MODELING BASAL GANGLIA ACTOR-CRITIC ARCHITECTURE FOR REINFORCEMENT LEARNING ON SPIKING NEURAL NETWORKS

A Thesis

by

Recep Buğra Uludağ

Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in the
Department of Computer Science

Özyeğin University
July 2024

MODELING BASAL GANGLIA ACTOR-CRITIC ARCHITECTURE FOR REINFORCEMENT LEARNING ON SPIKING NEURAL NETWORKS

Approved by:

Asst. Prof. İsmail Aktürk
Dept. of Computer Science
Ozyegin University

Prof. Neslihan Serap Şengör
Dept. of Electronics & Communication Engineering
Istanbul Technical Universtiy

Prof. Erhan Öztop
Dept. of Computer Science
Ozyegin University

Date Approved: July 22, 2024

To my family and teachers,

the architects of my future.

ABSTRACT

Neuromorphic computing, inspired by the intricate functionalities of the human brain, seeks to emulate its efficient computation through advanced models of neuron dynamics and synaptic interactions. A central challenge in this field is the development of effective learning mechanisms that extend beyond local synaptic changes, such as those described by Spike-Timing-Dependent Plasticity (STDP), to encompass more complex, system-wide processes. This research specifically focuses on the role of the basal ganglia and its use of dopaminergic reinforcement learning—a critical component often leveraged for its pivotal role in decision-making and reward-based learning in biological systems. By implementing and analyzing the basal ganglia's dopaminergic learning mechanisms on Spiking Neural Networks (SNNs), we aim to enhance the learning capabilities of these networks. Our study delves into the detailed computational modeling of these mechanisms, highlighting their potential to facilitate complex learning tasks such as sequence learning through reinforcement signals. The insights gained from this research contribute to a deeper understanding of neural processing and learning in artificial systems, marking a step forward in the quest for more sophisticated, biologically-inspired computing architectures.

ÖZETÇE

İnsan beyninin karmaşık işlevlerinden esinlenen nöromorfik hesaplama, gelişmiş nöron dinamikleri ve sinaptik etkileşim modelleri aracılığıyla beynin etkin hesaplama yeteneğini taklit etmeyi amaçlamaktadır. Bu alandaki temel bir zorluk, Vuru Zamanı Bağımlı Esneklik (İng. STDP) gibi yerel sinaptik değişiklikleri tanımlayan öğrenme mekanizmalarını, daha karmaşık, sistem genelinde süreçleri kapsayacak şekilde genişletmektir. Bu araştırma özellikle, karar verme ve ödül temelli öğrenmede kritik bir rol oynayan bazal ganglionların ve dopaminerjik pekiştirmeli öğrenmenin rolüne odaklanmaktadır. Bu çalışma, Vurulu Sinir Ağları (İng. SNN) üzerinde bazal ganglionların dopaminerjik öğrenme mekanizmalarını uygulayarak ve analiz ederek, bu ağların öğrenme yeteneklerini artırmayı amaçlamaktadır. Çalışmamız, bu mekanizmaların detaylı hesaplamalı modellemesine derinlemesine dalarken, pekiştirme sinyalleri aracılığıyla dizi öğrenme gibi karmaşık öğrenme görevlerini kolaylaştırma potansiyelini vurgulamaktadır. Bu araştırmadan elde edilen içgörüler, yapay sistemlerde nöral işleme ve öğrenme üzerine daha derin bir anlayış katkıda bulunmakta ve daha sofistike, biyolojik süreçlerden ilham almış hesaplama mimarileri için ileriye doğru bir adım niteliğindedir.

ACKNOWLEDGEMENTS

I am deeply grateful to my professors, whose expertise and guidance have been invaluable throughout this research. Their support and insights have been pivotal in navigating the complexities of this project.

I am also grateful to my colleagues for their collaboration and the hard work we shared throughout this journey. Working alongside such dedicated individuals has been both a privilege and a source of inspiration.

Special thanks to the Intel Neuromorphic Lab team for their assistance and support with the Loihi platform. Their readiness to help and provide solutions whenever challenges arose has been instrumental in the progress and success of our work.

To all who contributed, your support has made this research possible and enriched the experience immeasurably. Thank you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

In this chapter, we embark on an exploration of neuromorphic computing, a paradigm that seeks to harness the computational principles of the human brain. We begin by situating neuromorphic computing within the broader context of computational models, highlighting its potential to address current computational challenges. Following this, we delineate the research problem our work aims to tackle and introduce the proposed solution that leverages dopaminergic learning in the basal ganglia for reinforcement learning on Spiking Neural Networks (SNNs). The chapter concludes by underscoring the significance of this research, not only for advancing neuromorphic computing, but also for its broader implications in the field of artificial intelligence.

1.1 Introduction to Neuromorphic Computing

The science of computation has been rapidly advancing since the advent of computers. Relying on Von Neumann architectures, modern computers are capable of many great achievements that previous generations could only dream of. However, the era of computers based on Von Neumann Architecture is beginning to reach its limits as the separation of data and computation introduces a bottleneck in the transfer of information [6]. Furthermore, CMOS technology, which is dependent on silicon devices, also encounters a constraint on power consumption and heat generation [7]. In the face of such challenges, there are two obvious routes the engineers and scientists can take: either embark on a new computing paradigm, or squeeze the juice out of every possible optimization and technique for the current methodology. The engineers who have decided on the former journey must seek inspiration for novel architectures, and nature is one of the best at creating highly abstract and efficient systems.

The human brain has evolved well over millennia and is the most power-efficient and powerful computation device [8]. One of the marvelous things humans can do is generalize a novel problem and design an algorithm to solve said problems, meaning the human brain is an excellent function approximator [9]. This feature of the human brain has also caught the interest of many engineers as they have employed structural insights from the human nerve cell organizations to come up with Artificial Neural Networks (ANNs) [10]. While this concept of ANNs would flourish, the current challenges, such as the *Power Wall*, forced engineers to look for efficient ways of computing [7]. Then, the idea of figuring out how the human brain achieves its computational efficiency was discussed. The human brain uses approximately 20 Watts of power for its daily use [8]. One of the main contributors to this efficiency is how information is handled. There are approximately 86 Billion nerve cells in the human brain, but each neuron connects to a thousand other neurons on average [11]. This shows that the interconnect in the human brain is highly sparse. Moreover, nerve cells communicate using electrochemical pulses known as *spikes*. Spikes are identical, with slight noises and variations. Also, because of the way the human brain designed its communication scheme for the spikes, little spiking activity is required to transfer information. Thus, spikes are sparse in time as well. This structural and temporal sparse communication scheme (spike code) is the fundamental reason for brain power efficiency. However, there is one more facet of the computation in the brain, and that is how the information is represented and stored. At the same time, a conventional computer needs separate blocks such as CPU and memory. The human brain stores the data in its nerve cells, particularly on synapses that connect the cells [12] [13]. This way, there is no notion of distinct memory, and computation-memory separation is eliminated. This type of computation is known as *compute in memory* [14] [15], and the human brain is an excellent example. This bypasses the data transfer bottleneck. Along with it, the human brain is highly parallelized in its

microcircuits and, therefore, can handle massive amount of computations. When all of the above optimizations come together, we start to understand the root causes of computational and power efficiency of the human brain.

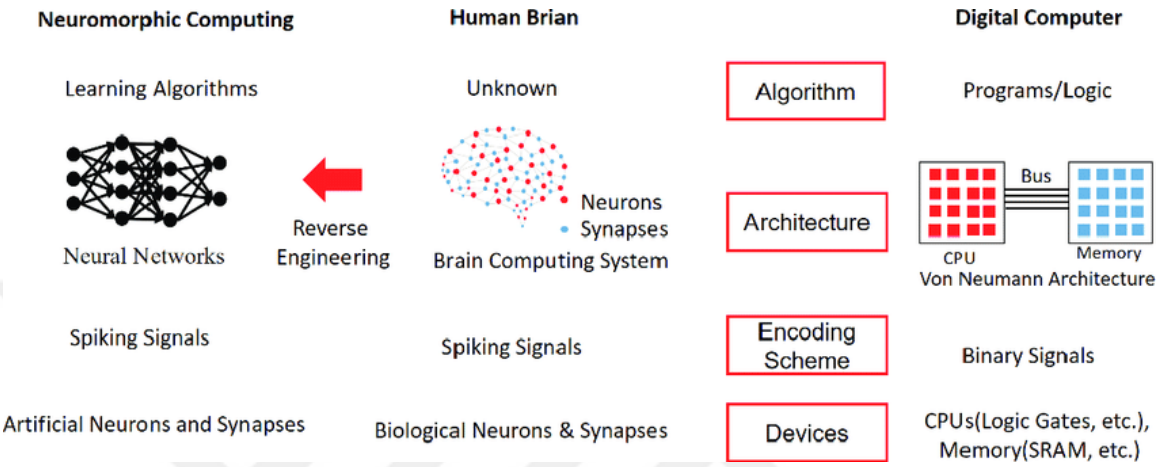
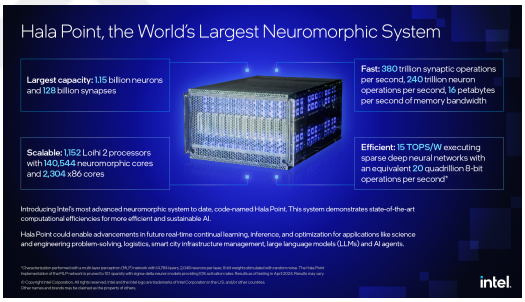


Figure 1: Computation stack comparison showing the differences between neuromorphic vomputing, the human brain, and Von Neumann architectures. [1]

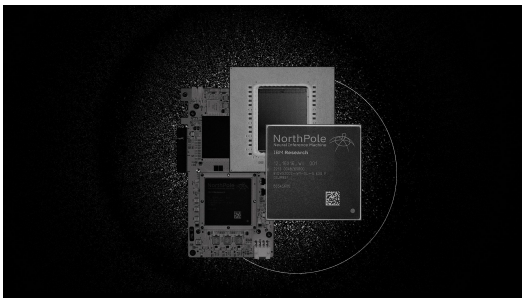
Neuromorphic computing or brain-like computing aims to employ these fundamental processes of the human brain to achieve better computation in terms of functionality and efficiency. Comparison of the computation stack between such paradigm and current digital computers is illustrated in Figure 1. The most essential of these strategies is the spiking behavior when introduced with the brain-like structure of neural networks, which results in what is known as Spiking Neural Networks (SNNs) where the non-linear activation function is the Heaviside function [16]. This activation function, however, is not enough on its own, and some neuronal dynamics need to be applied to take advantage of the temporal dynamics of the spiking behavior. This is introduced by membrane potential, a decayed integration of weighted spikes [17]. This integration over time and decay opens up the time dimension for computation, whereas, in ANNs, this weighted sum is calculated for each time frame. Still, there is no notion of propagating this value in time, so when looking from an ANN perspective, every neuron showcases implicit features of recurrency. This temporal dynamic

allows neuromorphic computing to sparsify connections and the load of these communications. This sparsity results in highly compressible communication, and with the spikes being essentially binary, they do not require high bandwidth, resulting in very low power computation. There are various ways neuromorphic computing can exploit these features. However, they can be mainly considered under three categories: hardware [18], software [19], and sensors [20] [21]. Neuromorphic hardware provides a substrate where SNNs can be mapped to utilize their power efficiency; various architectures have been developed so far, the most well-known ones include Loihi by Intel [22] [23], TrueNorth by IBM [24] and SpiNNaker from University of Manchester [25] as shown in Figure 2.

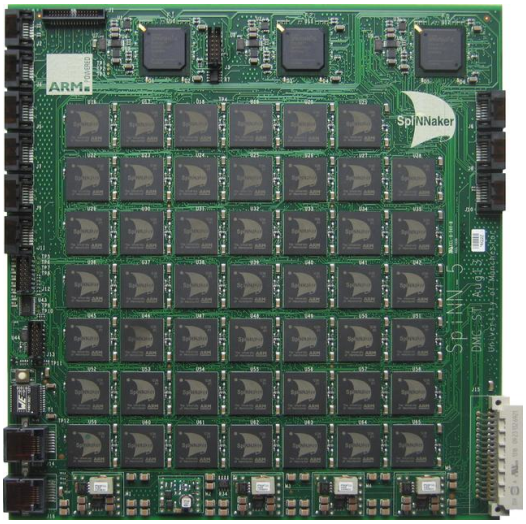
Intel Loihi2 [23]



Intel Hala Point [26]



IBM Northpole [24]



SpiNNaker [25]

Figure 2: Advanced neuromorphic chips/platforms.

Each of these architectures provides its advantages and disadvantages regarding programmability, scalability, efficiency, and complexity. On the software side, we have neuromorphic frameworks that provide simulation environments to configure and test SNNs, which allow for highly flexible designs in terms of complexity. Most well-known of these programs are Lava [27], snnTorch [28], Brian2 [29], SpikingJelly [30], Nengo [31] and NIR [32]. Each framework provides unique optimizations and ease of computation in certain neuromorphic subfields. Lastly, there are neuromorphic sensors [20]; as the SNNs are inherently designed to work with temporal data, our current static datasets do not closely match the dynamics of neuromorphic computing. Therefore, new sensors that temporally encode information in terms of spikes are created by inspired by their mechanism in the human body, such as the retina and cochlear; these devices are known as event-based as they only encode the change in stimuli; therefore, are highly sparse and can achieve much higher time resolution.

1.2 Current State of Neuromorphic Computing and Challenges

Neuromorphic Computing and SNNs have proven capable in many fields ranging from brain-computer interfacing, robotics, and computer vision to computational neuroscience [33]. While showing huge success on temporal datasets and achieving low power/latency computation, neuromorphic computing has not been an end-to-all solution or readily applicable for mainstream applications. Various issues must be addressed before SNNs can be mainstream, like how Deep Neural Networks have taken the spotlight [34] [35]. These roadblocks consist of issues including scalability in the hardware; even though the brain is very sparse in its connections, it is still far more advanced than our 3D CMOS fabrication processes and can connect an average of a thousand neurons to a single neuron. This fan-in is still hard to achieve when point neuron designs concern hardware. There are various methodologies to mitigate this problem. Still, they all introduce significant decision-making regarding

the architectural design of said hardware, as interconnect designs inevitably affect the synaptic micro-architecture as well [36] [37] [38]. Apart from scaling limitations, current neuromorphic algorithms still need to be matured [39]. There are various techniques to train SNNs [28], and all software frameworks target a different side of this computation stack, so there is an obvious need for holistic software frameworks that generalize well.

1.3 Learning in SNNs: A Dual Approach

If we dive more deeply into algorithms for learning in SNNs, we can see two consensus on how to approach learning in SNNs, which stem from the researchers' background. People from the neuroscience and physics fields take the founding mentality of neuromorphic computing into their minds and bring insights from brain and biological learning rules to adapt them to the neuromorphic field. Engineers and computer scientists, recognizing the practical value of neuromorphic computing as an evolution of ANNs, are adapting existing learning algorithms for rapid deployment on SNNs. This mainly involves backpropagation, which is hard to implement on SNNs as the Heaviside function, which produces the thresholding and spiking behavior, is not differentiable. Therefore, various roundabouts of backpropagation have been developed along with other algorithms that also consider temporal dynamics of spike-code [40] [41] [42].

1.4 Biologically Plausible Learning Rules

Various bio-realistic learning rules are built upon Hebbian learning and Long Term Potentiation (LTP) / Long Term Depression (LTD) [43] [44]. Even then, there are problems with these learning rules as they all encompass unsupervised local synaptic update rules. While this type of learning explains circuit-level dynamics, it shows inadequate performance in broader networks. To define learning on a global scale, we can turn to the brain for inspiration and examine learning in specific brain regions.

Most prominent among these regions are the Cerebral Cortex, Basal Ganglia (BG), and Cerebellum, which implement unsupervised learning, reinforcement learning, and supervised learning, respectively [45] [46]. While their internal computation and learning have been understood to a degree, their interactions have generally been studied in isolation.

A deep dive into these neural circuits directs our attention to the mechanisms of the Basal Ganglia, the center for reinforcement learning [47]. This brain region is central to decision-making, and a better understanding of it may lead to the development of bio-realistic reinforcement learning in SNNs.

1.5 Research Aim and Potential Contributions

Our exploration delves into the intricate neural circuits of the brain, particularly spotlighting the Basal Ganglia (BG) as the cornerstone of reinforcement learning [46]. Renowned for its pivotal role in decision-making processes, a nuanced comprehension of the BG enlightens us on the fundamentals of cognitive function. It paves the way for pioneering advancements in bio-realistic reinforcement learning for SNNs. This thesis is dedicated to unraveling the computational intricacies of the BG, dissecting how dopamine neurons intricately encode rewards and punishments, and elucidating the representation of state-action pairs derived from environmental interactions which are key dynamics for reinforcement learning.

By constructing a robust framework for reinforcement learning within SNNs, this research aims to contribute to neuromorphic computing. The potential applications of such a framework are vast, from enhancing robotic autonomy to refining the intelligence of smart devices. We aim to understand reinforcement learning in biological systems and create a foundation for it in SNNs, paving the way for biologically plausible and efficient computational models.

CHAPTER II

LITERATURE REVIEW

The quest to bridge the gap between biological intelligence and artificial computing has long fascinated researchers, leading to the emergence of neuromorphic computing as a frontier in this exploration [34] [35]. This paradigm, inspired by the intricate workings of the human brain, seeks not only to replicate its computational efficiency but also to understand the underlying principles that enable complex behaviors and learning capabilities.

This literature review aims to navigate the vast research landscape surrounding neuromorphic computing, explicitly focusing on neuron models, reinforcement learning mechanisms, the anatomy of the Basal Ganglia (BG), and the role of dopamine in learning. Given the interdisciplinary nature of this field, spanning neuroscience, computational neuroscience, and machine learning, a wide array of literature forms the backbone of the discussion herein.

We begin by exploring various neuron models fundamental to understanding neural dynamics and implementing them in Spiking Neural Networks (SNNs). The review then focuses on the anatomy of the BG, a critical brain region involved in reinforcement learning and decision-making. Finally, we discuss the role of dopamine, a neurotransmitter that plays a pivotal role in encoding reward signals in the BG.

As we traverse these topics, the review also identifies gaps and opportunities for further research, setting the stage for the subsequent chapters. The ultimate goal of this literature review is to provide a coherent foundation that not only contextualizes the research problem but also underscores the significance of the proposed study in contributing to the advancement of neuromorphic computing and our understanding

of learning mechanisms within the BG.

2.1 Neuron Models in Neuromorphic Computing

Neuromorphic computing seeks to emulate the neural processing mechanisms of the brain to develop computing systems capable of complex, adaptive behaviors. At the heart of these systems are neuron models, mathematical formulations replicating biological neurons' electrical activity.

Neuron models in neuromorphic computing serve to bridge the gap between biological neural processes and computational algorithms. These models are critical for understanding and implementing neural dynamics in hardware or software simulations. Their complexity can vary significantly, from simple integrative units to comprehensive models that meticulously capture the biophysical properties of neurons.

2.1.1 Spectrum of Neuron Models

The range of neuron models spans from computationally efficient, simplistic models to bio-realistically accurate, computationally complex ones [48] [49]. On one end of the spectrum, we have models like the Integrate-and-Fire [50] and its variations, which abstract neurons to essential elements that accumulate inputs until a threshold is reached, triggering an output spike. These models, while computationally lightweight, offer limited biological fidelity. On the other end, there are detailed models like the Hodgkin-Huxley and its extensions [51–54], which provide intricate representations of the ion channel dynamics governing neural excitability. While offering high biological realism, these models demand significant computational resources, especially when simulating large networks.

2.1.2 The Izhikevich Neuron Model

Among the myriad of neuron models, the Izhikevich model stands out for striking a balance between biological realism and computational simplicity. Introduced by Eugene Izhikevich in 2003, this model was designed to replicate the rich firing patterns observed in biological neurons while maintaining computational efficiency [55].

2.1.2.1 Model Equations

The model is succinctly described by a system of ordinary differential equations:

$$\frac{dv}{dt} = 0.04v^2 + 5v + 140 - u + I \tag{1}$$

$$\frac{du}{dt} = a(bv - u) \tag{2}$$

where v represents the membrane potential, u is a recovery variable that accounts for the activation of K^+ ionic currents and inactivation of Na^+ ionic currents, and I is the synaptic or injected current. The parameters a, b, c, and d are constants that determine the neuron's firing properties. When the membrane potential reaches the apex of a spike at 30mV, it is reset according to:

$$\text{if } v \geq 30 \text{ mV, then } \begin{cases} v \leftarrow c, \\ u \leftarrow u + d. \end{cases} \tag{3}$$

These equations define the dynamic behavior of the Izhikevich neuron model. The resetting condition ensures that after a spike is generated (when v reaches or exceeds 30 mV), the membrane potential v is reset to c, and the recovery variable u is increased by d, which simulates the refractory period of a neuron after firing an action potential.

2.1.2.2 Patterns and Behaviors

The Izhikevich model is capable of reproducing a wide range of spiking and bursting patterns observed in cortical neurons, including chattering, fast spiking, intrinsically bursting, and more [56]. The model's versatility comes from the adjustable parameters a, b, c, and d, which can be tuned to emulate specific neuron types or behaviors.

2.1.2.3 Advantages and Applications

Its computational efficiency and ability to simulate diverse neural behaviors make the Izhikevich model particularly appealing for large-scale simulations of neural networks and for implementation in neuromorphic hardware. The model's relative simplicity allows for real-time simulations of complex systems, opening new avenues for research into brain-like computing and the development of brain-inspired algorithms.

2.2 The Basal Ganglia

The Basal Ganglia (BG) are a group of nuclei in the brain that play a critical role in coordinating and learning movement [57], as well as various cognitive and emotional functions. Historically, research on the BG began with a focus on their involvement in motor control [58], mainly due to their association with Parkinson's disease and other motor disorders [59, 60]. Early studies highlighted the BG's role in regulating and fine-tuning voluntary movements, as evidenced by the severe motor deficits observed in disorders like Parkinson's and Huntington's disease [61].

In the latter half of the 20th century, research expanded to explore the BG's contributions to cognitive processes and reinforcement learning [62] [63]. This paradigm shift was driven by advancements in neuroimaging, neurophysiology, and computational modeling, which revealed the BG's involvement in a broader array of functions beyond motor control. These advancements have allowed researchers to understand better the complex interplay between the BG and other brain regions, highlighting

their role in decision-making, habit formation, and procedural learning.

2.2.1 Anatomy of The Basal Ganglia

The BG comprise several interconnected nuclei, each with distinct functions and contributions to the overall circuitry. These nuclei include the striatum (caudate nucleus, putamen, and nucleus accumbens), globus pallidus (external and internal segments), subthalamic nucleus, and substantia nigra (pars compacta and pars reticulata) [2]. Here, we describe each component and its functional significance (Figure 3).

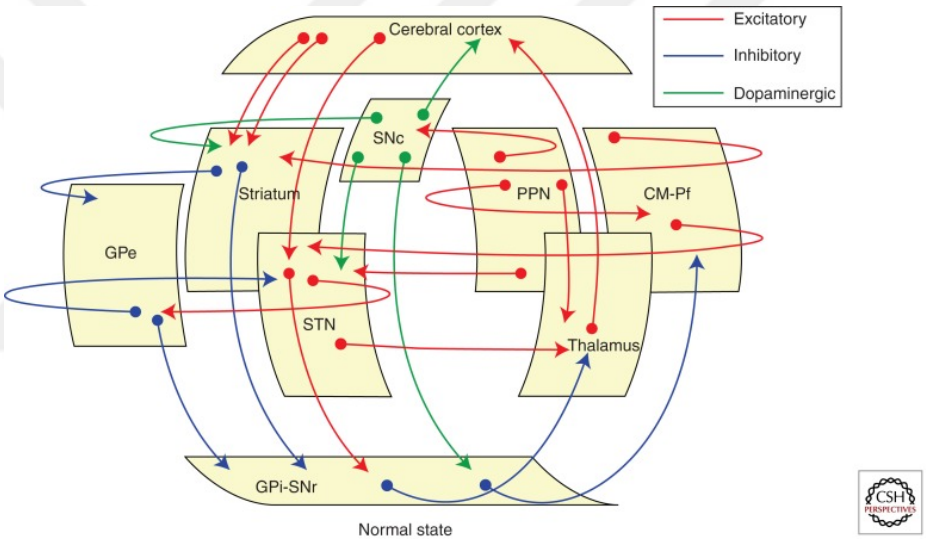


Figure 3: Basal ganglia circuits. Abbrevations: Substantia Nigra (SN), Globus Pallidus (GP) ,Subthalamic nucleus (STN), centromedian-parafascicular (CM-Pf). pedunculopontine nucleus (PPN) [2].

2.2.1.1 Striatum

It is the most significant component of the BG and serves as the primary input nucleus. It is divided into the dorsal and ventral striatum [64]:

Dorsal Striatum: Comprising the caudate nucleus and putamen, the dorsal striatum receives extensive excitatory inputs from the cerebral cortex, particularly the motor and premotor areas, and the thalamus. It plays a pivotal role in motor control, action selection, and specific cognitive functions.

- *Caudate Nucleus*: Located adjacent to the lateral ventricles, the caudate nucleus is involved in various higher-order cognitive functions such as learning and memory.

- *Putamen*: Situated lateral to the caudate nucleus, the putamen is primarily associated with regulating movements and influencing various motor skills.

Ventral Striatum: Including the nucleus accumbens and olfactory tubercle, the ventral striatum is integral to reward processing and reinforcement learning [65]. It receives inputs from the limbic system and the prefrontal cortex.

- *Nucleus Accumbens*: Divided into the core and shell regions, the nucleus accumbens is crucial for processing reward and reinforcement, influencing motivation and goal-directed behavior.

- *Olfactory Tubercle*: Involved in the sense of smell and reward processing.

The striatum contains two main types of neurons: *Medium Spiny Neurons (MSNs)* and *Interneurons*. MSNs are GABAergic projection neurons and constitute about 95% of the neuronal population in the striatum [66]. They are divided into two classes based on their projection targets and receptor expression: i) *D1-type MSNs* which express dopamine D1 receptors and are part of the direct pathway, facilitating movement by projecting to the GPi and SNr; ii)*D2-type MSNs* which express dopamine D2 receptors and are part of the indirect pathway, inhibiting movement by projecting to the GPe.

On the other hand, there are various types of *interneurons* that modulate the activity of MSNs through neurotransmitters such as GABA, acetylcholine, and somatostatin. Notable types of interneurons include *Parvalbumin-positive (PV) interneurons* that are fast spiking interneurons providing strong inhibitory control over MSNs; and *Cholinergic interneurons* that release acetylcholine and modulate the striatal output by influencing both MSNs and other interneurons.

The striatum is further organized into two distinct compartments: the matrix and the striosomes (also known as patches). These compartments are characterized by differences in their neurochemical composition, connectivity, and function, providing an additional layer of complexity to the striatal organization.

The *matrix compartment* constitutes most of the striatal volume, encompassing about 85% of the striatum [67]. The matrix receives dense excitatory inputs from the sensorimotor and association cortices, implicating it in motor and cognitive processes. The projection neurons within the matrix primarily target the external globus pallidus (GPe) and the substantia nigra pars reticulata (SNr), integrating and modulating cortical information crucial for action selection and movement regulation.

The *striosome compartment*, accounting for about 15% of the striatum, are interspersed within the matrix-like islands [67]. These compartments are enriched in mu-opioid receptors and substance P, giving them a distinct neurochemical profile. Striosomes receive inputs from the limbic regions, including the prefrontal cortex and amygdala, highlighting their role in reward-related processing. The projection neurons within the striosomes predominantly target the substantia nigra pars compacta (SNc), influencing the dopaminergic system and thereby playing a critical role in modulating reward and motivational states [65] [68].

The compartmentalization of the striatum into matrix and striosomes underscores its intricate structure and functional specialization, enabling the striatum to process diverse types of information for the orchestration of complex behaviors.

2.2.1.2 Globus Pallidus

The globus pallidus is divided into two segments [69]:

External Segment (GPe): Involved in the indirect pathway, the GPe receives inhibitory input from the D2-type MSNs of the striatum and projects to the subthalamic nucleus (STN) and other parts of the BG. It plays a critical role in modulating

movement and maintaining motor control.

Internal Segment (GPi): Part of the direct pathway, the GPi receives inhibitory input from the D1-type MSNs and provides major inhibitory output to the thalamus and brainstem. It is crucial for the inhibition of unwanted movements and the facilitation of intended movements.

2.2.1.3 Subthalamic Nucleus (STN)

It is a small, lens-shaped nucleus that receives excitatory inputs from the cerebral cortex and inhibitory inputs from the GPe [70]. It sends excitatory projections to the GPi and SNr, vital in regulating the BG output. The STN is integral to the indirect pathway and controls movement by modulating the activity of the GPi and SNr.

2.2.1.4 Substantia Nigra

The substantia nigra is divided into two parts [71]:

Pars Compacta (SNc): It contains dopaminergic neurons that project to the striatum, modulating the activity of MSNs. Dopamine release from the SNc enhances the direct pathway via D1 receptors and inhibits the indirect pathway via D2 receptors, facilitating movement initiation and execution.

Pars Reticulata (SNr): It functions similarly to the GPi, sending inhibitory outputs to the thalamus and brainstem. It plays a critical role in controlling eye movements and other motor functions.

2.2.1.5 Functional Pathways

The BG operates through three primary pathways that have opposing effects on movement [72]:

- *Direct Pathway*: Facilitates movement by disinhibiting the thalamus. This pathway involves D1-type MSNs projecting to the GPi and SNr, which inhibit

the thalamus less, allowing for increased cortical activity and movement execution.

- *Indirect Pathway*: Inhibits movement by increasing thalamic inhibition. This pathway involves D2-type MSNs projecting to the GPe, which inhibits the STN. The STN then sends excitatory signals to the GPi and SNr, leading to increased inhibition of the thalamus and reduced cortical activity, thus inhibiting movement.

- *Hyperdirect Pathway*: Provides a rapid means of inhibiting movement. This pathway involves direct excitatory projections from the cortex to the STN, sending excitatory signals to the GPi and SNr, increasing their inhibitory output to the thalamus and thus quickly suppressing movement.

Dopaminergic inputs from the SNc modulate the balance between these pathways [73] [74]. Dopamine release enhances the direct pathway and suppresses the indirect pathway, promoting movement initiation and execution.

2.2.1.6 Additional Circuitry and Connections

The BG is also interconnected with other brain regions, forming complex circuits that contribute to their diverse functions:

- *Corticostriatal Connections*: The cortex provides the primary excitatory inputs to the striatum, forming the corticostriatal pathway. This pathway is crucial for integrating sensory, motor, and cognitive information.

- *Thalamostriatal Connections*: The thalamus sends excitatory inputs to the striatum, providing additional sensory and motor information.

- *Nigrostriatal Pathway*: Dopaminergic projections from the SNc to the striatum modulate the activity of MSNs, influencing the balance between the direct and indirect pathways.

The BG also interacts with the limbic system, including the hippocampus and amygdala, contributing to emotional regulation and memory processes [75].

2.2.2 Functional Dynamics and Mechanisms of the Basal Ganglia

The BG mainly functions as an *action selection* mechanism [76]. As an agent (animal) has many command systems and limited resources such as cognitive circuitry, working memory, and motor pathways shared by these command systems, an agent should have a way of allocating these resources to programs or actions that need them. This type of action control mechanism needs to be able to select an action among many competing actions depending on both intrinsic and extrinsic stimuli and should learn from experience to optimize resource allocation. Such a mechanism should suppress competing actions while favoring and allowing control of the selected action [76].

