frontiers in Neural Circuits

ORIGINAL RESEARCH
published: 21 June 2017
doi: 10.3389/fncir.2017.00045

A Biologically Plausible Architecture of the Striatum to Solve Context-Dependent Reinforcement Learning Tasks

*Sabyasachi Shivkumar, Vignesh Muralidharan and V. Srinivasa Chakravarthy **

Computational Neuroscience Lab, Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai, India

Basal ganglia circuit is an important subcortical system of the brain thought to be responsible for reward-based learning. Striatum, the largest nucleus of the basal ganglia, serves as an input port that maps cortical information. Microanatomical studies show that the striatum is a mosaic of specialized input-output structures called striosomes and regions of the surrounding matrix called the matrisomes. We have developed a computational model of the striatum using layered self-organizing maps to capture the center-surround structure seen experimentally and explain its functional significance. We believe that these structural components could build representations of state and action spaces in different environments. The striatum model is then integrated with other components of basal ganglia, making it capable of solving reinforcement learning tasks. We have proposed a biologically plausible mechanism of action-based learning where the striosome biases the matrisome activity toward a preferred action. Several studies indicate that the striatum is critical in solving context dependent problems. We build on this hypothesis and the proposed model exploits the modularity of the striatum to efficiently solve such tasks.

Keywords: striatum, basal ganglia, context dependent learning, striosomes and matrisomes, self organizing maps, modular reinforcement learning

OPEN ACCESS

Edited by:
Michael M. Halassa, New York University, United States

Reviewed by:
Masahiko Takada, Kyoto University, Japan
Ann M. Graybiel, Massachusetts Institute of Technology, United States

****Correspondence:***
V. Srinivasa Chakravarthy
schakra@iitm.ac.in

Received: *22 March 2017*
Accepted: *08 June 2017*
Published: *21 June 2017*

Citation:
Shivkumar S, Muralidharan V and Chakravarthy VS (2017) A Biologically Plausible Architecture of the Striatum to Solve Context-Dependent Reinforcement Learning Tasks. Front. Neural Circuits 11:45. doi: 10.3389/fncir.2017.00045

INTRODUCTION

In order to understand the role of the striatum within the basal ganglia (BG) circuit, it is essential to understand the rich and complex microcircuitry of this structure. It is well-known that the striatum has a modular architecture, containing specialized input-output structures called the "striosomes" and regions of the surrounding matrix called the "matrisomes" (Graybiel et al., 1991). The striosomes are known to receive limbic inputs and send their projections to the substantia nigra pars compacta, a midbrain dopaminergic nucleus, whereas the matrisomes mostly receive sensorimotor and associative inputs and project to downstream BG nuclei (Graybiel et al., 1994). The cortico-striatal connectivity seems to show a divergence property, where there is spread of connections coming from the cortex to the striatum followed by a convergence at the level of the globus pallidus (GP; Graybiel et al., 1994). There have also been suggestions that the striatum constructs low dimensional representations of the cortical states via the cortico-striatal projections (Bar-Gad et al., 2000, 2003). Indirect evidence for this comes from experiments which indicate hebbian like learning in cortico-striatal projections (Charpier and Deniau, 1997).

Therefore, the striatum has the cellular and molecular machinery to possibly construct such reduced representations of cortical states. These facts about striatal microanatomy lead us to believe that the striatum could build representations for several state and action spaces.

Anatomically the striosome-matrisome complex has a center-surround structure (Graybiel et al., 1991), and the proposed computational architecture for the striatum is inspired by this fact. Studies investigating the projection of prefrontal areas to the striosomes show specificity to certain cortical areas (Eblen and Graybiel, 1995). These cortical projections to anterior striosomes are mostly from frontal regions like the orbitofrontal cortex, anterior insula, and the anterior cingulate cortex (Eblen and Graybiel, 1995) which could very well represent the task or state space (Wilson et al., 2014). The matrisome which receives more sensorimotor information would well represent the action space (Flaherty and Graybiel, 1994). In classical reinforcement learning (RL) literature, the expected reward signal in a given state is called the value function (Sutton and Barto, 1998). The striosomes are known to have reciprocal projections to both the ventral tegmental area (VTA) and the substantia nigra pars compacta (SNc) and thus would receive the prediction error signal from these midbrain nuclei, which can serve as a reinforcement signal that aids in the computation of the state value function (Granger, 2006; Wall et al., 2013). On the other hand the action representations perhaps evolve at the level of matrisomes, and get mapped on to action primitives at the level of GPi (Pasquereau et al., 2007). Thus, using the reward information from the environment and the representations built in the striatum, the BG can learn to perform reward based decision making tasks.

This functional organization and the modularity of the striatum has been hypothesized to perform context dependent tasks (Amemori et al., 2011). Mulitple spatio-temporal contexts could then be mapped to different striatal modules. This leads to the distribution of context information to different modules, a facet of modular reinforcement learning (Kalmár et al., 1999). We then consider the selection of the module appropriate to a given context to be driven by a responsibility signal, which is a function of the uncertainty in the environment. Uncertainty in the environment from previous approaches has been represented by reward variance (Balasubramani et al., 2015). Since change in context leads to increased uncertainty, reward variance could help identify this change.

In the current study, we propose a hierarchical self-organizing structure to model the striosome-matrisome compartments. Self-organizing maps (SOMs) have been used to represent high-dimensional information in 2-D sheets of neurons (Kohonen, 1990). The striosome and the matrisome layers are both modeled as a double SOM layer, consisting of Strio-SOM and Matri-SOM respectively, where a single Strio-SOM neuron has projections to the surrounding Matri-SOM neurons. The activity of the Matri-SOM is mapped to action primitives via the direct and indirect pathways of the BG to perform action selection. The reward information from the environment is utilized by the Strio-SOM to bias the surrounding Matri-SOM activity toward a preferred action. This provides a biologically plausible way of carrying out action based Q-learning (Sutton and Barto, 1998) and is a novel feature of our model. This model has been tested on standard grid-world problems.

The model has been extended to cater to problems with varying contexts (changing reward locations). Different striatal modules map different contexts and tonically active neurons (TANs; Apicella, 2007) aid in module selection This selection is driven by the risk (reward variance) in the environment which is used to calculate the responsibility signal (Amemori et al., 2011) for a particular module. We have tested this model on grid-world problems with varying reward distributions and the model is able to solve these problems efficiently.

METHODS

Modeling the Microanatomy of the Striatum

We have proposed an architecture consisting of two layers of SOMs as a method for mapping center-surround structures seen in the striatum (**Figure 1A**). This architecture is used to model striosomes and matrisomes which map the state space and action space, respectively.

The first layer called Strio-SOM models the striosomes and maps the state space. The second layer activated by the Strio-SOM is called the Matri-SOM which models the matrisomes and maps the action space (**Figure 1B**).

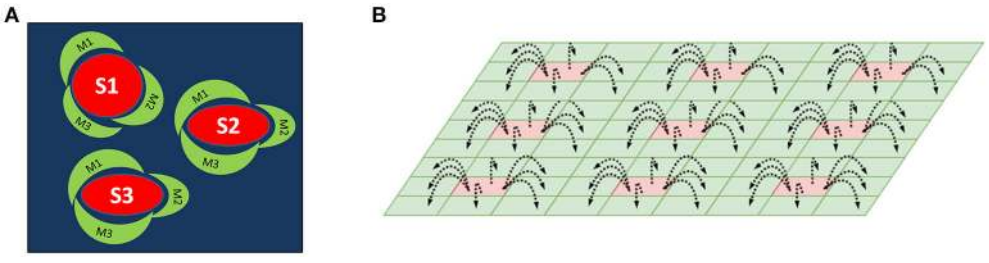

FIGURE 1 | (A) A schematic of the striosome-matrisome center surround mapping in the striatum. The red structures represent the striosomes and the surrounding green structures represent the matrisomes. **(B)** A Schematic of the layered SOM structure modeling the striosomes and matrisomes. The Strio-SOM (Red) represents the striosomes and the Matri-SOM (Green) represents the matrisomes; each Strio-SOM neuron has projections to the surrounding Matri-SOM neurons.

