A Computational Substrate for Self-Organizing Biologically-Plausible AI

Anonymous submission

Abstract

The divergence between ANNs and biological neural computation represents a fundamental challenge in developing truly adaptive, efficient, and robust AI systems. While modern deep learning achieves remarkable performance through global optimization, it lacks the inherent plasticity, continual learning capacity, and computational power characteristic of biological neural networks. We address this gap by introducing PAULA (Predictive Adaptive Unsupervised Learning Agent), a formal model of a biologically-plausible computational substrate that synthesizes principles from cellular neuroscience, computational neuroscience, and synaptic plasticity into a unified mathematical framework.

PAULA implements the neuron as a dynamical system engaged in continuous real-time adaptation aimed at modeling its immediate local environment leveraging predictive coding theory. We employ representation learning as a mechanism to encode information in characteristic dynamical regimes that the network enters and maintains, naturally making it a modality-agnostic framework.

Neurons in PAULA are implemented as directed graphs, enabling spatiotemporal computation during dendritic integration. This architecture is built on a synergy between vector-based synapses (signal, plasticity, and modulation); a two-stage, backpropagation-free local learning rule combining Hebbian gating with predictive error magnitude; and dynamic homeostatic metaplasticity based on activity history and neuromodulatory state. This system exhibits local and global stability through homeostatic regulation, a key property of biological neural networks.

We validate our model's computational capabilities through systematic experimentation on MNIST digit classification using networks of 1-144 adaptive PAULA neurons. We design a pipeline that separates the adaptive network from a decoder that reads only membrane potentials and spike timings, ensuring that learning occurs entirely through local rules.

The results reveal computational richness at both single-unit and network levels. A single neuron achieves 38.1% accuracy on 10-class MNIST, far exceeding chance level. Time series and contour plots show that different digit classes drive the neuron into distinct dynamical regimes characterized by unique membrane potential trajectories and learning window dynamics, forming differentiable attractor landscapes.

At the network level, we uncover architectural principles through controlled experiments on connectivity sparsity. Networks with 25% sparse connectivity achieve 86.8% accuracy, consistently outperforming both 50% sparse (81.8%) and fully dense (78.6%) configurations. Dense networks exhibit homeostatic saturation, preventing functional specialization. Sparse networks maintain diverse homeostatic setpoints, enabling functional clusters to emerge. This sparsity scaling law demonstrates that representation quality requires architectural constraints on information flow, unifying principles of biological cortical connectivity and deep hierarchical ANN architectures.

We demonstrate direct correspondence between emergent network structure and behavioral performance, providing interpretability through neuroscience-appropriate population analysis. The network is resistant to traditional ablation studies and displays robustness to information noise and structural changes. Temporal dynamics reveal evidence integration analogous to cognitive decision-making, with correct answers appearing in top-3 predictions early and achieving top-1 status given more simulation time, demonstrating dynamical inference where confidence builds through sustained activity.

Our framework has immediate implications for multiple domains. For AI, it offers a biologically-grounded alternative to backpropagation for credit assignment, addressing continual learning through inherent homeostatic mechanisms. The sparse, event-driven architecture suits neuromorphic implementation, potentially unlocking substantial energy efficiency gains. For neuroscience, the model provides testable predictions about plasticity integration, dendritic computation, and network-level emergent properties.

This work demonstrates methodology that bridges top-down cognitive requirements with bottom-up biophysical mechanisms. The framework enables theoretican convergence across disciplines, including computational psychiatry predictions: autism and schizophrenia may represent opposite ends of a network connectivity spectrum, with dense networks requiring extended temporal integration (predicting repetitive behaviors) and sparse networks requiring enhanced global neuromodulation (predicting dopamine treatment efficacy).

This research establishes a validated computational substrate for advaicing the development of more robust, adaptive, and scalable artificial intelligence systems. By treating neurons as dynamical systems rather than static functions, we unlock computational capacities that emerge from temporal dynamics and homeostatic regulation. The framework provides both a foundation for neuromorphic engineering and a tool for neuroscience research, exemplifying the bidirectional exchange between AI and neuroscience.