The BG comprises many topographic channels connected to various cortical areas representing these command systems [77]. It also receives many intrinsic input afferents from across the cortex, signifying the internal and external states of the agent. These channels all output through their population area in the SNr and have lateral inhibiting connections from the STN to inhibit each other. When an action is deemed important by having more saliency than other competing actions, the SNr output of that channel is inhibited, allowing the disinhibition of the thalamic connections for that action, thereby realizing the selected action. Simultaneously, lateral connections increase inhibition by the SNr on other channels.

2.2.2.1 Go/No-Go Mechanisms

The Go/No-Go model describes how the BG facilitates or inhibits actions [78]. It involves two pathways:

- *Go Pathway (Direct Pathway)*: This pathway facilitates the initiation of desired movements. D1-type MSNs in the striatum project to the GPi and SNr, inhibiting these structures and thereby reducing their inhibitory output to the

thalamus. This disinhibition of the thalamus allows for increased activity in the motor cortex, promoting movement execution.

- *No-Go Pathway (Indirect Pathway)*: This pathway inhibits undesired movements. D2-type MSNs in the striatum project to the GPe, which in turn inhibits the STN. The STN sends excitatory signals to the GPi and SNr, increasing their inhibitory output to the thalamus. This increased inhibition reduces activity in the motor cortex, preventing movement execution.

Dopaminergic inputs from the SNc modulate the balance between these pathways. Dopamine enhances the Go pathway via D1 receptors and suppresses the No-Go pathway via D2 receptors, facilitating movement initiation and execution while inhibiting competing actions.

2.2.2.2 Reinforcement Learning

Reinforcement learning in the BG is driven by dopaminergic signaling, which encodes reward prediction errors (i.e., the difference between expected and actual outcomes) [3, 79–81]. Dopaminergic neurons in the SNc and VTA release dopamine in response to unpredicted rewards or expectations of rewards. This phasic dopamine release strengthens synapses in the direct pathway (Go). It weakens synapses in the indirect pathway (No-Go), reinforcing actions that lead to positive outcomes and discouraging those that do not.

Reward Prediction Error: When an unexpected reward is received, dopamine levels increase, enhancing synaptic plasticity in the Go pathway and promoting the repetition of actions that led to the reward. Conversely, when an expected reward is not received, dopamine levels decrease, enhancing synaptic plasticity in the No-Go pathway and discouraging the action .

Learning and Habit Formation: Over time, actions that consistently lead to rewards become habitual. The BG contributes to the transition from goal-directed

behavior, which is flexible and responsive to reward contingencies, to habitual behavior, which is automatic and less sensitive to reward changes [82].

These mechanisms highlight the BG's role in learning from environmental feedback and adapting behavior accordingly. By facilitating the selection of appropriate actions and reinforcing behaviors that lead to positive outcomes, the BG is essential for efficient and adaptive decision-making.

2.2.2.3 Decision-Making

In addition to motor control and reinforcement learning, the BG is involved in higher-order decision-making processes [83]. The BG integrates information from various cortical and subcortical regions to evaluate different actions' potential benefits and costs. This integration supports complex decision-making tasks, such as selecting the most beneficial strategy in uncertain environments.

2.2.2.4 Habit Formation

Habit formation is another critical function of the BG. Repeated reinforcement makes certain behaviors automatic and requires less cognitive effort to execute. This shift from goal-directed to habitual behavior is mediated by the striatum, with the dorsal striatum playing a pivotal role in habitual responses and the ventral striatum in goal-directed actions [82].

In summary, the BG's involvement in action selection, Go/No-Go mechanisms, reinforcement learning, and habit formation underscores their importance in both motor and cognitive functions. By integrating sensory, contextual, and reward information, the BG facilitates adaptive behavior and efficient decision-making by coordinating the direct, indirect, and hyperdirect pathways.

2.2.3 Computational Models of the Basal Ganglia

Computational modeling of the BG has been instrumental in deciphering their functions and in exploring their potential applications in neuromorphic computing and robotics [84] [85]. Early models focused on the BG's role in motor control, but more recent efforts have expanded to encompass their cognitive functions. One of the most influential models is the Gurney-Prescott-Redgrave (GPR) model, which provides a comprehensive framework for understanding BG function in action selection and has led to many other models that expand it in various ways [86–90].

GPR Model: The GPR model synthesizes anatomical and functional data into a coherent computational framework [91]. It formalizes the concept of action selection using a network of parallel channels representing different actions. Each channel's activity reflects the salience of the corresponding action, with the most salient action being selected through disinhibition in the output nuclei (GPi and SNr). The model incorporates the direct and indirect pathways, with dopamine modulating the activity of these pathways to influence action selection.

Dopamine's Role: In the GPR model, dopamine plays a critical role in modulating the activity of striatal neurons. D1 receptor activation enhances activity in the direct pathway, promoting action selection, while D2 receptor activation suppresses activity in the indirect pathway, inhibiting competing actions. This dual modulation by dopamine is essential for balancing the facilitation and inhibition of actions.

Extensions and Applications: The GPR model has been extended and adapted to incorporate new biological data and to explore its implications for neuromorphic systems [91]. These extensions include models of the wider thalamocortical loop, large-scale spiking neuron models, and robotics and artificial intelligence applications. For example, the GPR model has been used to control robotic behavior, demonstrating its utility in real-world applications [92] [93].

Actor-Critic Models of Basal Ganglia: The BG network can be examined

under the reinforcement learning paradigm, exhibiting an architecture that closely resembles the actor-critic model [94] [95]. In this framework, the actor component comprises the matrix of the striatum, the GPe, STN, and the SNr/GPi. This actor component is responsible for the policy function and action selection.

On the other hand, the critic component encodes the reward signal and is facilitated by the striosome compartment of the striatum [96]. The critic component encodes the value function through the nigrostriatal connections, influencing the actor's decision-making process based on the expected rewards. This integration allows the basal ganglia to efficiently learn and perform action selection based on the reinforcement learning principles.

2.3 Dopaminergic Learning and Dopamine Encoding

Dopamine, a neurotransmitter crucial for various neural processes, plays a central role in learning, motivation, and movement [97]. It is primarily produced in two main areas of the brain: the SNc and the VTA. From these regions, dopaminergic neurons project to various brain parts, including the striatum, prefrontal cortex, and limbic system, forming several key pathways.

The substantia nigra, part of the midbrain, gives rise to the nigrostriatal pathway, where dopamine neurons project to the striatum [98]. This pathway is essential for motor control and is significantly affected by Parkinson's disease. The VTA, another midbrain structure, originates from the mesolimbic and mesocortical pathways. The mesolimbic pathway projects to the nucleus accumbens in the ventral striatum and is crucial for reward processing and reinforcement learning [62] [99]. The mesocortical pathway, projecting to the prefrontal cortex, involves higher cognitive functions such as decision-making and executive control.

Dopaminergic circuits that project to the BG are particularly important [100]. Dopamine modulates the activity of BG circuits, influencing action selection and

learning processes. This modulation occurs through the dopaminergic neurons' tonic and phasic firing patterns, which correspond to baseline dopamine levels and bursts of dopamine release in response to salient or rewarding stimuli, respectively.

2.3.1 Dopamine Encoding: From Saliency to Reward Prediction

Dopamine, a neurotransmitter crucial for various neural processes, has been the focus of extensive research due to its central role in learning, motivation, and movement [101] [79]. Historically, dopamine's significance was first highlighted in the context of Parkinson's disease, which led to deeper explorations of its functions in the brain. Over time, research expanded beyond movement disorders, uncovering dopamine's pivotal role in reinforcement learning and decision-making processes.

Initially understood within the narrow confines of motor function, dopamine's broader relevance to neural processing and behavior gradually came to light through the latter half of the 20th century. Pioneering studies began associating dopamine with reward, formulating key hypotheses regarding its role in learning and motivation [62].

2.3.1.1 Saliency and Reward Prediction Error Hypotheses

Two foundational hypotheses have shaped our understanding of dopamine's function [81]: the saliency hypothesis and the reward prediction error hypothesis . The saliency hypothesis posits that dopamine signals the importance or saliency of external stimuli, aiding attention and decision-making. This view was later complemented and expanded by the reward prediction error hypothesis, which suggests that dopamine neurons encode the difference between expected and received rewards, a mechanism at the heart of reinforcement learning.

2.3.1.2 Development of the Reward Prediction Error Hypothesis

The reward prediction error hypothesis, grounded in the principles of temporal difference learning [102], gained empirical support through neurophysiological studies showing that dopaminergic neuron firing correlates with unexpected rewards, as illustrated in Figure 4. This hypothesis has been further refined to describe how dopamine signals encode reward and prediction errors, thereby driving adaptive behavior and learning [3].

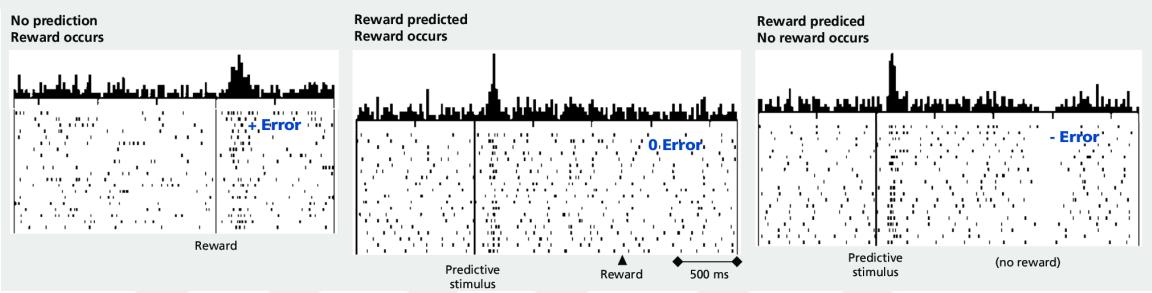


Figure 4: Dopamine Reward Prediction Encoding. Novel reward (left), predicted and received reward (middle), and omission of predicted reward (right) [3].

The hypothesis states that the dopamine neurons encode the difference between the expected and received reward. In this mechanism, the dopamine neurons react strongly to unpredicted novel rewards, which is the primary reinforcer. As the agent interacts with the environment and the task, the critic learns the predictive cues of the reward. This results in the expectation of the reward to shift the reward signal encoded by the DA neurons to the time of learned reward-predicting stimuli. If the resulting sequence of choices brings a reward, this reward is not encoded as it was expected. If this reward is omitted, the DA neurons are inhibited, resulting in below-baseline activity. Therefore, utilizing this concept of long-term planning or effective learning of a long sequence of action is possible as in a chain of reward-predicting cues depicted in Figure 5 [79] [103].

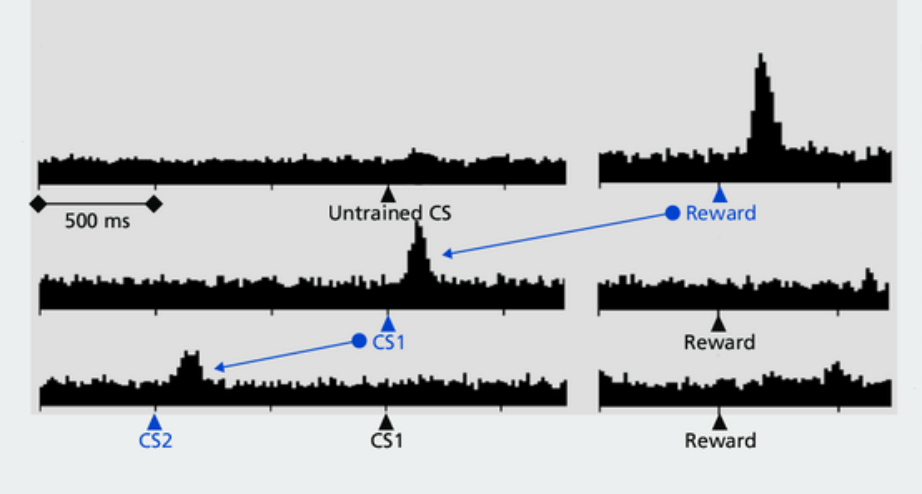


Figure 5: The back-shift of reward signal to learned reward predictive cues (CS: Conditioned Stimulus) in time as the critic learns (time axis flowing from top to bottom) [3].

The reward prediction error encoded by dopamine neurons can be mathematically described as [96]:

$$\begin{aligned} \hat{R} &= R + \Delta DA \\ \Delta DA &= \gamma P(t) - P(t-1) \end{aligned} \tag{4}$$

where ΔDA represents the change in dopaminergic activity, R is the received reward, $\hat{R}$ is the expected reward, γ is the discount factor, and $P(t)$ is the predicted value of a stimulus. This simple yet powerful equation encapsulates how dopamine influences learning through feedback mechanisms.

2.3.1.3 Beyond Reward: Additional Complexity and Functions

Recent research has revealed additional complexities in dopamine's role, identifying diverse subpopulations of dopaminergic neurons with distinct functions, ranging from movement to cognitive processes [101]. This diversity suggests a more nuanced role of dopamine in neural computation, extending beyond the classical reward-centric framework. Moreover, dopamine encodes other aspects of tasks and behavior, such

as effort, motivation, and the temporal aspects of rewards, indicating its involvement in a broader spectrum of decision-making processes [101].

Dopamine's encoding of saliency and reward prediction error represents foundational principles in our understanding of neural processing and learning. The evolving view of dopamine, from a neurotransmitter associated with motor control to a critical player in complex cognitive functions, underscores the dynamic interplay between neurochemical signaling and behavior. As research continues to unravel the multifaceted roles of dopamine, our models of neuromorphic computing and artificial intelligence stand to benefit from more profound insights into this versatile neurotransmitter's functions.

CHAPTER III

IMPLEMENTATION OF LEARNING NEURON MODELS

This chapter presents a comprehensive methodology for implementing a bio-realistic model of the Basal Ganglia (BG) on Spiking Neural Networks (SNNs) realized on neuromorphic substrates. The following outlines the systematic approach taken, from implementing the Izhikevich neuron model on the Loihi2 chip to integrating learning rules and developing learning neuron models. We cover the architecture of the Loihi2 chip, the detailed process of adapting the Izhikevich model for this hardware, the implementation of synaptic and dopaminergic dynamics, and the application of spike-timing-dependent plasticity (STDP) and neuromodulated learning rules.

3.1 Intel Loihi2 Architecture

The second-generation neuromorphic hardware by Intel, named Loihi2, along with the accompanying software framework LAVA, represents a significant advancement in neuromorphic computing. This fully digital neuromorphic chip features a many-core network-on-chip architecture, where each core, referred to as a neurocore, operates asynchronously while communicating through global synchronization policies [22] [104]. The router architecture's core address mappings facilitate essential fan-in and fan-out scalability and graded spike support, making the system robust and versatile for various computational tasks.

One of the distinguishing features of the Loihi2 chip, which significantly contributed to its selection for this research, is its programmability through LAVA framework. Each neurocore is equipped with microcircuits that allow extensive customization, enabling the implementation of complex neuron models and learning algorithms. The detailed structure of an individual neurocore is illustrated in Figure 6.

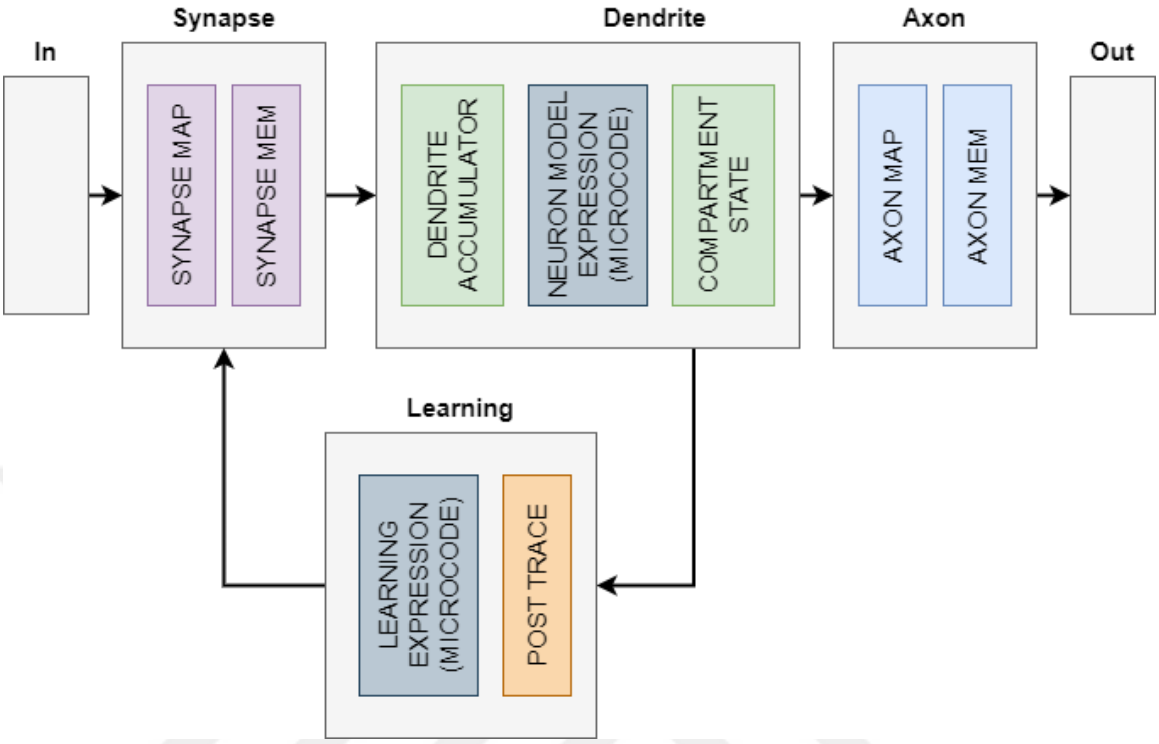


Figure 6: Structure of an individual neurocore in Loihi2 [4].

3.1.1 Neurocore Architecture

The neurocores of the Loihi2 chip contains an internal SRAM, known as MPDS, which stores all necessary information for the functioning of the microcircuits. Critical to our work is the dendrite microcircuit, which is responsible for calculating neuron dynamics. These dynamics are categorized into two types: hard-coded neuron models that are supported on-chip, allowing for highly efficient computations, and micro-coded neuron models, which are defined through micro-code and executed by neuron threads on the microcircuit.

3.1.2 LAVA Software Framework

Lava is a versatile software framework designed for developing neuro-inspired applications and deploying them on neuromorphic hardware like Intel's Loihi 2. It provides a modular, composable, and extensible architecture, allowing the creation of complex neural networks. The framework is built around the concept of Processes and

ProcessModels. Processes define the internal variables and input/output of neural components, while ProcessModels handle the hardware-specific implementation details and describe the behaviour.

Lava supports both CPU simulation and execution on neuromorphic chips, offering flexibility in development and deployment. The framework ensures that learning rules can be implemented without modifications across different platforms. This compatibility allows seamless transition from CPU-based simulations to hardware execution on Loihi 2.

One of the key advantages of using Lava is its ability to optimize code for specific hardware, enhancing performance and efficiency. While CPU simulations provide a flexible and less constrained environment for development, execution on neuromorphic hardware like Loihi 2 offers faster computation and lower power consumption, crucial for large-scale neural network applications. However, hardware-specific constraints might require additional considerations in ProcessModel implementations.

Overall, Lava bridges the gap between high-level neural network design and low-level hardware execution, facilitating the development of advanced neuromorphic computing applications.

3.1.3 Neuron Dynamics and Micro-Code Execution

Neuron threads utilize four 64-bit registers for specific purposes and several sequential processing elements (PEs). Each micro-code segment, or pass, translates into a single run of these PEs. At the start of each pass, the necessary memory structure (compartmental state; Cx) belonging to the neuron, which holds the neuron states, is loaded onto the register. Sub-circuits manage the writing of the program segment (pass), which contains the micro-code instructions onto the PEs. By sequentially executing the entire micro-code, pass by pass, neuron dynamics are realized, and spikes are instantiated through specialized instructions.

As there is a single memory-related register on the threads, only a single compartmental state can be manipulated and used simultaneously. Fortunately, two registers can store custom variables defined in the micro-code and maintain their value between passes.

3.1.4 Compartmental States and Fixed-Point Representation

Compartmental states are 64-bit memory blocks in SRAM and are grouped according to their neuron associations. Each neurocore manages a certain number of neurons, which is determined by the compiler's network mapping onto the chip. These neurons have state variables stored in the compartmental states and are sequentially grouped in the SRAM. Neuron threads, of which there are four per neurocore, process all compartmental states iteratively. Loihi2 utilizes fixed-point representation for ease of computation, with variables defined as either 8, 16, or 24 bits, signed or unsigned. Each compartmental state must be 64-bits long and fully utilized.

3.1.5 Memory and Custom Programming

Each Loihi2 neurocore has constraints on its SRAM, defining the storage capacity for neuron-related variables such as SynMem (holding synaptic weights and addresses), AxMem (routing information for axons), Cx states, and micro-circuit configuration fields. The upper limit for Cx states is 4095, which can also be allocated for other processes. Custom programming supports various processes necessary for networks, such as injectors, extractors, monitoring tools, inputs, spike generators, or any logic implemented through micro-code.

3.1.6 Instruction Set and Arithmetic Operations

Each micro-code instruction is executed by sequential PEs, which can require up to two PEs depending on complexity. These instructions support basic integer arithmetic and logic operations, except division. The flexibility of the Loihi2 architecture does

not end with the dendrite block; the learning engine, another microcircuit under each neurocore, can also be programmed with custom learning rules.

3.1.7 Learning Engine

The Loihi2 chip features a programmable learning engine that supports a variety of learning rules, including spike-timing-dependent plasticity (STDP) and more complex neuromodulated learning rules like the 3-factor (3F) learning rule. The key components of this learning engine are:

- *Tags, Weights, and Delays*: The learning engine uses three primary variables: tags, weights, and delays. These variables are manipulated by custom learning rules to adjust synaptic strengths and timings.

- *Microcode Programming*: Learning rules are implemented using microcode, allowing precise control over learning dynamics. Each microcode segment called a pass, is executed sequentially by neuron threads, processing synaptic and neuronal state variables.

- *Epoch-Based Updates*: For efficiency, trace and synaptic variable updates proceed in learning epochs with a length of t_{epoch} time steps. Within an epoch, spike events are recorded, but trace and synaptic variable updates are computed and applied at the end of the epoch, ensuring efficient processing without impacting learning accuracy as long as spikes are infrequent within each epoch.

- *Sum-of-Products Learning Rules*: Learning rules are specified in sum-of-products form, where each product consists of trace variables multiplied by scaling factors and dependency operators. This structure allows complex learning dynamics to be efficiently implemented.

- *Dependency Operators*: These operators condition the evaluation of products

on the presence of pre- or post-synaptic spikes during the past epoch or evaluate products unconditionally every other epoch. They determine when trace variables are evaluated during an epoch.

- *Scaling Factors*: Each product is associated with a scaling factor given in mantissa/exponent form, allowing precise adjustment of learning rule parameters.

- *Traces*: Traces are low-pass filtered versions of spike trains used in online implementations of STDP. Loihi2 provides a set of two pre-synaptic traces and three post-synaptic traces per synapse, with dynamics given by:

$$\frac{dT(t)}{dt} = -\frac{T(t)}{\tau} + \sum \delta(t - t_i) \tag{5}$$

where $T(t)$ is the trace value, τ is the decay time constant, and $\delta(t-t_i)$ represents the occurrence of a spike at time t_i.

The programmability of the learning engine on Loihi2, combined with its support for custom microcode and flexible trace dynamics, allows for a wide range of learning rules to be implemented efficiently.

3.2 Implementation of the Izhikevich Neuron Model on Loihi2

The choice of the Izhikevich neuron model for this study stems from its balance between biological realism and computational efficiency. The model captures a wide range of neuronal behaviors with a relatively simple set of equations, making it ideal for neuromorphic implementations where computational resources are at a premium.

The differential equations of Izhikevich model's fast and slow dynamics were introduced in the Equations 1-2-3. These equations are then transformed into their discrete form using the Euler-Maruyama method [105], resulting in the following;

$$
\begin{aligned}
v[t+\Delta t] &= v[t] + (0.04v[t]^2 + 5v[t] - u[t] + I)\Delta t \\
u[t+\Delta t] &= u[t] + a(bv[t] - u)\Delta t \\
v[t] > v_p &\Rightarrow v[t+\Delta t] = c, u[t+\Delta t] = u[t] + d
\end{aligned}
\tag{6}
$$

In addition to the soma equations, synaptic dynamics are crucial in modeling the interactions of various neurotransmitters, neuromodulators, and synaptic currents. Key neurotransmitters involved are γ-Aminobutyric acid (GABA), glutamate, acetylcholine, and dopamine, each contributing to synaptic conductance with distinct time constants [72]. The combined synaptic inputs drive the membrane potential. The level of granularity at which these conductances are modeled depends on the desired biological plausibility and synaptic complexity. There are three primary approaches to handling these dynamics [106]:

1. **Individual Neurotransmitter Conductances:** Each neurotransmitter conductance is computed as a unique trace, reflecting complex chemical pathways and communications. This approach is mainly used when bio-plausibility is a primary concern, and neural computation depends on the specific neurotransmitters.

2. **Combined Excitatory and Inhibitory Currents:** Excitatory and inhibitory neurotransmitters are grouped into two currents, resulting in a simplified dynamic model.

3. **Simple Conductance Dynamics:** All incoming potentiations are integrated into a single current with a decaying nature. This approach is used in our model and is represented by the following equation:

$$
I = I_{const} + I_{syn}(1 + \beta \Delta dop) \tag{7}
$$

$$\frac{dI_{syn}}{dt} = -\frac{I_{syn}}{\tau_{syn}} + \sum (w_{ij}\delta(t - t_f^{(i)})) \tag{8}$$

The current feeding the neuron consists of two components: the constant current, which implies background activity from other neural structures, and the driving current, denoted as the synaptic current. Dopamine typically has a neuromodulatory effect on the synaptic weight and stimulates synaptic current. This effect can be excitatory or inhibitory, depending on the receptors located on the synapse. These effects are mediated through dopamine conductance. However, since our model uses simple conductance dynamics, this effect is translated into a scaling effect on the synaptic current. The coefficient β represents the receptor in the synapse and defines how the current changes according to dopamine levels, while δ_{dop} represents the dopamine level relative to the baseline dopamine. This dopamine scaling allows us to simulate dynamics similar to those observed in Medium Spiny Neurons (MSNs) of the striatum, which express different dopamine receptors.

The drive current is expressed as the integration of synaptic input over the weights of the connections. It exhibits a decaying behavior, allowing the neuron to retain a notion of short-term memory of past inputs. Similar to how the neuron model equations were discretized, we convert the continuous-time synaptic equations into their discrete form using the Euler-Maruyama method, resulting in Equation 9.

$$I_{syn}[t + \Delta t] = \alpha I_{syn}[t] + \sum (w_{ij} f_i(t)) \tag{9}$$

As the Loihi2 works in discrete time-steps which can be termed algorithmic time-steps we take δt as $0.125ms$ which is a natural power of 2, this makes the computation on hardware that manipulates binary number more convenient to use, allowing us to exploit the available shift operations on the PEs of neuron threads. The decay factor $\alpha = 1 - \Delta t/\tau_{syn}$ reduces the importance of information encoded in a spike as time passes.

A high-level look at a neuron and how input spikes get processed sequentially by conductance dynamics and soma neuron model equations can be seen in Figure 7.

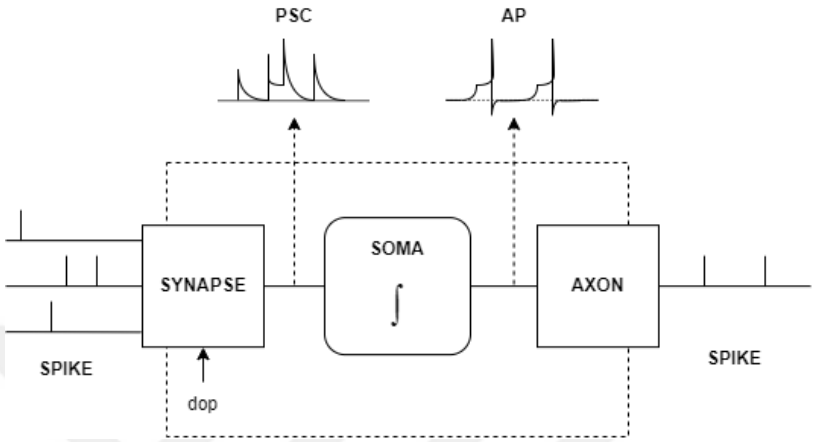


Figure 7: Single point neuron diagram of Izhikevich neuron model [4]

The final analytical solution of all the neuron equations provides a foundation for building our logic, expressing this solution in micro-code to implement it on Loihi2. The micro-code logic necessitates binary operations on the Cx state variables, requiring a standard representation for these variables. We employ fixed-point representation, which aligns well with the neurocore architecture. The precision of the fixed point, chosen based on the range and sensitivity of the variables, is set at 12 bits. As previously mentioned, the Cx state holds dynamic variables of either 8, 16, or 24 bits. Utilizing 12 bits as fraction bits, we calculate the necessary range for the variables and determine the optimal bit width, as presented in Table 1.

Table 1: Memory allocation of model parameters in compartment states.

Variable	Size (bits)	Cx State Index	Variable	Size (bits)	Cx State Index	Variable	Size (bits)	Cx State Index
a	16	0	c	24	1	v	24	2
b	16	0	d	24	1	u	24	2
I_{syn}	24	0	Δ_{dop}	16	1	I_{const}	16	2

Given the constraint of accessing only one Cx state at a time for computation and the requirement to fill all 64 bits of a Cx state with variables, careful consideration must be given to variable mapping. The primary factor in deciding this mapping is the convenience of usage, leveraging spatial locality. Variables frequently used together should be mapped to the same Cx state. Conflicts with this rule result in wasted passes to transfer variables using the neuron thread register, thereby wasting potential register space that could otherwise be used for computation.

Before adapting the implementation to neurocore-specific micro-code, we must express it as an algorithmic flow. This flow incorporates the discrete Izhikevich model equations and synaptic dynamics as primary concerns. Figure 8 depicts the detailed algorithmic flow.

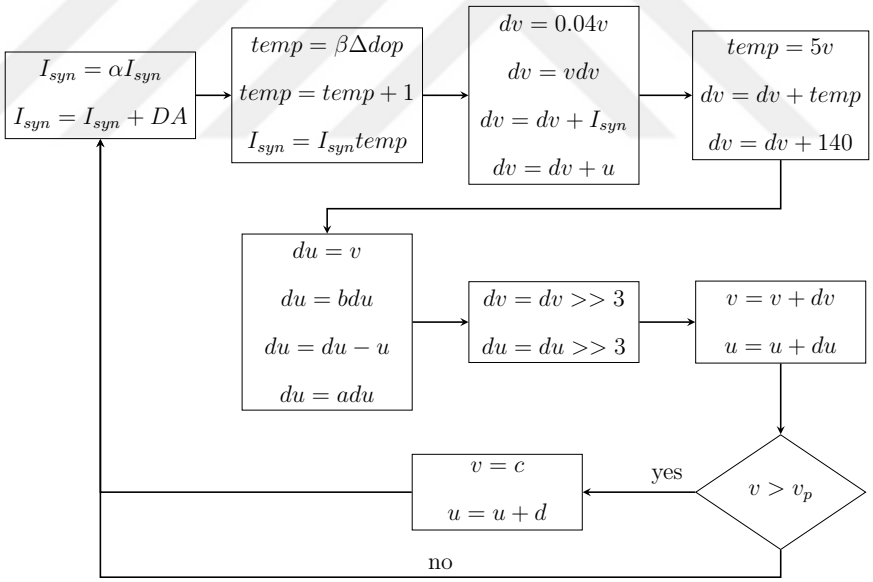


Figure 8: Flow diagram of operations needed to implement the discrete form of the Izhikevich model. Each primitive operation (e.g., addition, multiplication, and shift-right) is implemented via corresponding instruction in the microcode.