In order to map the state space, we have a Strio-SOM of size $m_1 \times n_1$. If s is a state vector, the weights of the Strio-SOM (W^S) are of dimension $m_1 \times n_1 \times \dim(s)$, where $\dim(s)$ stands for the dimension of the state vector s. Similarly, to map the action space, we have a Matri-SOM of size $m_2 \times n_2$. If a is an action vector, the weights of all the Matri-SOMs (W^M) are of dimension $m_1 \times n_1 \times m_2 \times n_2 \times \dim(a)$ as each neuron in the Strio-SOM is connected to a Matri-SOM.

The activity for a neuron n in the Strio-SOM for a state input s is given in Equation (1).

$$X^S{}_{[n]} = \exp\left(\frac{-||W^S{}_{[n]} - s||_2{}^2}{\sigma_S{}^2}\right) \quad (1)$$

where $[n]$ represents the spatial location of the neuron n and σ_S controls the sharpness of the neuron activity. The complete activity of the Strio-SOM (X^S) is the combination of individual activity of all the neurons. The neuron with the highest activity ("winner") for a state s is denoted by $n_s{}^*$.

Similarly, the activity for a neuron n in the Matri-SOM for an action input a in a state s is given in Equation (2).

$$X^M{}_{[n_s{}^*][n]} = \exp\left(\frac{-||W^M{}_{[n_s{}^*][n]} - a||_2{}^2}{\sigma_M{}^2}\right) \quad (2)$$

where σ_M controls the sharpness of the neuron activity. The complete activity of the Matri-SOM corresponding to neuron $n_s{}^*$ ($X^M{}_{[n_s{}^*]}$) is the combination of individual activity of all the neurons in the Matri-SOM corresponding to n_s^*. The neuron with the highest activity ("winner") for an action a in a state s is denoted as $n_{s,a}{}^*$.

The weight of a neuron n in the Strio-SOM for a state input s is updated according to the following rule (Equation 3)

$$W^S_{[n]} \leftarrow W^S_{[n]} + \eta_S . \exp\left(\frac{-||[n] - [n_s{}^*]||_2{}^2}{\sigma_S{}^2}\right).(s - W^S_{[n]}) \quad (3)$$

The weight of neuron n in the Matri-SOM for an action input a in a state s is updated according to Equation (4).

$$W^M_{[n_s{}^*][n]} \leftarrow W^M_{[n_s{}^*][n]} + \eta_M . \exp\left(\frac{-||[n] - [n_{s,a}{}^*]||_2{}^2}{\sigma_M{}^2}\right).(a - W^M_{[n_s{}^*][n]}) \quad (4)$$

Reinforcement Learning in Basal Ganglia

The striatum model developed in the previous section was useful in developing representations for states and actions. In this section, we incorporate the striatum model in a BG model and apply the model to standard reinforcement learning tasks. A schematic diagram of the model is given in **Figure 2**.

Let us assume that the animal is in a state s. The activity of the striosomes gives us the representation of the state in the striatum. In our model, the activity of the striosomes is given as the activity of the neurons in the Strio-SOM where the activity of a single neuron is given by Equation (1). Thus, the activity is of dimension $m_1 \times n_1$.

This activity of the Strio-SOM projects to the SNc and represents the value for the state s in our model (Equation 5).

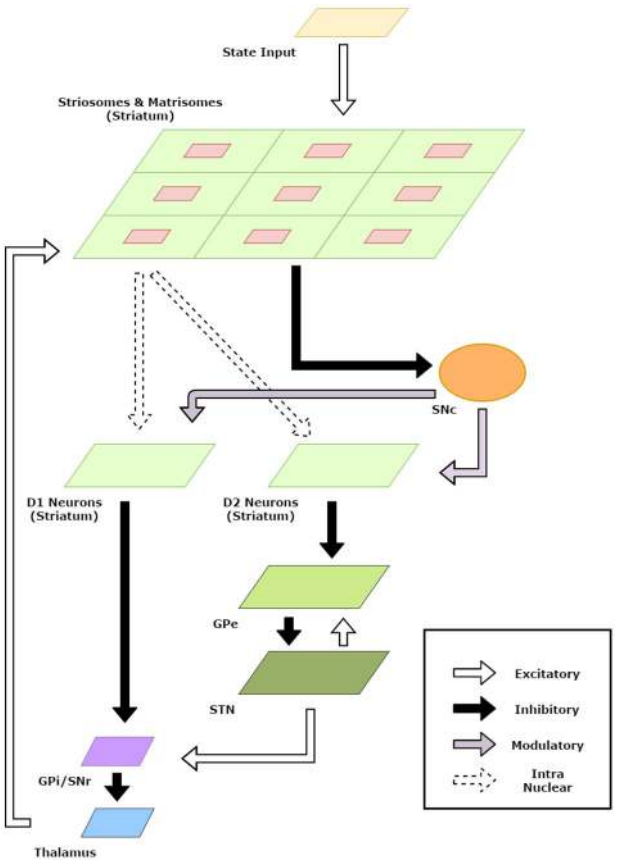

FIGURE 2 | Schematic Diagram for the basal ganglia model. The arrows indicate connections and their type. The component sizes are proportional to their dimensions. The feedback connections from the thalamus project the information about the action chosen back to the striatum.

These weights from the striatum to SNc ($W^{Str \to SNc}$) are trained using the signal from SNc which is representative of Temporal Difference (TD) error (δ) (Equation 6). The TD error is calculated as $\delta = r + \gamma V(s') - V(s)$ where s' is the new state after taking action a (Equation 19), r is the reward obtained and γ is the discount factor.

$$V(s) = \sum_{\forall n} W^{Str \to SNc}{}_{[n]} X^S{}_{[n]} \quad (5)$$

$$\Delta W^{Str \to SNc}{}_{[n]} = \eta^{Str \to SNc} \delta X^S{}_{[n]} \quad (6)$$

where $V(s)$ represents the value for state s, $\eta^{Str \to SNc}$ is the learning rate for $W^{Str \to SNc}$.

The representation for the various actions the agent in state s can perform is given by the activity of the matrisomes surrounding the corresponding striosome neuron for the state. In our model, this is given by the activity of the neurons of the Matri-SOM corresponding to the neuron with the highest activity in the Strio-SOM ($n_s{}^*$) where the activity of a single neuron in the Matri-SOM is given in Equation (2). Thus, the activity is of dimension $m_2 \times n_2$. The action input a is given as feedback input from the thalamus to the striatum (**Figure 2**).

The activity of Matri-SOM neurons is further tuned by the connections between the neurons in the Strio-SOM and the Matri-SOM ($W^{S\to M}$). These connections are also trained using TD error as above using the Matri-SOM activity for the action (a) chosen, as follows:

$$Y^{M}{}_{[n_s^*][n]} = \alpha X^{M}{}_{[n_s^*][n]} + (1-\alpha) W^{S\to M}{}_{[n_s^*][n]} X^{S}{}_{[n_s^*]} \quad (7)$$

$$\Delta W^{S\to M}{}_{[n_S^*][n]} = \eta^{S\to M} \delta X^{M}{}_{[n_S^*][n]} \quad (8)$$

where α controls the contribution of the action and the lateral connections to the activity of the Matri-SOM and $\eta^{Str\to SNc}$ is the learning rate for $W^{S\to M}$. Choosing a low value of α and low initial weights for $W^{S\to M}$ ensures that the activity is driven by the action representation initially and then driven by the lateral weights once the $W^{S\to M}$ have been trained sufficiently. The Strio-SOM/Matri-SOM weights ($W^{S\to M}$) are thresholded and normalized by their sum to ensure stability.

