Neurocomputing 548 (2023) 126377

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Reinforcement learning in a spiking neural model of striatum plasticity

Álvaro González-Redondo [a,*], Jesús Garrido [a], Francisco Naveros Arrabal [a], Jeanette Hellgren Kotaleski [b], Sten Grillner [b], Eduardo Ros [a]

[a] Research Centre for Information and Communications Technologies (CITIC-UGR), University of Granada, Calle Periodista Rafael Gómez Montero 2, E18071 Granada, Spain
[b] Kungliga Tekniska Högskolan, SE-100 44 Stockholm, Sweden

ARTICLE INFO

Article history:
Received 26 February 2022
Revised 16 May 2023
Accepted 22 May 2023
Available online 26 May 2023

Keywords:
Striatum
Reinforcement learning
Spiking neural network
Dopamine
Eligibility trace
Spike-timing-dependent plasticity

ABSTRACT

The basal ganglia (BG), and more specifically the striatum, have long been proposed to play an essential role in action-selection based on a reinforcement learning (RL) paradigm. However, some recent findings, such as striatal spike-timing-dependent plasticity (STDP) or striatal lateral connectivity, require further research and modelling as their respective roles are still not well understood. Theoretical models of spiking neurons with homeostatic mechanisms, lateral connectivity, and reward-modulated STDP have demonstrated a remarkable capability to learn sensorial patterns that statistically correlate with a rewarding signal. In this article, we implement a functional and biologically inspired network model of the striatum, where learning is based on a previously proposed learning rule called spike-timing-dependent eligibility (STDE), which captures important experimental features in the striatum. The proposed computational model can recognize complex input patterns and consistently choose rewarded actions to respond to such sensorial inputs. Moreover, we assess the role different neuronal and network features, such as homeostatic mechanisms and lateral inhibitory connections, play in action-selection with the proposed model. The homeostatic mechanisms make learning more robust (in terms of suitable parameters) and facilitate recovery after rewarding policy swapping, while lateral inhibitory connections are important when multiple input patterns are associated with the same rewarded action. Finally, according to our simulations, the optimal delay between the action and the dopaminergic feedback is obtained around 300 ms, as demonstrated in previous studies of RL and in biological studies.

1. Introduction

Animals learn to choose actions among many options by trial and error, thanks to the feedback provided by sparse and delayed rewards. Reinforcement learning (RL) serves as a theoretical framework for an agent, a system that acts based on received feedback, to learn to map situations to actions. This state-action mapping aims to maximize the performance of actions, mainly (but not exclusively) considering how rewarding or punishing the consequences of the actions are [52]. The basal ganglia (BG), a group of forebrain nuclei, are posited to play a critical role in action-selection based on RL [22,21,25,23]. However, the roles of recent findings, such as striatal spike-timing-dependent plasticity (STDP) models and striatal asymmetrical lateral connectivity, remain unclear. Investigating these interactions could improve our comprehension of the BG's role in RL, potentially leading to the development of more efficient bio-inspired reinforcement learning agents.

This study aims to explore the impact of homeostatic mechanisms and asymmetric lateral inhibitory connections on action-selection in the striatum. We use the RL framework to gain insights into the neural basis of decision-making and contribute to more biologically plausible basal ganglia models. Our model stands out from previous models in several ways: it does not require a critic or extra circuitry for a temporal difference signal, thereby simplifying the model and reducing computational complexity; additionally, it employs a spiking neural network with spike-time pattern representation that adapts well to varying pattern complexities in the pattern classification layer.

We propose a functional, biologically inspired striatum network model that incorporates dopamine-modulated spike-timing-dependent eligibility (STDE, [24] and asymmetric lateral connectivity [6]. This model improves upon existing striatum models by integrating homeostatic mechanisms, asymmetric lateral inhibitory connections, and the STDE learning rule, capturing essential experimental features found in the striatum.

* Corresponding author.
E-mail address: alvarogr@ugr.es (Á. González-Redondo).

https://doi.org/10.1016/j.neucom.2023.126377
0925-2312/

In this article, we present a model that effectively processes complex input patterns in the context of reinforcement learning. We conduct multiple analyses to assess the interaction between the learning rule, homeostatic mechanisms, and lateral inhibitory connectivity patterns. By incorporating these elements, we strive to develop a comprehensive and biologically plausible striatum model that offers valuable insights. Our study examines the individual and combined effects of these factors, shedding light on the unique topology of the striatum network and its role in reinforcement learning tasks.