The neurocore adaptation of this algorithmic flow considers the constraints on neuron threads, such as the pass structure, limited registers, restricted access to Cx state variables, and limited bitwidth. The implementation begins with a pass

to calculate the postsynaptic current (PSC), which involves the weighted sum of connecting inputs, represented by the term $\sum (w_{ij}\delta(t - t_f^{(i)}))$. This term is realized through synapse and dendritic accumulator (DA) micro-circuits and is stored in the DA variable. When creating connection processes in the Lava framework, the bit precision for expressing weights is initially set to a default of 6 bits. An additional 4 bits are added when defining the weights. In the micro-code, the precision is further shifted left by 2 bits to achieve the common ground of 12 bits shared by other variables. A similar approach is applied to I_{const} to ensure the variable range is expressed without sensitivity errors. Out of the 16 bits allocated to I_{const}, 10 bits are considered fractions instead of the 12-bit weight choice to ensure sufficient binary bits for the variable range. This discrepancy is later addressed using shift operations to equalize the fractions inside the micro-code.

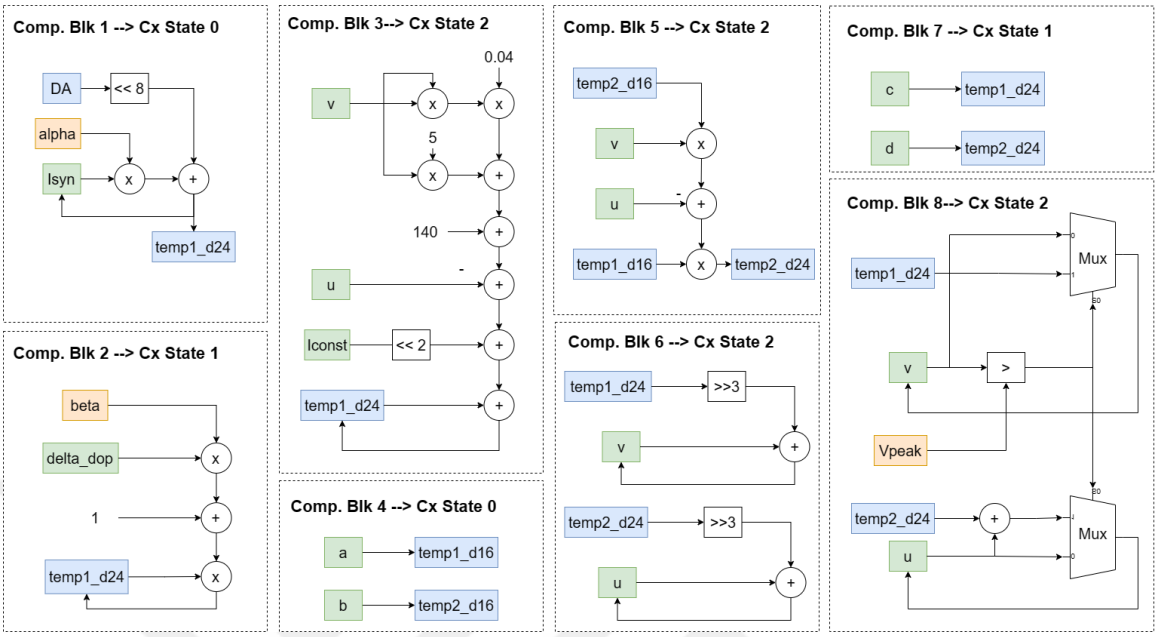


Figure 9: Flow diagram of discrete computation of the Izhikevich neuron on Loihi2. Each computational block executes a limited number of instructions and accesses a single Cx state. Computational blocks are executed sequentially, starting from Comp. Blk 1 to 8. Variables in green boxes are stored in the Cx state, blue boxes represent the allocation of the register, and variables in orange boxes are constants. The same constant value is used by all neurons in a given neuron group [4].

Incorporating all constraints and techniques, the algorithmic flow that defines the synaptic dynamics and neuron model is transformed into the Loihi2 dendrite micro-circuit micro-code, realizing the Izhikevich model with simple conductance dynamics and dopamine modulation on hardware. The final implementation is depicted in Figure 9.

In our earlier research, we extensively validated the implementation of the Izhikevich neuron model against various subcortical neuron types by applying constant current and comparing the resulting membrane voltage trends to those reported by Izhikevich [4]. We also analyzed the impact of hardware-specific constraints, such as

those on the Loihi2, on neuron behavior by comparing floating-point CPU simulations with fixed-point implementations on both the Loihi2 and CPU. Additionally, we compared the efficiency of Izhikevich and Leaky Integrate-and-Fire (LIF) neurons regarding power consumption, latency, and resource allocation whose details were published in our previous work [4].

3.3 Implementation of the Learning-capable Izhikevich Neuron Model

We extend our model by implementing Spike-Timing-Dependent Plasticity (STDP) in Izhikevich neurons. This is followed by the introduction of neuromodulated learning using a three-factor (3F) learning rule, incorporating Izhikevich neurons capable of reward-modulated STDP (R-STDP). Subsequently, we observe the learning on a network scale using a classical conditioning experiment. Finally, we simulate a fourth factor, inspired by the biochemistry of dopamine, to showcase the capabilities of four-factor (4F) learning.

3.3.1 STDP Izhikevich Model on Loihi2/CPU

The learning rules apply to both CPU and Loihi2 implementations without any modifications, as Lava replicates the behavior of the Loihi2 learning engine. Consequently, the learning rule can be seamlessly used on CPU and Loihi2 platforms.

The aspect that requires updating is the neuron behavior defined in the ProcessModel, as the trace calculations need to be performed within the ProcessModel, which is hardware-dependent. Therefore, each platform (CPU/Loihi2) uses its code to describe trace equations. This implementation is more straightforward and subjected to fewer constraints on the CPU side, so we will initially implement it on the CPU side.

In CPU implementations the inherited function *run_spk* runs at every timestep on the CPU and is considered a substep in the Loihi2 processing phases. In Loihi2,

an algorithmic timestep comprises specific time phases [107]:

- *Pre-Management Phase:* During this phase, the available CPUs on the Loihi2 chip execute pre-management code (if available) before neuron calculations are executed on neurocores. This allows direct control of variables in the neurocores, the application of control logic, and manipulation of the NoC.
- *Spike Phase:* Neuron calculations are performed by neurocores, axons send resulting spikes, and neuron variables are updated. This step realizes the handshake protocol for overall synchronization.
- *Learning Phase:* During this phase, the learning engine applies the learning rule, and the resulting weight updates are reflected in the MPDS SynMem. This phase occurs only when every t_{epoch} has passed.
- *Post-Management Phase:* Similar to pre-management, this phase occurs at the end of the timestep, allowing fine-grain control over the execution flow.

In the ProcessModel, the *run_spk* function replicates the Spike phase of the Loihi2. Within this function, we can define the functionality executed by the neurocores or, in this context, the CPU. The traces are managed in this phase, and the variables used in the learning engine are calculated here via code (CPU) or micro-code (Loihi2). This addition to our previous implementation, which only included synaptic dynamics and neuron equations, now also encompasses calculating the learning rule variables.

The exact implementation applies to Loihi2, where trace calculations are necessary. However, additional architectural details must be addressed. In addition to the Cx states that hold neuron-related dynamic variables, we have the Dendritic Accumulator Memory (UCDA). The variables y_{epoch0} and t_{spike} are crucial in this memory compartment.

Every learning process inherits variables $y1$ and $y2$ through the Learning Neuron Process. The learning rule defines the post-trace as $y1$ and the third factor, which will

be introduced later, as $y2$. These variables are processed within the code, utilizing y_{epoch0} for the post-trace and writing the value to $y1$. This addition necessitates modifying the algorithmic flow, corresponding to two passes: i) a pass for calculating the decayed trace and adding the impulse, and ii) a pass for writing these effects to the appropriate variable, addressing bit effects related to fixed-point representation on Loihi2. The t_{spike} variable acts as a flag set every time a spike occurs, forming the time-dependency variables in the learning rule that express pre- and post-spike timings.

The addition to the algorithmic flow and the newly structured pass system can be seen in Figure 10 and Figure 11.

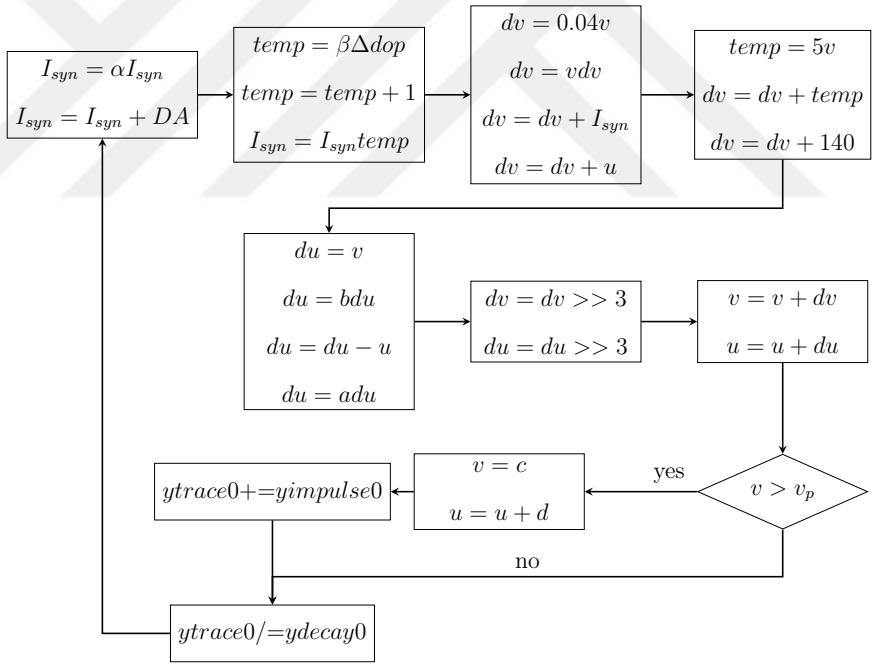


Figure 10: Flow diagram of operations needed to implement trace calculations in conjunction with the discrete Izhikevich model for STDP learning on Loihi2. Showcasing the integration of synaptic dynamics and learning traces.

The learning rule is provided to the neuron model as a parameter, encompassing all necessary constants for trace calculation. The values of $ytrace0$ and $yepoch0$ are

stored in the Cx states and UCDA memory, respectively. Due to Loihi2's lack of support for dynamic division, the decay value is precomputed and embedded into the microcode as a constant. The impulse magnitude upon spike detection is handled similarly.

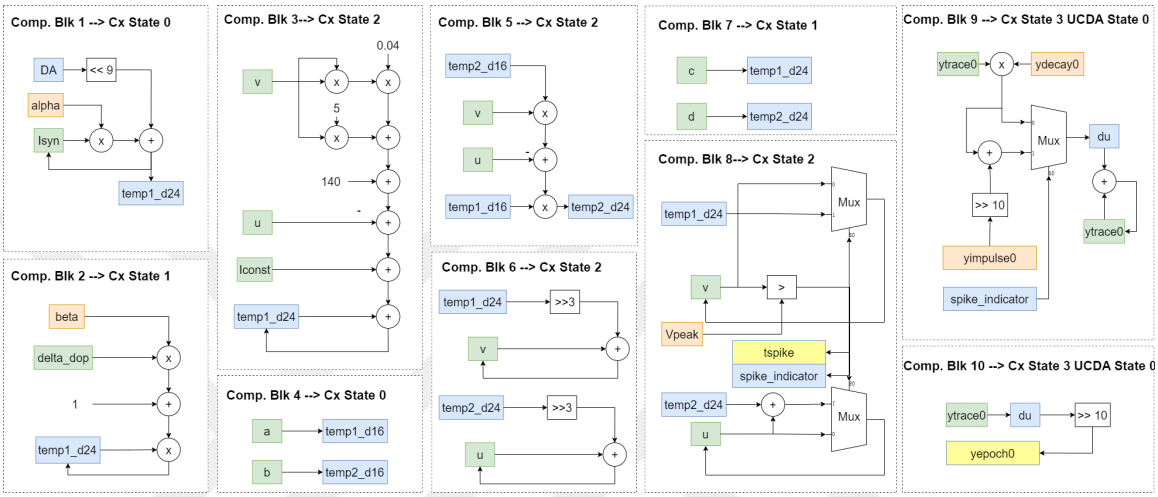


Figure 11: Flow diagram of discrete computation of the Izhikevich neuron on Loihi2 extended for STDP trace calculations. Blue boxes represent the allocation of the register, and variables in orange boxes are constants, while yellow boxes are UCDA memory for learning engine variables.

One critical piece of information about trace calculation on Loihi2 is that the learning engine treats *yepoch*0 as an integer with no fractional bits. Consequently, the decaying operation must be performed on a higher-precision variable to avoid rounding up to the same integer value. Therefore, the bit width for *ytrace*0 has been selected as 16 bits, with 10 bits allocated for precision. The shifting operations in the microcode are used to perform the calculations at a higher precision before rounding the number to a lower precision.

Two neurons were created to test the learning rule: a pre-neuron connected to a post-neuron, both modeled as regular spiking Izhikevich neurons. The connection strength between the neurons was initialized to zero. The pre-neuron was stimulated

with a random single spike input, while the post-neuron was driven to fire using a constant current. The learning rule was validated by examining trace information, spike timings, and weight changes to demonstrate its accuracy and functionality.

3.3.2 R-STDP Izhikevich Model on Loihi/CPU

The 3F (three-factor) learning rule enhances the traditional spike-timing-dependent plasticity (STDP) model by incorporating a dynamic term representing a neuromodulator (R) [108]. This addition refines the weight update mechanism, as detailed in Equation 10:

$$\Delta W = R \cdot E \tag{10}$$

$$\frac{dE}{dt} = -\frac{E}{\tau} + \text{STDP}(\text{pre}, \text{post}) \tag{11}$$

The neuromodulator term introduces another trace dynamic. In classical reward-based STDP models, the reward signal functions as a gating mechanism, typically the input given through the reward connection without a dynamic component. Our model incorporates a more biologically realistic approach by adding a baseline to the trace and implementing an exponential decay akin to dopamine dynamics. The Python implementation of this dopaminergic 3F calculation is provided in Listing 1 in the appendix.

The Python code illustrating the third-factor calculation on the CPU is provided in the appendix (see Listing 1). To replicate this behavior on the Loihi2, additional micro-code passes calculate *yepoch*1, which holds the reward trace value, and write it into $y2$, the third factor for the learning engine. As the eligibility trace calculation by the learning engine encountered unknown problems, we do not include the Loihi2 implementation of the R-STDP rule and Izhikevich correspondence in this study.

Equations 10 and 11 are implemented using a sum-of-products form, as provided in the appendix (see Listing 2).

For testing the R-STDP Izhikevich models, we used a similar setup to the STDP case with occasional reward signals. The eligibility trace, spike times, traces, reward signal, and weight changes were observed to verify the learning rule's efficacy. These observations and verifications are detailed in Chapter 5.

3.3.2.1 Classical Conditioning Experiment

The classical conditioning experiment demonstrates how stimuli can be associated with rewards through consistently rewarding a stimulus. This experiment follows the setup described by Izhikevich [109]. The primary alteration in our setup is the method of input current delivery to the neurons. While Izhikevich uses a random current representing thalamic input to cortical neurons, we achieve a $1Hz$ firing rate of the cortical neurons by overdriving them with a single super-threshold spike.

Another significant modification is the duration of the signals. The biological constants for synaptic eligibility suggest time constants of up to 1 second for eligibility traces and 0.2 seconds for dopamine signals. Replicating these durations in the Lava is infeasible due to constraints in the Loihi2 learning engine architecture, which Lava inherits even on the CPU. The decay variable responsible for the exponential decay is too small to be represented within the number representation allowed, with the exponential field ranging from -9 to 7. This constraint limits the maximum time constant to $140ms$. Consequently, we redesigned the task, adjusting the eligibility trace time constant , dopamine time constant, and reward delay to achieve comparable weight potentiation.

The network simulation runs for 15 minutes of simulated time with $1ms$ timestep.

The network comprises 800 excitatory regular spiking neurons with plastic connections and 200 non-plastic fast-spiking inhibitory neurons, with a connection probability of 0.1. The background firing rate is $1Hz$. This network receives input from 100 stimuli groups containing 50 stimuli neurons. We record the average weight of the connections made by the neurons in the stimuli group.

3.3.3 Dopamine Dependency - 4 Factor Learning

The classical conditioning experiment by Izhikevich highlights the challenge of distractor stimuli benefiting from reward signals. To address this, we introduce a fourth factor, dopamine receptivity, which adds a temporal component to the neuromodulatory signal based on the dynamics of Calcium-Calmodulin-Dependent protein kinase II (CAMKII) autophosphorylation [5]. This receptivity factor, modeled as a Gaussian-shaped delay, ensures that only stimuli occurring within a specific time window around the reward influence synaptic strengths, reducing the impact of distractors.

We modified the classical conditioning setup to include 100 stimuli, each with its eligibility trace, dopamine level, and synaptic weight. Stimuli were randomly presented at intervals, and weights were updated based on the product of dopamine, eligibility traces, and receptivity. The experiment was conducted for a simulated hour, with synaptic weights, eligibility traces, dopamine levels, and receptivity recorded over time.

For the experiment, a simplified setup was used where the nuances of spiking neurons were overlooked, and no real neurons were used. The network ran two versions: one with the dopamine receptivity factor added to the weight change equation and one without. Stimuli were presented randomly, with some eliciting rewards after a delay. The weight changes, eligibility traces, dopamine levels, and receptivity were recorded over time.

CHAPTER IV

MODELING THE ACTOR-CRITIC ARCHITECTURE OF BASAL GANGLIA

The Basal Ganglia (BG) plays a pivotal role in action selection and reinforcement learning, making it an ideal subject for modeling neural mechanisms underlying these processes. Following the validation of our neuron model, we progress to a higher level of organization, focusing on the intricate network of nuclei within the BG. This chapter outlines the incremental development of a BG model, beginning with a simple Go/No-Go task and culminating in a complex sequence learning task integrated with a reinforcement learning framework.

Our approach involves a incremental methodology to build and validate the BG model. Initially, we model the fundamental Go/No-Go task, a classic paradigm to study BG-mediated action gating. This foundational task serves as a baseline to ensure the network can gate actions appropriately. Upon successful implementation, we expand the model to tackle sequence learning [110], incorporating a dynamic environment where state information is encoded onto a cortex-stimulating neuron group. The expanded model consists of two approaches: first, a no-channel model that embodies the idea of a holistic BG is implemented. Second, a segregated multi-channel BG model is developed to facilitate the action selection process.

To validate the action selection mechanism within the BG, we initially employ artificial strengthening of connections between specific stimuli groups and corresponding action channels. This step confirms that the network can perform correct action selection based on input stimuli. The next phase involves developing a temporal-difference (TD) learning-based dopaminergic critic [102]. This subcircuit encodes the dopamine

signal, reflecting the reward prediction error. By incorporating this critic, we aim to create a robust model where dopamine modulates synaptic plasticity, guiding the network towards optimal action selection.

Finally, we integrate the actor and critic components to form a comprehensive BG model capable of learning and performing sequence tasks. The actor network represents the action selection mechanism, while the critic network provides the reward signals necessary for learning. We refine the model through iterative training and validation to ensure it accurately replicates the biological processes involved in BG-mediated learning and action selection.

This structured approach, progressing from simple action gating to complex sequence learning, allows us to build and validate a bio-realistic model of the BG systematically. Each step is thoroughly tested to ensure fidelity to biological processes, ultimately providing a robust framework for understanding the neural basis of reinforcement learning within the BG.

4.1 Basal Ganglia Circuit

In this section, we focus on the specific mechanics of our BG circuit, the neuron model, and the complexity involved. These elements define inductive biases and establish the foundational structure of our network.

The structure illustrated in Figure 12 forms the bases of our BG model. This section analyzes the design choices and the rationale behind their selection.

For modeling the BG, we adopt a top-down approach and consider the global brain circuitry. The BG primarily receives input from the cortex, encompassing almost the entire cortical region, and the thalamus. It is posited that the thalamus serves as a relay station and performs its own computation on sensory information to encode signals. Additionally, the limbic system contributes to the input received by the BG. The BG processes these inputs and generates outputs via the thalamus, which are

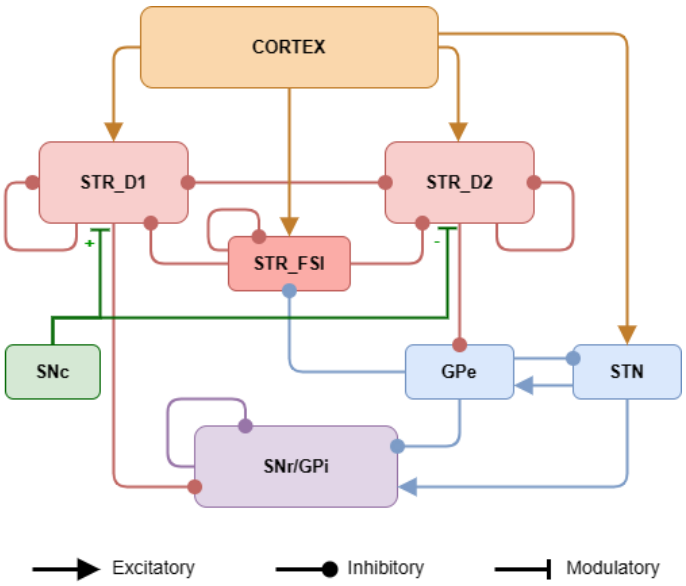


Figure 12: The diagram of the BG model. STR-D1: D1 receptor expressing MSNs, STR-D2: D1 receptor expressing MSNs, STR-FSI: striatum fast-spiking neuron, STN: subthalamic nucleus, GPe: globus pallidus external, SNc: substantia nigra pars compacta, SNr/GPi: substantia nigra pars reticulata/globus pallidus internal [4].

then relayed back to the cortex for execution or form a closed loop without involving execution regions, thereby enabling internal simulation. This entire loop is referred to as the Cortico-Ganglia-Thalamo-Cortical loop [111].

In our model, we have replaced the six-layered cortex and other input regions with a Poisson Spike Generator (PSG) neuron group to simulate the output layer of the cortex onto the striatum. This inductive bias simplifies the environment's input to the agent. Typically, the cortex is responsible for learning environmental patterns and refining sensory data into high-level abstract information about states. It allows for better representation of the environment by identifying what is relevant and efficiently vectorizing this information, effectively serving as an encoder of state information. Since we explicitly define the environment and state space, we can design the state stimulus to contain the most relevant information. This approach enables us to bypass the cortex's complex information processing and learning. The final parameters for the PSG group mimicking the cortex and its connections are detailed in Table 2, the

G_Ctx1 group inputs to Striatum and its interneurons, while G_Ctx2 group inputs to STN.

Table 2: Poisson Spike Generator Groups and Connection Parameters

Presyn. Group	#Neuron	Postsyn. Group	Weight	Conn. Prob.	Freq. (Hz)
G_Ctx1	300	STR_D1	5	0.05	10
		STR_D2	5	0.05	
		STR_FSI	2.5	0.15	
G_Ctx2	2	STN	6	0.2	7

Another inductive bias pertains to the output side of the loop. Since it is unnecessary to complete the loop biologically, we can exclude the thalamus and manually read the BG output to select the action. Our chosen environment requires discrete actions, making spike encoding of action information in time non-essential for our setup, thus bypassing the thalamus. In the biological system, the thalamic loop relays selected information back to the command systems in the cortex, creating a positive feedback loop to add persistence [112]. This persistence reinforces the chosen action, preventing dithering between closely competing inputs. However, our model does not implement this mechanism, as we rely on extrinsic state information rather than intrinsic command system saliency.

The peripherals of the BG discussed above. In the following, the internal structure of the BG model is explored. As previously discussed, the input nuclei of the BG is the Striatum. There are three main aspects to consider when analyzing the Striatum: its anatomical structure and internal nuclei, the topographical organization of the neurons, and the neuron types along with their functionality within the Striatum.

As mentioned in Chapter 2, the striatum is divided into Dorsal and Ventral Segments. Despite the functional differences between these regions, they are interconnected, and their intricate circuitry allows for cohesive algorithmic behavior. Various models either consider or ignore this anatomical structure. Our model does not separate the ventral and dorsal striatum as their primary separation, based on their connectivity to different regions, is irrelevant due to our simplified cortex model. Thus, in our model, the striatum is anatomically homogeneous.

From a topographical perspective, the striatum is not homogeneous and comprises two distinct structures: the patches of neuron compartments known as striosomes (patches) and the surrounding matrix. While this distinction is important, it is not considered in our Go/No-Go model as we are currently not focusing on learning. This differentiation will become crucial later when we examine the critic and actor as separate components.

The last important detail about the striatum is the neuron types it consists of. The interneurons of striatum includes neurons that regulate the behavior of the striatum by projecting internally [113]. These interneurons vary according to the chemical neurotransmitters they express. However, as the synaptic dynamics in our model do not consider individual neurotransmitter conductances, we can consolidate all different interneurons under the category of Fast Spiking Interneurons.

One critical detail about Medium Spiny Neurons (MSNs) is their distinctive structure and functionality. MSNs possess dendritic spines that connect to numerous cortico-striatal connections, with each MSN forming approximately 10,000 synapses. The chemical dynamics between the intracellular and extracellular environments are governed by receptor composition, enabling MSNs to exhibit up and down states [114].

In the down state, the excitability of MSNs is low, rendering them less responsive to non-coordinated background firings that fail to generate sufficient conductance on the dendritic spine. However, coordinated cortical activity can lead to sudden

excitation in the spine, altering conductivity and producing a significant difference between sub-threshold and super-threshold behavior. When an MSN transitions to the up state, its excitability peaks, resulting in a burst of activity [115].

This filtering mechanism allows MSNs to remain inactive during non-specific background noise but becomes highly responsive to relevant stimuli that encode state information from the cortex. Our model simplifies this behavior by representing MSNs as regular spiking neurons, deviating from their natural bursting behavior. While this simplification overlooks the inherent filtering and robust representation capabilities of MSNs, it aligns with our current approach of manually selecting actions and inputting stimuli.

The original nature of MSNs, with their capacity for discrete bursts of activity, enhances the robustness of state-action representations by filtering and packaging information into discrete quanta [115]. Although we overlook this detail, we plan to introduce the bursting nature of MSNs when developing a more sophisticated stimuli representation for the state space. This adaptation allows us to leverage the inherent filtering and robust representation capabilities of MSNs more effectively.

Following our examination of the striatum, we now consider the other processing nuclei and the output nuclei of the BG. In this context, it is important to note that we are combining the Globus Pallidus internal (GPi) and Substantia Nigra pars reticulata (SNr) into a single functional unit. This decision stems from the perspective that GPi and SNr serve similar roles in our model, primarily functioning as output nuclei. Consequently, this simplification allows us to streamline the model without compromising its functional integrity.

The final deviation from biological accuracy in our model concerns the tri-phasic behavior of the output nuclei. This phenomenon is crucial for the proper functioning of state-action quanta in the BG and has been documented extensively in the literature [72, 116]. The tri-phasic behavior refers to the changes in the firing rate of the

output nuclei under the influence of the three distinct pathways within the BG, each operating on different timescales.

The hyperdirect pathway, which takes input directly from the cortex to the STN, excites the SNr/GPi with the fastest transmission speed, causing a sudden increase in the firing rate of SNr/GPi. This rapid excitation initiates the quanta, functioning similarly to a preamble in an information packet. It suppresses all competing motor actions, effectively disrupting ongoing activity.

Following this, the inhibition of the selected action is relieved via the direct pathway. Due to its longer transmission delay, this occurs after the initial excitation from the hyperdirect pathway. Finally, the indirect pathway, which has the longest delay, comes into effect. It provides a closing excitation that returns the low firing rate induced by the direct pathway back to the baseline. This sequential activity constitutes the tri-phasic behavior, inherently dependent on these delay mechanisms.

In our model, however, we do not incorporate these timing intricacies. Consequently, we cannot observe the clear separation of quanta characteristic of the tri-phasic behavior. Despite this, the omission does not undermine the validity of our model for the intended simulations and analyses. The focus remains on the essential functionalities that can be achieved without the detailed timing mechanics of the biological system.

The design choices for the remaining aspects of the model, such as nuclei neuron types and connectivity schemes, are determined by network parameters rather than inductive biases. The specific parameters for neuron types in each nucleus are detailed in Table 3, while the connection parameters between nuclei and the sizes of the nuclei are provided in Table 4.

Table 3: Parameters used in the Izhikevich neuron models for the BG Circuit

Group	a	b	c	d	I_{const}	β
STR_D1	0.02	0.2	-65	8	-3.5	0.6
STR_D2	0.02	0.2	-65	8	-3.5	-0.6
STR_FSI	0.1	0.2	-65	2	-3.5	0
GPe	0.1	0.585	-65	4	7	0
STN	0.005	0.265	-65	2	5	0
GPi/SNr	0.005	0.32	-65	2	12	0

Table 4: Connectivity Parameters of BG Circuit

Presyn. Group	#Neuron	Postsyn. Group	Weight	Conn. Prob.
STR_D1	1255	STR_D1	-0.982	0.1
		STR_D2	-0.982	0.1
		GPi/SNr	-1.500	0.15
STR_D2	1255	STR_D1	-0.982	0.1
		STR_D2	-0.982	0.1
		GPe	-1.000	0.15
STR_FSI	84	STR_D1	-0.982	0.1
		STR_D2	-0.982	0.1
		STR_FSI	-0.982	0.1
GPe	46	STR_FSI	-1.429	0.05
		GPe	-1.429	0.25
		STN	-1.429	0.25
		GPi/SNr	-1.000	0.25
STN	14	GPe	1.000	0.5
		GPi/SNr	1.000	0.5
GPi/SNr	27	GPi/SNr	-1.000	0.25

4.2 Go/No-Go Task Implementation

The Go/No-Go task is simulated by observing the effect of dopamine on BG pathways, and the simulation time is divided into three distinct phases. During the first phase of the simulation, the dopamine signal is suppressed below the baseline. This low dopamine level increases activity in D2 neurons, as dopamine typically causes inhibitory currents in these neurons. Consequently, the inhibition of dopamine releases the inhibition on D2 neurons while reducing the excitation in D1 neurons. In the second phase of the experiment, dopamine is stabilized at the baseline, indicating no stimuli or expectation of stimuli, during which we expect the BG to return to its resting conditions. The final phase of the experiment induces a Go signal by increasing dopamine levels above baseline, causing dynamics that reverse the effects observed during the initial No-Go phase.

We replicated the model on a CPU simulation using the Lava framework as the final test. We compared the performance of the two networks running on different hardware platforms (CPU and Loihi2) by analyzing the differences in firing rates. This comparison allowed us to assess the consistency and reliability of the model across different computational environments, ensuring the validity of our implementation.

4.3 Actor-Critic Model

The Actor-Critic model is a foundational framework in reinforcement learning, combining two distinct components: the actor, responsible for selecting actions, and the critic, which evaluates the actions by predicting rewards [96]. This section explores implementing an actor-critic model within the BG, leveraging dopaminergic signaling pathways to modulate learning and action selection. By integrating these components, the model aims to replicate the adaptive learning mechanisms observed in biological systems, providing a robust platform for studying complex neural dynamics and decision-making processes.

