based neuron evolution), static representation inspection (embedding, weight, and activation tracking), and functional validation (controlled ablation). We apply this to CNNs of varying depth and complexity, including three-layer MLP-Net, ResNet-18, VGG-16, and ResNet-50 trained on diverse datasets.

Our findings reveal patterns consistent with Darwinian dynamics across CNNs. In shallow layers, representational variation is high, with neurons displaying noisy and irregular trajectories. Middle layers intensify selective dynamics, filtering neurons that fail to sustain adaptive displacement. Deep layers culminate in selective retention: a compact set of specialized neurons consolidates into high-utility manifolds, while others fade into functional irrelevance. Ablation studies further support this interpretation, showing robustness under moderate perturbation and sharp collapse once the selected subset is disrupted. Together, these results

suggest that CNNs achieve both robustness and specialization through internal selection processes consistent with Darwinian principles.

## 2 Related Work

### 2.1 On Neural Networks Analysis

A large body of work has investigated how neural networks form and consolidate internal structure, spanning pruning, representational similarity, loss geometry, and interpretability. Pruning studies demonstrate that overparameterized models contain trainable sparse subnetworks, with the Lottery Ticket Hypothesis Frankle & Carbin (2019) and its extensions Liu (2019); Sanh (2020); Lee (2019); Evci (2020); Morcos (2019) showing that subnetworks can be identified via sensitivity measures Lee (2019), dynamic rewiring Evci (2020), or transfer across tasks Morcos (2019). Representation analyses such as SVCCA Raghu (2017) and CKA Kornblith (2019) reveal convergent layerwise structures, while neural tangent kernel theory Jacot (2018) and deep linear dynamics Saxe (2014) provide analytic descriptions of training. Geometric studies show low-loss mode connectivity Garipov (2018); Draxler (2018) and neural collapse phenomena Han (2022), connecting optimization to generalization. Interpretability methods including Network Dissection Bau (2017), TCAV Kim (2018), Integrated Gradients Sundararajan (2017), and SHAP Lundberg & Lee (2017) further expose concept-level features, while symmetry and re-basin analyses Ainsworth (2023) link parameter permutations to solution geometry. Finally, work on large-batch training Keskar (2017) and dynamical isometry Pennington (2017) elucidates how optimization biases shape solution quality. Taken collectively, these perspectives highlight redundancy, convergence, and selection-like pressures in neural networks, aligning with our Darwinian view of neuron-level competition.

### 2.2 Neuron Darwinian

The conceptual foundation for Darwinian mechanisms in neural systems was established by Edelman, whose theory of neuronal group selection framed brain function as the result of variation among neuronal populations, selective reinforcement of functional circuits, and the inheritance of stable connectivity patterns over developmental and experiential timescales Edelman (1987). Building on this biological paradigm, recent advances in artificial neural networks embed analogous variation–selection processes at multiple computational scales, challenging the dominance of gradient-only optimization. Du et al. cast late-epoch backpropagation-trained networks as "ancestral genomes" and apply differential evolution to offspring models, selecting fitter variants to reduce overfitting and accelerate inference in large-scale vision settings Du (2024). At the neuron level, NeuroFS introduces a synaptic plasticity–inspired mechanism that dynamically prunes and regrows input neurons during training, enabling networks to adapt structure on the fly within strict sparsity constraints Zahra (2023). In dynamical systems, Czégel et al. demonstrate Darwinian neurodynamics in reservoir computing: reservoir activity patterns are imperfectly copied between units, and fitter configurations are preferentially selected, enabling unsupervised emergence of combinatorial problem-solving capabilities Czégel (2021). Spiking architectures benefit from similar evolutionary processes: Shen et al. propose NeuEvo, which evolves excitatory–inhibitory circuit patterns under spike-timing–dependent plasticity, achieving strong CIFAR-10 and ImageNet performance Shen (2023). At the architectural level, Shafiee et al. encode connectivity as heritable "DNA," evolving compact yet competitive offspring networks Shafiee (2018). More recently, Chen et al. introduce OPNP, a gradient-sensitivity–based pruning framework that selects neurons and parameters to enhance out-of-distribution robustness—again mirroring evolutionary pressure for generalization Chen (2023). Collectively, these approaches demonstrate a convergent trend: embedding variation–and–selection mechanisms across synaptic, dynamical, and structural scales in neural systems to achieve adaptability, sparsity, and improved generalization beyond what gradient descent alone affords. Our work extends this trajectory by introducing a neuron-level temporal analysis framework, where activation trajectories across training are quantified to distinguish "survived" and "eliminated" neurons, providing direct empirical evidence for Neural Darwinism within modern deep learning architectures.

### 2.3 Neuron Trajectory

Recent work has begun to focus on understanding neuron trajectories, i.e., the evolution and influence of individual neuron activations or weights across layers and time, as a lens for interpretability and generalization. Fu et, al. formalize learning trajectories during training and derive generalization bounds dependent on the complexity of these trajectories Fu (2023). Pesme and Flammarion analytically characterize gradient-flow trajectories in 2-layer diagonal networks, showing how paths traverse successive saddles before converging to minimal -norm solutions Pesme & Flammarion (2023); and Ahn spotlight on threshold neurons links the "edge of stability" training dynamics to emergence of threshold-like activations, elucidating trajectory-based neuron behavior Ahn (2023). Mechanistic interpretability research has also tracked neuron- or head-level trajectories through networks. Conmy et, al. propose the ACDC method to automatically extract circuits—i.e., neuron-activation paths—employing trajectory-based subgraph discovery Conmy (2023). Complementing this, Syed et, al. use attribution patching along activation trajectories to unearth causal subcircuits in transformer activations Syed (2024). These methods illuminate how information flows along neuron trajectories during inference and how specialized paths—neuron trajectories—mediate specific computations. Graph-based sequence forecasting approaches, such as AMAG, repurpose trajectory modeling techniques originally applied to biological neurons—forecasting future unit activity—demonstrating that neuron activity trajectories can be explicitly modeled and predicted Li (2023). In spiking networks, trajectory-inspired frameworks optimize spike-based neuron firing patterns, effectively shaping activation trajectories to reduce spiking load while preserving performance Shi (2024); Shen (2024). In summary, these lines of inquiry—from gradient-flow theory and threshold-emergence phenomena to subcircuit extraction and spiking dynamics—position neuron trajectories as a unifying construct that bridges training dynamics, interpretability, and functional behavior in neural models.

## 3 Method

We formalize neuron evolution during training as a continuous-time dynamical system driven by both optimization gradients and intrinsic information-theoretic pressures. Intuitively, we treat each neuron as an evolving agent whose state is not only determined by its parameters but also by how it responds to data and gradients. This perspective allows us to study neural computation through the lens of dynamical systems and evolutionary selection Saxe (2014); Mei (2018); Chizat & Bach (2018).

Let a neural network $f_\theta : \mathcal{X} \to \mathcal{Y}$ consist of layers $\{L_k\}_{k=1}^{D}$, where layer $L_k$ contains neurons $\{a_i^{(k)}\}_{i=1}^{n_k}$. Each neuron is parameterized by a weight vector $w_i^{(k)} \in \mathbb{R}^{d_{k-1}}$, bias $b_i^{(k)} \in \mathbb{R}$, and activation function $\sigma$. Its activation at time $t$ is:

$$a_i^{(k)}(x,t) := \sigma\left(\langle w_i^{(k)}(t), h^{(k-1)}(x,t)\rangle + b_i^{(k)}(t)\right), \tag{1}$$

where $h^{(k-1)}$ is the output from $L_{k-1}$ and $h^{(0)} = x$. Thus, activations evolve jointly with weights and reflect both optimization and stochastic fluctuations Schoenholz (2017); Poole (2016).

### 3.1 Neuron Evolutionary Dynamics System (NEDS)

To make this evolution explicit, we introduce the *neuron state vector*, which concatenates its trainable parameters, average activity, gradient statistics, and information-theoretic descriptors:

$$\psi_i^{(k)}(t) := \left[w_i^{(k)}(t),\ b_i^{(k)}(t),\ \mu_i^{(k)}(t),\ g_i^{(k)}(t),\ \mathcal{I}_i^{(k)}(t)\right]. \tag{2}$$

Here:

$$\mu_i^{(k)}(t) = \mathbb{E}_{x\sim\mathcal{D}}\left[a_i^{(k)}(x,t)\right], \tag{3}$$

$$g_i^{(k)}(t) = \mathbb{E}_{x\sim\mathcal{D}}\left[\frac{\partial \mathcal{L}(x)}{\partial a_i^{(k)}}\right], \tag{4}$$

$$\mathcal{I}_i^{(k)}(t) = \text{differential entropy of } a_i^{(k)}(x,t). \tag{5}$$

The evolution of each neuron is then modeled as a differential equation:

$$\frac{d}{dt}\psi_i^{(k)}(t) = \mathbf{F}_\theta^{(k)}\left(\psi_i^{(k)}(t), \mathcal{D}, \mathcal{L}\right), \tag{6}$$

where $\mathbf{F}_\theta^{(k)}$ captures the joint effect of gradient descent updates and intrinsic representational dynamics Bonnabel (2013). This abstraction allows us to borrow tools from dynamical systems theory to analyze stability, convergence, and diversity of neurons Achille & Soatto (2018b).

**Assumption 3.1** (Smooth and Bounded Dynamics)**.** The parameter trajectory $\theta(t)$ is $C^1$ in $t$, $\mathbf{F}_\theta^{(k)}$ is Lipschitz in $\psi$, and there exist constants $B_g, B_a > 0$ such that for all $t \in [0, T]$:

$$\|g_i^{(k)}(t)\| \leq B_g, \quad \mathrm{Var}[a_i^{(k)}(t)] \leq B_a.$$

Furthermore, the trajectory length $\mathcal{L}_i^{(k)}$ defined in equation 7 is finite as $T \to \infty$.

**Assumption 3.2** (Gaussian Activation Approximation)**.** For entropy estimation, neuron activations are approximately Gaussian: $a_i^{(k)}(x,t) \overset{\text{approx}}{\sim} \mathcal{N}(\mu, \sigma^2)$.

*Remark* 3.3. While Gaussianity may not strictly hold for ReLU-family nonlinearities, empirical validation indicates the entropy–variance relationship remains approximately monotonic, allowing equation 5 to be a consistent proxy for variability Amjad (2021).

### 3.2 Trajectory-Based Evolutionary Fitness

We now quantify the "fitness" of a neuron through its trajectory in state space. The trajectory

$$\Gamma_i^{(k)} := \{\psi_i^{(k)}(t) \mid t \in [0, T]\}$$

records how the neuron evolves during training. From this path we extract three complementary quantities:

1. **Arc length** $\mathcal{L}_i^{(k)}$ (Eq. 7): measures the cumulative representational movement of a neuron. A long arc length indicates that the neuron undergoes substantial representational change rather than remaining stagnant.