The main contributions and findings of this work are:

- A functional and biologically inspired network model of the striatum that integrates dopamine-modulated STDE, homeostatic mechanisms, and asymmetric lateral inhibitory connectivity, providing a more comprehensive and biologically plausible representation of the striatum's function.
- Analysis of the role of homeostatic mechanisms in making learning more robust and facilitating recovery after rewarding policy swapping.
- Investigation of the importance of lateral inhibitory connections when multiple input patterns are associated with the same rewarded action.
- The use of a spiking neural network with spike-time pattern representation that scales well with different pattern complexity, making the model suitable for a wide range of reinforcement learning tasks.
- Demonstration that the optimal delay between action and dopaminergic feedback occurs around 300 ms, which is consistent with previous reinforcement learning and biological studies.
- A model that does not require a critic, simplifying the learning process and reducing the need for additional circuitry.

1.1. Basal ganglia circuitry and striatal connectivity in decision making

The BG network is composed of several structures, grouped in inputs [being the striatum the best known, and populated by medium spiny neurons (MSN)], intermediate layers [the external segment of the globus pallidus (GPe), and the substantia nigra pars compacta (SNc)] and output [substantia nigra pars reticulata (SNr)]. The information flows segregated through the BG circuits [9,42]. It has been proposed that the BG process a large number of cognitive streams or channels in parallel [23], each of them representing a feasible action to be performed [51]. According to recent research, this segregation through the entire cortico-BG-thalamic loop shows a very high specificity, down to almost neuron-to-neuron level [30,11]. Thus, it seems feasible to impact behavior at different levels of detail. However, with the current biological evidence it is not exactly known how the activation of a channel maps to the corresponding behavior and we just assume here that these channels involve a decision making process.

The striatum, as the primary input of the basal ganglia, connects to the SNr via direct and indirect pathways, which are traditionally thought to promote and inhibit behavior, respectively. Each pathway crosses the striatum through different subpopulations of MSNs, expressing dopamine receptors D1 for the direct pathway and D2 for the indirect pathway. Recent genetic and optical studies on striatal circuits have allowed for testing classical ideas about the functioning of this system, but new models are needed to better understand the role of the striatum in learning and decision-making [8].

1.2. Spiking neural networks: learning, reward modulation, and striatal connectivity

In recent decades, the use of biologically plausible computational models composed of spiking neurons able to learn a target function has demonstrated being increasingly successful [53,54]. These models use discrete-time events (spikes) to compute and transmit information. As the specific timing of spikes carry relevant information in many biological contexts, these models are useful to understand how the brain computes at the neuronal description level. Combined with the use of local learning rules, these models can be implemented in highly efficient, low-power, neuromorphic hardware [44]. Within this framework, learning from past experiences can be achieved using the STDP learning rule, a synaptic model featuring weight adaptation that has been observed in both biological systems [33] and the BG [10]. The STDP also was demonstrated to be competitive in unsupervised learning of complex pattern recognition tasks [37,16]. The complexity of the patterns comes from their statistically equivalent activity level and from being immersed within a noisy stream of hundreds or thousands of inputs. These studies shown that an oscillatory stream of inputs reaching a population of spiking neurons enables a target post-synaptic neuron equipped with STDP to detect and recognize the presence of repetitive current patterns [37]. The added oscillatory drive performs a current-to-phase conversion: the neurons that receive the most potent static current will fire the first during the oscillation cycle. This mechanism locks the phase of the spike time, facilitating the recognition of the previously presented patterns.