This section presents various models depicting the network at different levels of complexity and biological plausibility. Each model incorporates different inductive biases and introduces unique dynamics, resulting in varying degrees of task accuracy. By examining these dynamics and validating the results, we can discuss the merits and functions of each method. The models explored are as follows:

- *No-Channel, DA Sampled, Not Receptor Specific Actor-Critic Model:* This model considers the BG as an integrated network without distinct channels, based on the hypothesis that the topographically segregated channels of the BG can emerge naturally as the actor learns. Dopaminergic neuron firing rates of the critic are sampled to generate the reward signal, resulting in a binary reward signal where a thresholding mechanism rewards the actor and critic only for significant outcomes. Without a defined baseline for BG dopaminergic input, the D1 and D2 receptor-specific learning rule is simplified to a basic R-STDP rule.

- *Multi-Channel, Receptor Specific Actor-Critic Model:* This model adheres to empirical evidence about the BG, introducing segregated channels [77]. With a baseline for the DA signal, the model utilizes baseline-referenced D1/D2 receptor-specific and opposite learning rules, closely aligning with observed biological processes. Additionally, the task structure includes rest periods between trials, providing more apparent separation between trials.

These models facilitate a comprehensive examination of the BG's role in reinforcement learning, offering insights into how various configurations and learning rules influence the network's performance and adaptability.

4.3.1 Sequential Learning Task and Environment Dynamics

BG is responsible for learning valuable stimulus-action pairs, but these stimuli must originate from an external source, typically the cortex. The cortex processes external

stimuli and the internal model of the environment into a high-dimensional output vector, providing a rich, high-level abstraction representation of stimuli. However, in our BG setup, we simplified this process, as discussed in previous sections, using a direct stimulus approach akin to Izhikevich's classical conditioning experiment. Before delving into the specifics of stimuli, we should introduce the task and environment.

Our task involves a simplified version of sequential learning with three stimulus-action pairs inspired by [117]. The states/stimuli are [A, B, C], and the actions are [X, Y, Z]. The correct sequence for obtaining a primary reward is A-X => B-Y => C-Z. Any incorrect action selection resets the environment to the beginning of the sequence, providing zero reward. Learning begins from the end of the sequence with the C-Z association and progressively increases in complexity.

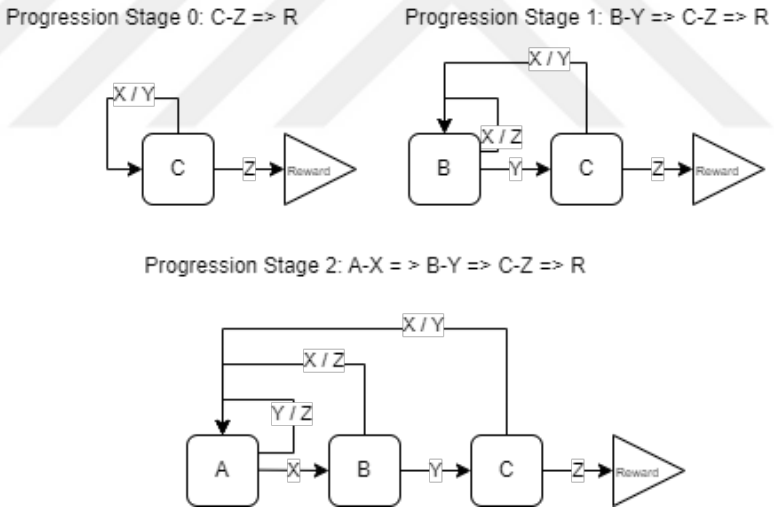


Figure 13: Sequence learning task represented as a state machine. The task is implemented in a backward fashion. Starting from Progression Stage 0, the stage increases at every threshold of correctly known sequences

In our model, these states are encoded as stimulus neuron groups that activate in a coordinated manner on the Cortex Neurons (simulated by PSG). The number of neurons in a stimulus group depends on the Cortex PSG used and the firing rates of the D1 and D2 groups.

Information quanta are believed to be processed every 200-300ms in the human brain, corresponding to the periodic nature of eye saccades. Similarly, our classical conditioning experiment used a stimulus interval between 100ms and 300ms, averaging 200ms. Our current task and environment adhere to this observation, with state-action pairs processed at 200ms intervals. Figure 13 illustrates the resulting task state machine.

The task structure remains similar across various models. The only difference is that in the most basic No-Channel Actor-Critic Model, there is no idle time to separate trials, which results in the D2 neuron of the actor influencing different trials, which has been later resolved with a rest time to separate trials.

4.4 No-Channel (DA Sampled, Not Receptor Specific) Actor-Critic Model

This model introduces the most simplified dynamics and does not adhere to various biological mechanisms. These are taken as a design choice and explained with their reasoning:

- In this model, the BG matrix is not segregated into channels. The rationale behind this design constraint is to observe if channel dynamics can naturally develop inside a holistic ganglia. Therefore, this model uses the actor introduced in the Go/No-Go task and separates the SNr output into three action selection bins.

- Although all critic mechanisms are the same in each model, the connectivity scheme between DA neuron outputs differs. This model generates the DA signal by an artificial non-neuron-like process that takes combined DA neuron firings and their moving averages. This average is thresholded to produce a binary DA signal with a refractory period, which induces a big dopamine spike.

4.4.1 Actor Model

This model's actor architecture is similar to the Go/No-Go BG model but slightly different. As in this holistic view, the SNr output is separated into decision bins, and a more robust decision-making scheme requires larger SNr populations. This implementation's new neuron parameters and connectivity scheme are given in Table 5 and Table 6. The architecture is illustrated in Figure 14.

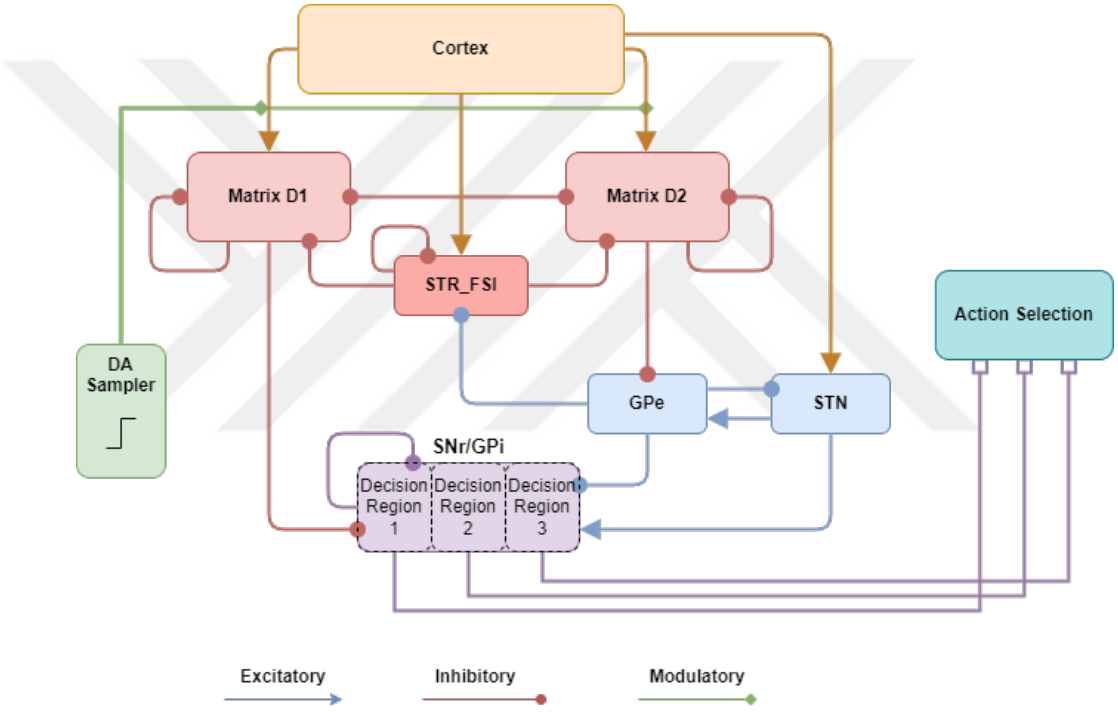


Figure 14: The holistic BG actor model with separated SNr decision regions. The dopaminergic input to the system comes from the DA sampler process.

Table 5: Parameters for Izhikevich Neurons in the Basal Ganglia Circuit for No-Channel Actor-Critic Model.

Group	a	b	c	d	I_{const}
strD1	0.02	0.2	-54	0.6	-8
strD2	0.02	0.2	-54	0.6	-7
strFSI	0.1	0.2	-65	8	-3.5
STN	0.005	0.265	-65	2	5
GPe	0.005	0.585	-65	4	7
SNr	0.005	0.32	-65	2	12

As can be seen, the major difference between this model and the Go/No-Go model is the MSNs of the striatum. These neurons have been modified to mimic their biological counterparts' up and down states and have been readjusted to be bursting neurons.

Table 6: Connectivity Parameters for connections in the Basal Ganglia Circuit for No-Channel Actor-Critic Model. All inhibitory and excitatory connection weights have been scaled by 10 and 20, respectively.

Presyn. Group	#Neurons	Postsyn. Group	Weight	Conn. Prob.	Variance
strD1	1255	strD2	-0.0982	0.1	0.2
strD1	1255	SNr	-0.075	0.15	0.2
strD2	1255	GPe	-0.1	0.15	0.2
strD1	1255	strD1	-0.0982	0.1	0.2
strD2	1255	strD1	-0.0982	0.1	0.2
strD2	1255	strD2	-0.0982	0.1	0.2
strFSI	84	strD1	-0.0982	0.1	0.2
strFSI	84	strD2	-0.0982	0.1	0.2
strFSI	84	strFSI	-0.0982	0.1	0.2
GPe	46	STN	-0.1429	0.25	0.2
GPe	46	SNr	-0.05	0.25	0.2
GPe	46	GPe	-0.1429	0.25	0.2
GPe	46	strFSI	-0.1429	0.05	0.2
SNr	243	SNr	-0.0111	0.25	0.2
STN	14	GPe	0.05	0.5	0.2
STN	14	SNr	0.05	0.5	0.2
Cortex	3000	strD1	0.1875	0.05	0.2
Cortex	3000	strD2	0.1875	0.05	0.2
Cortex	3000	strFSI	0.125	0.15	0.2
Cortex	2	STN	0.3	0.2	0.2

4.4.2 Critic Model

The adaptive critic in the BG operates within the striosome compartments, where D1 and D2 neurons receive stimuli from the Cortex. These neurons influence DA neurons in the SNc and VTA through two pathways. The side-loop pathway involving the STN provides immediate excitatory input to DA neurons. The direct pathway uses

inhibitory GABA connections from MSNs to DA neurons, causing prolonged inhibition. These pathways generate the Temporal Difference (TD) or Reward Prediction Error (RPE) signal, as demonstrated by [5].

This scheme is illustrated in Figure 15. The striosome neurons' output defines the value function in RL, with immediate excitation represented as $\gamma P(t)$ and prolonged inhibition as $-P(t-1)$. Together, these elements provide effective reinforcement, supported by primary reinforcement directly from the environment.

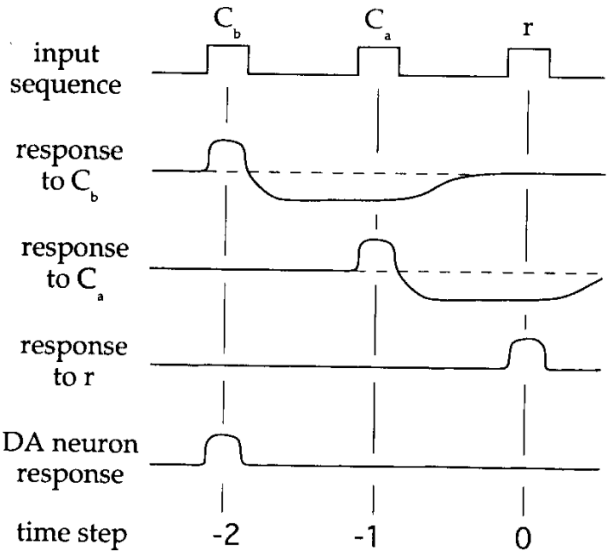


Figure 15: The critic model behavior and excitation/inhibition from the side-loop and direct pathway creating the critic dynamic [5].

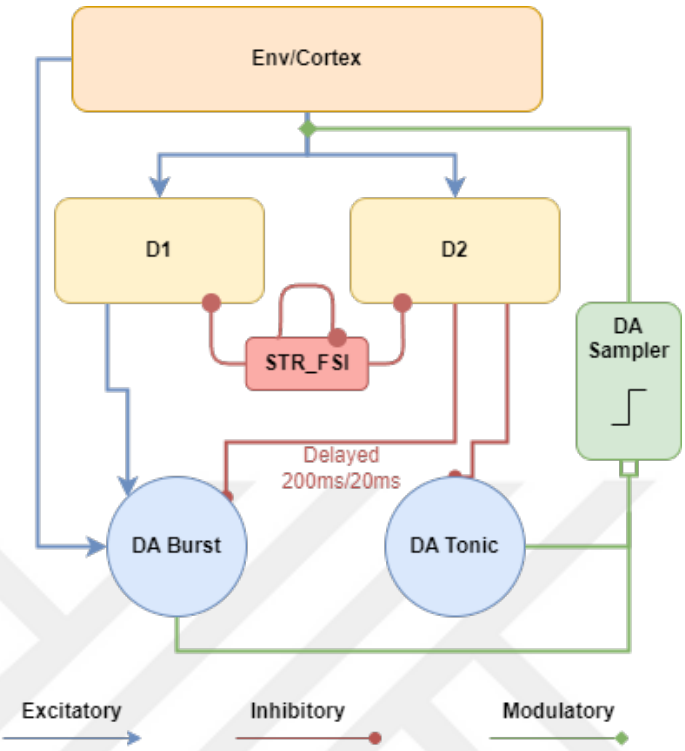


Figure 16: Critic model diagram showing neuron populations and connections. The inhibitory connections projecting from D2 neurons have $20ms$ and $200ms$ delays to inhibit primary reward, pre-inhibit excitation, and propagate earlier predictions, respectively. The environment and Cortex have been shown together, but the primary reward connection to burst neurons comes from the environment in reality.

The architecture involves two groups of dopamine neurons: tonic (baseline) and phasic (bursting), D1/D2 striosome MSNs, and cortical input for stimuli. The neuron groups and their parameters are given in Table 7.

Table 7: Parameters for Izhikevich neurons for critic component in No-Channel Actor-Critic Model.

Group	a	b	c	d	I_{const}
STR_D1	0.02	0.2	-54	0.6	-1.5
STR_D2	0.02	0.2	-54	0.6	-1
STR_FSI	0.1	0.2	-65	8.0	-3.5
DA_tonic	0.002	0.2	-65	2	0
DA_burst	0.02	0.2	-54	0.6	-4

In this model, the side-loop is simplified to be directly excitatory with a single connection, excluding the STN, as the excitation mechanism through the side-loop is unclear in [5]. Delayed sparse connections with $200ms$ and $20ms$ delays propagate earlier prediction values over inhibitory synapses, mimicking prolonged inhibition. The connection parameters of this model are provided in Table 8.

Table 8: Connectivity parameters for the critic in the No-Channel Actor-Critic Model. Inhibitory and excitatory weights are scaled by 10 and 20, respectively, with 0.2 variance for all except the primary reward to DA_burst group.

Presyn. Group	#Neurons	Postsyn. Group	Weight	Conn. Prob.	Delay (ms)
Ctx_D1D2	2000	STR_D1	0.04	0.1	0
Ctx_D1D2	2000	STR_D2	0.04	0.1	0
STR_FSI	100	STR_D1	-0.0982	0.1	0
STR_FSI	100	STR_D2	-0.0982	0.1	0
STR_FSI	100	STR_FSI	-0.5	0.1	0
STR_D1	1000	DA_burst	0.25	0.1	0
STR_D2	1000	DA_tonic	-1	0.1	200/20
STR_D2	1000	DA_burst	-1	0.1	200/20
Ctx_DA	2000	DA_tonic	0.05	0.1	0
Primary_Reward	1	DA_burst	20 (not scaled)	1.0	0

The D1 neurons are now considered part of the side-loop and provide excitatory input to phasic DA neurons, while D2 neurons inhibit this group with a delay. The tonic groups can only be suppressed with the D2 neurons and cannot be excited. The primary reinforcement is connected to phasic DA neurons via an excitatory connection. The combined spiking of burst and tonic DA neurons forms the dopamine signal. This architecture is modeled in Figure 16.

4.4.3 Actor-Critic Model

The actor and critic are combined using the modulatory connections from the DA sampler to create feedback using effective reward on the cortico-striatal connections of both critic and actor. The final architecture is illustrated in Figure 17.

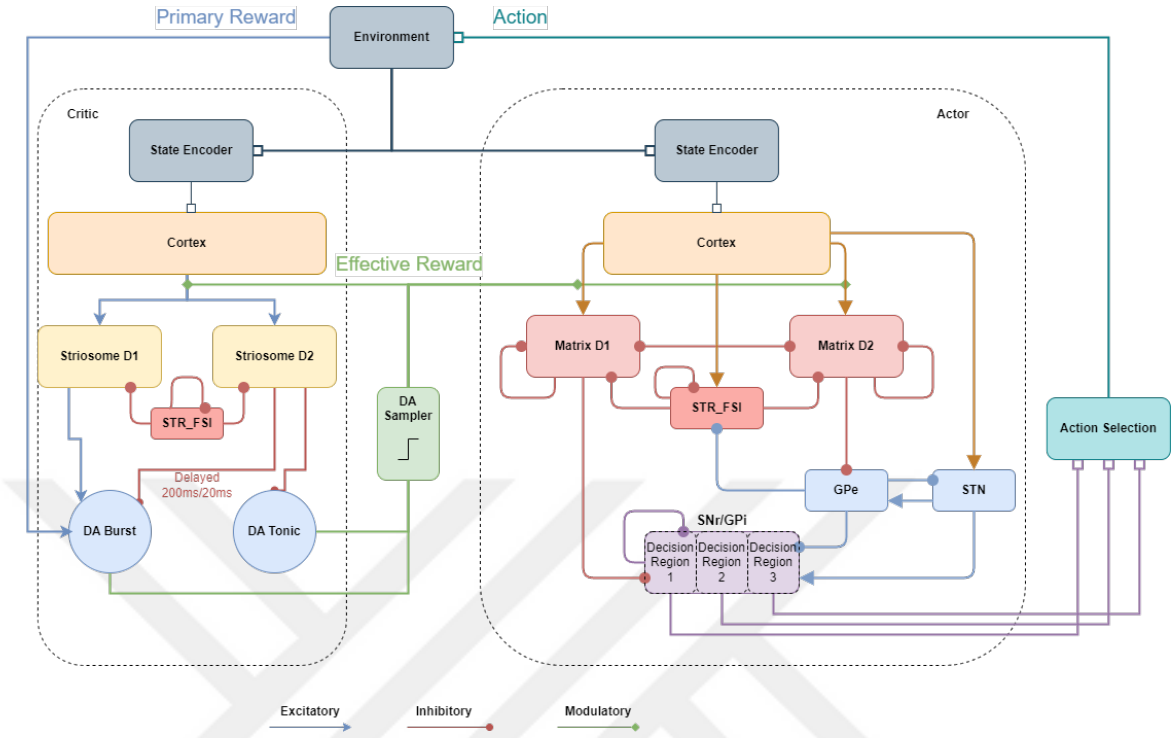


Figure 17: Combined Actor-Critic model for No-Channel Actor-Critic BG Model. The square connections define out of network connections.

The network has run on the mentioned task with specific parameters in Table 9. It is important to note that the trial depicts a single step of environment and action selection. This run setup also does not include trial separation.

Table 9: Network parameters used for the No-Channel Actor-Critic BG Model experiment.

Parameter	Value	Description
num_trials	100	Number of steps (action selection) for environment
interval	200 ms	Interval between steps
idle_time	200 ms	Idle time at the beginning of the experiment
dt	1 ms	Time step
vrblt	0.1	Stochasticity parameter
inh_w_scale	10	Inhibitory weight scaling factor for all inhibitory connections
ex_w_scale	20	Excitatory weight scaling factor for all excitatory connections
num_states	3	Number of states
num_actions	3	Number of actions
neurons_per_group	25	Stimulus neurons per stimuli group
total_neurons_critic	2000	Total neurons for the critic cortex
total_neurons_actor	3000	Total neurons for the actor cortex
progression_threshold	10	Progression threshold for the environment
Ctx_frequency_D1D2	1	Cortex frequency for D1/D2 neurons in critic
Ctx_frequency_DA	1	Cortex frequency for DA neurons in critic
on_hit_DA	0.5	Dopamine strength per sampled DA effective reward
observation_length	20 ms	Observation period
Reward_Delay	20 ms	Primary reward delay for successful trial after the stimuli
window_size	10 ms	Window size for DA sampler
threshold	2.75 Hz	Threshold for DA sampler
refractory_period	100 ms	Refractory period for DA sampler

4.5 Multi-Channel (Receptor Specific) Actor-Critic Model

This model introduces topographical projections and the resulting segregated channels in the BG. The dopaminergic connections are directly projected onto plastic connections. Therefore, fine-grained DA concentrations are possible. Baseline dopamine can be provided to implement baseline-referenced D1/D2 receptor-specific learning

rules. The model also features the separation of trials with an idle time.

4.5.1 Actor Model

The BG, organized hierarchically and somatotopically, select actions by processing topographically localized inputs from the cortex in the dorsal striatum. These inputs form channels representing possible actions. The channels compete, with the most salient action facilitated and others inhibited, ensuring efficient decision-making. This system allows compatible actions to be executed simultaneously while preventing non-compatible actions and maintaining precise and adaptive motor control.

In this model, the transition is made from a single Go/No-Go BG to a multi-action BG by utilizing a channel structure. Three No-Go circuits connected in parallel, with cross-channel STN-SNr connections, are implemented. These circuits receive inputs from shared nuclei, representing extrinsic stimuli from the environment as encoded by the cortex. Following stimulus introduction, bursts of activity in D1 and D2 neurons lead to changes in the firing rates of the output nuclei. The decision is made by analyzing the firings in the decision bins $20ms$ after stimulus introduction, with the channel showing the least activity selected as the action chosen by the BG.

Table 10: Parameters for Izhikevich neuron models in a multi-action Basal Ganglia circuit channel.

Group	a	b	c	d	I_{const}
STR_D1	0.02	0.2	-54	0.6	-8
STR_D2	0.02	0.2	-54	0.6	-7
STR_FSI	0.1	0.2	-65	8	-3.5
GPe	0.005	0.585	-65	4	7
STN	0.005	0.265	-65	2	5
GPi/SNr	0.005	0.32	-65	2	12

D1 and D2 neurons are modeled as bursting Izhikevich neurons to achieve the

required information processing, with parameters defined in Table 10. Lateral inhibition between channels is established by connecting the STNs of each channel to all other SNr nuclei and GPe. A high-saliency input excites D2 neurons, which then release the inhibition of GPe on the STN. This increases STN excitation, increasing activity in all competing channels, and creating a surround excitation effect.

The connectivity parameters of the multi-action BG circuit are outlined in Table 11. Connections from the Cortex (channel) are replicated for each channel. Figure 18 illustrates the final multi-channel BG model.

Table 11: Connectivity Parameters of the multi-action BG Circuit.

Presyn. Group	#Neuron	Postsyn. Group	Weight	Conn. Prob.	Variance
STR_D1	1255	STR_D2	-0.0982	0.1	0.2
STR_D1	1255	GPi/SNr	-0.3	0.15	0.2
STR_D1	1255	STR_D1	-0.0982	0.1	0.2
STR_D2	1255	STR_D1	-0.0982	0.1	0.2
STR_D2	1255	STR_D2	-0.0982	0.1	0.2
STR_D2	1255	GPe	-0.2	0.15	0.2
STR_FSI	84	STR_D1	-0.0982	0.1	0.2
STR_FSI	84	STR_D2	-0.0982	0.1	0.2
STR_FSI	84	STR_FSI	-0.0982	0.1	0.2
GPe	46	STR_FSI	-0.1429	0.05	0.2
GPe	46	STN	-0.1429	0.25	0.2
GPe	46	GPi/SNr	-0.1429	0.25	0.2
GPe	46	GPe	-0.1429	0.25	0.2
STN	14	GPe	0.05	0.5	0.2
STN	14	GPi/SNr	0.05	0.5	0.2
GPi/SNr	27	GPi/SNr	-0.1429	0.25	0.2
Cortex (Channel)	3000	STR_D1	0.25	0.05	0.2
Cortex (Channel)	3000	STR_D2	0.25	0.05	0.2
Cortex (Channel)	3000	STR_FSI	0.125	0.15	0.2
Cortex (Channel)	2	STN	0.3	0.2	0.2
STN (Cross-channel)	14	GPi/SNr	0.05	0.5	0.1
STN (Cross-channel)	14	GPe	0.05	0.5	0.1

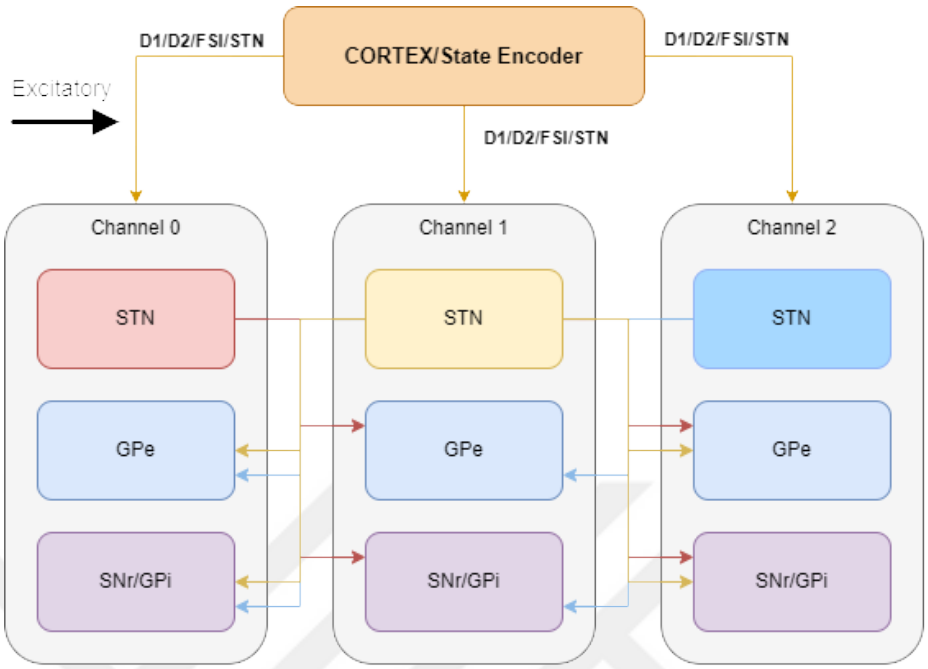


Figure 18: Multi-Channel BG actor network. Extrinsic stimuli encoded by the Cortex/State Encoder project to D1/D2/FSI/STN of every channel, while STNs of each channel project to other GPe and SNr nuclei.

A backward learning scheme [117], is employed, defining the dopamine signal as the primary reinforcing reward from the environment, akin to the saliency dopamine hypothesis [81]. This network uses R-STDP Izhikevich neurons with sparse plastic connections between the Cortex and D1/D2 MSNs. The state-action loop runs continuously, injecting the current state of the environment through stimuli groups in the cortex every $200ms$, except for the initial interval to account for transient effects. After $20ms$ of activity, the decision groups in SNr/GPi are observed, and the least active group is selected as the chosen action. When sufficient correct selections are made, the following sequence element in the backward chain is introduced until the entire sequence is learned, monitored by observing the environment's reward signals.

The sequence stimuli are inputted in order to verify the model's action selection capability. Connection weights from the stimuli group to D1 and D2 neurons for the

correct stimuli-action channel are set to three times the average of the other connections, simulating learned weights in action selection. Observing the SNr activity $20ms$ after the stimuli input determines the action.

4.5.2 Critic Model

The critic component of this architecture is similar to earlier model with slight differences. This model introduces baseline-referenced learning rules. The critic striosome neurons are also obliged to D1 neuron rules as the network needs to lower the prediction of a reward in the case of omission of the reward.

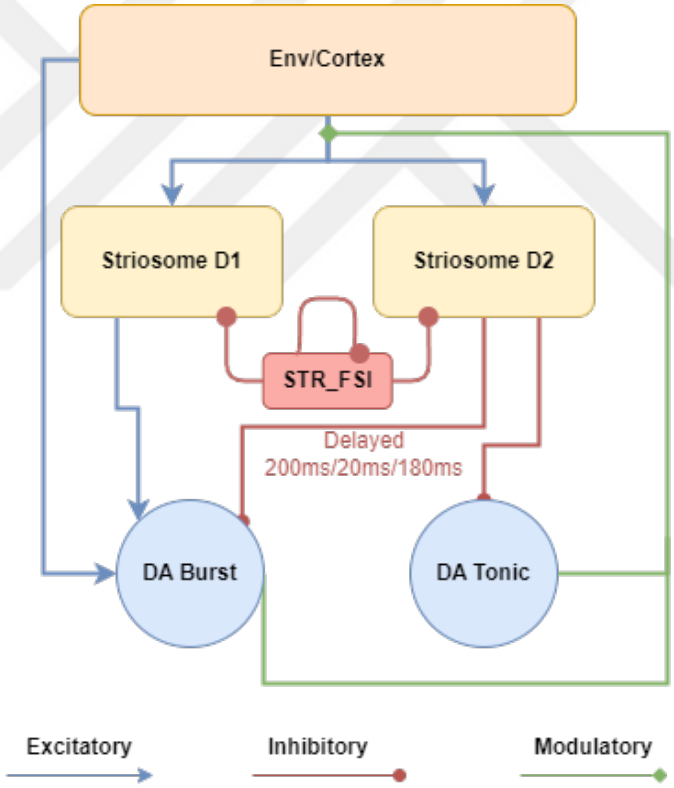


Figure 19: Critic model diagram for Multi-Channel Actor-Critic Model.

The architecture for this network is illustrated in Figure 19. The connection parameters are given in Table 12. The neuron parameters are the same as the previous model; therefore, they are not mentioned.

Table 12: Connectivity Parameters for Connections in the Critic Circuit for Multi-Channel Actor-Critic Model.

Presyn. Group	#Neurons	Postsyn. Group	Weight	Conn. Prob.	Delay (ms)	Variance
strFSI	100	strD1	-0.0982	0.1	0	0.2
strFSI	100	strD2	-0.0982	0.1	0	0.2
strFSI	100	strFSI	-0.5	0.1	0	0.2
Ctx_D1D2	2000	strD1	0.5	0.1	0	0.2
Ctx_D1D2	2000	strD2	0.5	0.1	0	0.2
strD1	1000	DA_burst	0.5	0.1	0	0.2
strD2	1000	DA_tonic	-1	0.1	200/180/20	0.2
strD2	1000	DA_burst	-1	0.1	200/180/20	0.2
Ctx_DA	2000	DA_tonic	1	0.1	0	0.2

The network is tested on four setups to showcase the reward expectation shifting to earlier reward-predicting cues. The full sequence of stimuli is introduced to the network in $200ms$ intervals with $200ms$ idle time at the beginning of the experiment. In the first setup, the stimuli connections between the cortex and D1/D2 neurons are not strengthened, resulting in a strong reaction to the primary reward by dopamine neurons. To simulate the learning of the reward-predicting cue, the stimuli neuron connections to striatum neurons are sequentially strengthened from the backward stimuli first. This setup allows observing the critic dynamics, bringing the reward expectation forward as the critic learns the reward-bringing stimuli.