2. **Final-stage stochasticity** $\mathcal{S}_i^{(k)}$ (Eq. 8): quantifies how unstable the neuron remains near convergence. Stable neurons are desirable, while persistently fluctuating ones are typically pruned.

3. **Integrated entropy** $\mathfrak{H}_i^{(k)}$ (Eq. 9): measures how much diversity of information the neuron maintains throughout training. High entropy suggests richer representational capacity Quétu (2024); Spadaro (2023).

Formally:

$$\mathcal{L}_i^{(k)} := \int_0^T \left\| \frac{d\psi_i^{(k)}(t)}{dt} \right\|_2 dt, \tag{7}$$

$$\mathcal{S}_i^{(k)} := \frac{1}{\delta} \int_{T-\delta}^T \left\| \frac{d\psi_i^{(k)}(t)}{dt} \right\|_2^2 dt, \tag{8}$$

$$\mathfrak{H}_i^{(k)} := \int_0^T \mathcal{I}_i^{(k)}(t)\, dt. \tag{9}$$

These factors combine into the neuron's *evolutionary fitness*:

$$\Phi_i^{(k)} := \alpha \cdot \mathcal{L}_i^{(k)} - \beta \cdot \mathcal{S}_i^{(k)} + \gamma \cdot \mathfrak{H}_i^{(k)}, \quad \alpha, \beta, \gamma > 0. \tag{10}$$

Intuitively, a neuron is "fit" if it explores sufficiently diverse states, settles into stable dynamics, and avoids redundancy Molchanov (2017); Fang (2023).

### 3.3 Selection and Survival Criteria

To link fitness to survival, we define thresholds relative to population statistics:

**Definition 3.4** (Survived Neuron)**.** Neuron $i$ in layer $k$ is *survived* if:

$$\Phi_i^{(k)} \geq \mathbb{E}_j[\Phi_j^{(k)}] + \lambda \cdot \mathrm{SD}(\Phi_j^{(k)}), \quad \lambda > 0.$$

This creates an evolutionary-like selection pressure, where only the most informative and stable neurons persist Han (2015); Frankle & Carbin (2019); Morcos (2019).

Moreover, the following lemma shows that neurons which remain highly unstable while simultaneously losing entropy inevitably collapse to vanishing fitness, predicting their elimination.

**Lemma 3.5** (Instability Predicts Elimination)**.** *If $\mathcal{S}_i^{(k)} \geq \delta_{\max}$ and $\frac{d}{dt}\mathcal{I}_i^{(k)}(t) < 0$ for $t \in [T-\delta, T]$, then:*

$$\lim_{T\to\infty} \Phi_i^{(k)} = -\infty.$$

### 3.4 Theoretical Analysis

**Theorem 3.6** (Fitness Threshold Implies Gradient–Variance Contribution)**.** *Let*

$$\Delta_i^{(k)} := \mathbb{E}_{x\sim\mathcal{D}}\left[\left(\frac{\partial \mathcal{L}(x)}{\partial a_i^{(k)}}\right)^2 \cdot \mathrm{Var}[a_i^{(k)}(x)]\right].$$

*Under Assumptions 3.1 and 3.2, there exist constants $\tau, \kappa > 0$ such that:*

$$\Phi_i^{(k)} \geq \tau \quad \Rightarrow \quad \Delta_i^{(k)} \geq \kappa.$$

This result bridges our trajectory-based measure with a classical signal-to-noise criterion, showing that neurons with high fitness necessarily contribute to meaningful gradient–variance interactions Achille & Soatto (2018a); Martens (2020).

### 3.5 Multilayer Coupled Dynamics

At the layer level, survival is not independent. Let $\Psi^{(k)}(t) = [\psi_i^{(k)}(t)]_{i\in\mathcal{N}_k}$ be the joint state of all neurons in layer $k$. We define the *inter-layer coupling operator*:

$$\mathcal{C}_{k\to k+1}(t) := \left[\frac{\partial \psi_j^{(k+1)}(t)}{\partial \psi_i^{(k)}(t)}\right]_{i,j}.$$

Its Frobenius norm quantifies total sensitivity of layer $k+1$ states to layer $k$. The *layer influence matrix* is:

$$\mathbf{M}_{k,l}(t) = \begin{cases} \|\mathcal{C}_{k\to l}(t)\|_F, & |k-l| = 1, \\ 0, & \text{otherwise.} \end{cases}$$

We define the *Darwinian flow energy* Zheng (2025):

$$\mathcal{E}_{\mathrm{Darwin}} := \sum_{k=1}^{D}\sum_{l=1}^{D}\int_0^T \mathbf{M}_{k,l}(t)\cdot\phi\Big(\mathrm{KL}(\rho^{(k)}(t)\,\|\,\rho^{(l)}(t))\Big)\,dt,$$

where $\rho^{(k)}(t)$ is the activation distribution in $L_k$ and $\phi$ is convex.

**Theorem 3.7** (Coupled Survival Principle). *If $\Phi_i^{(k)}(t)$ is Lipschitz in $t$ and $\mathbf{M}_{k,k+1}(t) \geq \mu > 0$ almost everywhere, then:*

$$\min_{j \in \mathcal{N}_{k+1}} \sum_{i \in \mathcal{S}^{(k)}} \|\mathcal{C}_{k \to k+1}^{(i,j)}(t)\| > \epsilon \quad \Rightarrow \quad \frac{|\mathcal{S}^{(k+1)}|}{|\mathcal{N}_{k+1}|} \geq \eta(\epsilon, \tau_k, \tau_{k+1}) > 0.$$

**Theorem 3.8** (Global Convergent Specialization). *If $\mathcal{E}_{Darwin} \geq \epsilon > 0$ and all $\Phi_i^{(k)} \in C^1([0,T])$, then:*

$$\lim_{T \to \infty} \frac{\#\{i : \Phi_i^{(k)} < \tau_k\}}{|\mathcal{N}_k|} = 0 \quad \forall k.$$

Overall, the Neuron Evolutionary Dynamics System (NEDS) provides a principled framework to study representational dynamics under Neural Darwinism. Neurons are no longer seen as static units with fixed importance, but as evolving entities competing for survival through their trajectory length, stability, and entropy. This formalism both explains empirical neuron pruning phenomena and predicts inter-layer propagation of specialization Raghu (2017); Jacot (2018).

## 4 Experiments

We designed a series of experiments to examine whether CNNs exhibit dynamics consistent with Neural Darwinism, and how such processes shape robustness and representational specialization. Our analysis proceeds in two complementary strands. First, we conduct ablation experiments on a CNN trained on MNIST to quantitatively assess representational resilience under progressive neuron removal. Second, we perform dynamic trajectory analyses across multiple CNN architectures and datasets—VGG-16 on CIFAR-100, and ResNet-50 on Tiny-ImageNet—within the framework of the Neuron Evolutionary Dynamics System (NEDS), with additional experiments on a three-layer MLP-Net with MNIST and ResNet-18 with CIFAR-10 provided in the Appendix. These experiments share a common methodology—tracking neuron activations, weights, and representational trajectories—while progressively scaling the model depth and dataset complexity. Across all settings, neurons are categorized into survived, eliminated, and other groups based on their long-term representational stability, providing a unified lens for comparing functional contributions across architectures and scales.

### 4.1 Ablation Experiment

We conducted ablation experiments on a CNN trained on MNIST in order to test the resilience of its internal representations under progressive neuron removal. The results are summarized in Figure 1. In the unperturbed network (0% ablation), accuracy reaches 99.3%, and the t-SNE projection reveals tight, well-separated clusters for each digit class, demonstrating a highly structured and linearly separable latent space. When 30% of the neurons are ablated, the accuracy remains essentially unchanged at 99.0%, and the clusters in the t-SNE embedding preserve their compactness and separation, indicating that the representational geometry is only minimally disturbed. This strongly suggests that the network possesses a large degree of representational redundancy. At 60% ablation, accuracy decreases slightly to 98.3%, and the clusters in the t-SNE space begin to expand and partially overlap, particularly at their boundaries. Although separability is degraded, the global structure of the representation is still preserved, implying that the network reallocates representational burden to the remaining subset of neurons. A qualitatively different figure emerges at 90% ablation: accuracy collapses to 64.9%, and the t-SNE projection shows the complete dissolution of the cluster structure, with digit classes intermingled in a disorganized cloud. To summarize, these results provide direct evidence for a Darwinian view of neural representations. Up to moderate levels of ablation, redundant or weakly integrated neurons are eliminated while the core representational structure is maintained, preserving both accuracy and geometric separability. However, once the ablation encroaches upon the Darwinianly selected subset of neurons that are critical for maintaining task-relevant structure, both accuracy and representation quality collapse. This pattern demonstrates that artificial neural networks exhibit precisely the mixture of robustness and selectivity predicted by Neural Darwinism: multiple neuronal assemblies initially compete to encode overlapping information, but only a small, stabilized ensemble ultimately sustains discriminative capacity under extreme perturbation.

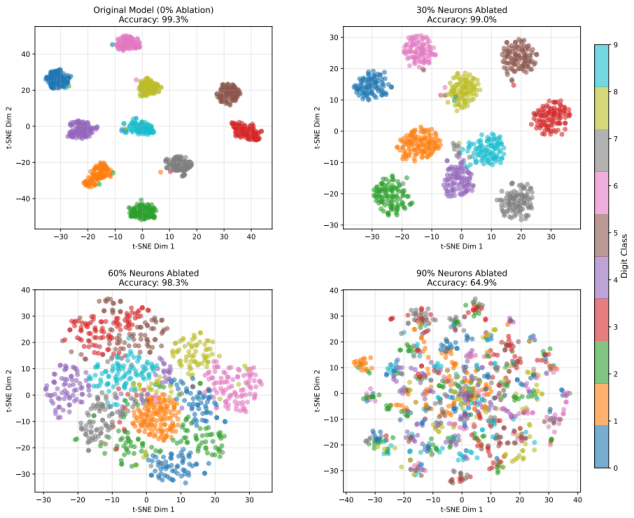


Figure 1: Ablation Experiment on MNIST.