However, STDP-based learning systems tend to use statistical correlations to strengthen synaptic connections, resulting in the selection of the most frequent patterns at the expense of the most rewarding [16]. Thus, the STDP rule can be modified to drive the learning of patterns that statistically correlate with a reward signal [31,32]. In biological systems, unexpected rewards signal relevant stimuli during learning by releasing dopamine (DA). More specifically, the reward signal is linked to the phasic modulation of dopaminergic neurons in the SNc and ventral tegmental area [47], that sends reinforcement signals to the striatal neurons. These rewards do not need to happen instantly after the relevant stimulus; they can be delayed seconds, resulting in the distal reward and temporal credit assignment problems. In [31,32], the authors suggest a reward-modulated STDP rule that enables a neuron to detect rewarded input patterns lasting milliseconds, even if the reward is delayed by seconds, by using the so-called eligibility trace. Also, based on the eligibility trace, [24] developed a synaptic learning rule called Spike-Timing-Dependent Eligibility (STDE) based on physiological data that captures many features found in the biological MSN of the basal ganglia. This model is more flexible than the previous STDP-like rules as different learning kernels can be used depending on the amount and type (reward or punishment) of reinforcement received. Although the authors did not include some important BG features like the GPe nucleus or a cortico-striatal loop, their model successfully learned to select an action channel driven by stronger cortical input, based only on the timing of the input and the reward signal.

Another relevant feature of the striatum is its connectivity. [6] proposed a model of asymmetric lateral connectivity in the striatum that tries to explain how different clusters of striatal neurons interact and which role they play in information processing. This model accounts for the *in vivo* phenomenon of co-activation of sub-populations of D1 or D2 MSNs, which seems paradoxical as each subpopulation projects to behaviorally opposite pathways (direct and indirect, respectively). This structured connectivity pattern is determined by lateral inhibition between neurons that belong to the same channel and between neurons within different

channels but accounting for the same receptor type (D1 or D2). The authors also include asymmetrical connections with more intensive intra-channel inhibition from D2 to D1 neurons than in the opposite direction. This pattern resulted in synchronized phase-dependent activation between MSN D1 and D2 neuron groups that belong to different channels.

1.3. Contribution

All the previous ideas are important pieces of the process of goal-oriented learning but further research is required as their respective roles and how they complement each other are still not well understood. The combination of the STDE rule within a network with asymmetrically structured lateral inhibition has not been studied before, and some relevant conclusions emerge from this specific study. In this article, we design and study a functional and biologically inspired model of the striatum. Our approach is based on spike time representation of complex input patterns and integrates dopamine modulated STDE and asymmetric lateral connectivity, among other mechanisms. This model learns to select the most rewarding action to complex input stimuli through RL. The proposed model has been demonstrated to be capable of recognizing input patterns relevant for the task and consistently choosing rewarded actions in response to that input. We performed numerous analyses to measure and better understand the interaction between the learning rule with homeostatic mechanisms and the lateral inhibitory connectivity patterns. By measuring the single and combined effects of these factors in the learning process, we want to shed light on how the particular topology of the striatum network facilitates the resolution of RL tasks.

2. Methods

Aiming to implement a RL framework in a biologically plausible striatum model, we started designing a task where the agent has to learn how to map different input patterns into actions based on the reward signal delivered by the environment. We implemented a network model of the striatum capable of learning this task. This system behaves like a RL agent and can solve action-selection tasks.

The methods section is structured as follows: we first define the neuron and synapse models, input pattern generation, and networks structures used in our experiments. Then we describe the experimental design used with the network model and how we measure its learning capability. In Supplementary Materials we also explain both a previous experiment and a simpler model we made to test the viability of the combination of oscillatory inputs, STDE and homeostatic rules that we employed in the final network model.

2.1. Computational models

2.1.1. Neuron models

We used conductance-based versions of the Leaky-Integrate and Fire (LIF) neuron model [17] as it is computationally efficient and captures certain biological plausibility. We use this model in every layer of the network, but with different parameters. We classify the neuron types according to the layer they belong to: cortical neurons for the input, striatal neurons (divided in two subpopulations according to which DA receptor express, D1 or D2) for the learning layer, and action neurons for the output. There is also a dopaminergic neuron that receives the rewards and punishments. The parameters used for each type were manually tuned to obtain reasonable firing rates. For the cortical neurons we used a number of spikes per input cycle (with 8 cycles per second) close to [37] and [16] (see details about the input protocol in Section 2.1.2). For the striatal neurons, we tuned the parameters to obtain a mean firing rate of around one spike per second to be within biological ranges [39] but with activity peaks of two or three spikes per input cycle (16–24 spikes per second). The action neurons (an integrative population that outputs the agent's behavior) are tuned to fire every input cycle if they receive enough stimulation from its channel (at least two more spikes from D1 neurons than D2 neurons each cycle). The dopamine neuron was tuned to have a firing range from 50 to 350 spikes per second, with these unrealistic values chosen for performance (instead of simulating a bigger dopaminergic population). The parameters used for each neuron type are shown at supplementary table 1.