4.5.3 Actor-Critic Model

The final model combines the actor and critic models with the dopaminergic connections. The effective reward calculated by dopamine neurons modulates both Cortico-Striatal connections in the actor channels and the critic itself. Figure 20 illustrates the final model architecture.

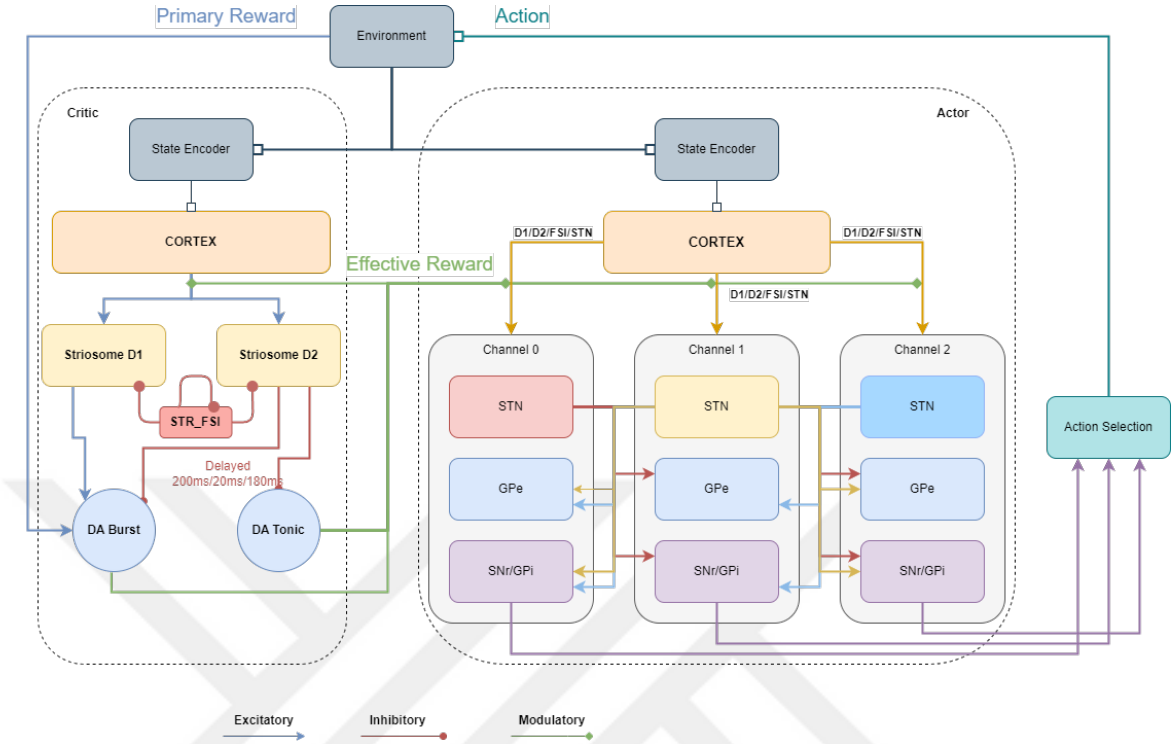


Figure 20: The overall architecture BG model including Environment, Multi-Channel Actor and Critic. The state information and selected action are not spike encoded and are, therefore, depicted with open square connections.

The DA synapses provide single excitation on the DA concentration per DA spike, which is modeled as on_hit_DA variable, showcases dopamine's strength, and is similar to the learning rate. The learning rule also has been adapted to include a baseline and D1/D2 receptor dynamics and is provided in Listing 3 in appendix. Table 13 states the learning rule parameters for the learning connections.

Table 13: Learning rule parameters for Multi-Channel Actor-Critic BG circuit. Max weights consider the weights after scaling. All learning rules use learning rate, A_+, A_- variables as 1,1 and -1 respectively.

Rule	Connection	Max Weights	Base	Pre/Post τ	Elig τ
RSTDP_D1	(Actor) Ctx to D1	2^4	2^{-6}	20	0.0078 (140ms)
RSTDP_D1	(Actor) Ctx to D2	2^4	2^{-6}	20	0.0078 (140ms)
RSTDP_D2	(Actor) Ctx to D2	2^4	2^{-6}	20	0.0078 (140ms)
RSTDP_Basic	(Critic) Ctx to D1	2^4	2^{-6}	20	0.0078 (140ms)
RSTDP_Basic	(Critic) Ctx to D2	2^4	2^{-6}	20	0.0078 (140ms)

The critic network employs the RSTDP_D1 learning rule for D1 and D2 neurons. This approach is essential because these neurons directly encode the value function of the stimuli, thus necessitating the reinforcement of correct connections. The RSTDP_D2 learning rule would be more appropriate in scenarios where network connectivity utilizes disinhibition. However, this network's current side-loop and prolonged inhibition mechanisms do not follow disinhibition and deviate from strict biological realism. This design choice allows for a more straightforward implementation while capturing the essential computational principles required for reinforcement learning in the critic network. The network has been run with setup parameters stated in Table 14.

Table 14: Network parameters used for the Actor-Critic BG Circuit experiment.

Parameter	Value	Description
num_trials	50	Number of steps (action selection) for environment
interval	200 ms	Interval between steps
idle_time	200 ms	Idle time at the beginning of the experiment and in between sequences
dt	1 ms	Time step
vrblt	0.1	Stochasticity parameter
inh_w_scale	10	Inhibitory weight scaling factor for all inhibitory connections
ex_w_scale	20	Excitatory weight scaling factor for all excitatory connections
num_states	3	Number of states
num_actions	3	Number of actions
neurons_per_group	100	Stimulus Neurons per stimuli group
total_neurons_critic	2000	Total neurons for the critic cortex
total_neurons_actor	3000	Total neurons for the actor cortex
progression_threshold	5	Progression threshold for the environment
Ctx_frequency_D1D2	0.1	Cortex frequency for D1/D2 neurons in critic
Ctx_frequency_DA	1	Cortex frequency for DA neurons in critic
on_hit_DA	0.01	Dopamine strength per spike for the actor (halved for the critic)
observation_length	20 ms	Observation period
Ctx_frequency	1	Cortex frequency for D1/D2 neurons in actor
Ctx_frequency_STN	1	Cortex frequency for STN neurons in actor
Reward_Delay	20 ms	Primary reward delay for successful trial after the stimuli

4.6 Putting It Altogether

This chapter has outlined the comprehensive process of developing a bio-realistic BG model, showcasing its application in reinforcement learning and action selection tasks. The stepwise approach, beginning with the fundamental Go/No-Go task and advancing to complex sequence learning tasks, provides a robust framework for understanding and replicating the neural mechanisms underlying BG functions.

The initial focus on the Go/No-Go task allowed for validation of the BG's capability in action gating, a critical function in neural processing. This foundation paved

the way for the integration of sequence learning tasks, where the model demonstrated its ability to handle dynamic environments and complex stimuli-action pairs. Organized channels within the BG facilitated sophisticated action selection processes, emphasizing the model's scalability and adaptability.

Incorporating a dopaminergic critic provided the necessary reward prediction error signals, enabling the model to learn from environmental feedback effectively. The detailed connectivity parameters and neuron model specifications ensured that the simulations closely mirrored biological processes, enhancing the model's fidelity and reliability.

Integrating actor and critic components into a unified BG model exemplifies the potential of neuromorphic computing platforms, such as Intel's Loihi2, to simulate complex neural circuits. Additionally, the comparison with CPU simulations validated the consistency and robustness of the model across different computational environments.

This chapter's findings contribute significantly to the field of neuromorphic computing, demonstrating the potential for creating bio-realistic neural networks that can perform intricate learning and decision-making tasks. The methodologies and insights presented herein pave the way for future advancements in understanding and replicating the neural basis of reinforcement learning, ultimately bridging the gap between biological and artificial neural systems.

CHAPTER V

EVALUATION AND RESULTS

This chapter evaluates the learning neuron models and BG circuits. We begin by showcasing the performance of the learning networks in a classical conditioning experiment. This is followed by an evaluation of the action selection process and the reward expectation shift to validate the model's actor and critic components. Finally, we present the results of different actor-critic architectures in a sequence learning task, demonstrating the integrated performance of the entire system.

5.1 Evaluation of Learning Neuron Models

This section provides results of separate Izhikevich neuron models that employ Spike-Timing-Dependent Plasticity (STDP), and Reward-Based Spike-Timing-Dependent Plasticity (R-STDP), respectively, demonstrating weight update of synapses between pre-/post-synaptic neurons.

5.1.1 STDP Implementation

The results of implementing the Izhikevich neuron model employing STDP on the Loihi2 platform are presented in this section. Figures 21, 22, 23, and 24 illustrate the spike timings, trace calculations, weight adjustments, and the overall STDP graph, respectively.

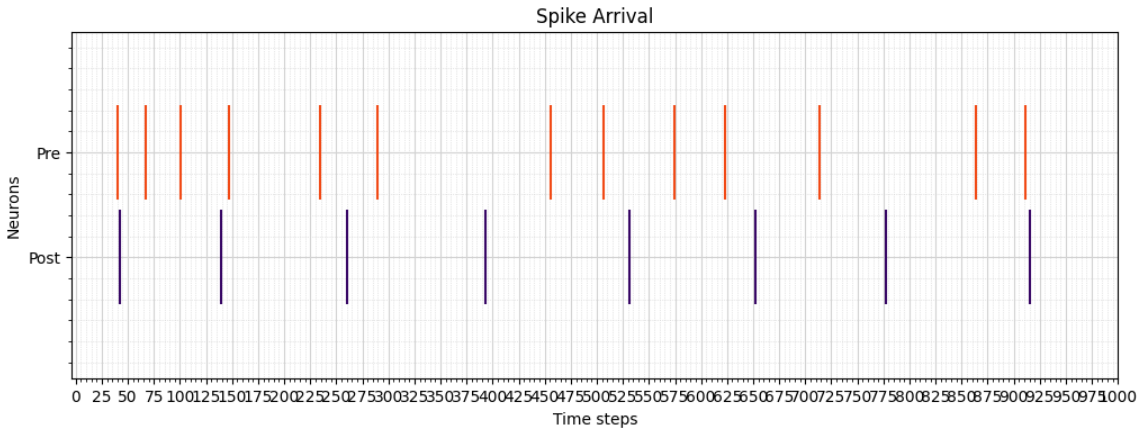


Figure 21: Pre/Post spike timings recorded during STDP implementation on Loihi2.

Figure 21 shows the pre- and post-synaptic spike timings recorded during the STDP process. These spike timings are crucial for calculating STDP traces, directly influencing synaptic weight adjustments. The connection process obtains pre- spike timings while post- spike timings are set during micro-code with $tspike$ variable. Spike timings create dependencies $x0$, $y0$ defined in the learning rule in Listings 2 and 3 that are provided in appendix.

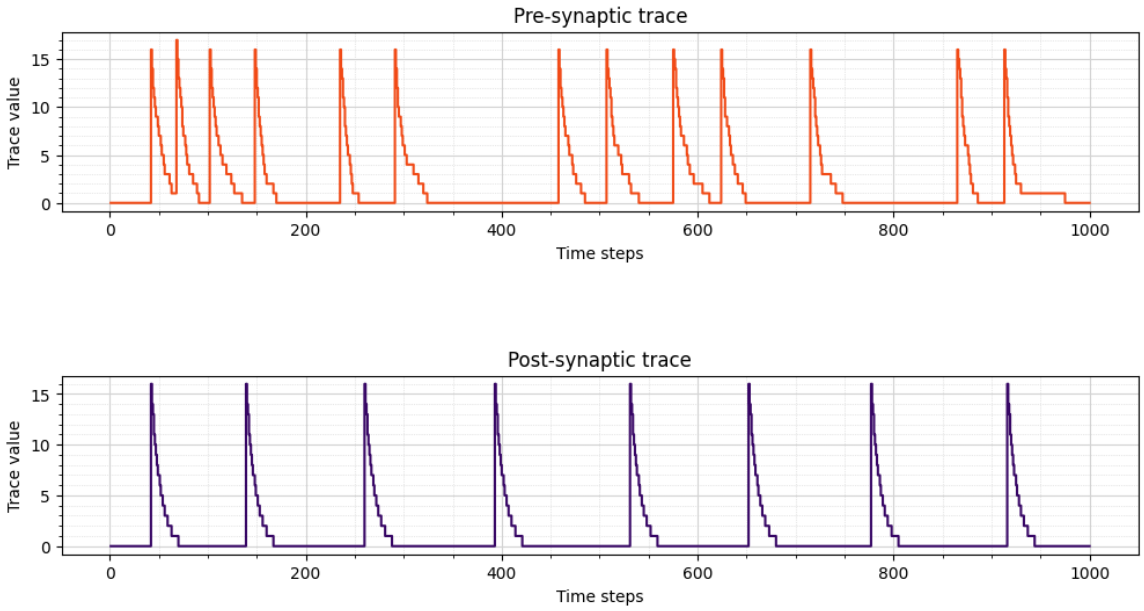


Figure 22: Trace calculations during STDP implementation on Loihi2 with a kernel magnitude of 16 per spike and time constant of $10ms$.

Figure 22 illustrate the pre-synaptic traces (top) and post-synaptic traces (bottom). These traces are calculated based on the impulse magnitude and the trace time constant defined in the learning rule. On the Loihi2 platform, the learning engine uses these calculated values directly. However, due to the fixed-point representation and the integer-based trace values, the precision of the traces is somewhat limited. This limitation makes it challenging to manipulate the trace values with high precision, affecting the granularity of synaptic weight changes. In CPU implementation, stochastic rounding is applied after the decay operation. At the same time, high precision multiplication followed by precision reduction was used to create the same effect on the Loihi2 micro-code.

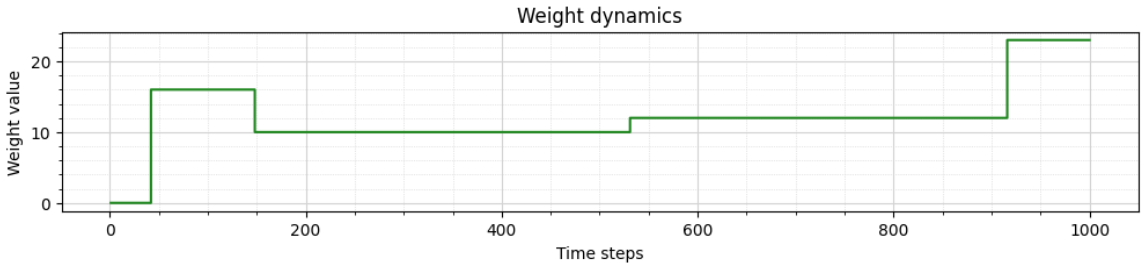


Figure 23: Weight changes according to STDP rule on Loihi2.

Figure 23 shows the changes in synaptic weights resulting from the STDP rule. The weight changes are computed by multiplying the pre- and post-synaptic traces at the dependencies on the learning engine (pre/post spike times). The resulting weight changes adhere to the STDP rule, demonstrating that the synaptic weights between neurons adjust according to the relative timing of spikes.

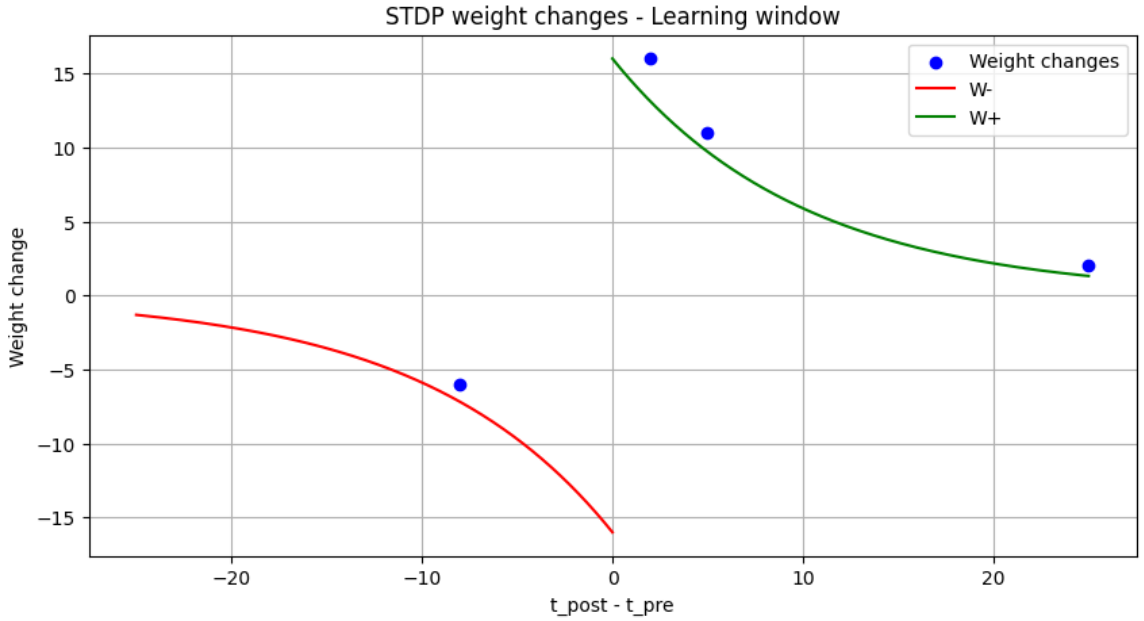


Figure 24: STDP graph illustrating coincident firings and weight changes according to relative timing between pre- and post- spikes.

Figure 24 illustrates the overall STDP graph, showing weight changes based on the pre- and post-synaptic spike timings. Deviations from precise values stem from

Loihi2's use of discrete time steps and fixed-point representation. Despite these deviations, the STDP rule effectively adjusts synaptic weights, demonstrating Loihi2's capability to perform learning tasks that mimic biological synaptic plasticity.

The results indicate that while fixed-point representation and discrete computation steps on Loihi2 introduce some limitations, the platform still successfully implements complex learning rules, such as STDP. The traces and weight adjustments align well with expected outcomes, showcasing neuromorphic hardware's potential to simulate neural plasticity effectively.

5.1.2 R-STDP Implementation

This section presents the results of implementing the Izhikevich neuron model employing R-STDP on the Loihi2 platform. The Figures 25, 26, 27, 28, and 29 illustrate the impact of reward signals on learning (focusing on the spike timings), trace calculations, eligibility traces, reward traces, and synaptic weight adjustments, respectively.

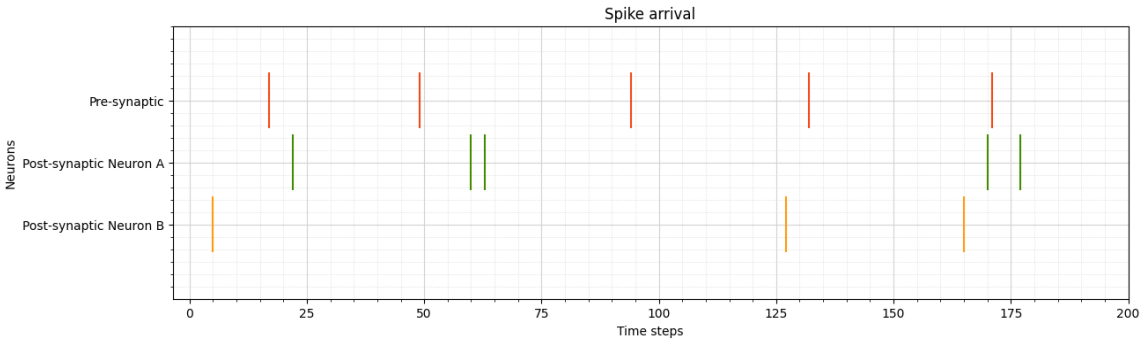


Figure 25: Pre/Post spike timings recorded during R-STDP implementation. Neuron A is forced to fire after the pre-synaptic neuron, and neuron B is to fire before, demonstrating potentiation and depression, respectively.

Figure 25 shows the pre- and post-synaptic spike timings recorded during the R-STDP process. These timings are critical for the subsequent calculation of eligibility traces and the influence of reward signals on synaptic adjustments. Similar to STDP implementation, the spike timings set dependency variables.

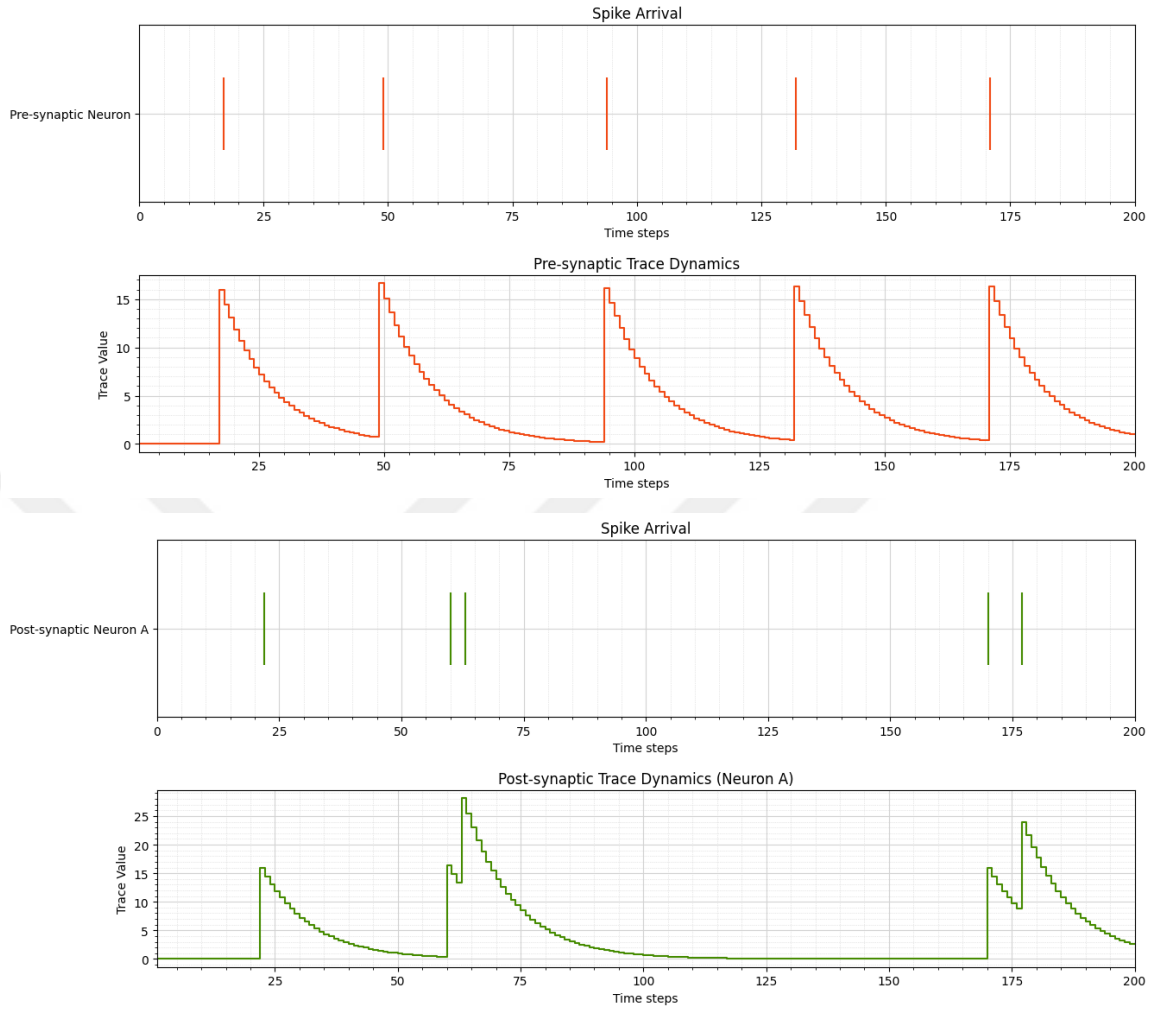


Figure 26: Trace calculations during R-STDP implementation with $10ms$ time constant and 16 kernel magnitude per spike.

Figure 26 show the pre-synaptic trace (second from the top) and post-synaptic trace (bottom), respectively. These traces are calculated based on the pre-synaptic spike timing (top) and post-synaptic spike timing (second from the bottom), and they play a crucial role in determining the eligibility trace, which influences the weight adjustments in R-STDP.

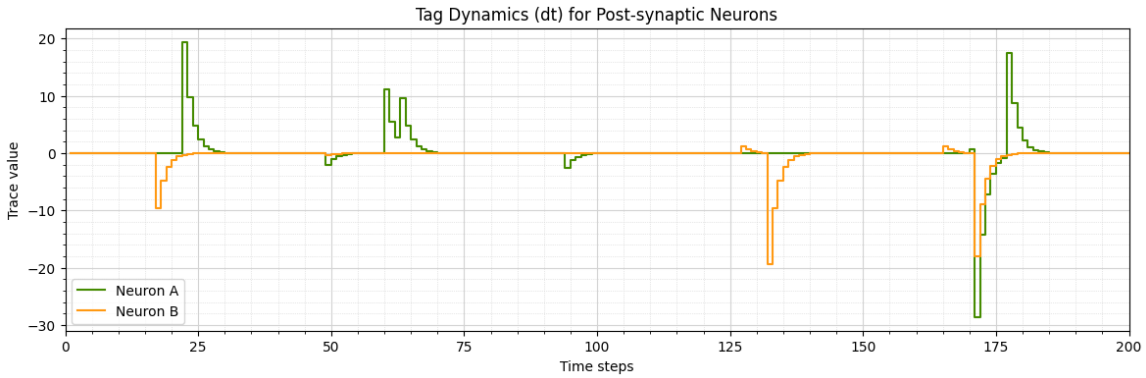


Figure 27: Eligibility trace calculated during R-STDP implementation.

Figure 27 illustrates the eligibility trace calculated from the pre- and post-synaptic firings. The eligibility trace represents the potential for synaptic weight change, influenced by the timing of spikes and the reward signal.

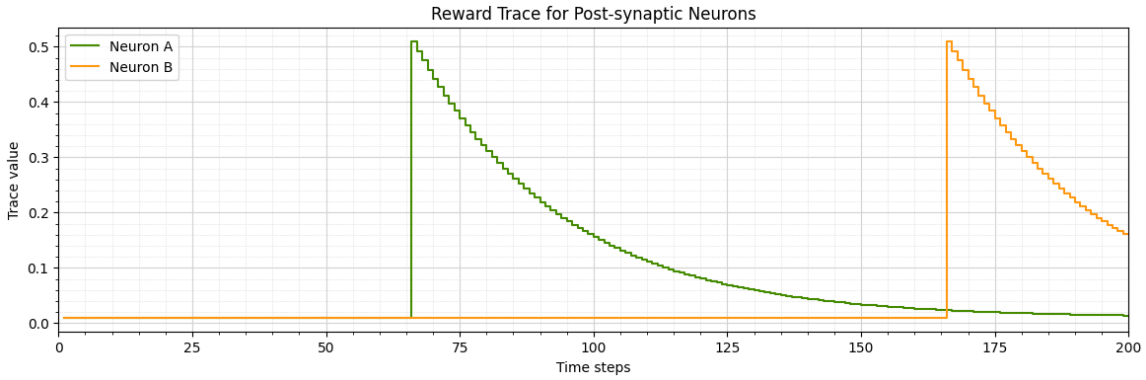


Figure 28: Reward trace calculated during R-STDP implementation with a time-constant of $28ms$ and 0.5 on_hit_DA level (refer to Listing 1 for definition).

Figure 28 shows the reward trace, which includes a baseline and exponential decay. This trace is crucial for modulating the synaptic weight changes based on the reward signal received, adding a third factor to the STDP learning rule (Listing 2).

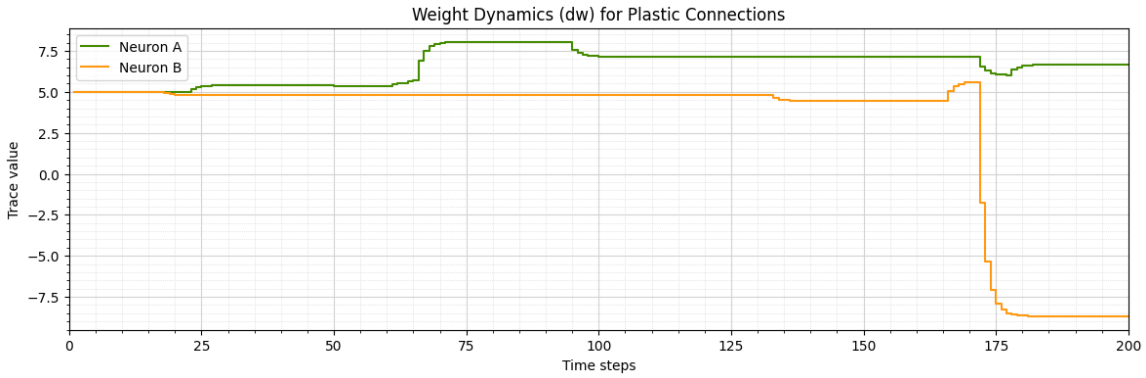


Figure 29: Synaptic weight changes observed during R-STDP implementation.

Figure 29 presents the changes in synaptic weights resulting from the R-STDP learning rule. The eligibility and reward traces influence the weight changes, showcasing how the reward signals modulate the learning process.

The R-STDP implementation on Loihi2 demonstrates how reward signals can enhance learning dynamics by adding a third factor to the traditional STDP rule. The pre- and post-synaptic traces, eligibility traces, and reward traces collectively influence the synaptic weight adjustments, allowing the network to learn from the timing of spikes and the reward signals.

These results highlight the capability of the Loihi2 platform to implement complex learning rules that mimic biological learning processes. Integrating reward signals into the STDP framework provides selective learning, which is not within the capabilities of 2-factor learning rules. The selectivity signal makes implementing reinforcement learning possible in Spiking Neural Networks.

5.2 Validation of Learning Neuron Models at Network Scale

In the following, the learning neuron models were validated at network scale through classical conditioning experiment. Then, the result of a 4-factor learning rule and dopamine dependency have been given.

5.2.1 Classical Conditioning Experiment

This experiment validates the 3F learning rule at network scale task of selectively reinforcing a single stimulus from a group of stimuli [109].

Figure 30 shows the average weight changes in response to stimuli. The connections corresponding to the target reward-bringing stimuli have been strengthened; however, the distinctions between the target and other stimuli are unclear. This behavior is likely due to the absence of a fourth factor, which is crucial for effective separation in the temporal domain. The subsequent section demonstrates the results of an enhanced learning method incorporating this fourth factor, illustrating its impact on stimulus separation and synaptic weight differentiation.

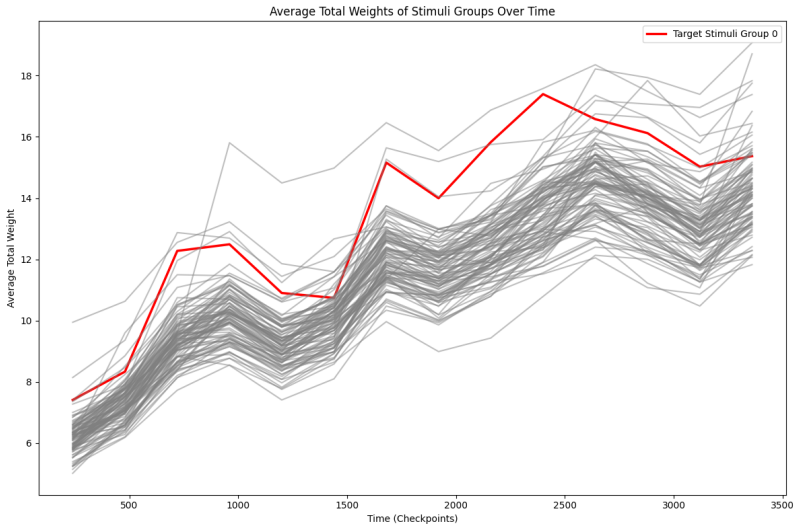


Figure 30: Average weight changes in the classical conditioning experiment.