## 4.2 VGG-16 on CIFAR-100

### 4.2.1 Dynamics Neuron Trajectory and Evolution Analysis

In the shallow layer of Figure 2, the dynamic PCA trajectory analysis reveals early indications of neuronal differentiation consistent with the principles of Neural Darwinism. Survived neurons—characterized by relatively higher activation levels and modestly higher weight magnitudes—tend to originate near the PCA origin at the start of training and progressively diverge along more extended and directionally consistent paths in activation space (Figure 2(a), top). Their trajectories exhibit sustained cumulative displacement over the training epochs (Figure 2(c), top), suggesting continued adaptation. Although the paths are often noisy and irregular, the outward spread indicates a gradual specialization process that may enable distinct low-level feature subspaces to emerge under task-driven gradient signals. By contrast, eliminated neurons generally follow more compact trajectories, remaining closer to the origin and displaying shorter cumulative displacements (Figure 2(a,c), top). Their temporal variance is lower and their trajectory curvature less pronounced, implying reduced representational change. The L2 weight norms of this group are on average slightly lower than those of survived neurons, but the distributions remain strongly overlapping (Figure 2(d), top). While gradient flow is not directly quantified, the limited representational mobility is consistent with the interpretation that these neurons receive weaker or less task-relevant updates during training. The neurons classified as "other" occupy an intermediate position. Their trajectories are more diffuse and less directionally stable (Figure 2(a), top), with cumulative lengths that are broadly comparable to those of survived neurons but accompanied by larger variance (Figure 2(c), top). Some display periods of outward displacement before stabilizing, while others remain closer to the origin throughout. This heterogeneity suggests that they represent a transitional population whose role is not firmly consolidated within the finite training horizon. Overall, these patterns support a local form of Neural Darwinism: within the shallow layer, a subset of neurons progressively differentiates and maintains higher representational activity, whereas others remain less engaged and gradually lose relative influence. The emergence of such divergence close to the raw input highlights that selection pressures may act from the earliest stages of learning.

In the middle layer—where hierarchical abstractions become more pronounced—the selective dynamics appear intensified relative to the shallow layer. PCA trajectories (Figure 2(a), middle) show that many survived neurons diverge from the origin early and continue outward with sustained displacement, though their paths remain noisy and variable. While most neurons cluster near the PCA origin, a modest subset of survived neurons extends into more distinct regions of the projection space, suggesting partial occupation of differentiated representational subspaces. Eliminated neurons, by contrast, display shorter or less stable trajectories: some show brief excursions before returning toward the origin, whereas others remain in intermediate posi-

tions without consistent outward drift. The "other" neurons again form a heterogeneous group, with some traveling considerable distances but frequently changing direction, and others staying confined near the origin. Quantitatively (Figure 2(c), middle), survived neurons accumulate the greatest trajectory lengths by the final epoch, though the margin over other groups is modest (approximately 0.3–0.4 units). In terms of weight evolution (Figure 2(d), middle), all neuron types exhibit monotonic L2 norm decay, with survived neurons showing a slightly slower decline and thus ending with marginally higher magnitudes. This suggests that survival is associated with maintaining relatively stronger synaptic weights, though the effect size is small. Collectively, the middle layer illustrates an intensification of competitive dynamics, where survived neurons maintain more persistent representational mobility, eliminated neurons adapt weakly or transiently, and the majority of units remain in flux without converging to stable roles.

In the deep layer—the final fully connected stage before classification—the rate of representational change appears increased, consistent with a late-phase consolidation process. Survived neurons continue to accumulate trajectory length (Figure 2(c), bottom), but at a quicker rate compared to earlier layers. In the PCA projection (Figure 2(a), bottom), these neurons drift outward from the origin and follow moderately directed paths, with curvature and displacement gradually increasing over time. This pattern indicates partial stabilization, consistent with their role in encoding higher-level, semantically richer features that require fewer adjustments once tuned. Weight magnitude curves (Figure 2(d), bottom) similarly show that survived neurons maintain slightly higher norms than eliminated and other neurons, though the separation remains limited. Eliminated neurons in the deep layer exhibit shorter cumulative trajectory lengths and modestly lower weight norms. While some early movement is evident, their displacement growth slows considerably, and their PCA positions remain relatively central, indicating constrained representational change. The "other" group again occupies an intermediate position, with moderate representational shifts and weight growth, suggesting residual but limited contribution to the final predictive function.

In summary, these observations align with a Neural Darwinism perspective in which neuronal survival reflects continued representational mobility and modestly stronger synaptic weights, while elimination corresponds to reduced or transient adaptation. Importantly, the presence of a large heterogeneous "other" group underscores that selection pressure operates continuously, and many neurons remain in transition rather than converging to stable roles. The progression from shallow to middle to deep layers reflects a gradual sharpening of selection, culminating in a smaller set of stabilized neurons in the deepest layer.

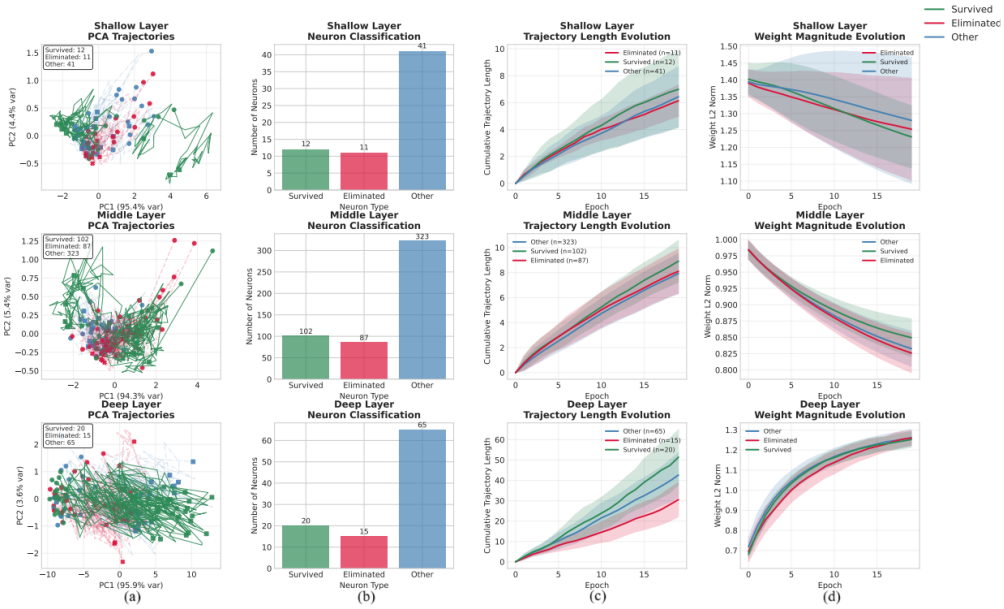


Figure 2: Dynamics Neuron Trajectory and Evolution Analysis on CIFAR-100.

#### 4.2.2 Static PCA and Activation Evolution

In the shallow layer, the final-epoch PCA projection in Figure 3 left shows that the first two principal components account for approximately 99% of the total variance (PC1: 95.4%, PC2: 4.4%), indicating that most inter-neuron activation variability can be represented in a low-dimensional subspace. Despite the limited receptive fields of early convolutional layers, survived neurons (green) occupy more peripheral regions of the PCA plane, with greater dispersion from the origin and from one another, suggesting a tendency toward differentiated feature sensitivities. By contrast, eliminated neurons (red) remain densely concentrated near the origin, reflecting low variance and limited representational differentiation. The activation evolution curves in Figure 3 bottom-left reinforce this observation: neurons with persistently higher activation norms tend to survive, while those with steadily declining norms move toward elimination. The distribution of survived neurons suggests diversity in low-level tuning—potentially edges or localized textures—that broadens the expressive basis available for subsequent layers. While the pattern is not definitive, it is qualitatively consistent with a threshold-like competitive process, in line with selection mechanisms hypothesized in Neural Darwinism.

In the middle layer, the PCA projection in Figure 3 middle explains roughly 99% of the variance (PC1: 94.3%, PC2: 5.4%). Here, survived neurons (green) are broadly distributed across the PCA space, often forming multiple partially separated groups, whereas eliminated neurons (red) cluster tightly near the origin. The other group (blue) occupies an intermediate band, positioned between the high-variance survived regions and the low-variance eliminated cluster. Activation evolution patterns (Figure 3 bottom-middle) reveal that survived neurons maintain high and relatively stable activation norms, eliminated neurons exhibit a consistent decline, and others remain at intermediate levels with mild fluctuations. The spread of survived neurons across the PCA space suggests an increasing degree of representational diversification at this stage, corresponding to the formation of mid-level abstractions. The non-random structure—characterized by local coherence within groups and broader separation between groups—indicates systematic partitioning of representational space. The central concentration of eliminated neurons, coupled with their declining activations, is consistent with redundancy or reduced gradient flow, whereas the transitional behavior of the other group may reflect delayed specialization.

In the deep layer, corresponding to the final fully connected stage, the PCA projection in Figure 3 right shows that the first two principal components explain about 99% of the variance (PC1: 95.9%, PC2: 3.6%). This high concentration of variance suggests a compressed and highly structured representational space, consistent with the role of this layer in integrating features for classification. Survived neurons are predominantly located in peripheral regions of the PCA plane, often grouped into small clusters. The activation trajectories in Figure 3 bottom-right show that survived neurons maintain higher and often increasing activation norms across training epochs, indicating sustained engagement in the final decision space. By contrast, eliminated neurons cluster near the PCA origin and exhibit consistently lower activation magnitudes and slower growth, suggestive of early functional deactivation. Other neurons occupy intermediate positions, with activation dynamics reflecting transient or weak selectivity that does not consolidate into either survival or elimination.

Overall, the three-layer comparison in Figure 3 highlights a consistent pattern: variance in activations is concentrated in a few dominant dimensions, survived neurons occupy more dispersed regions and sustain higher activity levels, while eliminated neurons remain near the origin with declining activations. The other group exhibits transitional characteristics, reflecting instability or incomplete specialization. The combined static and dynamic views are qualitatively consistent with a selection-based process in which functionally distinctive neurons persist and redundant ones fade, echoing principles of Neural Darwinism.

