



Dynamic routing in shallow, brain-inspired, spiking neural networks

Applicant: Brennen Hill

Proposed Supervisors: Dr James Knight, Prof Thomas Nowotny

Application for AI PhD studentship

University of Sussex

6th January 2026

Research Proposal

Abstract

Contemporary machine learning is dominated by Deep Learning (DL) architectures characterized by deep sequential hierarchies. This structure is increasingly questioned by neuroscience, which posits the Shallow Brain Hypothesis (SBH): a shallow, massively parallel architecture for cortical function. This proposal outlines a novel computational framework that integrates the efficiency of Spiking Neural Networks (SNNs) with the organizational principles of the SBH to realize complex cognitive functions, specifically focusing on latent, multi-step reasoning. I will employ the architectural concepts of Mixture of Experts (MoE), constrained by biologically plausible inductive biases to form dynamic processing pathways, replicating the brain's principle of using multiplexed subspaces to route activity dynamically across brain-wide networks. My technical strategy is based on the specific implementation of rigorous SNN training algorithms, particularly EventProp combined with loss shaping, designed to overcome instability issues in exact gradient computation in spiking networks. This project aims to synthesize neuroscientific theory and advanced neuromorphic engineering to form a step in the next-generation bio-inspired computational intelligence.

1 Introduction and Motivation

The remarkable success of current DL models largely relies on constructing deep, multi-layered hierarchies to progressively extract abstract features [30]. However, this deep hierarchy concept is being critically challenged by neuroanatomical evidence supporting the Shallow Brain Hypothesis [30]. The SBH suggests that the brain is not a deep stack of layers but rather a shallow, massively parallel system where hierarchical cortical processing is integrated with direct, reciprocal connections to subcortical areas like the thalamus and basal ganglia [30]. This architecture prioritizes speed, enabling fast, reflexive computation through shallow pathways, thus avoiding the sluggishness associated with long polysynaptic transmission in deep stacks [30].

This parallel organization offers an advantage for compositional generalization. Unlike deep monolithic networks where features are entangled across sequential layers, a shallow, routed system facilitates the flexible reuse of learned primitives. By dynamically recombining independent processing modules, the architecture can compose known skills to tackle novel problems [5]. This approach uses the mixture of experts paradigm, which was proposed to decompose complex tasks into sub-problems solved by local experts [18], and recently scaled to massive capacities in deep learning via sparse gating [29]. However, where modern engineering approaches often rely on high-dimensional routing tokens, I propose a biological routing mechanism akin to dynamic routing by agreement in capsule networks [27], but implemented through the temporal dynamics of spiking neurons rather than iterative scalar agreement. This aligns with the brain’s ability to navigate new environments by recomposing established motor and cognitive strategies, suggesting that shallow models are suited for tasks requiring composable logic.

While Predictive Coding (PC) resembles biological neural processing more closely than standard DL in its use of local error minimization, it typically shares the issue of relying on deep feedforward architectures. PC offers a powerful theoretical framework explaining how the brain minimizes internal and sensory prediction errors to achieve inference [31]. Yet, computational models of PC typically instantiate this via a deep, cortical hierarchy where prediction errors cascade through many levels [26, 3]. While PC networks utilizing arbitrary computation graphs can approximate backpropagation gradients using local Hebbian rules and offering a powerful biologically plausible learning scheme [23], integrating this functionality into the structurally shallow and parallel architecture of the SBH remains a challenge. I argue that for SNNs to achieve efficient and brain-like function, they should adopt the SBH as a structural prior, replacing deep stacks with a shallow arrangement capable of dynamic routing. The proposed model would investigate if high-level latent reasoning, the ability to perform multi-step inference usually attributed to deep vertical hierarchies, can instead emerge from the temporal dynamics and recurrent connections of a shallow, parallel substrate.

2 Research Objectives

The core scientific goal is to design, implement, and rigorously validate a Shallow SNN capable of complex, cognitive reasoning tasks.