5.2.2 Dopamine Dependency Experiment

The influence of dopamine receptivity on synaptic weight changes is demonstrated in Figures 31 and 32. The results indicate that the presence of dopamine receptivity significantly affects the modulation of synaptic weights. Specifically, dopamine receptivity is a gating mechanism for the 3rd factor, facilitating more sophisticated learning behaviors, including temporal dynamics where only specific stimuli benefit

from the reward signal. This ability can be further utilized in temporal tasks.

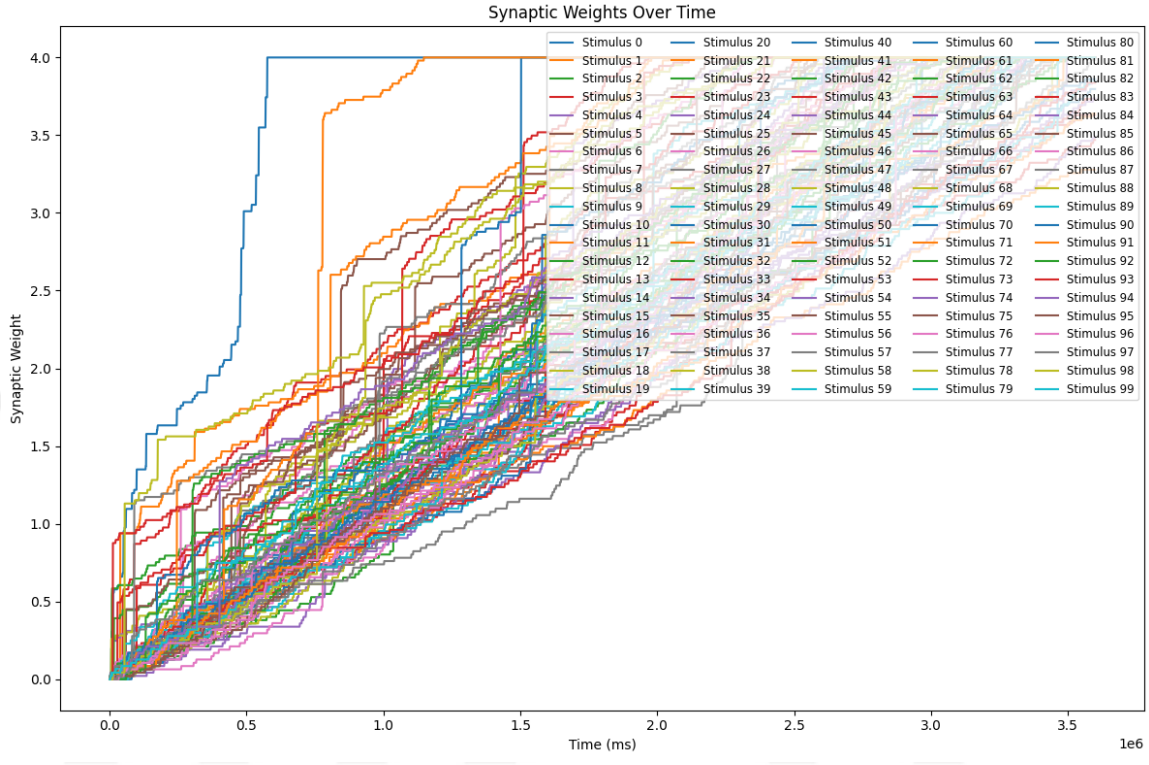


Figure 31: Weight changes without dopamine receptivity.

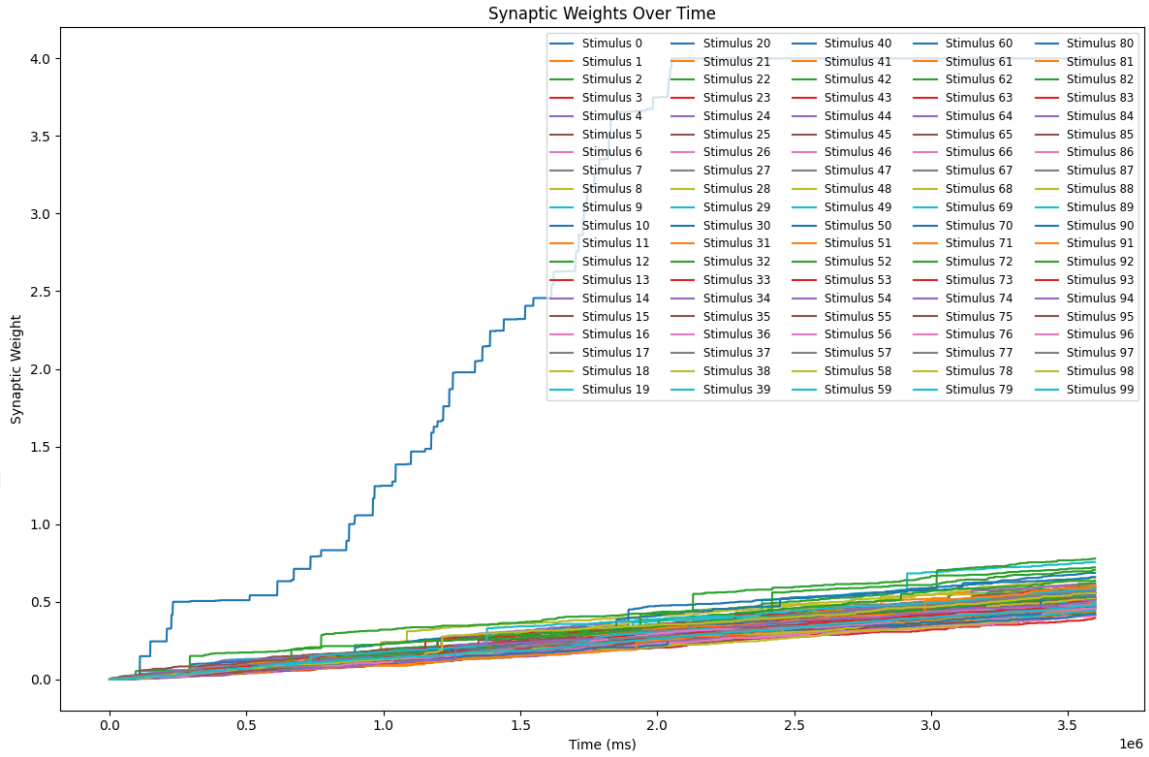
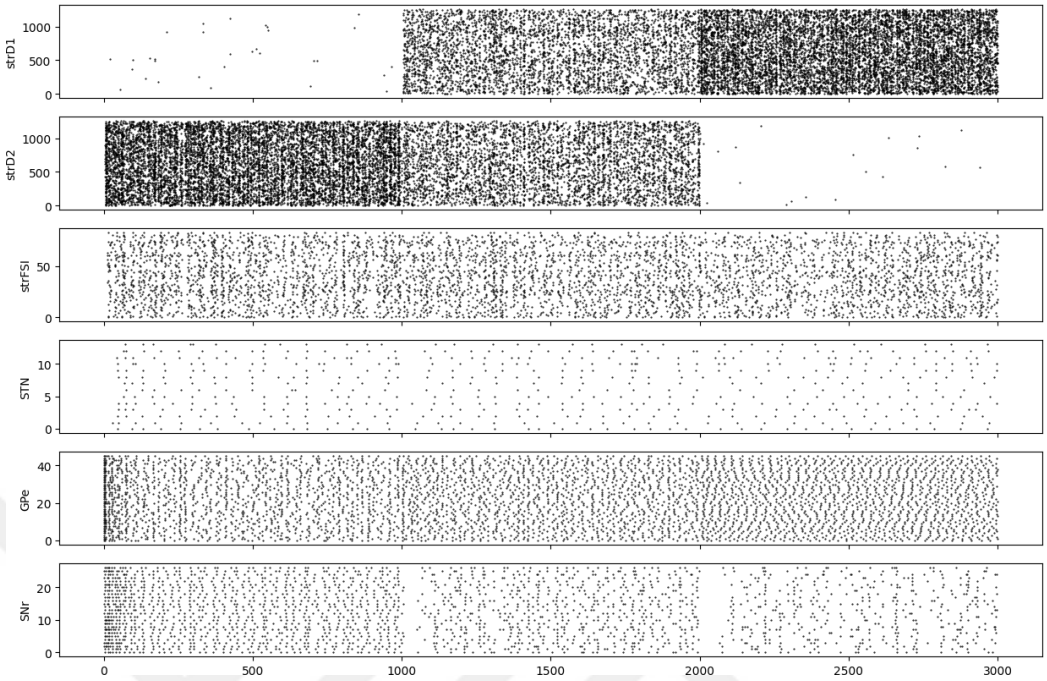


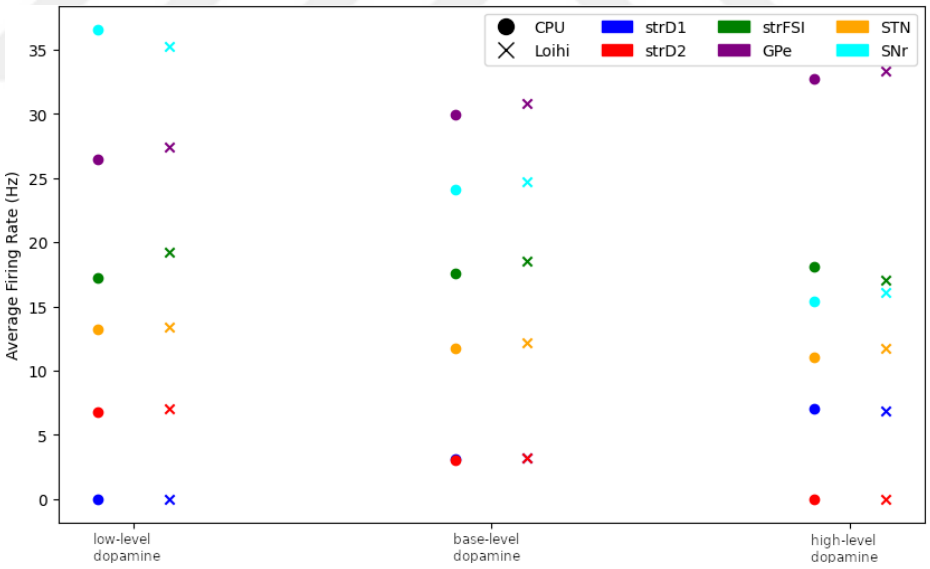
Figure 32: Weight changes with dopamine receptivity.

5.3 Evaluation of Basal Ganglia Circuit for Go/No-Go Task

Behavioral dynamics of the BG model were observed under different dopamine levels: below baseline (0-1000 ms), baseline (1000-2000 ms), and above baseline (2000-3000 ms). Figures 33 illustrate the simulation results.



(a)



(b)

Figure 33: Raster plots that depict the activity of neuron groups in the striatum (MSN D1 and D2, str_FSI), GPe, STN, and GPi/SNr) on Loihi2 under three conditions: low, base, and high dopamine levels (a). The average firing rates of neuron groups on Loihi2 and CPU simulation follow each other for different dopamine levels (b) [4].

5.4 Evaluation of No-Channel Basal Ganglia Model

The combined Actor-Critic model has been run for 100 steps. The combined DA moving average inputted to the DA sampler and the effective result sampled is shown in Figure 34.

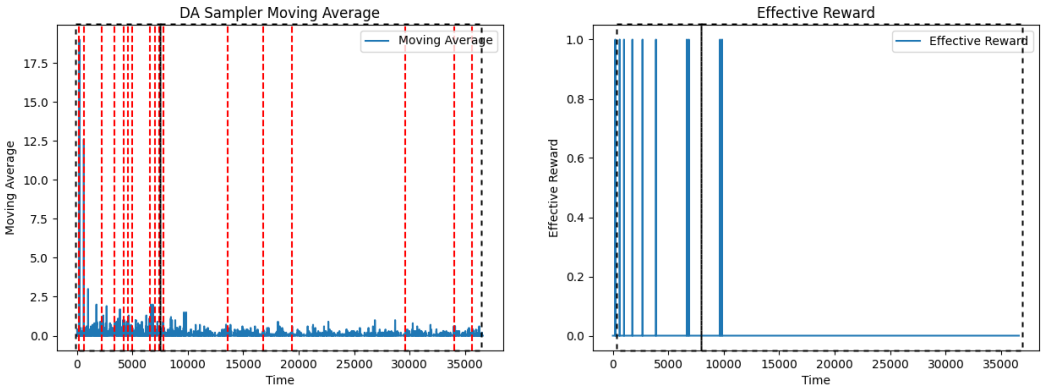


Figure 34: Combined DA neuron firings moving average by the DA sampler on the right. The red lines indicate the environment's primary reward. The effective reward is given on the right. The dashed box on the right of any plot depicts a single sequence element task (C-Z=>R), and the left one depicts two sequence element length task (B-Y=>C-Z=>R).

The sampling mechanism requires a threshold, as the D2 neurons cause the learned reward to bring less dopamine, the model outputs less DA firings; therefore, the thresholding causes no future effective reward to be released when critic cortico-striatal connections reach high levels to inhibit primary reward and the D1 excitation alone which causes the shifted reward is not enough to pass the threshold.

On top of this, reduced levels of firing, as the cortico-striatal connections strengthen, the critic can associate wrong stimuli neurons to D1 neurons, which results in an incorrect burst of dopamine, which explains the effective reward at around $10,000ms$. The network expects a reward caused by reward-predicting stimuli. However, the

predicted reward does not arrive. Unfortunately, this network does not employ prolonged inhibition; therefore, it cannot cause the wrongly associated connections to break, resulting in further disruption of critical behavior.

The network learns the first part of the sequence correctly, shown in the region around $4,500ms$ to $5,000ms$ and $6,500ms$ to $7,000ms$, as the interval between correct choices decreases and the correct choice is selected over many trials in series. The network, however, cannot accurately learn the next sequence element, as the dopamine shift with the effective reward was not realized. The resulting trial accuracy of the model is reported in Figure 35.

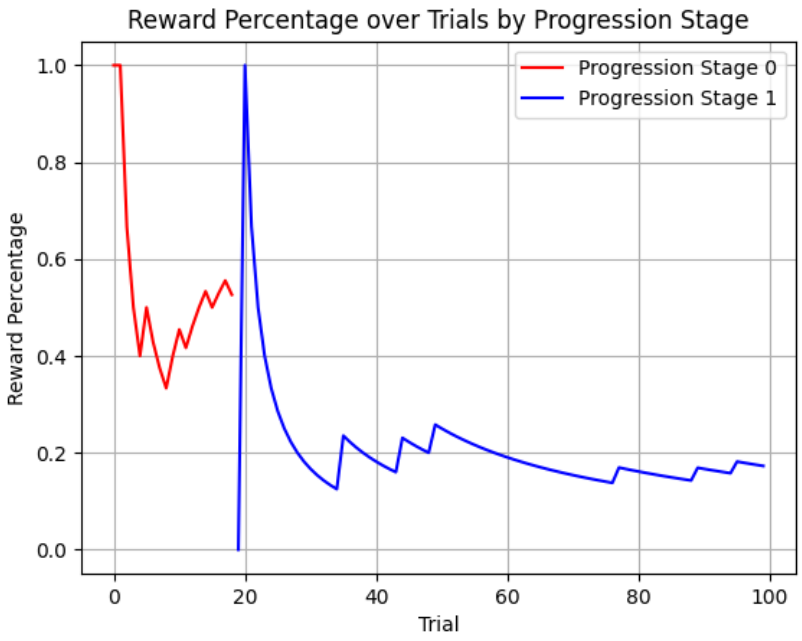


Figure 35: Actor-Critic model accuracy by progression stages.

The network shows initial results as the accuracy metric is stronger than the stochastic expected value. The poor performance on the two-length sequence task is attributed to learning impairment at the critic described earlier.

5.5 Evaluation of Multi-Channel Basal Ganglia Model

This section validates action selection on a learned actor component. Following the back-shift of expectation, the reward signal has been confirmed on the critic component. Lastly, the learning on the combined actor-critic model was evaluated.

5.5.1 Validation of Actor Model

The entire sequence (Stimuli A - Stimuli B - Stimuli C) has been inputted into the Multi-Channel BG circuit with correct stimuli-channel (Stimuli A/ Channel 1, Stimuli B/Channel 2, Stimuli C/Channel 3) weights strengthened. The model successfully chose the correct actions for each stimulus with an evident inhibition in the SNr activity. Figure 36 and Figure 37 show the fire rates of the BG nucleus.

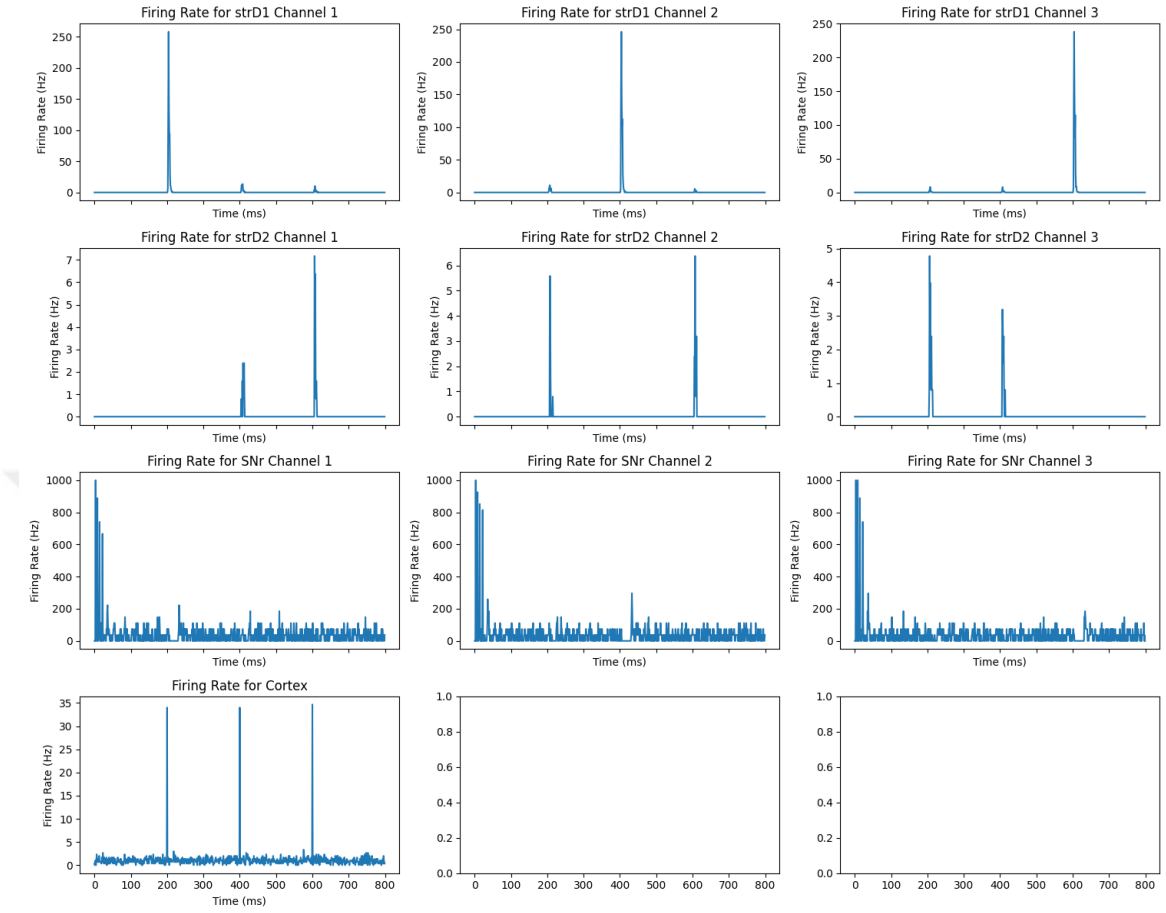


Figure 36: Spontaneous firing rate of Multi-Channel BG model with stimuli inputted at (Stimuli A), 400ms (Stimuli B), and 600ms (Stimuli C) marks.

The activity around the marks 200ms (Stimuli A), 400ms (Stimuli B), and 600ms (Stimuli C) depicts the selected channel. In this setup, while the connections to the D1 neurons have been strengthened, the same connections between stimuli and D2 of the corresponding channel have been weakened as the D2 receptors exhibit LTD on high dopamine levels during learning.

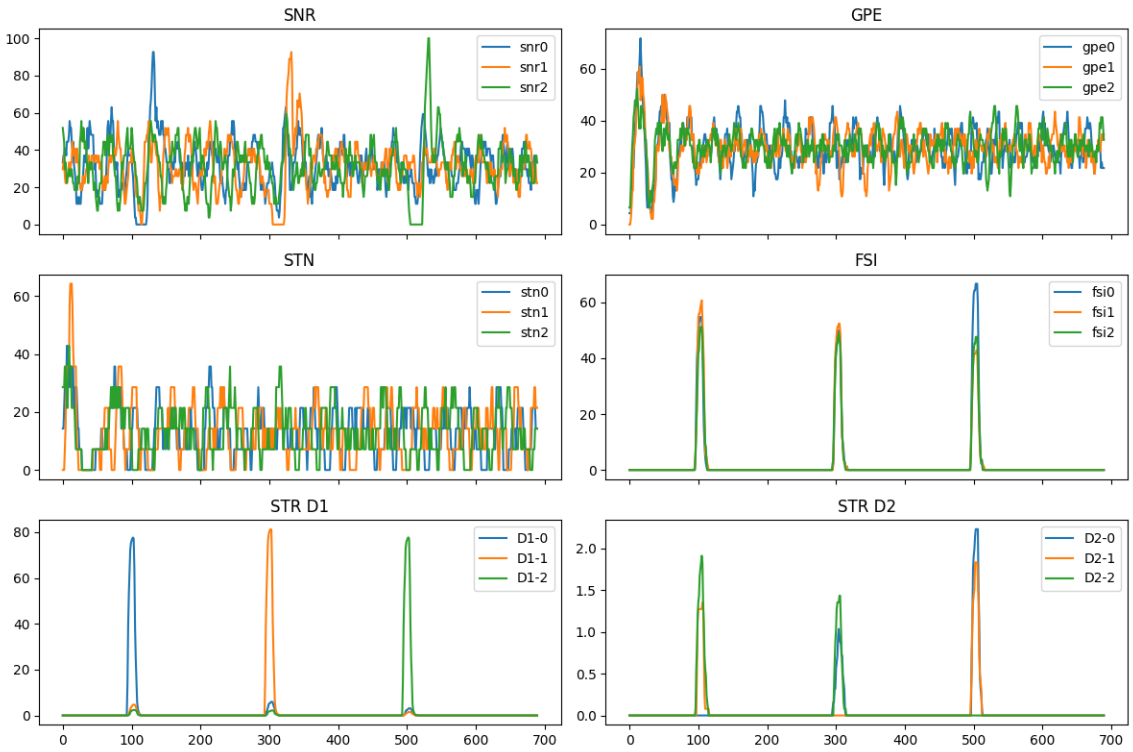


Figure 37: Moving average ($25ms$ windows) firing rate of Multi-Channel BG model with transient effect at $100ms$ cropped as SNr behaves erratically during initial setup.

The model has been run on a longer simulation (78 full sequence trials), and weights have been set to mimic the learned sequence. The actor selects the correct actions at each task point and showcases high accuracy. Figure 38 reports this long trial calculation results. While this ensures the model's ability to select an action, it should be noted that for this selection to occur, the network requires sharp separation in the input salience, which is indeed the case with artificial weights. However, to reach this separation during learning, additional network dynamics should be present, which is discussed in more detail in the Chapter 6.

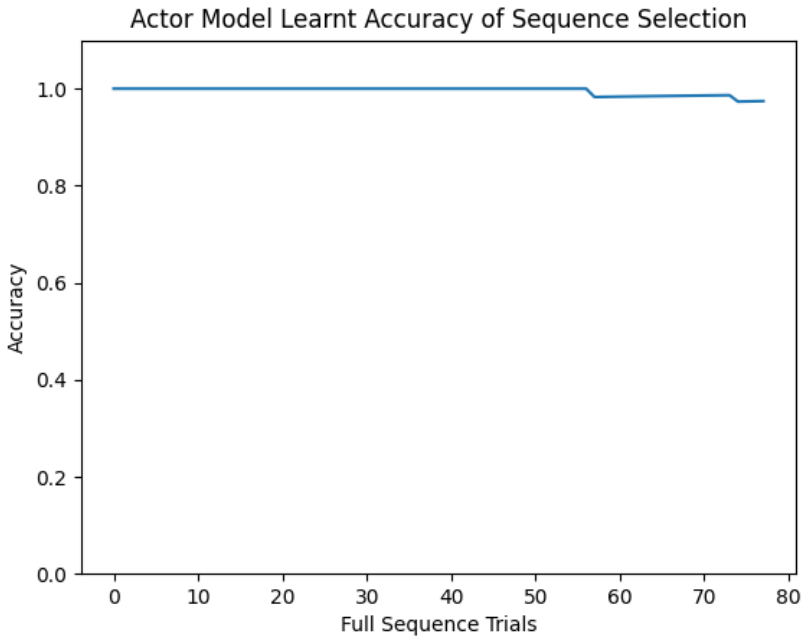


Figure 38: Multi-Channel Actor Model accuracy over 78 full sequence action selections with pre-learned weights. The model made only two mistakes in the sequence, resulting in over 97% accuracy.

5.5.2 Validation of Critic Model

The critic model's adaptive functionality is evaluated using pre-learned weights, which showcases the prediction function encoding of D1/D2 neurons and reward prediction error. The network has been evaluated under four setups.

The first setup (Figure 39) assumes the agent performs a novel task. When the state-action graph is undiscovered, the critic is expected to not evaluate any value for particular stimuli. As expected, when the entire sequence has been inputted to the critic, the critic does not react to the input nuclei (D1/D2 neurons). When a novel reward is received at the end of the sequence, the DA burst neurons replicate the phasic increase in DA concentration and fire a burst of dopaminergic spikes, increasing DA concentration above baseline.

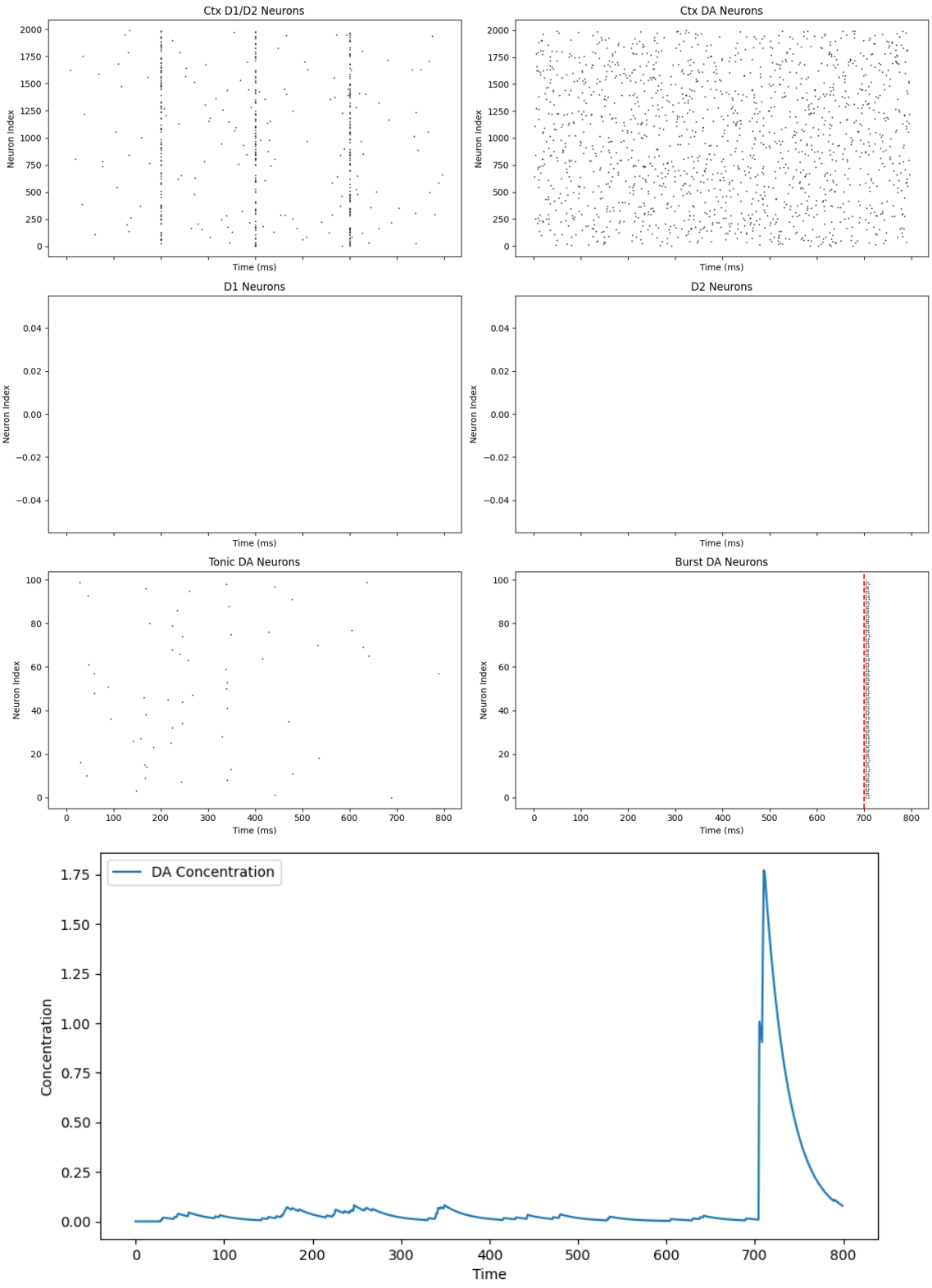


Figure 39: Critic network component neuron firings (top) and resulting DA concentration with 0.01 on_hit_DA spike strength. The red line indicates the primary reward.

As the primary reward encodes a novel reward, it is unexpected, resulting in an effective reward as dopamine rises. This DA signal reinforces the connections between firing D1/D2 neurons and the stimuli. Unlike the raster plot results, the DA signal modulates and stimulates connections. Therefore, D1 neurons start to fire with the incoming DA signal. This results in a correlation between stimuli and post-synaptic neurons, D1/D2 neurons, which causes higher excitation when the next time the same stimuli appear. As the critic gets subjected to the stimuli and reward pair, the D1 and D2 neurons start predicting the stimuli's value. The D1 neurons encode the value of reaching this stimulus, which makes up the reward expectation, and the D2 neurons encode the expected primary reward and suppress it. The final result (Figure 40 is that the burst DA neuron activation shifts to the time of stimuli appearance as in [3]).

The chaining of the reward allows for a higher concentration of DA response to coincide with eligibility raised at the time of stimuli. If the reward does not back-shift to earlier occurrences of stimuli, the network would have to learn either slowly or would not be able to learn any stimuli that have a shorter eligibility duration than the stimuli-reward time difference; in this setup, the eligibility tau is $140ms$ meaning it is not possible to learn a fourth stimuli in $200ms$ interval timing with only primary reinforcer.