2.1.2. Input and oscillatory drive

In the input generation procedure [37,16] we consider a trial as a segment of time of the simulation where we present some input stimuli to the network. The length of each trial is taken from a uniform random distribution between 100 and 500 ms. An input stimulus represents a combination of 2000 input current values conveyed one-to-one to a set of cortical neurons of the same size (Fig. 8A). An input *pattern* is a combination of current values which target precisely the same cortical neurons every time the input pattern is presented for the entire simulation. For every time bin, one or no pattern is presented. Only half of the cortical neurons (1000) are pattern-specific when presenting a specific pattern, while the other half receives random current values. The cortical neurons specific for each pattern are selected at the initialization. When no pattern is presented, all the cortical neurons receive random current values. Two thousand current-based LIF cortical neurons transform the input current levels into spike activity. These neurons have a firing rate between 8 to 40 spikes per second due to the sum of the input current values (ranged from 87% to 110% of the cortical neuron rheobase currents) and an oscillatory drive at 8 Hz feeding these neurons (with an amplitude of 15% of the rheobase current of the cortical neurons). This oscillatory drive turns the input encoding from analogical signal to phase-of-firing coding [37] by locking the phase of the cortical spikes within the oscillatory drive, as shown in Fig. 8B. By using these parameters, the cortical neurons fire between 1 and 5 spikes per cycle.

2.1.3. Spike-timing-dependent eligibility (STDE) learning rule

We implemented a version of the STDE learning rule [24], a phenomenological model of synaptic plasticity. This rule is similar to STDP, but the kernel constants are DA-dependant (that is, different values are defined for low DA and high DA values, and interpolated for DA values in-between, as shown in Fig. 1 and Supplementary Fig. 9Ai and Aii). STDE is derived from *in vitro* data and predicts changes in direct and indirect pathways during the learning and extinction of single actions. Throughout, we used the following parameters and procedures unless we specified otherwise. The kernel shape is defined by the parameters k_{DA}^{SPK} with $SPK \in \{+,-\}$ being the spike order pre-post for applying k_{DA}^{+} and post-pre for applying k_{DA}^{-}, respectively, and $DA \in \{hi, lo\}$ being the high- or low-DA cases, resulting in four parameters in total: $k_{hi}^{+}, k_{lo}^{+}, k_{hi}^{-}$ and k_{lo}^{-}. We obtained these learning kernel constant values by hand-tuning for both MSN D1 and D2 cases (see Supplementary Fig. 9 and supplementary Table 2). As in the classic STDP learning rule, the weight variation in STDE is calculated for every pair of pre- and post-synaptic spikes and decays exponentially with the time difference between the spikes (Fig. 1). We use time constants $\tau = 32$ms and the weights values are clipped to $[0, 0.075]$.

Our implementation of STDE uses elegibility traces that decay exponentially to store the potential weight changes, similarly to