Objective 1: Design a Shallow SNN Architecture Leveraging Multiplexed Subspaces

I will construct SNNs defined by a limited vertical depth, functionally organized to replicate the parallel, distributed communication observed in the brain [30]. This design is grounded in

the Neural Population Doctrine [28], which posits that neural computation is best understood through the trajectory of population activity along low-dimensional manifolds [6]. Consequently, I will exploit the principle of multiplexed subspaces (MS), whereby different dimensions within a neuronal population dynamically engage distinct, yet overlapping, subspace networks across the entire architecture [22]. This alignment mechanism dictates how information propagates, offering a flexible routing mechanism without relying on deep, fixed feedforward structures [22]. I will leverage the geometric insights gained from my VCNet research, where we modeled the information flow between cortical areas as transformations between manifolds [17] and from working in AI R&D where I devised a custom algorithm to reframe high-dimensional sensory data into lower-dimensional manifolds for embodied edge AI [12].

Objective 2: Implement Dynamic Routing via the Mixture of Experts Paradigm

To achieve fine-grained, context-aware functional connectivity, I will integrate the architectural solution of the Mixture-of-Experts (MoE) model [5]. Specifically, I will utilize a Heterogeneous MoE (HMoE) structure, where components (experts/regions) possess varied computational complexity (size/sparsity) [5]. Training should incorporate biologically relevant inductive biases, such as a regulatory routing cost that penalizes the use of larger experts (scaling inversely with task performance). This will extend my work on SAPIN, where I implemented structural plasticity to allow processing units to physically migrate and reorganize based on prediction error [15].

I will implement structural plasticity mechanisms to dynamically prune connections to underperforming experts while reinforcing active pathways. Leveraging frameworks for GPU-accelerated structural plasticity [21], this approach allows the network to discover sparse, efficient connectivity patterns essential for minimizing the computational cost of the MoE architecture. By adapting these principles of plasticity to weight-space routing combined with stochastic expert dropout [5], this constrained learning process yields a Mixture-of-Pathways (MoP) model whose emergent connectivity patterns mirror the brain’s dynamic interactions, such as those between cortical and subcortical pathways during skill acquisition [5].

Objective 3: Achieve Latent Reasoning in the Shallow Architecture The final objective is to realize high-level cognitive function by enabling Latent Reasoning (LR) within this shallow, dynamic architecture. My research on Intention Communication in multi-agent systems demonstrated that agents could learn latent world models to simulate future trajectories and communicate plans rather than raw percepts [16]. I will apply this experience to enable the SNN to perform multi-step inference entirely within the model’s continuous hidden state.

This latent processing circumvents the expressive bandwidth constraints imposed by discrete token generation inherent to explicit chain-of-thought models [38]. The LR capability, relying on either vertical (expanding computational depth through recurrence) or horizontal (increasing sequential capacity) temporal recurrence, is crucial for solving complex multi-step problems beyond simple classification [38].

3 Methodology

Simulation Environment: PyGeNN and mlGeNN The foundational simulation infrastructure will leverage the GPU-Enhanced Neural Networks (GeNN) framework via the convenient Python bindings provided by PyGeNN and mlGeNN [19, 20]. PyGeNN exposes the full functionality of the C++ GeNN library, allowing the definition of complex, arbitrary SNN topologies essential for modeling the shallow, highly interconnected MoP structure [20, 19]. The efficiency of GeNN’s parallel computation and optimized spike handling is critical for scaling the numerous simulations required for parameter search and ablation studies [19].

I will specifically utilize the recently introduced mlGeNN interface, designed for spike-based machine learning, enabling rapid prototyping and rigorous evaluation of complex recurrent SNN models [20]. My extensive experience in low-level system optimization at HRL Laboratories, where I architected a multi-pass compiler in Common Lisp for a custom quantum processor and engineered exact pattern-matching algorithms within the quic compiler, ensures I can maximize the utility of this framework [11, 10]. I will utilize a recently developed framework for flexible, GPU-accelerated structural plasticity in sparse SNNs [21]. This framework allows for the efficient simulation of the synaptic rewiring required for the dynamic routing topology, maintaining high simulation speeds even with changing connectivity matrices. My extensive experience in low-level system optimization at HRL Laboratories, where I architected a multi-pass compiler in Common Lisp for a custom quantum processor and engineered exact pattern-matching algorithms within the quic compiler, ensures I can maximize the utility of this framework [11, 10].