### 4.3 ResNet-50 on Tiny-ImageNet

#### 4.3.1 Dynamics Neuron Trajectory and Evolution Analysis

The dynamic PCA trajectories for the shallow layer (Figure 4(a), top) provide a temporal view of representational changes across training. Each trajectory reflects the evolution of a neuron's activation statistics in a low-dimensional PCA space. Survived neurons generally trace longer and more directionally consistent paths, suggesting progressive representational refinement and adaptation to task constraints. These

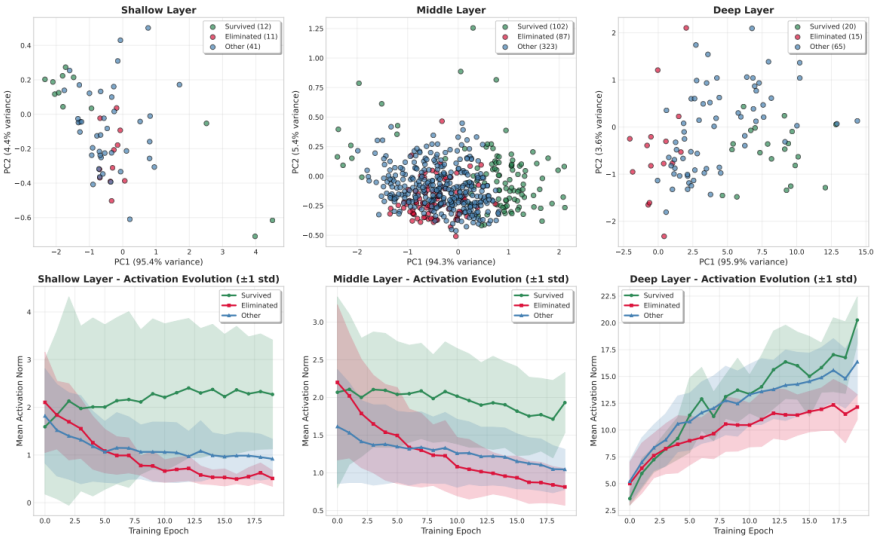


Figure 3: Static PCA and Activation Evolution on CIFAR-100.

trajectories tend to drift toward more structured regions of the PCA manifold, indicating a non-random reorganization that supports discriminative feature encoding. By contrast, eliminated neurons follow noticeably shorter, less exploratory trajectories that remain close to their initial locations in PCA space. This limited movement suggests functional stagnation, where neurons fail to develop distinctive representational roles, making them less competitive under resource-constrained optimization. Such stagnation is consistent with the early stages of Darwinian elimination, where less adaptive neurons gradually lose influence. Quantitative analysis reinforces these patterns. By the final epoch (Figure 4(c), top), survived neurons reach a median cumulative trajectory length of approximately 3.2 units, compared to 2.4 for eliminated neurons and around 2.3 for the "other" group. This indicates that sustained representational movement, rather than initial position, is associated with retention. Weight magnitude evolution (Figure 4(d), top) shows only minor differences across groups: eliminated neurons maintain slightly higher L2 norms than survived, with other neurons consistently lowest. The overall stability across training suggests that in shallow layers, synaptic resource allocation is relatively stable, with large-scale reallocation not yet evident.

The PCA trajectories for the middle layer (Figure 4(a), middle) capture a more pronounced divergence in representational dynamics across neuron types. Survived neurons traverse extended, often curved paths in the PCA space, largely oriented along PC1 (96.7% variance explained), with modest modulation along PC2 (3.1%). Although some trajectories exhibit partial rightward drift, clustering is weak and dispersion remains the dominant pattern. Eliminated neurons show substantially shorter displacements, remaining near their initialization points with fragmented paths. The intermediate "other" group exhibits moderate movement but does not match the sustained displacement of survivors. Trajectory length evolution (Figure 4(c), middle)) highlights this separation: by the end of training, survived neurons reach approximately 3.8 cumulative units, while eliminated neurons plateau near 2.8, with the "other" group is even lower. The gap is wider than in the shallow layer, underscoring that sustained representational plasticity becomes increasingly decisive at mid-level processing stages. Weight magnitude evolution (Figure 4(d), middle)) shows relatively stable rankings: eliminated neurons hold slightly higher norms than survived. The lack of pronounced growth for eliminated neurons—despite higher absolute values—suggests that strong initial parameterization was not matched by functional adaptation.

The dynamic PCA trajectories for the deep layer (Figure 4(a), bottom)) reveal the strongest differentiation in representational mobility. Survived neurons navigate long, structured arcs, reflecting continued refinement and consolidation of high-level semantic representations. These trajectories exhibit a clear convergence trend toward a more compact subregion of the PCA manifold, consistent with the emergence of attractor-like states that dominate the network's final decision space. Eliminated neurons, in contrast, show markedly shorter trajectories, with minimal displacement beyond early training epochs, indicating rapid stagnation.

Intermediate neurons display partial mobility but fail to achieve the sustained, directional movement observed in survivors. Trajectory length analysis (Figure 4(c), bottom)) accentuates this contrast: by the final epoch, survived neurons reach 7 cumulative units, while eliminated neurons remain near 4. This substantial gap indicates that extreme representational plasticity is a prerequisite for deep-layer survival. Weight magnitude evolution (Figure 4(d), bottom))) exhibit a global decay across all neuron types, converging toward lower norms over training. Survived and eliminated neurons follow nearly identical trajectories, with only a slight divergence at convergence, while other neurons stabilize at somewhat lower values. This suggests that in deeper layers, neuron differentiation is less pronounced in terms of synaptic strength, and survival is reflected more subtly in marginally higher residual weights.

Overall, these findings illustrate a progressive escalation of Darwinian dynamics across depth. In shallow layers, selection pressure is relatively permissive, with only subtle differences in trajectory and weight dynamics. In middle layers, divergence intensifies, as sustained plasticity becomes a critical factor for survival. In deep layers, selection culminates in large-scale consolidation, where only the most adaptive neurons persist to encode high-level abstractions. These results align with the three pillars of Neural Darwinism: variation (initially diverse representational behaviors), competition (divergent trajectory lengths under task pressure), and selective retention (resource amplification for neurons that maintain representational plasticity).

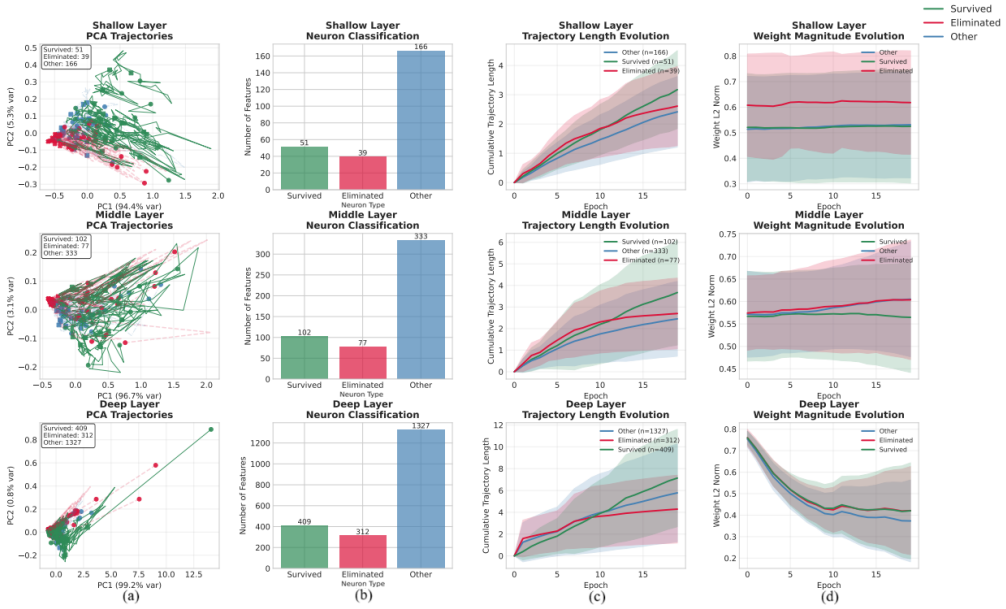


Figure 4: Dynamics Neuron Trajectory and Evolution Analysis on Tiny-ImageNet.

### 4.3.2 Static PCA and Activation Evolution

Figure 5 presents static PCA projections of final neuron states (top row) and mean activation norm trajectories (bottom row) across shallow, middle, and deep layer. In the shallow layer, PC1 explains 94.4% of variance while PC2 accounts for 5.3%, indicating that convergence is already dominated by a single representational axis. Survived neurons (green) occupy a moderately dispersed region offset from the origin, suggesting coordinated but not overly compact stabilization. Eliminated neurons (red) form a tight cluster near the lower-left quadrant, consistent with uniformly low activation magnitude. Other neurons (blue) lie in an intermediate zone, reflecting partial but incomplete adaptation. Activation dynamics confirm this structure: survived neurons maintain relatively high and stable norms, eliminated neurons exhibit monotonic decay toward near-zero activity, and other neurons follow an intermediate trajectory.

In the middle layer, PC1 accounts for 96.7% of variance and PC2 for 3.1%, indicating a stronger alignment to a single dominant direction compared to the shallow layer. Neurons distribute primarily along this axis: survivors occupy the central and positive range of PC1, reflecting sustained functional activity; eliminated

neurons cluster near the negative end of PC1, marking progressive silencing; and other neurons lie in between. Activation dynamics mirror this structure: survivors maintain consistently higher norms, eliminated neurons decay rapidly toward inactivity, and others exhibit moderate decline. These patterns suggest that competition in the middle layer becomes more directional, with survivors consolidating along the principal subspace while eliminated neurons are increasingly marginalized.

In the deep layer, PC1 captures 99.2% of the variance and PC2 only 0.8%, indicating an almost one-dimensional ordering of neuron states. Neurons concentrate into a dense central manifold dominated by "other" units, while eliminated neurons accumulate at the low-PC1 boundary and survived neurons extend outward along the positive-PC1 tail. Activation trajectories reinforce this separation: survivors rise rapidly in early epochs and stabilize at the highest activation norms, eliminated neurons decay swiftly toward silence, and others plateau at intermediate magnitudes. These dynamics suggest intensified axis-aligned selection, whereby survival is tied to displacement along the dominant representational axis.

Taken together, the progression across layers illustrates a Darwinian dynamic: initial variation, competitive decline of low-fitness units, and selective retention of survivors within compact, task-aligned manifolds. The increasing dominance of a single principal axis and the widening gap in activation dynamics demonstrate a layerwise intensification of selective pressures, culminating in deep-layer specialization.

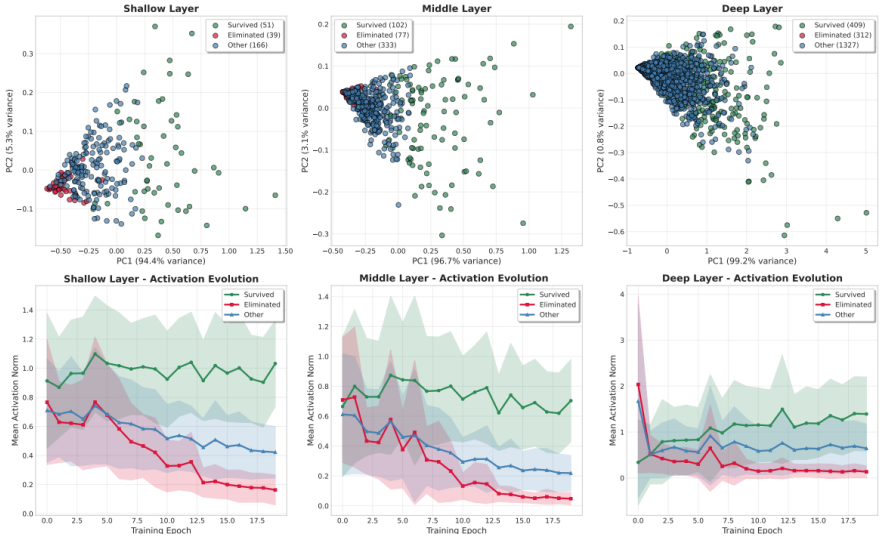


Figure 5: Static PCA and Activation Evolution on Tiny-ImageNet.