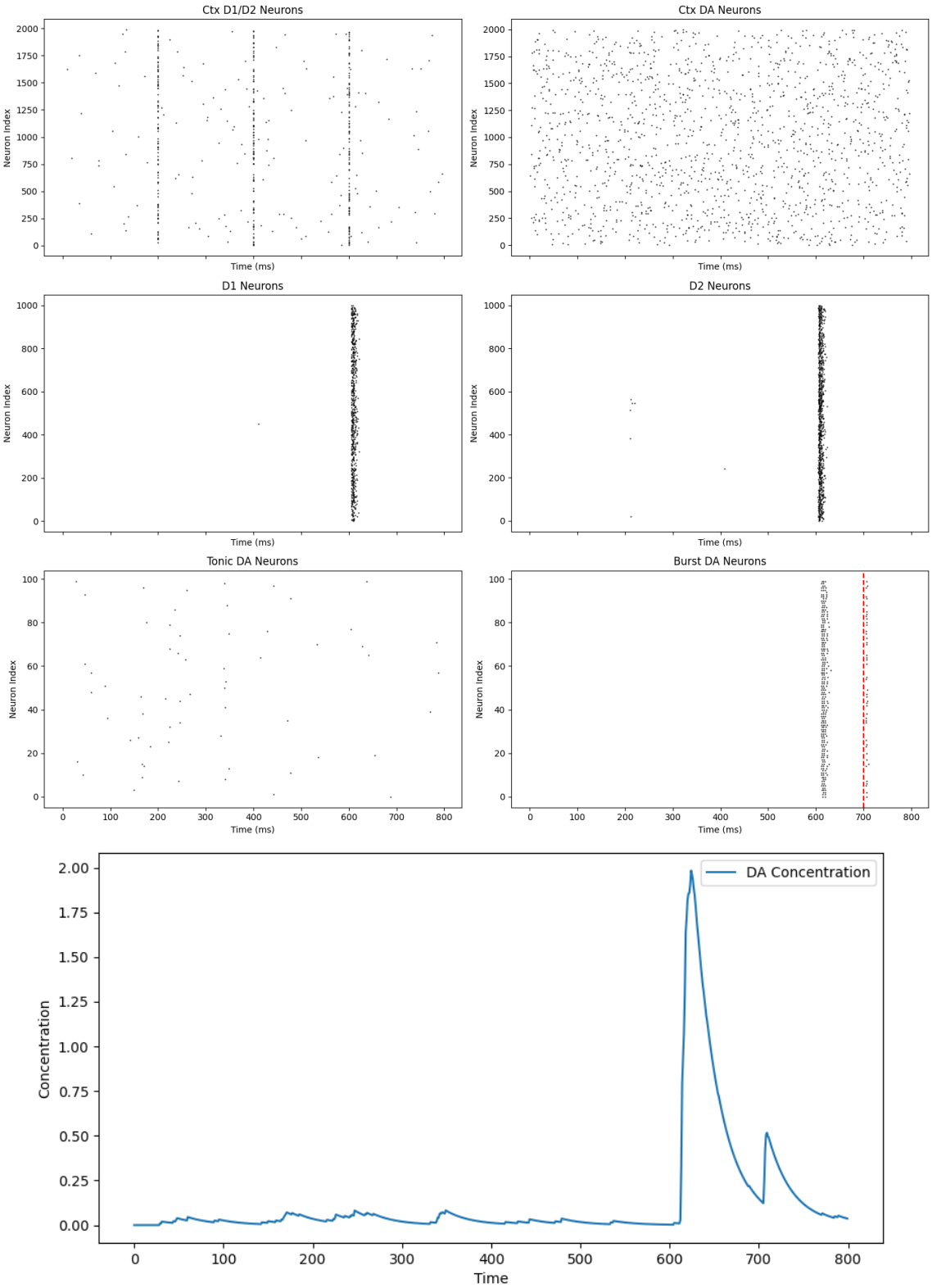


Figure 40: The critic associates the last sequence element (Stimuli C) with the reward. Resulting in an earlier response with a burst of dopamine.

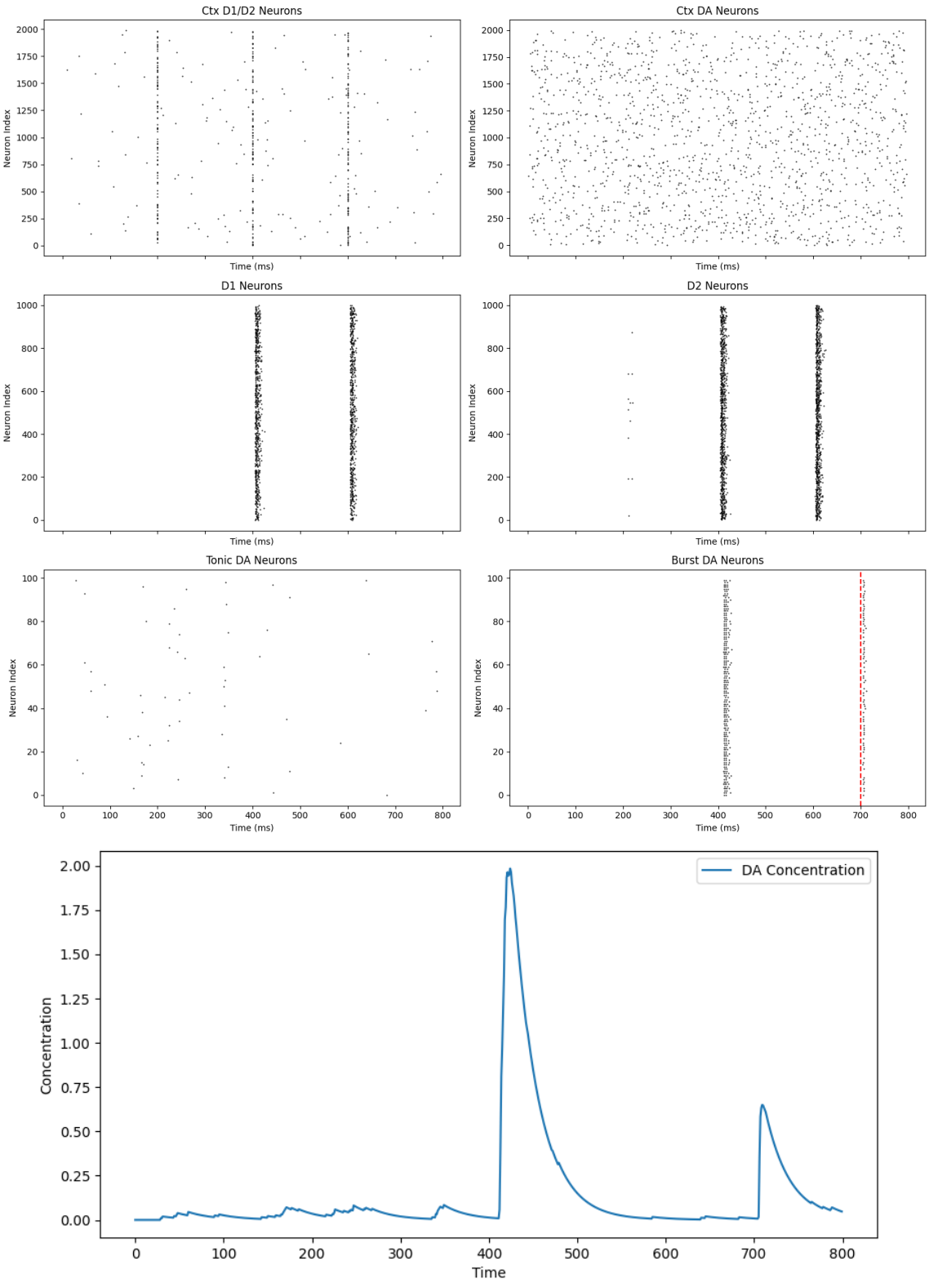


Figure 41: The negative of the current stimuli prediction (Stimuli B on D2 neurons) is propagated to the next stimuli prediction (Stimuli A on D1 neurons) by delays to suppress the previously encoded expectation of reward.

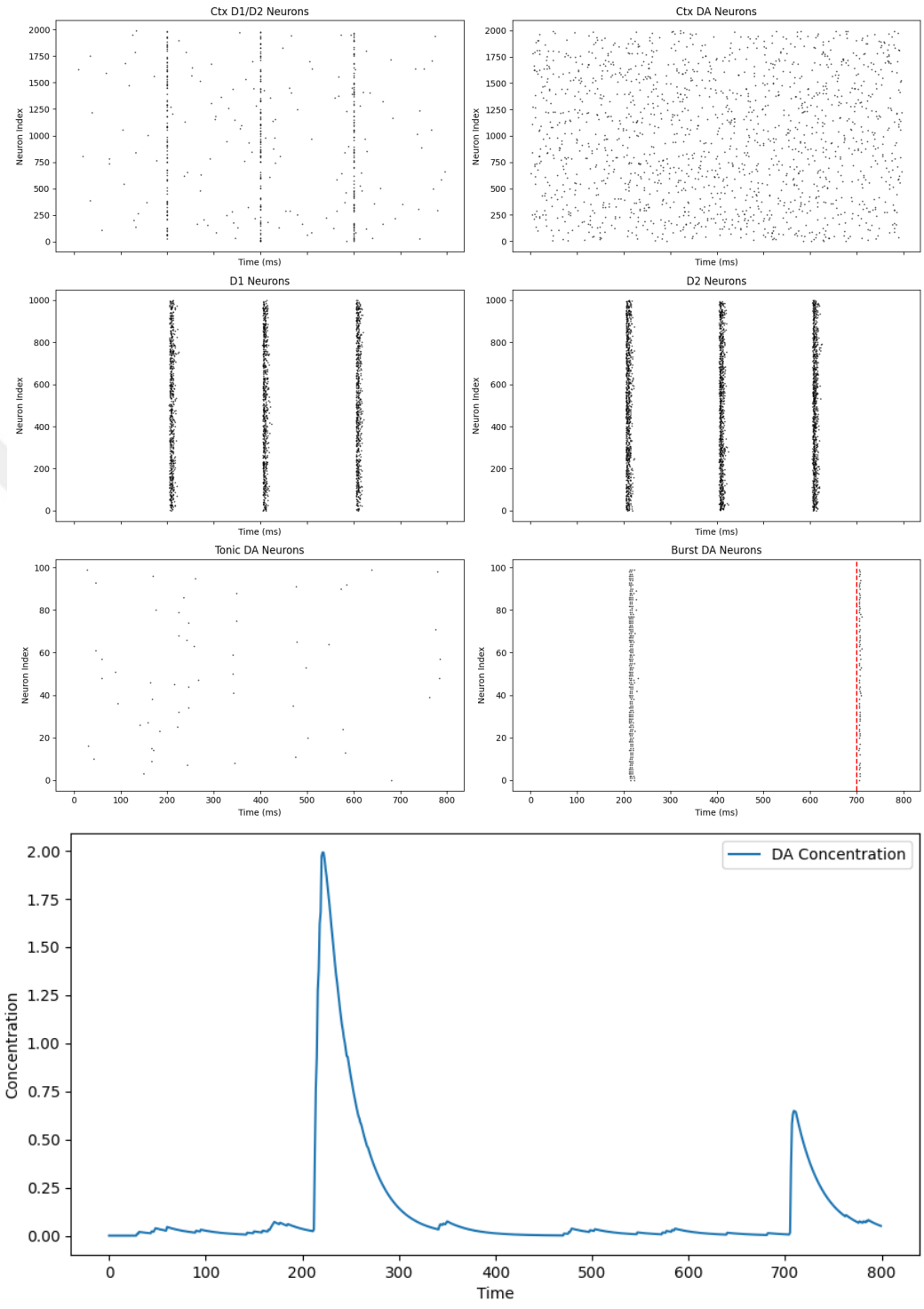


Figure 42: Entire sequence has been learned through the critic to initiate a reward expectation as the consequence of the learned sequence.

As shown in Figures 39, 40, 41 and 42, at the beginning of the experiment, the

critic model does not react to stimuli, analogous to a human who has no prior predictions about rewards in an unfamiliar environment. When an unexpected reward occurs, the primary reward signal, sourced from the environment, triggers a phasic increase in the firing rate of DA neurons. Over time, as the D2 neurons learn the value of the stimuli and associate it with the reward, the DA response shifts from reacting to the primary reinforcer to the stimuli itself. Consequently, the influence of the primary reinforcer is suppressed. Once the first element of the sequence is learned, the network focuses on learning subsequent sequence elements, continuing this process until the entire sequence is learned. This setup proves the critic network is capable of TD learning. The primary reward can be entirely suppressed by adjusting connection strength parameters (Figure 43) but left as a way to reinforce learned behaviors.

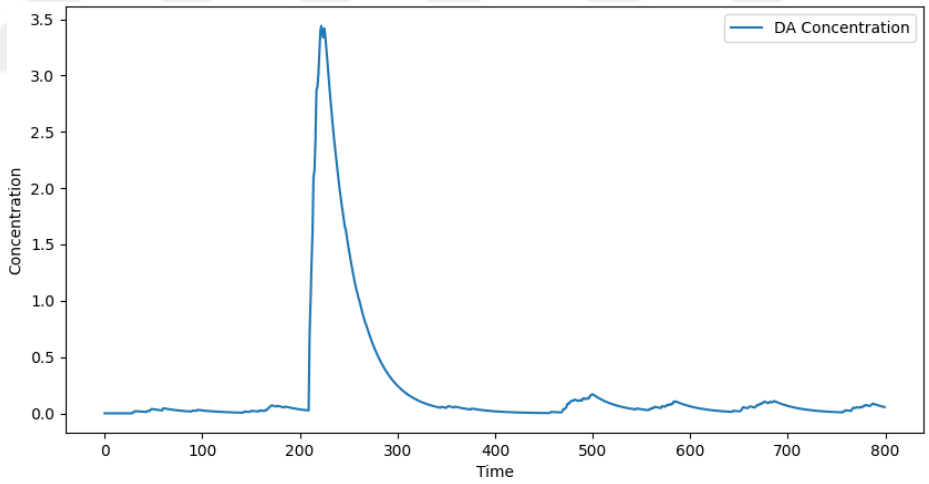


Figure 43: The connection strength of D2 to burst DA neurons can be adjusted to fully suppress the primary reinforcing DA signal in a learned scenario.

5.5.3 Validation of Combined Actor-Critic Model

The performance of the combined actor-critic model is evaluated, with results illustrated in Figure 44, showing task completion and reward maximization. As corticostriatal synapses associated with rewards are reinforced, the rewarding behavior is

repeated. However, due to the lack of balancing mechanisms, such as negative or below-baseline dopamine levels and homeostasis of neural activity, other connections also strengthen over time. This leads to the actor model's inability to distinguish between stimuli, ultimately reducing the network's effectiveness. Preliminary results suggest that further improvements are necessary while the network shows promise in effective learning.

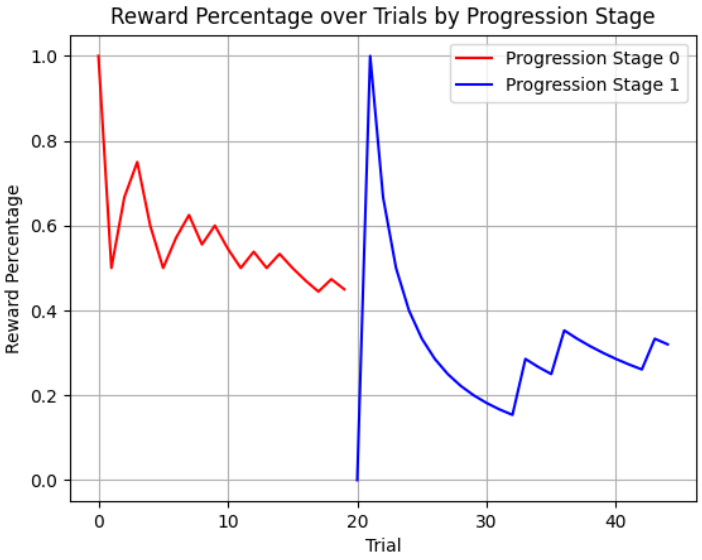


Figure 44: performance of combined actor-critic model.

5.6 Putting It Altogether

The learning neuron models are correctly implemented, with weight strengthening between neurons adhering to the established learning rules, both 2F STDP and 3F R-STDP. While these rules can facilitate localized learning, their application to global network setups, as demonstrated in the classical conditioning experiment, requires either more complex network behaviors to regulate neuronal firing and selectively amplify correct connections or higher-order learning rules such as 4F learning with reward receptivity, as shown in the dopamine receptivity experiment.

The Go/No-Go task has been successfully achieved in a single-channel Basal Ganglia (BG) model with dopamine modulation, demonstrating that the proposed model can selectively inhibit or permit certain actions.

Combining learning neurons with the striosome critic dynamics of the BG, a critic architecture has been proposed and evaluated on a sequence learning task. This evaluation has shown that a dual-pathway network can encode the current value function and apply prolonged inhibition to negate the earlier value function, thereby equating the TD learning error. Including the primary reward, this network can shift the reward to earlier predictors as these stimuli are learned. However, the proposed architecture does not directly implement prolonged inhibition but mimics the suppression of excitation through cleverly adjusted delays. While this approach can handle the excitation of subsequent cues, the lack of intermediate inhibition causes issues related to the omission of reward and below-baseline dopamine levels, eventually impairing learning. This issue may lie in the network's inability to separate weights, which will be discussed further in the Chapter 6.

The single-channel BG network provided two paths concerning the actor structure. Empirical evidence supports segregated channels for each commanding system and closed loops, but the anatomical structure of such loops raises concerns about generalizing this approach. Therefore, a new, No-Channel network structure independent of task structure has been proposed, along with a biologically plausible network. This actor structure has been simulated with weights of a learned agent to demonstrate the network's ability to correctly perform action selection under true learning conditions. While the model provided high accuracy, proving that action selection mechanisms work, a caveat (ambiguous selection) will be addressed in the Chapter 6.

The implemented networks have been combined with the critic component to demonstrate their learning capability, albeit with limited efficiency. Preliminary results from the experiments show that a fully learned actor and critic can manifest

action selection and reward prediction due to earlier cues, effectively solving sequence learning tasks. The network also shows preliminary success by providing higher accuracy than the stochastic expected value.

Overall, the results suggest that both actor-critic network architectures have potential applications in reinforcement learning, with improvements needed in some components and adjustments in hyperparameters.

CHAPTER VI

DISCUSSION ON FINDINGS AND OBSERVATIONS

This chapter discusses the limitations of the network that were identified, which are comprised mainly of the ambiguous selection process. Then, the missing dynamics of both neuron and network levels were discussed, leading to poor learning. Finally, another architecture involving the cortex has been proposed to apply temporal sequence memory as the policy component of reinforcement learning to the network.

6.1 Limitations of the Implemented Model

The task at hand has been solved with preliminary results through learning at the end of the implementation. However isolated, pre-learned validations for actor and critic components prove the model capable of the task. This contradiction points to poor dynamics when the network is learning. Some of these dynamics are non-existent in the network, while others are implemented with low fidelity. The subsequent parts will touch upon the low-performing parts in the model.

6.1.1 Indirect Pathway and Lateral Inhibition

The D2 receptor expressing MSNs in the striatum projects onto GPe-STN-SNr/GPi to create the indirect pathway of the BG. The sequence of inhibition and excitation leads to total excitation of the output nuclei under a highly salient motor command on the D2 input of a corresponding channel. As a result, the channel is excited, which means it is inhibited from getting selected. Utilizing STN-SNr cross-channel connections, this excitation of the output creates surround excitation to competing channels.

The first issue arises with the lack of excitation from the STN, as highly salient

input to D2 neurons causes GPe to be inhibited, which releases inhibition on STN to finally produce excitation to the SNr both on the native channel and competing channels. However, results from Figure 33 and Figure 37 illustrate that even when inhibition of GPe is successfully carried out by the increasing D2 MSNs firings, the disinhibition of STN is not strong enough to cause apparent excitation from STN.

In short, total lateral inhibition suffers across channels, requiring higher input saliency differences across direct pathways to recreate the competing balance as the indirect pathway is impaired. The balance between these nuclei is restored through reciprocal connections of GPe and STN, even if the connection strength is readjusted to focus on this STN rise. Therefore, STN's role in the indirect pathway and the lateral inhibition is cut.

6.1.2 Transmission Speed and Timing

One feature of the three pathways of BG is the relative timings of the signals. Direct pathway, as the name suggests, is direct. Requires less transmission delay to affect the output. The problem occurs when short latency action selection is concerned. The direct pathway selects an action. This winning channel should suppress the other channels, but the ambiguity and the selection already happened when the corresponding pathways challenge their respective counterparts' direct pathways. The lateral inhibition follows the indirect pathway and emerges late to this competition. The proposed network selects an action in a limited short window of time. The effect of lateral inhibition is late in resolving the competition.

Another issue is STN does not only create lateral inhibition. It also completes the indirect pathway. Local STN to SNr connections are faster than cross-channel STN to SNr connections. Therefore, lateral inhibition happens even after the whole encapsulation of the motor command selection is over. Thus, this lateral inhibition does not play a part in resolving the selection. So, one may wonder what role the late

inhibition actually plays. It has one definite answer which is the hyper-direct pathway. Whenever a new motor command is issued, the cortex must terminate all ongoing activity in the BG parts corresponding to the command set involving the latest selection. Utilizing STN lateral connections hyper-direct pathway increases excitation on all output nuclei, practically resetting the competition. Therefore, lateral inhibition is not used in the controlled inhibition of competing channels as proposed. This creates non-decisive selection and ambiguity in closely matching saliency commands.

6.1.3 Coexistence of Phasic (Temporal-Burst) and Tonic (Rate) Patterns

The striatum's medium spiny neurons (MSNs) receive afferent projections from Layer V pyramidal neurons of the cortex. These pyramidal neurons exhibit intrinsically bursting firing patterns, characterized by strong, short-duration pulses followed by tonic spiking. This dual-firing behavior allows cortical neurons to generate bursts of information signaling potential actions to the basal ganglia.

One notable feature of cortical columns is their ability to exhibit predictive states. During these states, individual cells within a column spike without causing the entire column to burst. However, when a new sequence begins in a cortical area, the entire column bursts, leading to activity distinct from the tonic firing seen in predictive states. In both of the proposed models, stimuli are encoded as single bursts of activity on top of a Poisson Spike Generator, thus requiring the striatum to handle both burst and rate-based cortical activities.

To simulate this cortical behavior, the striatal MSNs should effectively filter and respond to both types of cortical input. MSNs have two states: up and down. In the down state, neuron excitability is low, and low spiking activity in the cortex does not elicit significant responses. However, coordinated bursting activity from pyramidal neurons can transition MSNs to the upstate, where neuron conductance is higher, enabling successful information transmission to the striatum.

The original GPR [91] model assumes rate-based, long-duration excitation from the cortex, derived from predictive information associated with tonic cortical activity. This approach presents challenges when integrating short burst inputs from the cortex, as these bursts represent novel stimuli and are critical for accurate action selection.

In the proposed models, which extends the GPR framework, short burst inputs from the cortex do not provide clear separation and inhibition in the output nuclei resulting in ambiguous action selection. There are a few methods to mitigate this problem: A longer burst activity that adheres to the biology of the cortex, modifying neuron populations to work with both burst and rate-based inputs. A better representation of the burst activity of the cortex would involve a short pulse of high-frequency response. The single burst activity encoding a stimulus must be repeated to model such stimuli. While the Striatum MSNs have been modified to bursting neurons, the rest of the network should also be made to utilize burst information propagation.

6.2 Missing Low Level Processes: Synapse and Neuron Level Homeostasis

The proposed models encodes the stimulus on a pre-defined number of stimulus group neurons. As the cortical-striatal connections are sparse, the resulting stimulus to striatum connections is few. In biological neuronal dynamics, neurons try to achieve their preferred firing rates and avoid extreme deviations. To do so, numerous mechanisms ranging from synaptic level to network level [118] exist. If this homeostasis is breached in low frequency bound, a neuron might become unresponsive as its connections weaken and a neuron essentially becomes dead. Homeostasis prevents this end and distributes the representation onto many neurons to create richer and more robust connections.

6.3 Missing Model Level Dynamics

This section describes anticipated connectivity and network-level dynamics to ensure efficient learning.

6.3.1 Lateral Inhibition Between Striatum D1 MSNs

Contrary to action channel inputs coming from separate cortex regions, creating uncorrelated activity between stimuli and actions between channels, the proposed model inputs all channels with the same extrinsic stimuli. Therefore, the saliency in a channel is dependent on cortico-striatal connections from the cortex.

The actor validation ensures that input saliencies should differ with almost a magnitude of order for a precise selection. For this separation to happen, the cortex-channel connections should also be separated. Initializing from the unlearnt case, channel-wise strengths of connections are similar around a mean. This depicts similar saliency input to the ganglia. As mentioned before, the network does not reliably choose the correct channel; this can result in wrong D1 input-to-action pairings, ending with reinforcing wrong connections. However, this issue has been discussed before, so the following discussion assumes a fully reliable ganglia.

The problem is the separation of weights. As the correct input-action pairing brings back reward to reinforce connections, the competing channels are not exempt from the learning rule and dopamine, thus reinforcing them simultaneously. This creates an issue similar to that of the classical conditioning experiment.

The reward reinforces the highest saliency input as it has the most post-synaptic firings, which is the virtue of being the highest salience. As the experiment is repeated, the difference between weights deepens. Unfortunately, this dynamic cannot go on forever as there is a maximum boundary to what a weight can be. Therefore, solving such a problem requires suppressing the competing post-synaptic firings in D1 MSNs. Hence, a lateral inhibition between the D1 region can allow such learning.

Unfortunately, there is no such connection in the Basal Ganglia.

6.3.2 Persistence Signal and Close Loops

The organization of the Cortical-Basal Ganglia-Thalamo-Cortical loop suggests an architecture where segregated channels for each command system form a closed loop [112]. However, there are also regions of the cortex that project to all competing channels in a set. Our model showcased such a connection type and did not apply closed loops, as the thalamus was not implemented. One such unimplemented connection is the feedback onto the striatum and cortex. The thalamus feedbacks the output to positively project the striatum. Under such projections, the winning channel can excite its output to increase correlated pre- post- synaptic spike occurrences and reinforce correct stimuli-action channel pair connections.

6.3.3 Critic Dopamine Level

The previous sections put together a comprehensive discussion on the problems of the actor. Now, we turn to the dynamics of actor and dopamine signaling. Biologically, dopamine signaling refers to a baseline, and the signature of reward depends on this baseline. However, one such question is how neurons decide upon this baseline. This might involve higher-order adaptation mechanisms in the dopaminergic connections for the receiving neuron, which can form a base level for the incoming neuromodulator. The D1 and D2 receptors involve many biochemical processes to realize their functionality: changing synaptic weights depending on the relative dopamine signal. One problem in this representation is that negative dopamine signaling, defined by below baseline activity of the dopamine neurons, has a smaller range of values. The minimum of what can be represented with a single population group coding both negative and positive is limited by the group's firing. That minimum is no firing, which means zero dopamine level. On the other hand, positive representation has a greater range, with a possible practical limit on dopamine neurons' maximum firing.

The situation mentioned above can be balanced with longer inhibition times, resulting in dopamine levels staying down enough to counteract the phasic response. In the omission of reward expected by a reward-predicting cue, the primary reward cannot bring the excitation to offset the prolonged inhibition. In such a case, the inhibition cancels the earlier excitation, resulting in net zero dopamine levels. However, in the proposed networks, there is no long-term inhibition in the critic. Therefore, if the critic predicts reward by coming to a stage that should have been learned, a problem within the actor occurs to select a random choice. This behavior should not have been reinforced, but as the effective reward predicted causes a rise in dopamine levels, the incorrect action has been reinforced. This positive reinforcement should be actively suppressed by allowing the network to forget wrongly learned weights or punish wrong behavior. As explained, this dynamic is ensured by prolonged inhibition. In its absence, the wrong weights can be potentiated, causing further challenges in channel weight separation, which creates a self-feeding problem in the actor-critic dynamic.

In short, forgetting is crucial for efficient learning in neural networks [119]. In biological systems, neurons should forget outdated or irrelevant information to make room for new learning. In the critic model, the lack of active forgetting mechanisms can lead to persistent, incorrect synaptic associations. This is particularly problematic when the network has learned to expect a reward that does not materialize, as the erroneous expectation continues to influence network behavior.

Implementing forgetting mechanisms can significantly enhance the network's learning capabilities. Two potential approaches include:

- *Dynamic Baseline Adjustment:* Introducing a feedback mechanism to continuously readjust the baseline dopamine level, enabling the network to respond to deviations more effectively.

- *Adaptive Synaptic Scaling:* Employing synaptic scaling mechanisms that adjust

synaptic strengths based on overall activity levels, preventing runaway potentiation and ensuring balanced network activity.

6.3.4 Chunk Processing and Eligibility

Another point to consider is that multiple action selection queries might appear at the Ganglia in close proximity. In such a situation, the reward signal does reinforce both selection processes. This is similar to the problem faced by distractor stimuli in the classical conditioning experiment and requires temporal dynamics in the learning rule, such as dopamine receptivity introduced in one of our experiments. Figure 32 showcases how such clear separation might be possible under higher-order learning rules.

6.4 Missing Global Structure: Cortex

Looking back into Basal Ganglia's anatomy, the striatum takes topographical projections from motor command systems. The motor cortex projections feed individual channels with an action request. Therefore, the most significant conflict regarding the bio-plausibility of the network implemented is in the input scheme. So, the role of the cortex should be discussed.

The cortex works as a temporal memory that memorizes patterns inputted and categorizes temporal sequences of inputted sensory information and selected actions [120] [121]. As the input flows into the cortex, the hierarchical processing of the pattern causes invariant highly semantic information, keeps a world model, and predicts the changes in the environment.

The sequence of actions and the whole Markov Decision Process that constitutes the policy is recorded in the cortex [122]. At the same time, Ganglia enacts the policy and adjusts it by selecting rewarding actions that occur more frequently [123]. The cortex slowly forgets wrongly predicted sequences and reinforces the occurring pattern; in the task case, the state machine is defined in Figure 13.

One way to think of such a cortex is as an infinite state-action space graph. The cortex is a blank space to apply to the task at hand in its initial conditions. As the environment state stimuli flow into the cortex, a node defining the state of 'C' appears. The connecting edges are the actions. Edges are inputted into channels and stimuli, which is the state node 'C' fed onto channels. When the Ganglia selects an action, the resulting selection is also fed into the cortex, as the cortex updates the edge strength, which corresponds to the saliency of the edge projected onto the Ganglia. Therefore, the actual learning shifts onto the cortex, while the Ganglia's role is to adjust the saliency inputted using cortico-striatal connections to favor channels that result in dopamine release.

CHAPTER VII

CONCLUSION

In this thesis, we successfully implemented the Izhikevich neuron model on Loihi2 neuromorphic hardware. Building upon this, we integrated 2-factor (2F) Hebbian learning by implementing the Spike-Timing-Dependent Plasticity (STDP) learning rule on Loihi2. For more advanced 3-Factor (3F) learning rules, such as the Reward-Modulated STDP (R-STDP) based Izhikevich model, we initially realized them on a CPU to later transition to Loihi2. Additionally, we explored 4-factor (4F) learning rules inspired by biochemical processes, assessing their potential to enhance learning.

The implemented models were the foundation for constructing a reinforcement learning network model inspired by the neural structures of goal-directed learning in the Basal Ganglia (BG). Initially, we developed a BG model capable of performing Go/No-Go task, enabling selecting a single action. This model was subsequently adapted into a multi-action selection network, forming the actor component of the actor-critic model.