Neuron Models To capture the temporal dynamics inherent to latent reasoning and dynamic routing, the SNN will employ biologically plausible neuron models. Building on my experience directing the Wisconsin Neuromorphic Computing and NeuroAI Lab [33], I will prioritize models exhibiting rich internal dynamics, such as the Balanced Resonate-and-Fire (BRF) neuron [9]. RF neurons model damped or sustained subthreshold oscillations, enabling frequency extraction in the time domain. The BRF variant, incorporating a dynamic threshold and divergence boundary, ensures sparse spiking and fast, stable convergence, qualities crucial for energy-efficient training.

Alternatively, the Adaptive Leaky Integrate-and-Fire (ALIF) neuron [1] may be employed to incorporate long short-term memory-like temporal coding capabilities.

Training Algorithms: Loss Shaping and EventProp Learning will be driven by EventProp, an advanced SNN algorithm that computes the exact gradient for continuous-time SNNs by applying the adjoint method along with appropriate derivative jumps at spike times [35]. This choice represents a significant divergence from the current standard in neuromorphic engineering, which predominantly relies on Surrogate Gradient methods [24]. While surrogate approaches, such as SuperSpike [36], have successfully enabled gradient descent in SNNs by smoothing the non-differentiable spike generation function, they inherently rely on approximations that may decouple the backward pass from the precise temporal reality of the forward pass. In contrast, EventProp avoids these approximations entirely [35]. EventProp may prove generally faster for training SNNs than Backpropagation Through Time (BPTT); however, in large, sparsely connected networks with sparse activity, this advantage will likely be significantly amplified. This is because neither connectivity sparsity nor activity sparsity can be readily exploited by standard ML frameworks used for BPTT, whereas EventProp is inherently suited for event-based, sparse, and parallel computation [35, 25]. Standard loss functions like average cross-entropy can lead to unstable training in SNNs due to a spike deletion problem, where gradient descent inadvertently pushes synaptic weights across critical values, corrupting useful spiking activity [25]. To address this, I will apply the principle of loss shaping by extending EventProp to mathematically defined generalized loss functions (e.g., Lsum exp) [25]. This technique enhances gradient flow to hidden layers, enabling robust and high-accuracy learning in complex tasks [25]. My experience implementing custom loss functions to align foundational models in NanoGPT from scratch [14] and engineering a custom coarse-to-fine training hierarchy in HEFT [13] has prepared me to handle the mathematical rigor of exact gradient computation.

Evaluation Strategy: Composable and Latent Tasks To validate that shallow, routed architectures excel at compositional generalization, I will employ benchmarks that distinguish broad generalization from rote skill mastery. The Abstraction and Reasoning Corpus [4] will assess the model’s ability to synthesize programs from minimal examples, a domain where neurosymbolic hybrids currently outperform pure neural networks [2]. Complementing this, the RAVEN dataset [37] will test structural and analogical reasoning . Success on these metrics will confirm the architecture’s capacity to disentangle reasoning from perception by operating on structural representations rather than pixel statistics [37].

I will further evaluate the efficacy of modular routing in mitigating catastrophic forgetting

using Split-MNIST and Split-CIFAR. Consistent with the Shallow Brain Hypothesis, modular networks can retain atomic skills for future composition [32]. By routing tasks to distinct supermasks [34], the system aims to sequentially learn thousands of tasks by inferring the correct expert superposition. This approach eliminates the need for test-time task identity and demonstrates interference-free modular reuse.

Finally, latent reasoning will be validated in partially observable environments (e.g., Atari, MuJoCo). While approaches like DreamerV2 establish that agents can plan within a world model’s latent space [7], this project extends that capability to explicit continuous thought chains [8]. The model will perform multi-hop inference entirely within its continuous hidden state, effectively executing a breadth-first search in latent space, to maximize reasoning density before committing to discrete actions [8].