The matrisomes activity is projected to the direct and indirect pathways by the D1 and D2 neurons of the striatum. In our model, the Matri-SOM activity is modulated by a value difference signal (δ_V). If the agent goes from state $s^{(1)}$ to $s^{(2)}$, δ_V is the difference between the value of the two states, i.e., $\delta_V = V(s^{(1)}) - V(s^{(2)})$.

This value difference signal modulates the switching between the direct and indirect pathways and is thought to be represented by the dopamine signaled by SNc (Chakravarthy and Balasubramani, 2015). The activity of the D1 and D2 neurons are given in Equations (9, 10).

$$Y^{D1}{}_{[n]} = f(\lambda_{D1}\delta_V) Y^{M}{}_{[n_s^*][n]} \quad (9)$$

$$Y^{D2}{}_{[n]} = f(\lambda_{D2}\delta_V) Y^{M}{}_{[n_s^*][n]} \quad (10)$$

where f is a tanh non-linearity and λ_{D1} and λ_{D2} are the gains of the D1 and D2 neurons respectively. The indirect pathway consisting of the GPe and STN is modeled as network of coupled non-linear oscillators. The dynamics of these oscillators is highly dependent on the input, which constitutes the projections from the D2-expressing neurons of the striatum. The dynamics of GPe is given below:

$$\tau^{GPe} \frac{dX^{GPe}{}_{[n]}}{dt} = -X^{GPe}{}_{[n]} - \epsilon^{GPe} W^{GPe\to GPe}{}_{[n]} Y^{GPe}{}_{[n]} + W^{STN\to GPe}{}_{[n]} Y^{STN}{}_{[n]} + Y^{D2}{}_{[n]} \quad (11)$$

$$Y^{GPe}{}_{[n]} = \tanh(\lambda^{GPe} X^{GPe}{}_{[n]}) \quad (12)$$

where $W^{GPe\to GPe}$ are the lateral weights within the GPe, ϵ^{GPe} is the connection strength, $W^{STN\to GPe}$ are the connections between STN and GPe, and λ^{GPe} is a non-linear scaling parameter.

The STN layer in the model exhibits correlated activity suppressed for high striatal input, and uncorrelated oscillatory activity for low striatal inputs (Chakravarthy and Balasubramani, 2015). The uncorrelated oscillations of the STN are a key source of exploration for the agent. The dynamics of STN is given below:

$$\tau^{STN} \frac{dX^{STN}{}_{[n]}}{dt} = -X^{STN}{}_{[n]} + \epsilon^{STN} W^{STN\to STN}{}_{[n]} Y^{STN}{}_{[n]} - W^{GPe\to STN}{}_{[n]} Y^{GPe}{}_{[n]} \quad (13)$$

$$Y^{STN}{}_{[n]} = \tanh(\lambda^{STN} X^{STN}{}_{[n]}) \quad (14)$$

where $W^{STN\to STN}$ are the lateral weights within the STN, ϵ^{STN} is the connection strength, $W^{GPe\to STN}$ are the connections between Gpe and STN and λ^{STN} is a non-linear scaling parameter.

The D1 neurons of the striatum and the STN neurons project to the GPi leading to the convergence of the direct and indirect pathways in GPi. In the model, the number of GPi neurons equals number of actions [=dim(**a**)]. The weights $W^{D1\to GPi}$ and $W^{STN\to GPi}$ map the corresponding activities of D1 striatum and STN onto the GPi. The Matri-SOM activity (Y^{D1}) corresponding to the chosen action (a) (which comes via feedback) is used to train the two sets of weights, $W^{D1\to GPi}$ and $W^{STN\to GPi}$ using Hebb's rule. The output of GPi neurons are computed according to Equation (15), and the update for the weights $W^{D1\to GPi}$ and $W^{STN\to GPi}$ are done according to Equations (16, 17).

$$Y^{GPi}{}_{[n']} = W^{D1\to GPi}{}_{[n'][n]} Y^{D1}{}_{[n]} - W^{STN\to GPi}{}_{[n'][n]} Y^{STN}{}_{[n]} \quad (15)$$

$$\Delta W^{D1\to GPi}{}_{[n][n']} = \eta^{D1\to GPi} Y^{D1}{}_{[n]} X^{GPi}{}_{[n']} \quad (16)$$

$$\Delta W^{STN\to GPi}{}_{[n][n']} = \eta^{STN\to GPi} Y^{STN}{}_{[n]} X^{GPi}{}_{[n']} \quad (17)$$

The neurons in the GPi project to the thalamus. In our model, action selection takes place in the thalamus, following the integrator-race model (Bogacz, 2007) with thalamic neurons having self-exciting and mutually inhibiting interactions. The thalamic neuron that first crosses a threshold value (Y_{thresh}) determines the action. The thalamic neurons have low initial random activity which converge to a high activity for the chosen action and low values for the others. The dynamics of thalamic neurons is given as:

$$\dot{Y}^{Thal}_{[n]} = \sum_{n'\in Thal} W^{Thal}{}_{[n][n']} Y^{Thal}{}_{[n']} + Y^{GPi}{}_{[n]} \quad (18)$$

$$a = \{[n] : Y^{Thal}{}_{[n]} > Y_{thresh}\} \quad (19)$$

This action (a) chosen is carried out and the reward (r) is obtained. The action chosen is also projected back to the striatum to obtain the activity. Both the action and the reward are used for updates in Equations (6, 8, 16).

Reinforcement Learning in Environments with Multiple Contexts

Standard reinforcement learning techniques are suited for problems where the environment is stationary. However, in some tasks the environment suddenly changes and the agent has to adopt a policy suitable for the new environment. In such a case, the agent identifies the context either using a cue which is representative of the context or using its experience in the preceding trials. One of the techniques to solve problems of the

second category is the modular RL framework. In this method, the agent allocates separate modules to separate contexts. Each of the modules has its own copy of the environment in a particular context, represented by an environment feature signal (ρ). This copy is used to generate a responsibility signal, denoted by λ, which indicates how close the current context is to the one represented by the module. Thus, by identifying the module with the highest responsibility signal we can follow the policy developed in that module to solve the problem in an efficient manner.

Using the Striatal Modularity to Solve Modular Reinforcement Learning Tasks

The striatum model developed above forms the basic module capable of solving simple RL tasks. Multiple such modules in the striatum could then be exploited to tackle multi-context tasks using modular RL framework. A schematic of this extended model is given in **Figure 3**.

We believe that context selection happens at the level of the striatum and the context modulated activity is projected to the downstream nuclei of the BG for further processing. Thus, for clarity, we have expanded the intra-nuclear activity of the striatum in the model schematic (**Figure 3**). Supposing there are K modules denoted by M_1, M_2..., M_K. We now define the weights and activities in the previous sections for each module and denote $\{M_i\}$ with each term associated with module M_i. Thus, for a module m, the following variables undergo a change in notation: $X^S \rightarrow X^{S,\{m\}}$ (Equation 1), $X^M \rightarrow X^{M,\{m\}}$ (Equation 2), $W^S \rightarrow W^{S,\{m\}}$ (Equation 3), $W^M \rightarrow W^{M,\{m\}}$ (Equation 4), $V(s) \rightarrow V^{\{m\}}(s)$ (Equation 5), $W^{Str\rightarrow SNc} \rightarrow W^{Str\rightarrow SNc,\{m\}}$ (Equation 6), $X^M \rightarrow X^{M,\{m\}}$ (Equation 7), $W^{S\rightarrow M} \rightarrow W^{S\rightarrow M,\{m\}}$ (Equation 8).