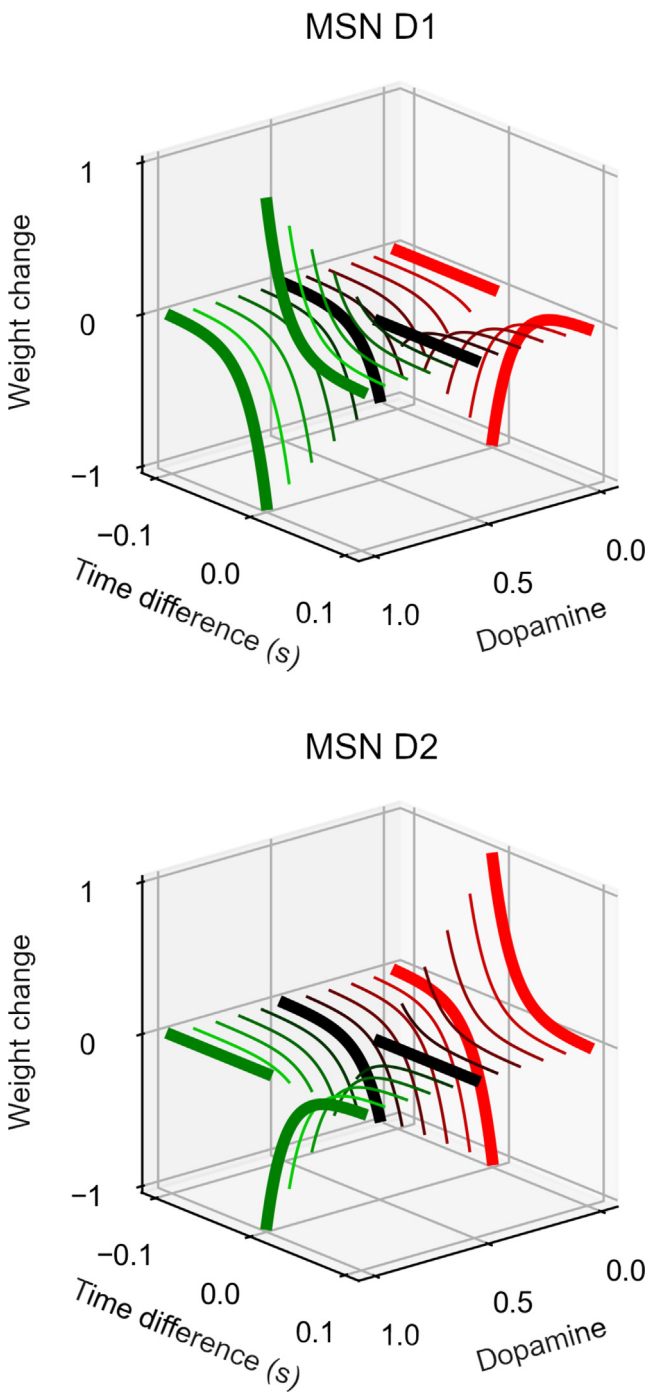

Fig. 1. Kernels used for STDE synapses of MSN D1 (top) and D2 (bottom), showing the weight change depending on the time difference between pre- and post-synaptic spikes and dopamine. Thick lines represent kernels at dopamine minimum, normal, and maximum values (red, black, and green, respectively). Thin lines are interpolations of these values.

[31]. Following [24] we have two different eligibility traces per synapse, c^+ and c^- for spike pairs with positive and negative timing respectively, updated for every pair of pre- and post-synaptic spikes at times t_j and t_i as in Eqs. (1) and (2):

$$\delta c^+ = \left(\alpha k_{hi}^+ + \overline{\alpha} k_{lo}^+\right) \cdot e^{\frac{t_j - t_i}{\tau_{eli}}} \text{ if } t_j \leqslant t_i \tag{1}$$

$$\delta c^- = \left(\alpha k_{hi}^- + \overline{\alpha} k_{lo}^-\right) \cdot e^{\frac{t_j - t_i}{\tau_{eli}}} \text{ if } t_j > t_i \tag{2}$$

with $\overline{\alpha} = 1 - \alpha$, α been a value dependent of DA that we define in Eq. 3, and τ_{eli} been the eligibility trace time constant with a value twice the length of the mean reward delay. Overall plastic change at a single synapse is then the sum of contributions from both c^+ and c^-, scaled by a learning rate factor $\eta = 0.002$.

The level of DA in the system is determined by one neuron that fires at high (and unrealistic) rates for computational simplicity, representing a population of neurons from the SNc. This neuron fires spontaneously at a baseline frequency of 200 Hz. The environment (i.e., the application of rewarding policies during the experiment) injects positive (or negative) current in the dopaminergic neuron when rewards (or punishments) are applied to the model, resulting in the firing rate of this neuron ranging between 50 Hz and 350 Hz. All plastic synapses share a global DA level d that decays exponentially with temporal constant $\tau_{da} = 20$ms. For each spike emitted by the dopaminergic neuron, d is increased by $\frac{1}{\tau_{da}}$ with 200-ms delay.