4 Conclusion

This research directly addresses the critical limitations of contemporary DL by substituting rigid, deep structures with a dynamic, shallow architecture inspired by the SBH. By unifying the structural efficiency of SNNs with advanced architectural concepts (MoP, MS) and rigorous learning algorithms (EventProp/Loss Shaping), I aim to demonstrate that complex, brain-like intelligence requires highly parallel organization rather than simply layer stacking. Success will yield SNN models that achieve superior performance on temporal tasks with minimal parameter counts and maximum energy efficiency, offering a vital pathway forward for neuromorphic computing and energy-constrained edge AI applications.

Bibliography

- [1] Guillaume Bellec, Darjan Salaj, Anand Subramoney, Robert Legenstein, and Wolfgang Maass. Long short-term memory and learning-to-learn in networks of spiking neurons, 2018. [vi](#)
- [2] Mikel Bober-Irizar and Soumya Banerjee. Neural networks for abstraction and reasoning: Towards broad generalization in machines, 2024. [vi](#)
- [3] Christopher L. Buckley, Chang Sub Kim, Simon McGregor, and Anil K. Seth. The free energy principle for action and perception: A mathematical review, 2017. [iii](#)
- [4] François Chollet. On the measure of intelligence, 2019. [vi](#)
- [5] Jack Cook, Danyal Akarca, Rui Ponte Costa, and Jascha Achterberg. Brain-like processing pathways form in models with heterogeneous experts, 2025. [iii](#), [iv](#)
- [6] Juan Alvaro Gallego, Matthew G. Perich, Lee E. Miller, and Sara A. Solla. Neural manifolds for the control of movement. *Neuron*, 94:978–984, 2017. [iv](#)
- [7] Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models, 2022. [vii](#)
- [8] Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space, 2025. [vii](#)
- [9] Saya Higuchi, Sebastian Kairat, Sander M. Bohte, and Sebastian Otte. Balanced resonate-and-fire neurons, 2024. [v](#)
- [10] Brennen A. Hill. Quantum circuit. <https://BrennenHill.com/quantum-circuit/>, 2023. [v](#)
- [11] Brennen A. Hill. Quantum compiler. <https://BrennenHill.com/quantum-compiler/>, 2024. [v](#)

- [12] Brennen A. Hill. Embodied AI. <https://BrennenHill.com/embodied-ai/>, 2025. iv
- [13] Brennen A. Hill. Heft: A coarse-to-fine hierarchy for enhancing the efficiency and accuracy of language model reasoning. *Preprint*, 2025. vi
- [14] Brennen A. Hill. LLM training. <https://BrennenHill.com/llm-training/>, 2025. vi
- [15] Brennen A. Hill. Structural plasticity as active inference: A biologically-inspired architecture for homeostatic control. In *Brain-Inspired Dynamics for Engineering Energy-Efficient Circuits and Artificial Intelligence*, 2025. iv
- [16] Brennen A. Hill, Mant Koh En Wei, and Thangavel Jishnuanandh. Communicating plans, not percepts: Scalable multi-agent coordination with embodied world models. In *Proceedings of NeurIPS 2025 Workshop on Scaling Environments for Agents*, 2025. Also presented at NeurIPS 2025 Workshop on Embodied World Models for Decision Making and NeurIPS 2025 Workshop on Optimization for Machine Learning. iv
- [17] Brennen A. Hill, Zhang Xinyu, and Timothy Putra Prasetyo. The geometry of cortical computation: Manifold disentanglement and predictive dynamics in vcnet. In *Proceedings of NeurIPS 2025 Workshop on Symmetry and Geometry in Neural Representations*, 2025. Also presented at NeurIPS 2025 Workshop on Interpreting Cognition in Deep Learning Models. iv
- [18] Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991. iii
- [19] James C. Knight, Anton Komissarov, and Thomas Nowotny. Pygenn: A python library for gpu-enhanced neural networks. *Frontiers in Neuroinformatics*, Volume 15 - 2021, 2021. v
- [20] James C Knight and Thomas Nowotny. Easy and efficient spike-based machine learning with mlgenn. In *Proceedings of the 2023 Annual Neuro-Inspired Computational Elements Conference*, NICE ’23, page 115–120, New York, NY, USA, 2023. Association for Computing Machinery. v
- [21] James C. Knight, Johanna Senk, and Thomas Nowotny. A flexible framework for structural plasticity in gpu-accelerated sparse spiking neural networks, 2025. iv, v
- [22] Camden J. MacDowell, Alexandra Libby, Caroline I. Jahn, Sina Tafazoli, and Timothy J. Buschman. Multiplexed subspaces route neural activity across brain-wide networks. *bioRxiv*, 2023. iv