We propose that in addition to the value of the state s, the activity of the Strio-SOM also projects to the SNc to represent the environment feature signal ($\rho^{\{m\}}$). The weights of these projections are denoted as $W_\rho^{Str\rightarrow SNc,\{m\}}$ and are trained using the signal from SNc which is representative of context prediction error (δ^*). The corresponding equations are given in Equations (20, 21). The context prediction error is calculated as $\delta^* = r - \rho^{\{m\}}(s)$

$$\rho^{\{m\}}(s) = \sum_{\forall n} {W_\rho}^{Str\rightarrow SNc,\{m\}}{}_{[n]} X^{S,\{m\}}{}_{[n]} \quad (20)$$

$$\Delta {W_\rho}^{Str\rightarrow SNc,\{m\}}{}_{[n]} = {\eta_\rho}^{Str\rightarrow SNc} \delta^* X^{S,\{m\}}{}_{[n]} \quad (21)$$

We believe that the selection of the appropriate module for the context is guided by the striatal interneurons. In our model, the activity of the interneurons represents the responsibility signal for each module, denoted by $\lambda^{\{m\}}$ for module m. In a given state s, the inter-neurons compete among themselves and the one with the highest λ chooses the module responsible for deciding the action in that state. Let the winning module in the state s be denoted by m^*. This module guides the projection to the direct and indirect pathway (Equations 9, 10) as given in

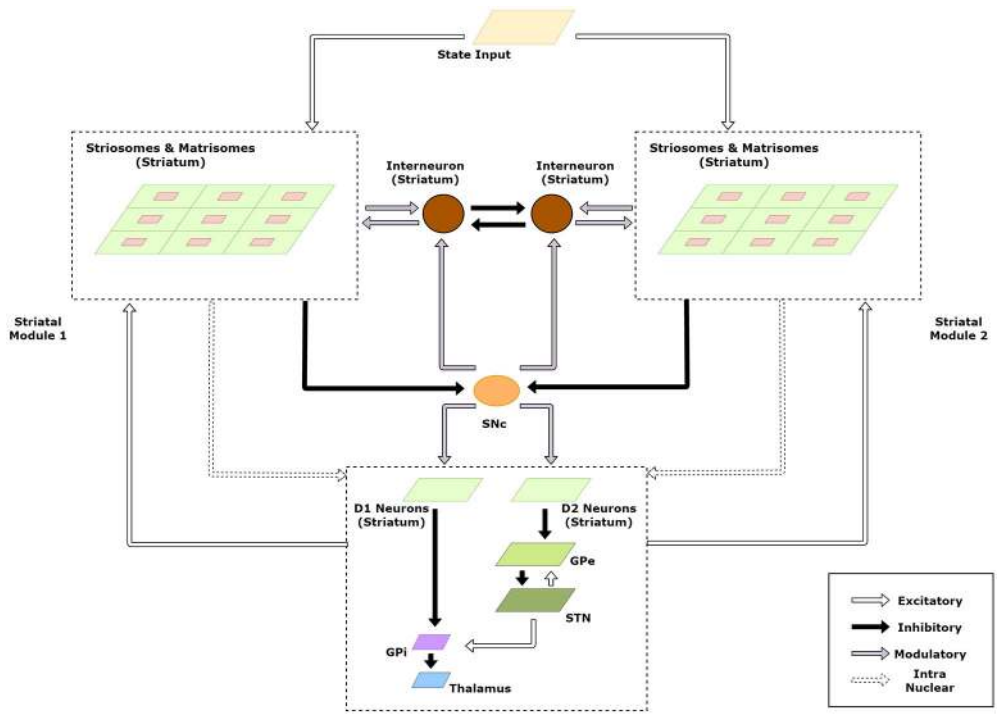

FIGURE 3 | Schematic of the extended model to handle modular RL tasks showing the case with two striatal modules. The state representations of the two modules are used to calculate their respective responsibilities which are then used by the striatal interneurons to choose the appropriate module.

Equations (22, 23).

$$Y^{D1}{}_{[n]} = f(\lambda_{D1}\delta_V)Y^{M,\{m^*\}}{}_{[n_s{}^*][n]} \quad (22)$$

$$Y^{D2}{}_{[n]} = f(\lambda_{D2}\delta_V)Y^{M,\{m^*\}}{}_{[n_s{}^*][n]} \quad (23)$$

Following this stage, the equations governing the signal flow are same as that in the previous section. The weight updates in the striatum are however done only to the module m^*.

The dynamics of the responsibility signal is given in Equation (24)

$$\dot{\lambda} = -\lambda - \alpha_\lambda(\delta^*)^2 \quad (24)$$

where α_λ controls the influence of context prediction error on the responsibility signal and $\delta*$ is the context prediction error.

RESULTS

Modeling the Microanatomy of the Striatum

We use a grid-world problem as a preliminary benchmark to test our model. The grid is of size 10×10 and the agent can take one of the four actions—up, down, right and left in a state. A reward is placed at one of the corners of the maze. The goal of the task is to make the model (agent) learn to reach this goal. We use the terms model and agent interchangeably in these sections since we use the model as a reinforcement learning agent in the various tasks. We used a 10×10 Strio-SOM to represent the state space and a 3×3 Matri-SOM, associated with each of the Strio-SOM neurons, for representing the action space.

In order to develop these representations, we make the agents explore various states and choose random actions in those states. Following this, we look at the neuron with the highest activity in the Strio-SOM for a particular state and the neurons with the highest activity for each action in the corresponding Matri-SOM for that state (**Figures 4A,B**). Upon looking at the combined Matri-SOM activity for all the actions, we observed predominantly two different configurations of the center-surround mapping (**Figures 4C,D**).

Reinforcement Learning in a Single Context Gridworld Task

The goal was placed at the top right of the grid as seen in **Figure 5A**. The agent received a reward of +20 when it reached the goal and 0 for all the other steps. At the beginning of an episode, the agent started at random and the episode ended when