## 5 Conclusion

This study provides empirical evidence that CNNs exhibit representational dynamics that are consistent with the principles of Neural Darwinism. Across architectures and datasets, we observe recurring signatures of variation, competition, and selective retention: neurons initially follow diverse representational trajectories, but only a subset sustains adaptive movement, stronger weight magnitudes, and higher activation norms. The ablation experiment highlights both robustness, arising from representational redundancy, and fragility, once the implicitly selected subset of critical neurons is disrupted. Layerwise analyses further suggest that selection pressure intensifies with depth, culminating in compact ensembles of specialized neurons that dominate high-level feature encoding.

These findings advance our understanding of representation learning by framing it not solely as gradient-driven optimization, but also as an emergent selection-like process operating at the neuron level. This dual perspective highlights how neural networks balance redundancy with specialization. Future work may investigate whether similar dynamics generalize to recurrent and transformer architectures, and explore implications for pruning, interpretability, and biologically inspired models of computation.

## References


Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(50):1–34, 2018a.

Alessandro Achille and Stefano Soatto. Information dropout: Learning optimal representations through noisy computation. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2897–2905, 2018b.

Kwangjun Ahn, et al. Learning threshold neurons via edge of stability. *Advances in Neural Information Processing Systems*, 36:19540–19569, 2023.

Samuel Ainsworth, et al. Git re-basin: Merging models modulo permutation symmetries. In *The Eleventh International Conference on Learning Representations*, 2023.

Rana Ali Amjad, et al. Understanding neural networks and individual neuron importance via information-ordered cumulative ablation. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12): 7842–7852, 2021.

Prithaj Banerjee, et al. Deep networks learn features from local discontinuities in the label function. In *The Thirteenth International Conference on Learning Representations*, 2025.

David Bau, et al. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6541–6549, 2017.

Silvere Bonnabel. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013.

Ivan Butakov, et al. Information bottleneck analysis of deep neural networks via lossy compression. In *The Twelfth International Conference on Learning Representations*, 2024.

Chao Chen, et al. Optimal parameter and neuron pruning for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 36:52293–52311, 2023.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

Lénaïc Chizat and Praneeth Netrapalli. The feature speed formula: a flexible approach to scale hyper-parameters of deep neural networks. *Advances in Neural Information Processing Systems*, 37:62362–62383, 2024.

Arthur Conmy, et al. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.

Dániel Czégel, et al. Novelty and imitation within the brain: a darwinian neurodynamic approach to combinatorial problems. *Scientific reports*, 11(1):12513, 2021.

Felix Draxler, et al. Essentially no barriers in neural network energy landscape. In *International conference on machine learning*, pp. 1309–1318. PMLR, 2018.

Guodong Du, et al. Impacts of darwinian evolution on pre-trained deep neural networks. In *2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 1907–1912. IEEE, 2024.

Gerald M Edelman. *Neural Darwinism: The theory of neural group selection.* Basic Books, 1987.

Utku Evci, et al. Rigging the lottery: Making all tickets winners. In *International conference on machine learning*, pp. 2943–2952. PMLR, 2020.

Gongfan Fang, et al. Depgraph: Towards any structural pruning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16091–16101, 2023.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019.

Jingwen Fu, et al. Learning trajectories are generalization indicators. *Advances in Neural Information Processing Systems*, 36:71053–71077, 2023.

Timur Garipov, et al. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31, 2018.

Song Han, et al. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28, 2015.

X.Y. Han, et al. Neural collapse under MSE loss: Proximity to and dynamics on the central path. In *International Conference on Learning Representations*, 2022.

Arthur Jacot, et al. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Nitish Shirish Keskar, et al. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.

Been Kim, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Simon Kornblith, et al. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Namhoon Lee, et al. SNIP: SINGLE-SHOT NETWORK PRUNING BASED ON CONNECTION SENSITIVITY. In *International Conference on Learning Representations*, 2019.

Jingyuan Li, et al. Amag: Additive, multiplicative and adaptive graph neural network for forecasting neuron activity. *Advances in Neural Information Processing Systems*, 36:8988–9014, 2023.

Zhuang Liu, et al. Rethinking the value of network pruning. In *International Conference on Learning Representations*, 2019.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020.

Song Mei, et al. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

Pavlo Molchanov, et al. Pruning convolutional neural networks for resource efficient inference. In *International Conference on Learning Representations*, 2017.

Ari Morcos, et al. One ticket to win them all: generalizing lottery ticket initializations across datasets and optimizers. *Advances in neural information processing systems*, 32, 2019.

Jeffrey Pennington, et al. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. *Advances in neural information processing systems*, 30, 2017.

Scott Pesme and Nicolas Flammarion. Saddle-to-saddle dynamics in diagonal linear networks. *Advances in Neural Information Processing Systems*, 36:7475–7505, 2023.

Ben Poole, et al. Exponential expressivity in deep neural networks through transient chaos. *Advances in neural information processing systems*, 29, 2016.

Victor Quétu, et al. The simpler the better: An entropy-based importance metric to reduce neural networks' depth. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 92–108. Springer, 2024.

Maithra Raghu, et al. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.

Victor Sanh, et al. Movement pruning: Adaptive sparsity by fine-tuning. *Advances in neural information processing systems*, 33:20378–20389, 2020.

A Saxe, et al. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *Proceedings of the International Conference on Learning Represenatations 2014*. International Conference on Learning Represenatations 2014, 2014.

Samuel S. Schoenholz, et al. Deep information propagation. In *International Conference on Learning Representations*, 2017.

Mohammad Javad Shafiee, et al. Deep learning with darwin: Evolutionary synthesis of deep neural networks. *Neural processing letters*, 48(1):603–613, 2018.

Guobin Shen, et al. Brain-inspired neural circuit evolution for spiking neural networks. *Proceedings of the National Academy of Sciences*, 120(39):e2218173120, 2023.

Hangchi Shen, et al. Rethinking the membrane dynamics and optimization objectives of spiking neural networks. *Advances in Neural Information Processing Systems*, 37:92697–92720, 2024.

Xinyu Shi, et al. Spikingresformer: bridging resnet and vision transformer in spiking neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5610–5619, 2024.

Gabriele Spadaro, et al. Shannon strikes again! entropy-based pruning in deep neural networks for transfer learning under extreme memory and computation budgets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1518–1522, 2023.

Mukund Sundararajan, et al. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Aaquib Syed, et al. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 407–416, 2024.

Atashgahi Zahra, et al. Supervised feature selection with neuron evolution in sparse neural networks. *Transactions on Machine Learning Research*, 2023(2), 2023.

Xiawu Zheng, et al. An information theory-inspired strategy for automated network pruning. *International Journal of Computer Vision*, pp. 1–28, 2025.

# A Appendix

## A.1 Notation and Preliminaries

To maintain consistency with the main text, we briefly recap key notations:

- Neural network $f_\theta : \mathcal{X} \to \mathcal{Y}$, layers $\{L_k\}_{k=1}^{D}$, where layer $k$ contains $n_k$ neurons indexed by $i$.
- Parameters of neuron $i$ at layer $k$: weights $w_i^{(k)}(t) \in \mathbb{R}^{d_{k-1}}$, bias $b_i^{(k)}(t) \in \mathbb{R}$, activation function $\sigma$.
- Activation:
$$a_i^{(k)}(x,t) := \sigma\left(\langle w_i^{(k)}(t), h^{(k-1)}(x,t)\rangle + b_i^{(k)}(t)\right). \tag{11}$$
- Neuron state vector (compound state):
$$\psi_i^{(k)}(t) := \left[w_i^{(k)}(t),\, b_i^{(k)}(t),\, \mu_i^{(k)}(t),\, g_i^{(k)}(t),\, \mathcal{I}_i^{(k)}(t)\right], \tag{12}$$
where
$$\mu_i^{(k)}(t) = \mathbb{E}_{x\sim\mathcal{D}}[a_i^{(k)}(x,t)], \quad g_i^{(k)}(t) = \mathbb{E}_{x\sim\mathcal{D}}\left[\frac{\partial \mathcal{L}(x)}{\partial a_i^{(k)}}\right],$$
and $\mathcal{I}_i^{(k)}(t)$ is the differential entropy of the activation.
- State evolution (ODE form, main text eq.(6)):
$$\frac{d}{dt}\psi_i^{(k)}(t) = \mathbf{F}_\theta^{(k)}\big(\psi_i^{(k)}(t), \mathcal{D}, \mathcal{L}\big). \tag{13}$$

Other quantities such as trajectory length $\mathcal{L}_i^{(k)}$, terminal stochasticity $\mathcal{S}_i^{(k)}$, integrated entropy $\mathfrak{H}_i^{(k)}$, and fitness $\Phi_i^{(k)}$ follow the main text definitions.

## A.2 Supplementary Technical Assumptions

We explicitly state additional mild assumptions needed for mathematical rigor and numerical stability. These assumptions clarify the hidden conditions of the main results.

**Assumption S1 (Smoothness, boundedness, and trajectory length)**
For each layer $k$, the vector field $\mathbf{F}_\theta^{(k)}(\psi, t)$ is locally Lipschitz in $\psi$ and measurable in $t$. There exist constants $B_g, B_a, B_\psi > 0$ such that for all $t \geq 0$:

$$\|g_i^{(k)}(t)\| \leq B_g, \quad \mathrm{Var}[a_i^{(k)}(t)] \leq B_a, \quad \|\psi_i^{(k)}(t)\| \leq B_\psi.$$

Moreover, the trajectory length $\mathcal{L}_i^{(k)}(T)$ is bounded for any finite $T$.

**Assumption S2 (Sub-exponential tails / sub-Gaussianity of activations)**
For all neurons $i, k$ and times $t$, the distribution of $a_i^{(k)}(x,t)$ over $x \sim \mathcal{D}$ is sub-Gaussian or at least has sub-exponential tails, enabling concentration bounds for sample estimators.