The critic model was developed to encode the reward prediction error dopamine signal, facilitating temporal difference learning [102]. We created comprehensive actor-critic models applied to a sequence learning task by integrating the actor and critic models. However, the model achieved only preliminary and sub-optimal results.

Despite the sub-optimal performance in sequence learning tasks, the BG model incorporating an action-selector actor and a reward prediction error critic shows promise. As software, network, and hardware limitations are addressed over time, this model should serve as a foundational framework for advancing the reinforcement learning paradigm within Spiking Neural Networks.

The insights and advancements from this study contribute to the growing body of knowledge in neuromorphic computing and bio-realistic neural network models, paving the way for future research and development in this domain.

APPENDIX

This appendix contains code listings relevant to the implementation of the third factor reward trace calculation and the R-STDP learning rules used in the model described in this thesis. These code snippets are written in the Lava Software Framework [27] and adhere to the requirements of the *Loihi2FLearningRule* and *Loihi3FLearningRule* classes, as well as trace calculations.

```
def calculate_third_factor_trace(self, s_graded_in: float) -> float:
    """Generate's a third factor reward traces based on
    graded input spikes to the R-STDP Izhikevich process.
    """
    baseline = self.DA_tonic
    DA_old = self.DA[:]
    self.DA[:] += -DA_old / self.DA_tau + (baseline / self.DA_tau) +
        self.on_hit_DA * s_graded_in
    return self.DA[:]
```

Listing 1: Third factor reward trace calculation in Lava Framework

The function *calculate_third_factor_trace* generates a reward trace based on graded input spikes. The dopamine (DA) level is updated using a combination of baseline DA and a graded input factor, which simulates the receiving of DA spikes on dopaminergic connections.

```
# Eligibility trace represented as dt
dt = f"{self.learning_rate} * {self.A_minus} * x0 * y1 + " \
     f"{self.learning_rate} * {self.A_plus} * y0 * x1 - " \
     f"u0 * t * {self.eligibility_trace_decay_tau}"
# Reward-modulated weight update
dw = f"({self.max_weights} - w) * u0 * t * y2 "
```

Listing 2: R-STDP learning rule (R-STDP_Basic)

The R-STDP learning rule implements a reward-modulated weight update mechanism. The eligibility trace (dt) is calculated using pre- and post-synaptic spike interactions, and the weight change (dw) is influenced by the reward signal, facilitating reinforcement learning.

```
class RSTDP_D1(Loihi3FLearningRule):
    ...
    dt = f"{self.learning_rate} * {self.A_minus} * x0 * y1 + " \
         f"{self.learning_rate} * {self.A_plus} * y0 * x1 - " \
         f"u0 * t * {self.eligibility_trace_decay_tau}"
    dw = f"({self.max_weights} - w) * u0 * t * (y2 - {self.baseline
       }) "
    ...
class RSTDP_D2(Loihi3FLearningRule):
    ...
    dt = f"{self.learning_rate} * {self.A_minus} * x0 * y1 + " \
         f"{self.learning_rate} * {self.A_plus} * y0 * x1 - " \
         f"u0 * t * {self.eligibility_trace_decay_tau}"
    dw = f"({self.max_weights} - w) * u0 * t * ({self.baseline} - y2
       ) "
    ...
```

Listing 3: RSTDP Learning Rule for D1 and D2 Neurons

These classes define the RSTDP learning rules for D1 and D2 neurons, crucial components in the striatum. The equations for dt and dw are tailored to reflect the distinct roles of the DA receptors the neuron types express in the reward-based learning process.

REFERENCES

[1] H. An, K. Bai, and Y. Yi, *The Roadmap to Realize Memristive Three-Dimensional Neuromorphic Computing System*. 10 2018. Accessed: 2024-06-11.

[2] J. L. Lanciego, N. Luquin, and J. A. Obeso, "Functional Neuroanatomy of the Basal Ganglia," *Cold Spring Harbor Perspectives in Medicine*, vol. 2, pp. a009621–a009621, Dec. 2012.

[3] W. Schultz, "Dopamine reward prediction error coding," *Dialogues in Clinical Neuroscience*, vol. 18, pp. 23–32, 03 2016.

[4] R. B. Uludağ, S. Çağdaş, Y. S. İşler, N. S. Şengör, and Aktürk, "Bio-realistic neural network implementation on Loihi 2 with Izhikevich neurons," *Neuromorphic Computing and Engineering*, vol. 4, p. 024013, June 2024.

[5] J. Houk, J. Adams, and A. Barto, "A Model of How the Basal Ganglia Generate and Use Neural Signals that Predict Reinforcement," *Models of Information Processing in the Basal Ganglia*, vol. Vol. 13, July 1995.

[6] W. A. Wulf and S. A. McKee, "Hitting the memory wall: implications of the obvious," *SIGARCH Comput. Archit. News*, vol. 23, p. 20–24, mar 1995.

[7] P. Bose, *Power Wall*, pp. 1593–1608. Boston, MA: Springer US, 2011.

[8] V. Balasubramanian, "Brain power," *Proceedings of the National Academy of Sciences*, vol. 118, no. 32, p. e2107022118, 2021.

[9] K. D. Harris, "Additive function approximation in the brain," 2019.

[10] E. Grossi and M. Buscema, "Introduction to artificial neural networks," *European journal of gastroenterology hepatology*, vol. 19, pp. 1046–54, 01 2008.

[11] S. Herculano-Houzel, "The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost," *Proceedings of the National Academy of Sciences*, vol. 109, no. supplement_1, pp. 10661–10668, 2012.

[12] Y. Wang, D. Liu, and Y. Wang, "Discovering the capacity of human memory," *Brain and Mind*, vol. 4, pp. 189–198, 01 2003.

[13] M. Mayford, S. A. Siegelbaum, and E. R. Kandel, "Synapses and Memory Storage," *Cold Spring Harbor Perspectives in Biology*, vol. 4, pp. a005751–a005751, June 2012.

[14] D. Patterson, T. Anderson, N. Cardwell, R. Fromm, K. Keeton, C. Kozyrakis, R. Thomas, and K. Yelick, "A case for intelligent ram," *IEEE Micro*, vol. 17, no. 2, pp. 34–44, 1997.

[15] O. Mutlu, S. Ghose, J. Gómez-Luna, and R. Ausavarungnirun, *A Modern Primer on Processing in Memory*, pp. 171–243. Singapore: Springer Nature Singapore, 2023.

[16] K. Yamazaki, V.-K. Vo-Ho, D. Bulsara, and N. Le, "Spiking neural networks and their applications: A review," *Brain Sciences*, vol. 12, no. 7, 2022.

[17] P. Dayan and L. Abbott, *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*, vol. 15. 01 2001.

[18] W. Zhang, B. Gao, J. Tang, P. Yao, S. Yu, M.-F. Chang, H.-J. Yoo, H. Qian, and H. Wu, "Neuro-inspired computing chips," *Nature Electronics*, vol. 3, pp. 371–382, July 2020. Publisher: Nature Publishing Group.

[19] P. Qu, L. Yang, W. Zheng, and Y. Zhang, "A review of basic software for brain-inspired computing," *CCF Transactions on High Performance Computing*, vol. 4, pp. 1–9, 03 2022.

[20] F. Liao, F. Zhou, and Y. Chai, "Neuromorphic vision sensors: Principle, progress and perspectives," *Journal of Semiconductors*, vol. 42, p. 013105, 01 2021.

[21] J.-N. Li and Y.-H. Tian, "Recent advances in neuromorphic vision sensors: A survey," *Jisuanji Xuebao/Chinese Journal of Computers*, vol. 44, pp. 1258–1286, 06 2021.

[22] M. Davies, N. Srinivasa, T.-H. Lin, G. Chinya, Y. Cao, S. H. Choday, G. Dimou, P. Joshi, N. Imam, S. Jain, Y. Liao, C.-K. Lin, A. Lines, R. Liu, D. Mathaikutty, S. McCoy, A. Paul, J. Tse, G. Venkataramanan, Y.-H. Weng, A. Wild, Y. Yang, and H. Wang, "Loihi: A neuromorphic manycore processor with on-chip learning," *IEEE Micro*, vol. 38, no. 1, pp. 82–99, 2018.

[23] Intel, "Intel unveils neuromorphic loihi 2 and lava software," 2021. Accessed: 2024-06-12.

[24] IBM, "Northpole: Ibm ai chip," 2022. Accessed: 2024-06-12.

[25] ResearchGate, "Scalable energy-efficient, low-latency implementations of trained spiking deep belief networks on spinnaker." Accessed: 2024-06-12.

[26] Intel, "Intel builds world's largest neuromorphic system," 2021. Accessed: 2024-06-12.

[27] "Lava: A software framework for neuromorphic computing." https://github.com/lava-nc/lava, 2024.

[28] J. K. Eshraghian, M. Ward, E. O. Neftci, X. Wang, G. Lenz, G. Dwivedi, M. Bennamoun, D. S. Jeong, and W. D. Lu, "Training spiking neural networks using lessons from deep learning," *Proceedings of the IEEE*, vol. 111, no. 9, pp. 1016–1054, 2023.

[29] M. Stimberg, R. Brette, and D. F. Goodman, "Brian 2, an intuitive and efficient neural simulator," *eLife*, vol. 8, p. e47314, aug 2019.

[30] W. Fang, Y. Chen, J. Ding, Z. Yu, T. Masquelier, D. Chen, L. Huang, H. Zhou, G. Li, and Y. Tian, "Spikingjelly: An open-source machine learning infrastructure platform for spike-based intelligence," 2023.

[31] T. Bekolay, J. Bergstra, E. Hunsberger, T. DeWolf, T. C. Stewart, D. Rasmussen, X. Choo, A. Voelker, and C. Eliasmith, "Nengo: a python tool for building large-scale functional brain models," *Frontiers in Neuroinformatics*, vol. 7, 2014.

[32] J. E. Pedersen, S. Abreu, M. Jobst, G. Lenz, V. Fra, F. C. Bauer, D. R. Muir, P. Zhou, B. Vogginger, K. Heckel, G. Urgese, S. Shankar, T. C. Stewart, J. K. Eshraghian, and S. Sheik, "Neuromorphic intermediate representation: A unified instruction set for interoperable brain-inspired computing," 2023.

[33] C. D. Schuman, T. E. Potok, R. M. Patton, J. D. Birdwell, M. E. Dean, G. S. Rose, and J. S. Plank, "A survey of neuromorphic computing and neural networks in hardware," *ArXiv*, vol. abs/1705.06963, 2017.

[34] D. V. Christensen, R. Dittmann, B. Linares-Barranco, A. Sebastian, M. L. Gallo, A. Redaelli, S. Slesazeck, T. Mikolajick, S. Spiga, S. Menzel, I. Valov, G. Milano, C. Ricciardi, S.-J. Liang, F. Miao, M. Lanza, T. J. Quill, S. T. Keene, A. Salleo, J. Grollier, D. Marković, A. Mizrahi, P. Yao, J. J. Yang, G. Indiveri, J. P. Strachan, S. Datta, E. Vianello, A. Valentian, J. Feldmann, X. Li, W. H. P. Pernice, H. Bhaskaran, S. Furber, E. Neftci, F. Scherr, W. Maass, S. Ramaswamy, J. Tapson, P. Panda, Y. Kim, G. Tanaka, S. Thorpe, C. Bartolozzi, T. A. Cleland, C. Posch, S. Liu, G. Panuccio, M. Mahmud, A. N. Mazumder, M. Hosseini, T. Mohsenin, E. Donati, S. Tolu, R. Galeazzi, M. E. Christensen, S. Holm, D. Ielmini, and N. Pryds, "2022 roadmap on neuromorphic computing and engineering," *Neuromorphic Computing and Engineering*, vol. 2, p. 022501, May 2022. Publisher: IOP Publishing.

[35] M. R. Ahmed and B. Sujatha, "A review on methods, issues and challenges in neuromorphic engineering," in *2015 International Conference on Communications and Signal Processing (ICCSP)*, pp. 0899–0903, 2015.

[36] A. R. Young, M. E. Dean, J. S. Plank, and G. S. Rose, "A review of spiking neuromorphic hardware communication systems," *IEEE Access*, vol. 7, pp. 135606–135620, 2019.

[37] A. Balaji, Y. Wu, A. Das, F. Catthoor, and S. Schaafsma, "Exploration of segmented bus as scalable global interconnect for neuromorphic computing," in *Proceedings of the 2019 Great Lakes Symposium on VLSI*, GLSVLSI '19, (New York, NY, USA), p. 495–499, Association for Computing Machinery, 2019.

[38] A. Balaji, P. K. Huynh, F. Catthoor, N. D. Dutt, J. L. Krichmar, and A. Das, "Neusb: A scalable interconnect architecture for spiking neuromorphic hardware," *IEEE Transactions on Emerging Topics in Computing*, vol. 11, no. 2, pp. 373–387, 2023.

[39] A. Mehonic and A. J. Kenyon, "Brain-inspired computing needs a master plan," *Nature*, vol. 604, pp. 255–260, Apr. 2022. Publisher: Nature Publishing Group.

[40] P. Panda, S. A. Aketi, and K. Roy, "Toward scalable, efficient, and accurate deep spiking neural networks with backward residual connections, stochastic softmax, and hybridization," *Frontiers in Neuroscience*, vol. 14, 2020.

[41] S. B. Shrestha and G. Orchard, "Slayer: Spike layer error reassignment in time," 2018.

[42] E. O. Neftci, H. Mostafa, and F. Zenke, "Surrogate gradient learning in spiking neural networks," 2019.

[43] N. Caporale and Y. Dan, "Spike timing–dependent plasticity: A hebbian learning rule," *Annual review of neuroscience*, vol. 31, pp. 25–46, 02 2008.

[44] R. C. Malenka and M. F. Bear, "LTP and LTD an embarrassment of riches," *Neuron*, vol. 44, pp. 5–21, 2004.

[45] K. Doya, "What are the computations of the cerebellum, the basal ganglia and the cerebral cortex?," *Neural Networks*, vol. 12, pp. 961–974, Oct. 1999.

[46] K. Doya, "Complementary roles of basal ganglia and cerebellum in learning and motor control," *Current Opinion in Neurobiology*, vol. 10, pp. 732–739, Dec. 2000.

[47] T. Buschman and E. Miller, "Goal-direction and top-down control," *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, vol. 369, 11 2014.

[48] R. Naud, N. Marcille, C. Clopath, and W. Gerstner, "Firing patterns in the adaptive exponential integrate-and-fire model," *Biological cybernetics*, vol. 99, pp. 335–47, 12 2008.

[49] C. Ganguly, S. S. Bezugam, E. Abs, M. Payvand, S. Dey, and M. Suri, "Spike frequency adaptation: bridging neural models and neuromorphic applications," *Communications Engineering*, vol. 3, no. 1, p. 22, 2024.

[50] L. Abbott, "Lapicque's introduction of the integrate-and-fire model neuron (1907)," *Brain Research Bulletin*, vol. 50, no. 5, pp. 303–304, 1999.

[51] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," *The Journal of Physiology*, vol. 117, no. 4, pp. 500–544, 1952.

[52] A. L. Hodgkin, A. F. Huxley, and B. Katz, "Measurement of current-voltage relations in the membrane of the giant axon of loligo," *The Journal of Physiology*, vol. 116, no. 4, pp. 424–448, 1952.

[53] A. L. Hodgkin and A. F. Huxley, "Currents carried by sodium and potassium ions through the membrane of the giant axon of loligo," *The Journal of Physiology*, vol. 116, no. 4, pp. 449–472, 1952.

[54] A. L. Hodgkin and A. F. Huxley, "The components of membrane conductance in the giant axon of loligo," *The Journal of Physiology*, vol. 116, no. 4, pp. 473–496, 1952.

[55] E. Izhikevich, "Simple model of spiking neurons," *IEEE Transactions on Neural Networks*, vol. 14, no. 6, pp. 1569–1572, 2003.

[56] E. Izhikevich, "Which model to use for cortical spiking neurons?," *IEEE Transactions on Neural Networks*, vol. 15, no. 5, pp. 1063–1070, 2004.

[57] J. Baladron, J. Vitay, T. Fietzek, and F. H. Hamker, "The contribution of the basal ganglia and cerebellum to motor learning: A neuro-computational approach," *PLOS Computational Biology*, vol. 19, pp. 1–29, 04 2023.

[58] H. J. Groenewegen, "The basal ganglia and motor control," *Neural Plasticity*, vol. 10, no. 1-2, p. 108384, 2003.

[59] T. Wu, J. Wang, C. Wang, M. Hallett, Y. Zang, X. Wu, and P. Chan, "Basal ganglia circuits changes in parkinson's disease patients," *Neuroscience Letters*, vol. 524, no. 1, pp. 55–59, 2012.

[60] F. Blandini, G. Nappi, C. Tassorelli, and E. Martignoni, "Functional changes of the basal ganglia circuitry in parkinson's disease," *Progress in Neurobiology*, vol. 62, no. 1, pp. 63–88, 2000.

[61] M. D. Humphries, J. A. Obeso, and J. K. Dreyer, "Insights into parkinson's disease from computational models of the basal ganglia," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 89, no. 11, pp. 1181–1188, 2018.

[62] S. Bariselli, C. Glangetas, S. Tzanoulinou, and C. Bellone, "Ventral tegmental area subcircuits process rewarding and aversive experiences," *Journal of Neurochemistry*, vol. 139, no. 6, pp. 1071–1080, 2016. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jnc.13779.

[63] C. Sibert, H. Hake, J. Laird, C. Lebiere, P. Rosenbloom, and A. Stocco, "The role of the basal ganglia in the human cognitive architecture: A dynamic causal modeling comparison across tasks and individuals," 05 2021.

[64] S.-Y. Chen, K.-M. Lu, H.-A. Ko, T.-H. Huang, J. H.-J. Hao, Y.-T. Yan, S. L.-Y. Chang, S. M. Evans, and F.-C. Liu, "Parcellation of the striatal complex into dorsal and ventral districts," *Proceedings of the National Academy of Sciences*, vol. 117, no. 13, pp. 7418–7429, 2020.

[65] M. R. Roesch, T. Singh, P. L. Brown, S. E. Mullins, and G. Schoenbaum, "Ventral striatal neurons encode the value of the chosen action in rats deciding between differently delayed or sized rewards," *Journal of Neuroscience*, vol. 29, no. 42, pp. 13365–13376, 2009.

[66] P. F. Durieux, S. N. Schiffmann, and A. de Kerchove d'Exaerde, "Targeting neuronal populations of the striatum," *Frontiers in Neuroanatomy*, vol. 5, 2011.

[67] K. R. Brimblecombe and S. J. Cragg, "The striosome and matrix compartments of the striatum: A path through the labyrinth from neurochemistry toward function," *ACS Chemical Neuroscience*, vol. 8, no. 2, pp. 235–242, 2017. PMID: 27977131.

[68] T. Yoshizawa, M. Ito, and K. Doya, "Reward-predictive neural activities in striatal striosome compartments," *eNeuro*, vol. 5, no. 1, 2018.

[69] N. Javed and M. Cascella, *Neuroanatomy, Globus Pallidus*. 06 2020.

[70] S. T. Kitai and H. Kita, "Anatomy and physiology of the subthalamic nucleus: A driving force of the basal ganglia," in *The Basal Ganglia II* (M. B. Carpenter and A. Jayaraman, eds.), (Boston, MA), pp. 357–373, Springer US, 1987.

[71] Y. Zhang, K. M.-H. Larcher, B. Misic, and A. Dagher, "Anatomical and functional organization of the human substantia nigra and its connections," *eLife*, vol. 6, p. e26653, aug 2017.

[72] H. Schroll and F. H. Hamker, "Computational models of basal-ganglia pathway functions: focus on functional neuroanatomy," *Frontiers in Systems Neuroscience*, vol. 7, 2013.

[73] T. Gilbertson and D. Steele, "Tonic dopamine, uncertainty and basal ganglia action selection," *bioRxiv*, 2020.

[74] N. A. Bratcher, V. Farmer-Dougan, J. D. Dougan, B. A. Heidenreich, and P. A. Garris, "The role of dopamine in reinforcement: Changes in reinforcement sensitivity induced by d1-type, d2-type, and nonselective dopamine receptor agonists," *Journal of the Experimental Analysis of Behavior*, vol. 84, no. 3, pp. 371–399, 2005.

[75] A. Buot and J. Yelnik, "Functional anatomy of the basal ganglia: Limbic aspects," *Revue neurologique*, vol. 168, pp. 569–75, 08 2012.

[76] P. Redgrave, T. Prescott, and K. Gurney, "The basal ganglia: a vertebrate solution to the selection problem?," *Neuroscience*, vol. 89, pp. 1009–1023, Apr. 1999.

[77] G. E. Alexander, M. R. DeLong, and P. L. Strick, "Parallel Organization of Functionally Segregated Circuits Linking Basal Ganglia and Cortex," *Annual Review of Neuroscience*, vol. 9, pp. 357–381, Mar. 1986. Publisher: Annual Reviews.

[78] J. Bahuguna, A. Aertsen, and A. Kumar, "Existence and control of go/no-go decision transition threshold in the striatum," *PLOS Computational Biology*, vol. 11, pp. 1–35, 04 2015.

[79] P. W. Glimcher, "Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis," *Proceedings of the National Academy of Sciences*, vol. 108, no. supplement_3, pp. 15647–15654, 2011.

[80] Y. Deng, D. Song, J. Ni, H. Qing, and Z. Quan, "Reward prediction error in learning-related behaviors," *Frontiers in Neuroscience*, vol. 17, 08 2023.

[81] M. Colombo, "Deep and beautiful. The reward prediction error hypothesis of dopamine," *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, vol. 45, pp. 57–67, Mar. 2014.

[82] C. A. Seger and B. J. Spiering, "A critical review of habit learning and the basal ganglia," *Frontiers in Systems Neuroscience*, vol. 5, 2011.

[83] N. Favila, K. Gurney, and P. Overton, "Role of the basal ganglia in innate and learned behavioural sequences," *Reviews in the neurosciences*, vol. 35, 07 2023.

[84] J. Zhang, J. Chen, S. Zhong, and H. Qiao, "A Bio-Inspired Integration Model of Basal Ganglia and Cerebellum for Motion Learning of a Musculoskeletal Robot," *Journal of Systems Science and Complexity*, vol. 37, pp. 82–113, Feb. 2024.

[85] T. J. Prescott, K. Gurney, F. Montes-Gonzalez, M. Humphries, and P. Redgrave, "The Robot Basal Ganglia:," in *The Basal Ganglia VII* (L. F. B. Nicholson and R. L. M. Faull, eds.), vol. 52, pp. 349–358, Boston, MA: Springer US, 2002. Series Title: Advances in Behavioral Biology.

[86] M. D. Humphries, R. D. Stewart, and K. N. Gurney, "A physiologically plausible model of action selection and oscillatory activity in the basal ganglia," *Journal of Neuroscience*, vol. 26, no. 50, pp. 12921–12942, 2006.

[87] M. D. Humphries and K. Gurney, "Network effects of subthalamic deep brain stimulation drive a unique mixture of responses in basal ganglia output," *European Journal of Neuroscience*, vol. 36, no. 2, pp. 2240–2251, 2012.

[88] C. Eliasmith, T. C. Stewart, X. Choo, T. Bekolay, T. DeWolf, Y. Tang, and D. Rasmussen, "A large-scale model of the functioning brain," *Science*, vol. 338, no. 6111, pp. 1202–1205, 2012.

[89] R. Bogacz and K. Gurney, "The basal ganglia and cortex implement optimal decision making between alternative actions," *Neural computation*, vol. 19, pp. 442–77, 03 2007.

[90] T. Prescott, F. Montes González, K. Gurney, M. Humphries, and P. Redgrave, "A robot model of the basal ganglia: Behavior and intrinsic processing," *Neural Networks*, vol. 19, pp. 31–61, Jan. 2006.

[91] M. Humphries and K. Gurney, "Making decisions in the dark basement of the brain: A look back at the gpr model of action selection and the basal ganglia," *Biological Cybernetics*, vol. 115, 07 2021.

[92] B. GIRARD, V. CUZIN, A. GUILLOT, K. N. GURNEY, and T. J. PRESCOTT, "A basal ganglia inspired model of action selection evaluated in a robotic survival task," *Journal of Integrative Neuroscience*, vol. 02, no. 02, pp. 179–200, 2003.

[93] B. Girard, D. Filliat, J.-A. Meyer, A. Berthoz, and A. Guillot, "Integration of navigation and action selection functionalities in a computational model of cortico-basal-ganglia–thalamo-cortical loops," *Adaptive Behavior*, vol. 13, no. 2, pp. 115–130, 2005.

[94] A. Araújo, I. C. Duarte, T. Sousa, J. Oliveira, A. T. Pereira, A. Macedo, and M. Castelo-Branco, "Neural inhibition as implemented by an actor-critic model involves the human dorsal striatum and ventral tegmental area," *Scientific Reports*, vol. 14, p. 6363, Mar. 2024. Publisher: Nature Publishing Group.

[95] P. Berthet, M. Lindahl, P. J. Tully, J. Hellgren-Kotaleski, and A. Lansner, "Functional Relevance of Different Basal Ganglia Pathways Investigated in a Spiking Model with Reward Dependent Plasticity," *Frontiers in Neural Circuits*, vol. 10, July 2016. Publisher: Frontiers.

[96] A. G. Barto, "Adaptive Critics and the Basal Ganglia," in *Models of Information Processing in the Basal Ganglia* (J. C. Houk, J. L. Davis, and D. G. Beiser, eds.), pp. 215–232, The MIT Press, Nov. 1994.

[97] W. Potjans, M. Diesmann, and A. Morrison, "An Imperfect Dopaminergic Error Signal Can Drive Temporal-Difference Learning," *PLOS Computational Biology*, vol. 7, p. e1001133, May 2011. Publisher: Public Library of Science.

[98] F.-M. Zhou and C. R. Lee, "Intrinsic and integrative properties of substantia nigra pars reticulata neurons," *Neuroscience*, vol. 198, pp. 69–94, Dec. 2011.

[99] K. T. Beier, E. E. Steinberg, K. E. DeLoach, S. Xie, K. Miyamichi, L. Schwarz, X. J. Gao, E. J. Kremer, R. C. Malenka, and L. Luo, "Circuit Architecture of VTA Dopamine Neurons Revealed by Systematic Input–Output Mapping," *Cell*, vol. 162, pp. 622–634, July 2015.

[100] S. N. Haber, "The place of dopamine in the cortico-basal ganglia circuit," *Neuroscience*, vol. 282, pp. 248–257, Dec. 2014.

[101] T. N. Lerner, A. L. Holloway, and J. L. Seiler, "Dopamine, Updated: Reward Prediction Error and Beyond," *Current opinion in neurobiology*, vol. 67, pp. 123–130, Apr. 2021.

[102] Y. Niv, M. Duff, and P. Dayan, "Dopamine, uncertainty and td learning," *Behavioral and brain functions : BBF*, vol. 1, p. 6, 06 2005.

[103] W. Schultz, P. Dayan, and P. R. Montague, "A neural substrate of prediction and reward," *Science*, vol. 275, no. 5306, pp. 1593–1599, 1997.

[104] M. Davies, A. Wild, G. Orchard, Y. Sandamirskaya, G. A. F. Guerra, P. Joshi, P. Plank, and S. R. Risbud, "Advancing neuromorphic computing with loihi: A survey of results and outlook," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 911–934, 2021.

[105] T. Faniran, "Numerical solution of stochastic differential equations," 05 2015.

[106] A. Roth and M. van Rossum, *Modeling Synapses*, vol. Vol. 06, pp. 139–159. 09 2009.

[107] W. G. Gomez, A. Pignata, R. Pignari, V. Fra, E. Macii, and G. Urgese, "First steps towards micro-benchmarking the lava-loihi neuromorphic ecosystem," in *2023 IEEE 16th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC)*, pp. 462–469, 2023.

[108] N. Frémaux and W. Gerstner, "Neuromodulated spike-timing-dependent plasticity, and theory of three-factor learning rules," *Frontiers in Neural Circuits*, vol. 9, 2016.

[109] E. M. Izhikevich, "Solving the Distal Reward Problem through Linkage of STDP and Dopamine Signaling," *Cerebral Cortex*, vol. 17, pp. 2443–2452, Oct. 2007.

[110] J. D. Sweatt, "Chapter 2 - studies of human learning and memory," in *Mechanisms of Memory (Second Edition)* (J. D. Sweatt, ed.), pp. 24–47, London: Academic Press, second edition ed., 2010.

[111] N. Foster, J. Barry, L. Korobkova, L. Garcia, L. Gao, M. Becerra, Y. Sherafat, B. Peng, X. Li, J.-H. Choi, L. Gou, B. Zingg, S. Azam, D. Lo, N. Khanjani, B. Zhang, J. Stanis, I. Bowman, K. Cotter, and H.-W. Dong, "The mouse cortico–basal ganglia–thalamic network," *Nature*, vol. 598, pp. 188–194, 10 2021.

[112] M. Humphries and K. Gurney, "A computational model of action selection in the basal ganglia: Thalamic and cortical interactions technical report aivru ref. no. 132," 11 1999.

[113] A. B. Muñoz-Manchado, C. Bengtsson Gonzales, A. Zeisel, H. Munguba, B. Bekkouche, N. G. Skene, P. Lönnerberg, J. Ryge, K. D. Harris, S. Linnarsson, and J. Hjerling-Leffler, "Diversity of interneurons in the dorsal striatum revealed by single-cell rna sequencing and patchseq," *Cell Reports*, vol. 24, no. 8, pp. 2179–2190.e7, 2018.

[114] J. C. Houk, J. L. Davis, and D. G. Beiser, *The Contribution of Cortical Neurons to the Firing Pattern of Striata! Spiny Neurons*, pp. 29–50. 1994.

[115] J. C. Houk, J. L. Davis, and D. G. Beiser, *The Contribution of Cortical Neurons to the Firing Pattern of Striatal Spiny Neurons*, pp. 29–50. 1994.

[116] K. Chakravarty, S. Roy, A. Sinha, A. Nambu, S. Chiken, J. Hellgren Kotaleski, and A. Kumar, "Transient Response of Basal Ganglia Network in Healthy and Low-Dopamine State," *eNeuro*, vol. 9, pp. ENEURO.0376–21.2022, Mar. 2022.

[117] R. E. Suri and W. Schultz, "Learning of sequential movements by neural network model with dopamine-like reinforcement signal," *Experimental Brain Research*, vol. 121, pp. 350–354, Aug. 1998.

[118] D. Thibault, N. Giguère, F. Loustalot, M.-J. Bourque, C. Ducrot, S. El Mestikawy, and L.- Trudeau, "Homeostatic regulation of excitatory synapses on striatal medium spiny neurons expressing the D2 dopamine receptor," *Brain Structure and Function*, vol. 221, pp. 2093–2107, May 2016.

[119] R. L. Davis and Y. Zhong, "The biology of forgetting—a perspective," *Neuron*, vol. 95, no. 3, pp. 490–503, 2017.

[120] J. Hawkins and S. Ahmad, "Why neurons have thousands of synapses, a theory of sequence memory in neocortex," *Frontiers in Neural Circuits*, vol. 10, 03 2016.

[121] S. Ahmad and J. Hawkins, "Untangling sequences: Behavior vs. external causes," 09 2017.

[122] D. Ballard, "Cortical memory," in *Brain Computation as Hierarchical Abstraction*, MIT Press, 2015.

[123] D. Ballard, "Programs via reinforcement," in *Brain Computation as Hierarchical Abstraction*, MIT Press, 2015.