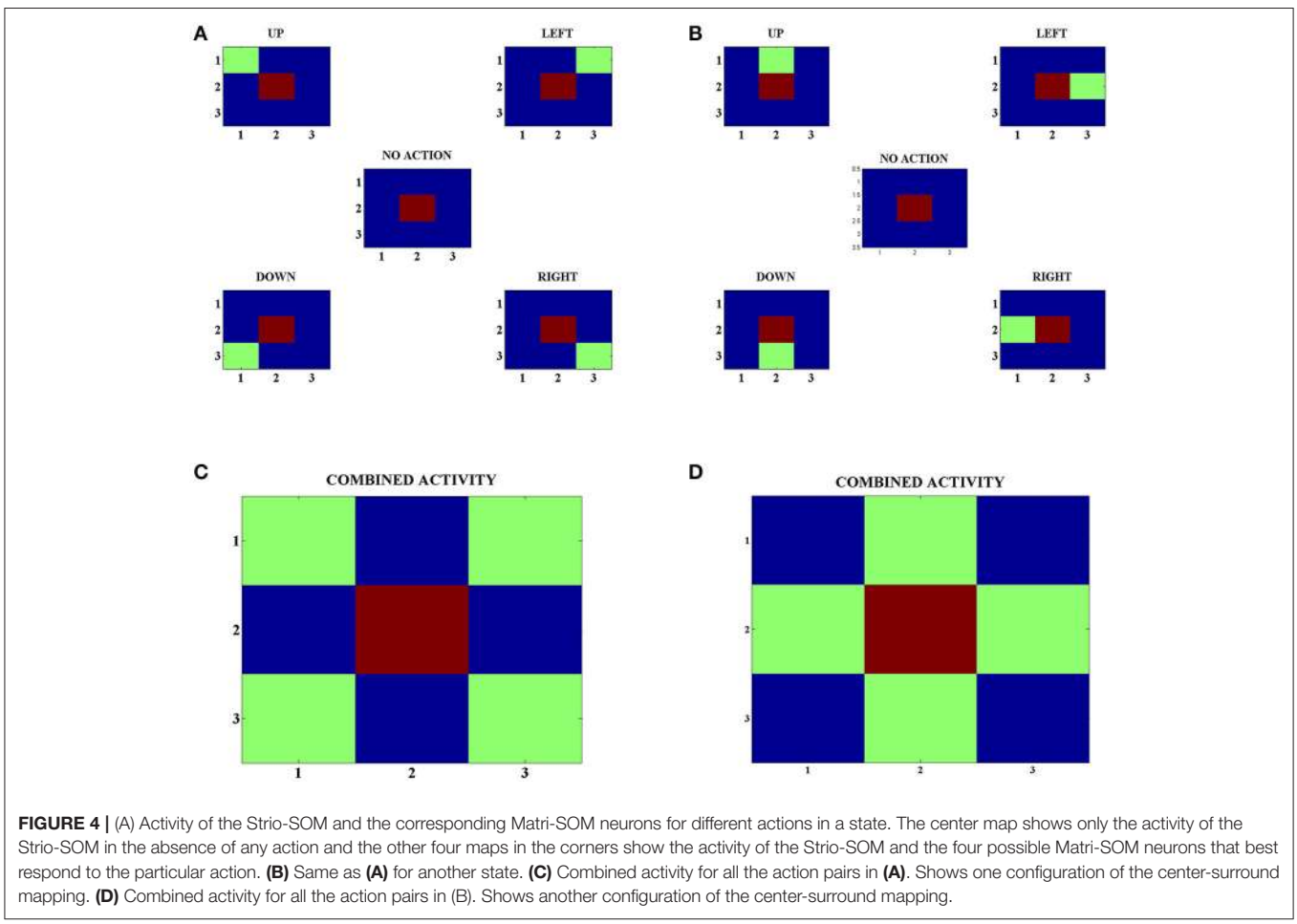

FIGURE 4 | (A) Activity of the Strio-SOM and the corresponding Matri-SOM neurons for different actions in a state. The center map shows only the activity of the Strio-SOM in the absence of any action and the other four maps in the corners show the activity of the Strio-SOM and the four possible Matri-SOM neurons that best respond to the particular action. **(B)** Same as **(A)** for another state. **(C)** Combined activity for all the action pairs in **(A)**. Shows one configuration of the center-surround mapping. **(D)** Combined activity for all the action pairs in (B). Shows another configuration of the center-surround mapping.

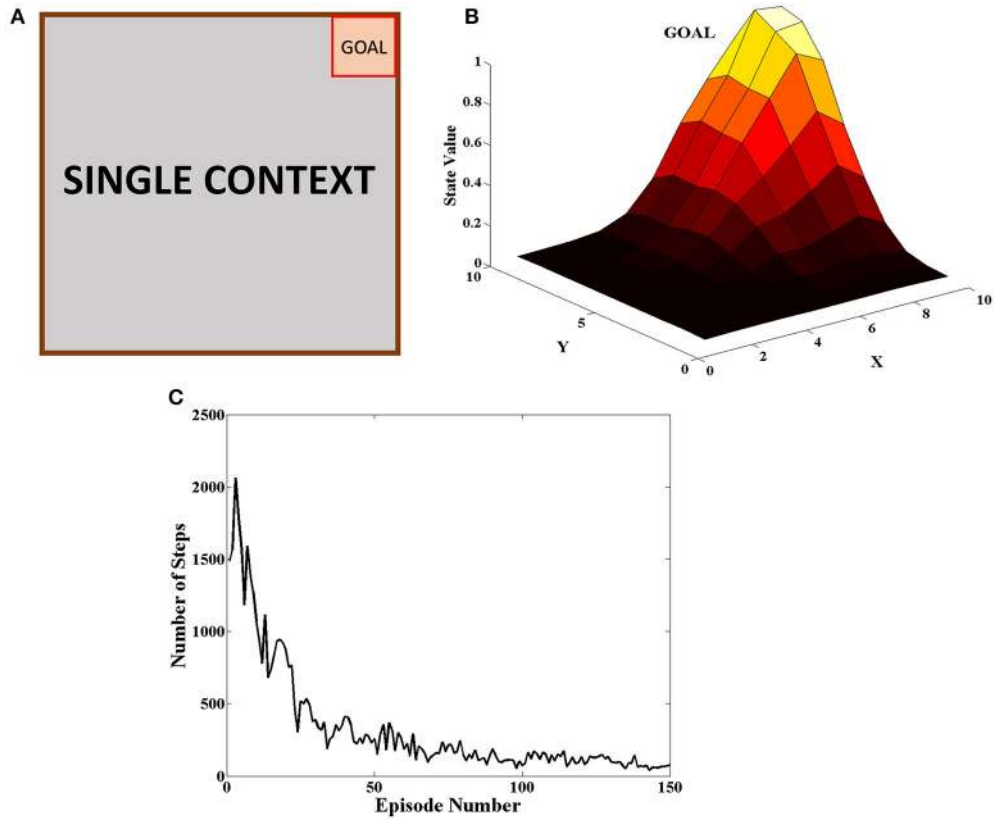

FIGURE 5 | (A) Schematic of the grid-world used in the task. A goal is located at the top right corner of the grid **(B)** State value map estimated by the agent at different spatial locations. We can see that the state value peaks at the goal location. **(C)** Plot of the Number of Steps taken by the agent in each episode averaged across 50 independent sessions. We see that the number of steps reduces as the agent learns across episodes.

the agent reached the goal or when it reached the upper limit on number of steps allowed in the episode. The agent carried on the task for 150 episodes. This procedure was carried out for 50 independent sessions and the mean number of steps to reach the goal in a particular episode was plotted in **Figure 5C**. The heat map of the state value function (Equation 5) estimated by the agent at different spatial locations is given in **Figure 5B** and peaks at the goal location. This combined with the fact that number of steps reduces as the episodes progress indicate that the agent is able to learn the single context task. The various parameter values for this task are given in **Table 1**.

Reinforcement Learning in a Multi-Context Grid-World Problem

In the multi context grid-world tasks, the agent had to reach the goal like the previous section but the goal location changed after a certain number of episodes. The goal was present either at the top right corner or at the bottom left corner as shown in **Figure 6A**. The goal was switched to the other location after 150 episodes. The task was carried out in 50 independent sessions with each session containing 900 episodes. The parameters used have the same values as given in **Table 1**. **Figure 6B** shows the value function (Equation 5) heat map and **Figure 6C** shows the environment feature signal (Equation 20) heat map estimated by the agent for the two contexts. We can observe that the agent is able to learn these values for both the contexts. **Figure 6D** shows the context chosen by the agent in different episodes and we can observe that the agent is able to switch context in sync with the switch in reward distribution. These results illustrate that the agent is able to successfully identify the context it is presently in, and complete the corresponding grid-world task.

The average number of steps required by the agent to reach the goal for each episode across 50 sessions is given in **Figure 7B**. The same plot for an agent with only a single module is given in **Figure 7A**. We can clearly see that the learning is more efficient for multi module agent as compared to the single module case. In order to quantify this improvement, we use two values to measure the agent's performance after a context switch. These are the peak number of steps to reach the goal after a context switch and the number of episodes for the number of steps needed to go below a certain threshold. We calculate these two values for each context switch in a session. These values are averaged across sessions and presented in **Figures 7C,D**, respectively. In both cases, we see that the multi module agent is better than the single module agent for solving the task. We use these measures to compare the model against experimental data in Brunswik (1939). Since we only have the average performance across sessions available in the reference, we calculate the corresponding values from our

TABLE 1 | Parameter values for single context and multi context tasks.