**Assumption S3 (Controlled Gaussian entropy approximation error)**
There exists a constant $C_{\text{gauss}} \geq 1$ such that for all neurons $i, k$ and times $t$,

$$\mathcal{I}_i^{(k)}(t) \leq \frac{1}{2}\log\left(2\pi e\,\mathrm{Var}[a_i^{(k)}(t)]\right) \leq \mathcal{I}_i^{(k)}(t) + \log C_{\text{gauss}}.$$

### A.3 Well-Posedness of the Continuous NEDS

Under Assumption S1, the vector field $\mathbf{F}_\theta^{(k)}$ is locally Lipschitz, thus by Picard–Lindelöf theorem, for any initial value $\psi_i^{(k)}(0)$ there exists a unique local solution. Boundedness and growth controls ensure global existence on finite intervals and continuous dependence on initial conditions and parameters.

### A.4 Detailed Proofs of Main Lemmas and Theorems

#### A.4.1 Lemma: Instability Predicts Elimination

**Lemma A.1** (Instability Predicts Elimination, rigorous form)**.** *Suppose there exist constants $\delta_{\max} > 0$, $c_I > 0$, and $\delta > 0$, such that for all sufficiently large $T$,*

1. *Terminal stochasticity satisfies $\mathcal{S}_i^{(k)}(T) \geq \delta_{\max}$.*
2. *The entropy derivative satisfies $\frac{d}{dt}\mathcal{I}_i^{(k)}(t) \leq -c_I < 0$ for all $t \in [T-\delta, T]$.*
3. *The trajectory length $\mathcal{L}_i^{(k)}(T)$ grows at most linearly with $T$.*

*Then,*

$$\lim_{T\to\infty} \Phi_i^{(k)}(T) = -\infty.$$

*Proof.* Recall the fitness function

$$\Phi_i^{(k)}(T) = \alpha\mathcal{L}_i^{(k)}(T) - \beta\mathcal{S}_i^{(k)}(T) + \gamma\mathcal{E}_i^{(k)}(T),$$

where $\mathcal{E}_i^{(k)}(T) = \int_0^T \mathcal{I}_i^{(k)}(t)dt$.

Since $\frac{d}{dt}\mathcal{I}_i^{(k)}(t) \leq -c_I < 0$ on every tail interval $[T-\delta, T]$, the integral entropy decreases at least by $c_I\delta$ each such window. If such intervals appear infinitely often as $T \to \infty$, then $\mathcal{E}_i^{(k)}(T)$ diverges to $-\infty$ linearly in $T$.

The terminal stochasticity term $-\beta\mathcal{S}_i^{(k)}(T)$ contributes a negative term bounded below by $-\beta\delta_{\max}$ each window.

The trajectory length term $\alpha\mathcal{L}_i^{(k)}(T)$ grows at most linearly and cannot offset the unbounded negative contribution from the integral entropy and terminal stochasticity terms.

Hence, $\Phi_i^{(k)}(T) \to -\infty$. □

#### A.4.2 Theorem: Fitness Threshold Implies Gradient-Variance Contribution

**Theorem A.2** (Quantitative lower bound on gradient-variance product)**.** *Under Assumptions S1–S3, there exist constants $\tau, \kappa > 0$ depending on $\alpha, \beta, \gamma$ and bounding constants, such that if*

$$\Phi_i^{(k)}(T) \geq \tau,$$

*then the following quantity is bounded below by $\kappa$:*

$$\Delta_i^{(k)} := \mathbb{E}_x\left[\left(\frac{\partial\mathcal{L}(x)}{\partial a_i^{(k)}}\right)^2 \mathrm{Var}[a_i^{(k)}(x)]\right].$$

*Proof sketch.* Using Assumption S3, differential entropy $\mathcal{I}_i^{(k)}(t)$ relates to the log-variance of activations with controlled approximation error.

The trajectory length $\mathcal{L}_i^{(k)}(T)$ can be lower bounded by the integrated norm of gradients $\|g_i^{(k)}(t)\|$.

Applying Cauchy–Schwarz inequality to integrals of the form

$$\int_0^T \mathbb{E}\left[\left(\frac{\partial \mathcal{L}}{\partial a_i^{(k)}}\right)^2\right] \mathrm{Var}[a_i^{(k)}(t)]dt,$$

and combining with the linear lower bounds on $\mathcal{L}_i^{(k)}(T)$ and $\mathcal{E}_i^{(k)}(T)$ implied by $\Phi_i^{(k)}(T) \geq \tau$, one obtains a strictly positive lower bound $\kappa$ on the time-averaged gradient-variance product $\Delta_i^{(k)}$. □

#### A.4.3 Theorem: Coupled Survival Principle

**Theorem A.3** (Coupled Survival Principle)**.** *Suppose that for some $\mu > 0$ and a subset $\mathcal{S}^{(k)} \subseteq \{1, \ldots, n_k\}$ of survived neurons at layer $k$, the layer-to-layer coupling matrix $\mathbf{M}_{k,k+1}(t)$ satisfies*

$$\sum_{i \in \mathcal{S}^{(k)}} \mathbf{M}_{k,k+1}(i,j)(t) \geq \epsilon > 0,$$

*for all neurons $j$ in layer $k+1$ and all sufficiently large $t$.*

*Then, there exists $\eta = \eta(\mu, \epsilon, \text{Lipschitz constants}) > 0$ such that at least an $\eta$ proportion of neurons in layer $k+1$ achieve high fitness (survival).*

*Proof sketch.* Positive lower bounds on coupling imply sustained energy inflow to downstream neurons. Via the Lipschitz continuity of the fitness function and the smoothness of the dynamics, survival of upstream neurons forces a positive measure of downstream neurons to cross the survival threshold.

Technical details involve integrating the coupled system over suitable time windows and applying compactness arguments. □

#### A.4.4 Theorem: Global Convergent Specialization

**Theorem A.4** (Global Convergent Specialization)**.** *If the total Darwinian flow energy $\mathcal{E}_{\text{Darwin}} \geq \epsilon > 0$ is bounded away from zero and the fitness functions $\Phi_i^{(k)}$ are sufficiently smooth and Lipschitz continuous, then as $t \to \infty$, the proportion of neurons with fitness below any fixed threshold tends to zero.*

*Proof sketch.* Construct a suitable Lyapunov function based on the sum over neurons of a decreasing convex function of their fitness values. The positive lower bound on Darwinian flow energy ensures the Lyapunov function decreases over time, implying convergence to the set of neurons with high fitness. LaSalle's invariance principle excludes non-convergent oscillations. □

### A.5 Discrete-Time Approximation and Relation to SGD

Actual training proceeds in discrete time steps, typically iterations or epochs. The continuous-time NEDS dynamics approximate the discrete SGD updates as follows:

- Discrete parameter update:
$$\theta_{t+1} = \theta_t - \eta_t \widehat{\nabla}_\theta \mathcal{L}(B_t; \theta_t),$$
where $B_t$ is the mini-batch at step $t$.
- For small learning rate $\eta_t$, the discrete updates approximate the stochastic differential equation
$$d\theta_t = -\mathbb{E}_x[\nabla_\theta \mathcal{L}(x; \theta_t)]dt + \sqrt{\eta_t}\Sigma(\theta_t)dW_t,$$
with $W_t$ Brownian motion and $\Sigma$ the noise covariance.

- Correspondingly, the neuron state differences

$$\Delta\psi_i^{(k)}(t) := \psi_i^{(k)}(t+1) - \psi_i^{(k)}(t)$$

  approximate $\frac{d}{dt}\psi_i^{(k)}(t)$.

- Therefore,

$$\mathcal{L}_i^{(k)} \approx \sum_t \|\Delta\psi_i^{(k)}(t)\|_2, \quad \mathcal{S}_i^{(k)} \approx \frac{1}{\delta}\sum_{t=T-\delta}^{T-1} \|\Delta\psi_i^{(k)}(t)\|_2^2, \quad \mathcal{E}_i^{(k)} \approx \sum_t \mathcal{I}_i^{(k)}(t).$$

Discrete estimation errors arise from step size, mini-batch noise, and finite sample effects.

### A.6 Numerical Estimation of Key Quantities

#### A.6.1 Mean activation $\mu_i^{(k)}$ and mean gradient $g_i^{(k)}$

Evaluate on a separate evaluation dataset $\mathcal{D}_{\text{eval}}$:

$$\mu_i^{(k)} = \frac{1}{|\mathcal{D}_{\text{eval}}|}\sum_{x\in\mathcal{D}_{\text{eval}}} a_i^{(k)}(x), \quad g_i^{(k)} = \frac{1}{|\mathcal{D}_{\text{eval}}|}\sum_{x\in\mathcal{D}_{\text{eval}}} \frac{\partial\mathcal{L}(x)}{\partial a_i^{(k)}}.$$

#### A.6.2 Variance $\mathrm{Var}[a_i^{(k)}]$

Estimated as the unbiased sample variance over $\mathcal{D}_{\text{eval}}$.

#### A.6.3 Differential Entropy $\mathcal{I}_i^{(k)}$

Three common estimators:

1. **Gaussian plug-in:**

$$\widehat{\mathcal{I}}_{\text{gauss}} = \frac{1}{2}\log\left(2\pi e\widehat{\mathrm{Var}}[a_i^{(k)}]\right).$$

   Fast but biased if distribution is non-Gaussian.

2. **Kernel density estimation (KDE):** Estimate density $\widehat{p}(z)$ via KDE and compute

$$\widehat{\mathcal{I}} = -\int \widehat{p}(z)\log\widehat{p}(z)dz.$$

3. **K-nearest neighbor (Kozachenko–Leonenko) estimator:** Uses neighbor distances among samples for nonparametric entropy estimation.

#### A.6.4 Trajectory length $\mathcal{L}_i^{(k)}$ and terminal stochasticity $\mathcal{S}_i^{(k)}$

Computed from saved parameter snapshots at each discrete step $t$:

$$\Delta\psi_i^{(k)}(t) = \|\psi_i^{(k)}(t+1) - \psi_i^{(k)}(t)\|_2,$$

then

$$\mathcal{L}_i^{(k)} = \sum_t \Delta\psi_i^{(k)}(t), \quad \mathcal{S}_i^{(k)} = \frac{1}{\delta}\sum_{t=T-\delta}^{T-1}\left(\Delta\psi_i^{(k)}(t)\right)^2.$$

### A.7 Additional Experiments on Three-layer MLP-Net with MNIST

#### A.7.1 Dynamics Neuron Trajectory and Evolution Analysis.

Figure 6(a), top shows the PCA-projected trajectories of shallow-layer neurons across training. Survived neurons (green) follow relatively long and directed paths, indicating sustained representational change. Their motion exhibits fewer reversals than eliminated neurons (red), which instead display short and irregular trajectories, often collapsing toward the origin. This contrast is reflected quantitatively in Figure 6(c), top, where cumulative trajectory length grows steadily for survived neurons. The weight dynamics in Figure 6(d), top reinforce this pattern: survived neurons exhibit increasing $L_2$ norms of incoming weights, whereas eliminated neurons remain almost flat, suggesting a gradual withdrawal of representational capacity. Collectively, these results indicate that even in the shallow layer, gradient descent implicitly differentiates between neurons that maintain sustained alignment with the loss signal and those that do not.