- [23] Beren Millidge, Alexander Tschantz, and Christopher L. Buckley. Predictive coding approximates backprop along arbitrary computation graphs, 2020. [iii](#)
- [24] Emre O. Neftci, Hesham Mostafa, and Friedemann Zenke. Surrogate gradient learning in spiking neural networks, 2019. [vi](#)
- [25] Thomas Nowotny, James P Turner, and James C Knight. Loss shaping enhances exact gradient learning with eventprop in spiking neural networks. *Neuromorphic Computing and Engineering*, 5(1):014001, January 2025. [vi](#)
- [26] Rajesh P. N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2:79–87, 1999. [iii](#)
- [27] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules, 2017. [iii](#)
- [28] Shreya Saxena and John P. Cunningham. Towards the neural population doctrine. *Current Opinion in Neurobiology*, 55:103–111, 2019. [iv](#)
- [29] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, 2017. [iii](#)
- [30] Mototaka Suzuki, Cyriel M. A. Pennartz, and Jaan Aru. How deep is the brain? The shallow brain hypothesis. *Nature Reviews Neuroscience*, 24(12):778–791, dec 2023. [ii](#), [iii](#)
- [31] Alexander Tschantz, Beren Millidge, Anil K Seth, and Christopher L Buckley. Hybrid predictive coding: Inferring, fast and slow, 2022. [iii](#)
- [32] Tom Veniat, Ludovic Denoyer, and Marc'Aurelio Ranzato. Efficient continual learning with modular networks and task-driven priors, 2021. [vii](#)
- [33] Wisconsin Neuromorphic Computing and NeuroAI Lab. People. <https://neuromorphic.cs.wisc.edu/people>, 2024. [v](#)
- [34] Mitchell Wortsman, Vivek Ramanujan, Rosanne Liu, Aniruddha Kembhavi, Mohammad Rastegari, Jason Yosinski, and Ali Farhadi. Supermasks in superposition, 2020. [vii](#)
- [35] Timo C. Wunderlich and Christian Pehle. Event-based backpropagation can compute exact gradients for spiking neural networks. *Scientific Reports*, 11(1), June 2021. [vi](#)

- [36] Friedemann Zenke and Surya Ganguli. Superspike: Supervised learning in multilayer spiking neural networks. *Neural Computation*, 30(6):1514–1541, June 2018. [vi](#)
- [37] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning, 2019. [vi](#)
- [38] Rui-Jie Zhu, Tianhao Peng, Tianhao Cheng, Xingwei Qu, Jinfa Huang, Dawei Zhu, Hao Wang, Kaiwen Xue, Xuanliang Zhang, Yong Shan, Tianle Cai, Taylor Kergan, Assel Kembay, Andrew Smith, Chenghua Lin, Binh Nguyen, Yuqi Pan, Yuhong Chou, Zefan Cai, Zhenhe Wu, Yongchi Zhao, Tianyu Liu, Jian Yang, Wangchunshu Zhou, Chujie Zheng, Chongxuan Li, Yuyin Zhou, Zhoujun Li, Zhaoxiang Zhang, Jiaheng Liu, Ge Zhang, Wen-hao Huang, and Jason Eshraghian. A survey on latent reasoning, 2025. [v](#)