Parameter	Value
Strio-SOM Dimension ($m_1 \times n_1$)	**10 × 10**
σ_S	0.01
η_S	0.4
γ	0.97
α	0.1
λ_{D1}	1
τ^{GPe}	3
ϵ^{GPe}	−0.01
λ^{GPe}	3
$\eta^{D1\rightarrow GPi}$	0.01
γ_{thresh}	1
α_λ	0.8
Matri-SOM Dimension ($m_2 \times n_2$)	**3 × 3**
σ_M	0.1
η_M	0.4
$\eta^{Str\rightarrow SNc}$	0.1
$\eta^{S\rightarrow M}$	0.1
λ_{D2}	−1
τ^{STN}	1
ϵ^{STN}	0.01
λ^{STN}	3
$\eta^{STN\rightarrow GPi}$	0.01
$\eta_\rho^{Str\rightarrow SNc}$	0.1

model and present these for the single module, multi module and the experimental case in **Figures 7E,F**, respectively. We can observe that multi module results have a similar trend to the experimental results as compared to the single module model, thus demonstrating that the BG could be using the modular architecture of the striatum to solve context switching tasks.

DISCUSSION

We have proposed a network model of BG incorporating a computational framework to capture the microanatomy of the striatum. Our model shares features with existing models of BG designed to solve reinforcement learning (RL) tasks. In addition to solving RL tasks, our model exploits the modularity of the striatum to solve tasks with varying reward distributions in multiple contexts.

Striosome-Matrisome Dynamics with Their Dopaminergic Projections

Our model is based on the assumption that striosomes map state information and matrisomes map action information. Earlier results suggest that the striosomes receive input from the orbitofrontal cortex (Eblen and Graybiel, 1995) known for coding reward related states (Wilson et al., 2014). Matrisomes receive connections from primary motor and somatosensory cortices and could have action representations (Flaherty and Graybiel, 1994), thereby supporting the assumptions of our model. Anatomical studies show that striosome medium spiny neurons (MSNs) project directly to SNc (Fujiyama et al., 2011; Lanciego et al., 2012; Smith et al., 2016). We believe that these projections could code for the state value of the agent as seen from the Strio-SOM to SNc connections in our model.

We propose that the striosome neurons influence the behavior of the surrounding matrisome neurons. Earlier results show that fast spiking interneurons (FSIs) and persistent and low-threshold spike (PLTS) interneurons are anatomically suitable candidates for this role since they branch across the patch and matrix (Gittis and Kreitzer, 2012). We believe that the dopaminergic projections to these interneurons (Bracci et al., 2002) could allow the striosome to bias the surrounding matrisome activity toward a preferred action. This is done by the thalamic feedback which drives the matrisome activity to the action chosen which is then reinforced by the prediction error signaled by the SNc. To our knowledge, this modulation (Equation 8) is a unique feature to our model and gives a biologically plausible mechanism to perform Q-learning. This is also supported by experiments which indicate that the striatum contributes to action selection by biasing its output toward the most desirable action (Samejima et al., 2005; Hikosaka et al., 2006).

Mapping Representations to Action Primitives

Striatal MSN recordings show that they encode action representations and are modulated by the expected reward for the actions (Isomura et al., 2013). Our model agrees with this as both the Matri-SOM D1 and D2 neurons represent the action space and are correspondingly modulated by the TD error which is representative of the expected reward. Experiments also show activity in the MSNs corresponding to the outcome of the chosen action (Kim et al., 2009). We believe again that this could be the signal required to bias the activity of the striatal MSN as seen in the model (Equation 8).

GPi forms the output nucleus of the BG and receives projections from Striatal MSNs through the direct and indirect pathways. Lesion studies show that GPi controls movement by inhibitory projections to the thalamus and lesioning GPi impairs motor responses (Baunez and Gubellini, 2010). Experiments also show that in the executive part of the task, the GPi activity is strongly related to the action performed (Pasquereau et al., 2007).

We propose that the connections from striatal D1 MSNs and STN to the GPi map the projections from action representations to action primitives. We believe that this mapping provides a flexible method to switch different action primitives for the same representations and vice versa, providing a plausible mechanism of adaptation in learning. Experiments show evidence of transformation of action information seen as higher degree of correlation in GPi activity as it passes from striatum to the GPi (Garenne et al., 2011).

Contextual Learning and Striatal Modularity

Contextual Learning refers to the ability of the agent to adapt and learn in different contexts. Some earlier operant conditioning experiments in such tasks have an explicit indication of contexts

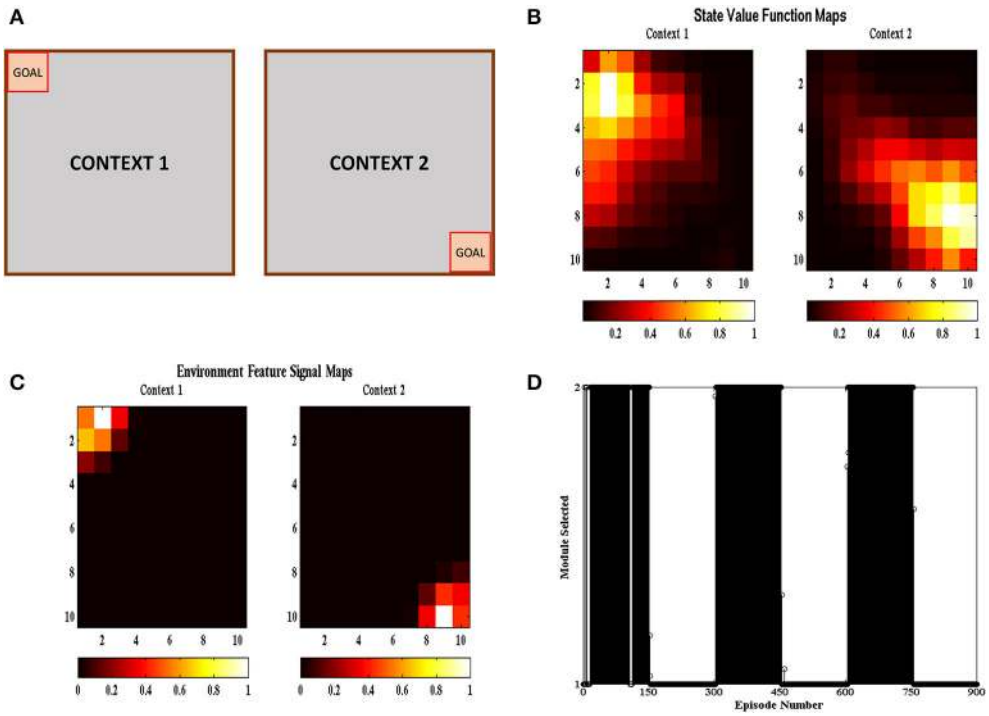

FIGURE 6 | (A) Schematic of the gridworld used in the task. A goal is switched between the top left and bottom right corner every 150 episodes. **(B)** State value map estimated by the agent at different spatial locations across different contexts. We can see that the state value peaks at the goal location corresponding to the context. **(C)** Environment Feature Signal maps estimated by the agent at different spatial locations across different contexts. We can see that the state value peaks at the goal location corresponding to the context. **(D)** Modules chosen by the agent at different episodes. We can see that the module chosen switched with change in context indicating that the agent is able to identify the context it is currently present in.

(different room or color for each context) using which the agent can choose its actions (Bouton and King, 1983; Bouton and Peck, 1989). In such tasks, the agent shows renewal upon context switching indicating a mechanism for context identification. Experimental results indicate that the BG encodes the context as well as the choices in those contexts (Garenne et al., 2011).

A recent study (Amemori et al., 2011) hypothesized that the modular architecture of the striatum makes it a suitable candidate for solving multi-context RL problems. We build on this by providing a computational neural model for the same. We describe the plausible correlates for computing the necessary variables for solving multi-context problems using a modular setting. The context prediction signal is very similar to a state value and we propose that neurons in the SNc code for this signal as well (Tobler et al., 2005). In our model this is represented by the projections from Strio-SOM to the SNc. There is also a need for a reward prediction variance signal or a risk signal. Dopamine in the midbrain is proposed to also represent the risk component in the environment (Schultz, 2010). In addition, it has been proposed that serotonin activity in the striatum correlates to risk or reward variance, just as dopamine codes for reward prediction error (Balasubramani et al., 2015).