In the middle layer (Figure 6(a), middle), the divergence becomes more pronounced. Survived neurons trace longer and more coherent trajectories, while eliminated neurons remain short and close to the origin. This is supported by Figure 6(c), middle, where the cumulative trajectory length of eliminated neurons grows at a substantially lower rate than that of survived neurons, already showing a marked slowdown by Epoch 2. Weight norms (Figure 6(d), middle) again show a separation, with growth for survived neurons and almost stagnation for eliminated ones. Compared to the shallow layer, the selective bottleneck appears stronger: neurons that fail to establish early alignment with the optimization signal are rapidly marginalized. This suggests that middle-layer neurons, receiving both bottom-up and top-down gradients, undergo more stringent selection toward functional specialization.

The deep layer presents a smaller sample size, but a similar trend is observable. As shown in Figure 6(a), bottom, survived neurons follow more extended trajectories, while the eliminated neuron remains nearly static. Correspondingly, trajectory length (Figure 6(c), bottom) and weight norm evolution (Figure 6(d), bottom) both indicate continued adaptation for survived neurons but not for the eliminated one. Although the limited number of neurons precludes strong statistical claims, the observed divergence suggests that selection pressures persist even near the output. Importantly, this implies that architectural proximity to the loss signal alone does not guarantee survival; functional alignment remains necessary.

Overall, Figure 6 highlights a consistent layer-wise pattern: shallow-layer neurons exhibit the earliest divergence, middle-layer neurons experience intensified selection with clearer separation between survived and eliminated groups, and deep-layer neurons—though fewer—still reflect selective retention. These results support the view that neuron survival is not imposed externally but emerges from the training dynamics, with selection pressures varying in strength across depth.

#### A.7.2 Static PCA and Activation Evolution

Figure 7 (top-left) presents the final-epoch PCA projection of first-layer neuron activations. Neurons categorized as survived occupy relatively dispersed regions, often farther from the origin, which correlates with higher activation magnitude and greater variance. Eliminated neurons cluster near the origin, suggesting low-output states with reduced contribution to the representational space. The majority of neurons fall into the "other" category, exhibiting intermediate positions without clear clustering, reflecting heterogeneous or drifting roles during training. The activation-norm trajectories (Figure 7, bottom-left) provide a temporal view of this differentiation. Survived neurons increase their average norm across epochs, indicating sustained engagement with learning signals. Eliminated neurons, in contrast, display a gradual decline toward low, stable norms, consistent with functional silencing. The "other" group remains in an intermediate range, suggesting partial adaptation without clear reinforcement or suppression.

In the middle layer (Figure 7, top-middle), the PCA projection reveals that eliminated neurons are shifted toward the positive-PC1 periphery, while survived neurons occupy a broader and more heterogeneous region spanning both central and peripheral zones. The activation trajectories (bottom-middle) sharpen this divergence: survived neurons exhibit a sustained rise in activation norm, whereas eliminated neurons remain suppressed with only marginal growth. Taken as a whole, these patterns suggest that selection-like dynam-

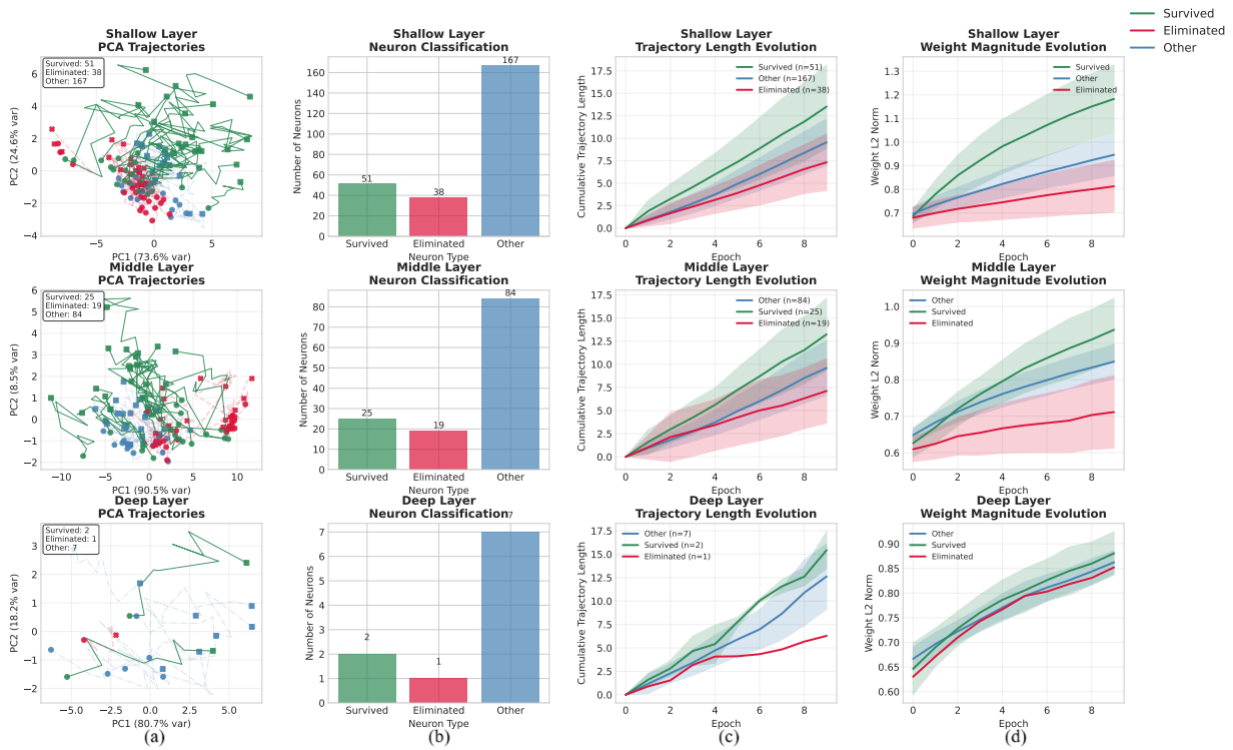


Figure 6: Dynamics Neuron Trajectory and Evolution Analysis on MNIST.

ics manifest most clearly in intermediate layers, where neurons are actively sorted into amplifying versus stagnant trajectories.

For the deep layer (Figure 7, top-right), the neuron count is small (only 2 survived and 1 eliminated), limiting statistical strength. The survived units exhibit higher final activation norms (bottom-right), whereas the eliminated unit declines toward a baseline. While this pattern resembles earlier layers, the small sample size precludes strong generalization.

Overall, the combination of static PCA projections and dynamic activation curves provides complementary evidence of neuron-level differentiation across depth. These results are consistent with the hypothesis that overparameterized networks allocate representational capacity unevenly, with some neurons reinforced while others become marginalized. However, the analyses are correlational and limited by dimensionality reduction and sample imbalance, particularly in deeper layers.

## A.8 Additional Experiments on ResNet-18 with CIFAR-10

### A.8.1 Dynamics Neuron Trajectory and Evolution Analysis

The shallow layer dynamic PCA trajectories (Figure 8(a), top) show that neuron activations in early convolutional layers—often assumed to encode low-level, generic features—already exhibit signs of representational divergence. Survived neurons tend to follow more stable and moderately directed paths in the PCA manifold, with reduced dispersion over training, suggesting a gradual consolidation toward more compact representational regions. In contrast, eliminated neurons display more irregular trajectories, with frequent directional changes and less coherence, indicating comparatively unstable representational roles.

This difference is also reflected in the cumulative trajectory length evolution (Figure 8(c), top): survived neurons maintain consistently higher cumulative movement compared to eliminated neurons, suggesting greater adaptability and sustained representational change across epochs. While the absolute gap is modest,

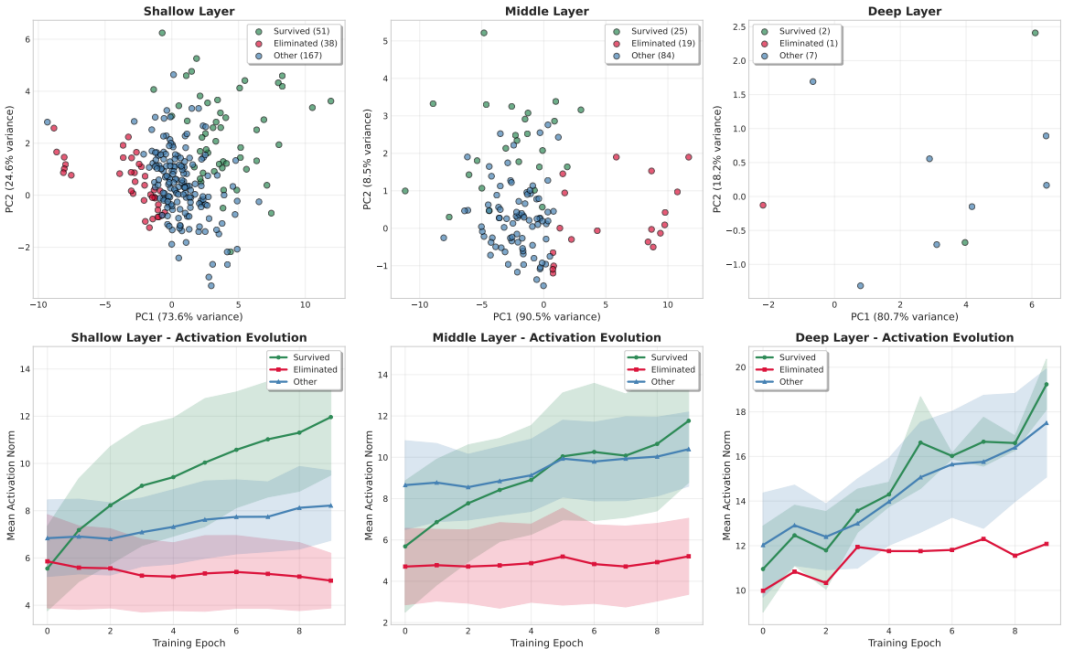


Figure 7: Static PCA and Activation Evolution on MNIST.

survived neurons display more continuous directional displacement, whereas eliminated neurons tend to plateau earlier, consistent with a potential stagnation of their representational contribution.