We propose that the module selection and switching in different contexts could be carried out by TANs. TANs exert a strong influence on striatal information processing and lesioning inputs to TANs impair learning after a change in reward distribution (Ragozzino et al., 2002). In our model, the TANs compete with each other and select the module appropriate for the task. Experiments support this hypothesis by showing that TANs can compete with each other using inhibitory connections similar to the model (Sullivan et al., 2008) and can cause widespread inhibition of MSNs by activating a GABAergic subpopulation (English et al., 2012). Another plausible method for context switching by TANs is by producing acetylcholine (ACh) which can inhibit targeted MSNs. Dynamic changes in ACh output in the medial striatum (Ragozzino and Choi, 2004) during reversal learning supports this claim.

Behavioral Observations

Several behavioral processes were also observed from the results of the experiments on the model. We saw in **Figure 6B** that the agent increases its down and right actions when the goal is placed at the bottom right corner. The agent thus exhibits acquisition (Graham and Gagné, 1940) since it strengthens certain actions over the others based on the reinforcement given. We saw in **Figure 6D**, that once the context has changed, the agent stops choosing the initial preferred response. This demonstrated extinction (Graham and Gagné, 1940) since the

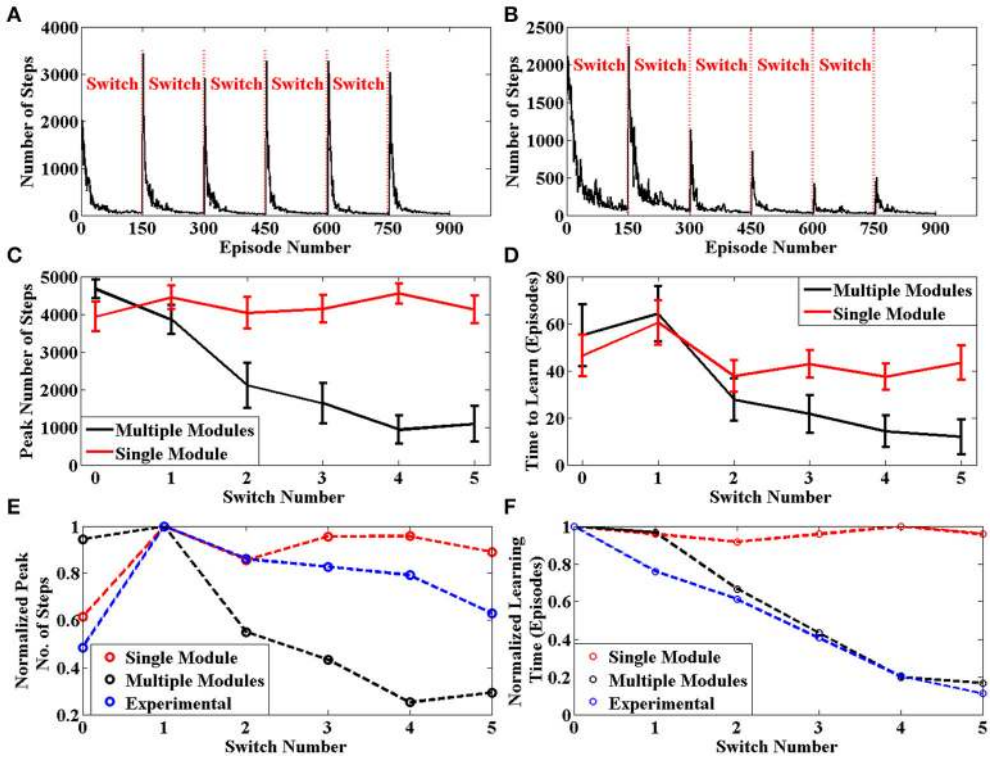

FIGURE 7 | (A) Plot of Number of Steps taken by the single module agent in each episode averaged across 50 independent sessions. We see that the agent needs to relearn after each context switch **(B)** Plot of Number of Steps taken by the multi module agent in each episode averaged across 50 independent sessions. We see that the agent efficiently switches modules after each context switch **(C)** Peak number of steps needed to reach the goal after a context switch averaged across 50 sessions. **(D)** Number of episodes for the number of steps required to reach the goal to go below a certain threshold averaged across 50 sessions **(E)** Peak value for the average number of steps needed to reach the goal after a context switch. The experimental values have been adapted from Brunswik (1939) **(D)** Number of episodes for the average number of steps required to reach the goal to go below a certain threshold. The experimental values have been adapted from Brunswik (1939).

behavior associated with a certain task gets elimininated when the reinforcement associated is removed. The experiments also indicate that the agent is able to show stimulus generalization and stimulus discrimination as the agent is able to distinguish between two different contexts which are two distinct stimuli (Till and Priluck, 2000). Also the value function peaks where the goal is given, therefore goals which are near each other will have similar value profiles. From **Figure 7B**, we saw that after two changes when the initial context reappears, the agent is able to bring back the policy learnt almost immediately exhibiting spontaneous recovery (Graham and Gagné, 1940) referring to the reappearance and faster relearning of a previously extinguished response.

LIMITATIONS AND FUTURE WORK

In a variable environment, there are two types of uncertainty—expected uncertainty/Risk which refers to the uncertainty even after full learning and unexpected uncertainty which is related to a sudden change in the environment. While the latter is tested in our model with the help of context switches, the rewards are certain and this makes the learning and module switching easier. However, the next step would be to look at harder problems where the rewards are also stochastic. In this case, the ability to detect change in context no longer remains trivial and would be an interesting problem to study.

The experimental validation in such tasks becomes very challenging due to the high number of states and trials required. While the grid world is a natural problem for testing RL frameworks, the number of trials continuously for a real animal is limited. Our model requires around 900 trials despite the various simplifying assumptions which is very taxing for the animal. Thus, there is a need to look at simpler tasks where we can test the model and the animal on various intricacies of the problem.

AUTHOR CONTRIBUTIONS

SS, VM: Conceived, developed the model and prepared the manuscript. VC: Conceived the model and prepared the manuscript.

REFERENCES

Amemori, K.-I., Gibb, L. G., and Graybiel, A. M. (2011). Shifting responsibly: the importance of striatal modularity to reinforcement learning in uncertain environments. *Front. Hum. Neurosci.* 5:47. doi: 10.3389/fnhum.2011.00047

Apicella, P. (2007). Leading tonically active neurons of the striatum from reward detection to context recognition. *Trends Neurosci.* 30, 299–306. doi: 10.1016/j.tins.2007.03.011

Balasubramani, P. P., Chakravarthy, V. S., Ravindran, B., and Moustafa, A. A. (2015). A network model of basal ganglia for understanding the roles of dopamine and serotonin in reward-punishment-risk based decision making. *Front. Comput. Neurosci.* 9:76. doi: 10.3389/fncom.2015.00076

Bar-Gad, I., Havazelet-Heimer, G., Goldberg, J. A., Ruppin, E., and Bergman, H. (2000). Reinforcement-driven dimensionality reduction-a model for information processing in the basal ganglia. *J. Basic Clin. Physiol. Pharmacol.* 11, 305–320. doi: 10.1515/jbcpp.2000.11.4.305

Bar-Gad, I., Morris, G., and Bergman, H. (2003). Information processing, dimensionality reduction and reinforcement learning in the basal ganglia. *Prog. Neurobiol.* 71, 439–473. doi: 10.1016/j.pneurobio.2003.12.001

Baunez, C., and Gubellini, P. (2010). Effects of GPi and STN inactivation on physiological, motor, cognitive and motivational processes in animal models of Parkinson's disease. *Prog. Brain Res.* 183, 235–258. doi: 10.1016/S0079-6123(10)83012-2

Bogacz, R. (2007). Optimal decision-making theories: linking neurobiology with behaviour. *Trends Cogn. Sci.* 11, 118–125. doi: 10.1016/j.tics.2006.12.006

Bouton, M. E., and King, D. A. (1983). Contextual control of the extinction of conditioned fear: tests for the associative value of the context. *J. Exp. Psychol. Anim. Behav. Process* 9, 248–265.