From a structural perspective, the weight magnitude evolution (Figure 8(d), top) indicates that the convolutional filters corresponding to survived neurons generally retain slightly higher L2 norms throughout training, while those of eliminated neurons remain lower. This trend is consistent with the interpretation that neurons contributing more strongly to gradient pathways receive relatively greater synaptic reinforcement, whereas others undergo gradual attenuation. Collectively, these results suggest that even shallow layers are subject to competitive dynamics, where only subsets of neurons demonstrating sustained utility remain functionally active.

The middle layers serve as a transitional zone between low-level and high-level representations, and this role is reflected in the diversity of neuron trajectory dynamics. As shown in the dynamic PCA projections (Figure 8(a), middle), neurons in these layers exhibit heterogeneous representational paths over training. Survived neurons tend to follow longer and more coherent trajectories, often traversing distinct regions of the PCA manifold, suggesting a gradual alignment with intermediate-level features. By contrast, many eliminated neurons show less coherent movement, with shorter and more irregular trajectories, though some maintain moderate displacement comparable to the "other" group.

The cumulative trajectory length curves (Figure 8(c), middle) provide quantitative support for these observations: on average, survived neurons reach greater cumulative lengths than eliminated or other neurons, reflecting more sustained representational plasticity. Eliminated neurons continue to grow but at a slower rate, with later signs of stagnation. A similar pattern is visible in the weight magnitude evolution (Figure 8(d), middle), where survived neurons exhibit slightly higher L2 norms than eliminated neurons. Although the difference is modest, its persistence across epochs indicates that neurons contributing more to the task tend to retain larger weight magnitudes. As a whole, these results suggest that the middle layers serve as a representational bottleneck where neurons undergo implicit selection, retaining those with flexible and task-relevant transformations.

In the deep layer, the contrast between neuron groups becomes more pronounced. As illustrated by the dynamic PCA trajectories (Figure 8(a), bottom), survived neurons follow long, smooth, and more aligned paths through representation space, frequently converging to structured low-dimensional subspaces. These

neurons appear to encode abstract, class-discriminative information that supports final classification. In contrast, eliminated neurons reveal short, noisy, and non-convergent trajectories, often stagnating or oscillating without clear direction, suggesting limited long-term utility.

This distinction is also evident in the trajectory length evolution (Figure 8(c), bottom), where survived neurons maintain the highest cumulative distances relative to eliminated neurons. These lengths reflect sustained representational change that tracks increasing class separability. Moreover, the variance among survived neurons is smaller, suggesting more constrained roles in the deep layer. The weight magnitude evolution (Figure 8(d), bottom) further highlights this separation: survived neurons retain high L2 norms, while eliminated neurons undergo progressive attenuation. The resulting divergence is strongest in this layer, consistent with stronger selective pressure as representations become more task-specific.

Overall, these findings are consistent with the framework of Neural Darwinism: across layers, neurons exhibit competitive dynamics shaped by their sustained utility. While shallow layers already show signs of divergence, the middle layers intensify selective processes, and the deep layers consolidate highly specialized neurons. The evidence from trajectory dynamics and weight evolution collectively supports the interpretation that representational selection operates hierarchically, shaping survival and elimination throughout the network.

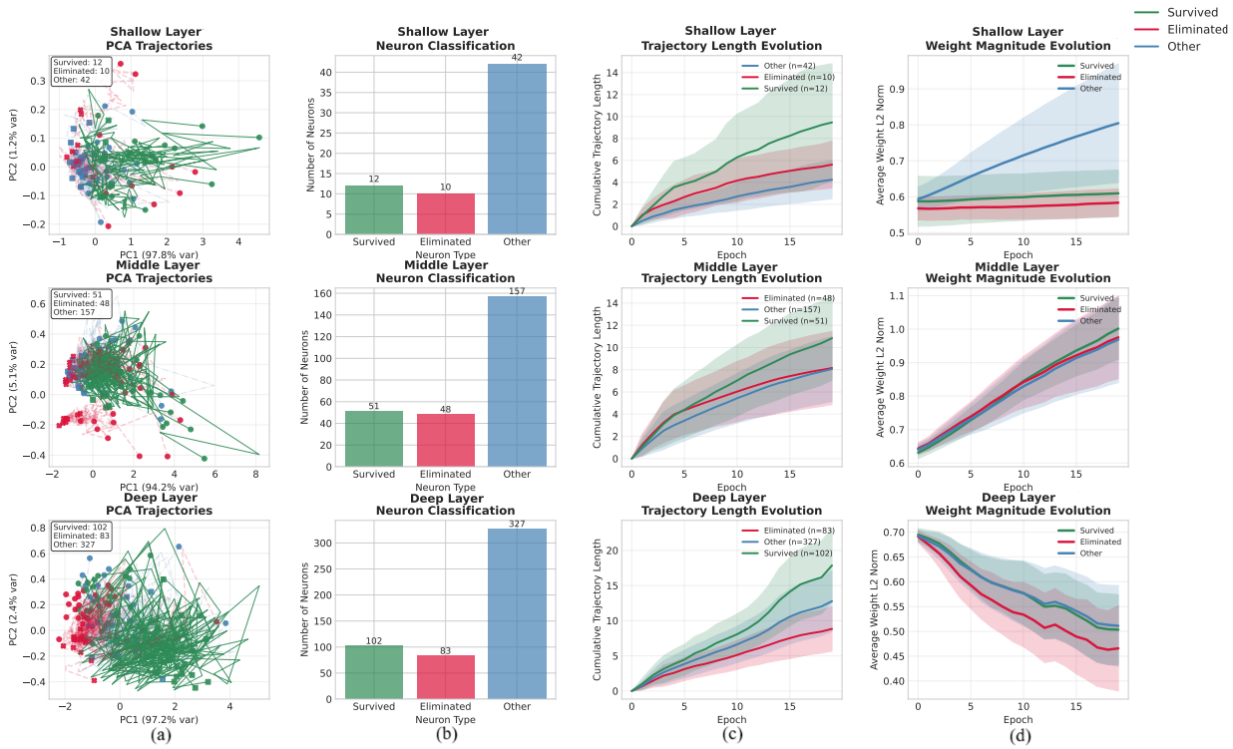


Figure 8: Dynamics Neuron Trajectory and Evolution Analysis on CIFAR-10.

### A.8.2 Static PCA and Activation Evolution

In Figure 9 left and bottom-left, the PCA projection (97.8% variance explained by PC1) shows that survived neurons occupy a relatively more compact region of the activation space, while eliminated neurons are scattered toward peripheral, low-density zones. Other neurons form a diffuse cloud spanning both regions. The activation evolution curves corroborate this structure: survived neurons sustain moderately higher activation norms with gradual stabilization, whereas eliminated neurons display persistently weak activations, and others remain intermediate. These patterns suggest that even at early layers—traditionally considered low-level feature extractors—there is already a degree of representational competition, consistent with the Neural Darwinism view that selection pressure operates from the outset of learning.

In Figure 9 middle and bottom-middle, the PCA embedding (94.2% variance explained by PC1) reveals a clearer differentiation than in shallow layers. Survived neurons cluster more tightly along dominant axes, while eliminated neurons are dispersed across orthogonal or low-density subspaces. Other neurons span an intermediate gradient, partially overlapping both groups. The activation dynamics mirror this structure: survived neurons maintain higher, stable activations, eliminated neurons steadily decline. These findings are consistent with the hypothesis that middle layers face stronger selective pressure, as they form an intermediate representational bottleneck where neurons must converge toward task-relevant manifolds to persist.

In Figure 9 right and bottom-right, in the final layer (97.2% variance explained by PC1), survived neurons are broadly distributed along the dominant axis but relatively compact along PC2, indicating alignment to a high-variance representational subspace. Eliminated neurons are concentrated in the lower-PC1 region, while others populate an intermediate zone overlapping both groups. The activation evolution curves reinforce this separation: survived neurons sustain the highest activation norms with relative stability, eliminated neurons remain consistently suppressed, and others occupy intermediate levels. Therefore, the static and dynamic views suggest that deep layers culminate the Darwinian competition, consolidating a high-utility representational manifold surrounded by marginal units.

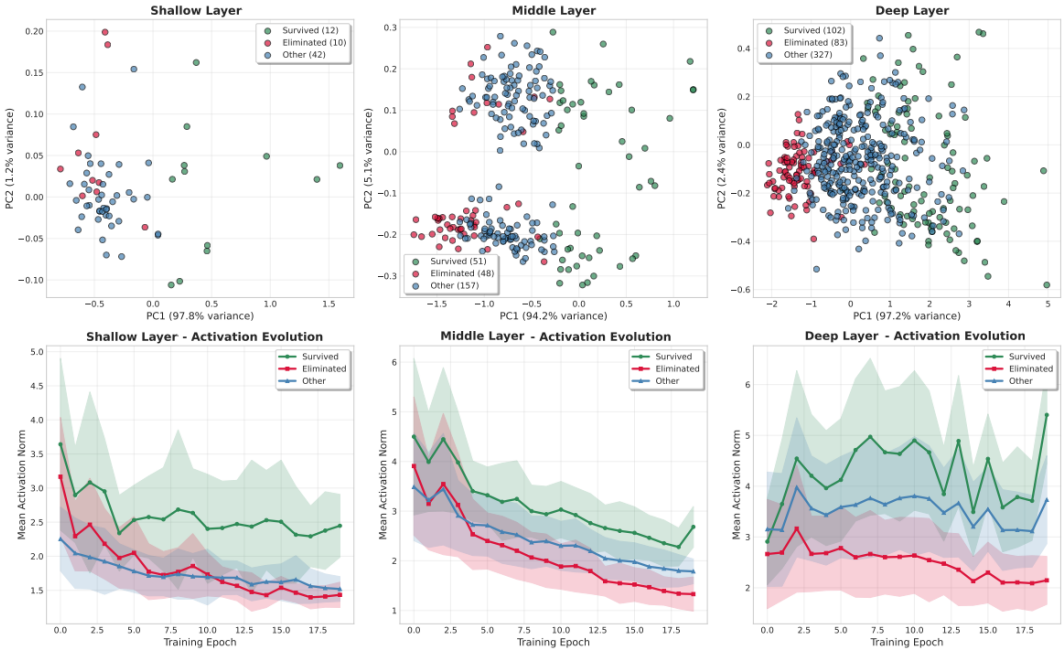


Figure 9: Static PCA and Activation Evolution on CIFAR-10.