Bouton, M. E., and Peck, C. A. (1989). Context effects on conditioning, extinction, and reinstatement in an appetitive conditioning preparation. *Anim. Learn. Behav.* 17, 188–198.

Bracci, E., Centonze, D., Bernardi, G., and Calabresi, P. (2002). Dopamine excites fast-spiking interneurons in the striatum. *J. Neurophysiol.* 87, 2190–2194. doi: 10.1152/jn.00754.2001

Brunswik, E. (1939). Probability as a determiner of rat behavior. *J. Exp. Psychol.* 25, 175. doi: 10.1037/h0061204

Chakravarthy, V. S., and Balasubramani, P. P. (2015). Basal ganglia system as an engine for exploration. *Encyclopedia Comput. Neurosci.* 315–327. doi: 10.1007/978-1-4614-6675-8_81

Charpier, S., and Deniau, J. (1997). *In vivo* activity-dependent plasticity at cortico-striatal connections: evidence for physiological long-term potentiation. *Proc. Natl. Acad. Sci. U.S.A.* 94, 7036–7040. doi: 10.1073/pnas.94.13.7036

Eblen, F., and Graybiel, A. M. (1995). Highly restricted origin of prefrontal cortical inputs to striosomes in the macaque monkey. *J. Neurosci.* 15, 5999–6013.

English, D. F., Ibanez-Sandoval, O., Stark, E., Tecuapetla, F., Buzsáki, G., Deisseroth, K., et al. (2012). GABAergic circuits mediate the reinforcement-related signals of striatal cholinergic interneurons. *Nat. Neurosci.* 15, 123–130. doi: 10.1038/nn.2984

Flaherty, A., and Graybiel, A. M. (1994). Input-output organization of the sensorimotor striatum in the squirrel monkey. *J. Neurosci.* 14, 599–610.

Fujiyama, F., Sohn, J., Nakano, T., Furuta, T., Nakamura, K. C., Matsuda, W., et al. (2011). Exclusive and common targets of neostriatofugal projections of rat striosome neurons: a single neuron-tracing study using a viral vector. *Eur. J. Neurosci.* 33, 668–677. doi: 10.1111/j.1460-9568.2010.07564.x

Garenne, A., Pasquereau, B., Guthrie, M., Bioulac, B., and Boraud, T. (2011). Basal ganglia preferentially encode context dependent choice in a two-armed bandit task. *Front. Syst. Neuroscience* 5:23. doi: 10.3389/fnsys.2011.00023

Gittis, A. H., and Kreitzer, A. C. (2012). Striatal microcircuitry and movement disorders. *Trends Neurosci.* 35, 557–564. doi: 10.1016/j.tins.2012.06.008

Graham, C., and Gagné, R. M. (1940). The acquisition, extinction, and spontaneous recovery of a conditioned operant response. *J. Exp. Psychol.* 26, 251. doi: 10.1037/h0060674

Granger, R. (2006). Engines of the brain: The computational instruction set of human cognition. *AI Mag.* 27, 15–32. Available online at: http://dl.acm.org/citation.cfm?id=1167633

Graybiel, A. M., Aosaki, T., Flaherty, A. W., and Kimura, M. (1994). The basal ganglia and adaptive motor control. *Science* 265, 1826–1826. doi: 10.1126/science.8091209

Graybiel, A., Flaherty, A., and Gimenez-Amaya, J.-M. (1991). "Striosomes and matrisomes," in *The Basal Ganglia III,* eds M. B. Carpenter, G. Bernardi, G. Di Chiara (New York, NY: Springer), 3–12.

Hikosaka, O., Nakamura, K., and Nakahara, H. (2006). Basal ganglia orient eyes to reward. *J. Neurophysiol.* 95, 567–584. doi: 10.1152/jn.00458.2005

Isomura, Y., Takekawa, T., Harukuni, R., Handa, T., Aizawa, H., Takada, M., et al. (2013). Reward-modulated motor information in identified striatum neurons. *J. Neurosci.* 33, 10209–10220. doi: 10.1523/JNEUROSCI.0381-13.2013

Kalmár, Z., Szepesvári, C., and Lörincz, A. (1999). Modular reinforcement learning. *Acta Cybernetica* 14, 507–522.

Kim, H., Sul, J. H., Huh, N., Lee, D., and Jung, M. W. (2009). Role of striatum in updating values of chosen actions. *J. Neurosci.* 29, 14701–14712. doi: 10.1523/JNEUROSCI.2728-09.2009

Kohonen, T. (1990). The self-organizing map. *Proc. IEEE* 78, 1464–1480.

Lanciego, J. L., Luquin, N., and Obeso, J. A. (2012). Functional neuroanatomy of the basal ganglia. *Cold Spring Harb. Perspect. Med.* 2:a009621. doi: 10.1101/cshperspect.a009621

Pasquereau, B., Nadjar, A., Arkadir, D., Bezard, E., Goillandeau, M., Bioulac, B., et al. (2007). Shaping of motor responses by incentive values through the basal ganglia. *J. Neurosci.* 27, 1176–1183. doi: 10.1523/JNEUROSCI.3745-06.2007

Ragozzino, M. E., and Choi, D. (2004). Dynamic changes in acetylcholine output in the medial striatum during place reversal learning. *Learn. Mem.* 11, 70–77. doi: 10.1101/lm.65404

Ragozzino, M. E., Jih, J., and Tzavos, A. (2002). Involvement of the dorsomedial striatum in behavioral flexibility: role of muscarinic cholinergic receptors. *Brain Res.* 953, 205–214. doi: 10.1016/S0006-8993(02)03287-0

Samejima, K., Ueda, Y., Doya, K., and Kimura, M. (2005). Representation of action-specific reward values in the striatum. *Science* 310, 1337–1340. doi: 10.1126/science.1115270

Schultz, W. (2010). Dopamine signals for reward value and risk: basic and recent data. *Behav. Brain funct.* 6:24. doi: 10.1186/1744-9081-6-24

Smith, J. B., Klug, J. R., Ross, D. L., Howard, C. D., Hollon, N. G., Ko, V. I., et al. (2016). Genetic-based dissection unveils the inputs and outputs of striatal patch and matrix compartments. *Neuron* 91, 1069–1084. doi: 10.1016/j.neuron.2016.07.046

Sullivan, M. A., Chen, H., and Morikawa, H. (2008). Recurrent inhibitory network among striatal cholinergic interneurons. *J. Neurosci.* 28, 8682–8690. doi: 10.1523/JNEUROSCI.2411-08.2008

Sutton, R. S., and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge: MIT press.

Till, B. D., and Priluck, R. L. (2000). Stimulus generalization in classical conditioning: an initial investigation and extension. *Psychol. Market.* 17, 55–72. doi: 10.1002/(SICI)1520-6793(200001)17:1<55::AID-MAR4>3.0.CO;2-C

Tobler, P. N., Fiorillo, C. D., and Schultz, W. (2005). Adaptive coding of reward value by dopamine neurons. *Science* 307, 1642–1645. doi: 10.1126/science.1105370

Wall, N. R., De La Parra, M., Callaway, E. M., and Kreitzer, A. C. (2013). Differential innervation of direct-and indirect-pathway striatal projection neurons. *Neuron* 79, 347–360. doi: 10.1016/j.neuron.2013.05.014

Wilson, R. C., Takahashi, Y. K., Schoenbaum, G., and Niv, Y. (2014). Orbitofrontal cortex as a cognitive map of task space. *Neuron* 81, 267–279. doi: 10.1016/j.neuron.2013.11.005

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.