Our implementation of STDE uses the linear mixing function α in Eq. (3), clipped to $[0, 1]$, to smoothly morph between kernels with low and high DA:

$$\alpha = \frac{d - d_{min}}{d_{max} - d_{min}} \tag{3}$$

where d_{min} and d_{max} are the minimum and maximum values of DA considered. We use this equation for computational efficiency instead of the Naka-Rushton function used in [24] (the authors also noted that this is not a requirement, as long as the mixing function was increasingly monotonic and saturating). The function is bounded to values of DA firing rate between 50 and 350 Hz, with the baseline at 200 Hz.

2.1.4. Homeostatic mechanisms

During learning, in some cases, the neurons can stop firing indefinitely due to a learning history leading to the wrong parameters. Neuron activity can also die by sudden changes in the reward policy, leaving the state of the synaptic weights ill (not representing any stimuli and not getting enough input to fire by chance). To recover neurons from this state, we added two different homeostatic mechanisms, one at the synaptic level and one at the neuron level. Although one or the other is enough to avoid the ill-states, we saw in our tests that we recovered faster and more reliably by using both.

The synapses implementing the STDE included a non-Hebbian strengthening in response to every pre-synaptic spike. For each arriving spike, the synaptic weight increases by $C_{pre} = \eta \cdot 4 \cdot 10^{-4}$. This non-Hebbian strengthening is added to enable the recovery of low-bounded synapses (e.g., after a rewarding policy switch). Although the rewarding policy does not change in the network experiment, this homeostatic mechanism also benefits the complete network model learning (more details in Section 5.2.2 and Supplementary Fig. 14).

In order to avoid neurons to become permanently silent during learning, we include adaptive threshold to our neuron models based on [15] according to the following equation:

$$\frac{dV_{th}}{dt} = -\frac{V_{th} - E_{leak}}{\tau_{th}} \tag{4}$$

where V_{th} represents the firing threshold at the current time, E_{leak} is the resting potential of the neuron, and τ_{th} is the adaptive threshold time constant. According to Eq. 4, in the absence of action potentials, the threshold progressively decreases towards the resting potential, facilitating neuron firing. When the neuron spikes, the firing threshold increases a fixed step proportional to the constant C_{th} as indicated in Eq. 5, making neuron firing more sparse.

$$\delta V_{th} = \frac{C_{th}}{\tau_{th}} \tag{5}$$

2.1.5. Striatum network model

The network model of the striatum (Fig. 3A) contains two channels (channel A and channel B, each one representing a possible action). Every channel contains two same-sized subpopulations (D1 and D2 neurons, respectively) of striatal-like neurons (in total, 16 neurons per channel) and one so-called *action neuron* that integrates excitatory activity from D1 neurons and inhibitory activity from D2 neurons. This design simplifies the biological substrate in which all MSN are inhibitory, but we implemented the network computation by considering the net effect of each neuron type on behavior. Biological MSN D1 neurons inhibit SNr, which promotes behavior, and MSN D2 neurons inhibit GPe, which, in turn, inhibit SNr with the total effect of decreasing behavior (Fig. 3A).

Our striatum model implements lateral inhibition within each MSN D1 population, within each MSN D2 population, between MSN D1 and MSN D2 populations within the same channel, and between the MSN populations associated with different action channels. Inspired by [6], we used an asymmetrical structured pattern of connectivity (Fig. 5E in [6], and adapted here in Fig. 2). Following this connectivity pattern, we added lateral inhibition between neurons that belong to the same channel and between those that belong to different channels but use the same dopaminergic receptor D1 or D2 (with stronger inhibition from D2 to D1 neurons than in the opposite direction). Since the small size of the network under study and the small weight of the D1 to D2 MSN connections, the overall contribution of these connections was neglectable, so we decided not to include them in our simulations as we see no significant impact on previous simulations.

The environment generates the reinforcement signal based on comparing the chosen and the expected action and then delivers it to the dopaminergic neuron. Rewards are excitatory, and punishments are inhibitory inputs to this neuron. The dopaminergic modulatory signal is global and delivered to every STDE connection from cortical layer to striatal neurons (Fig. 3A). It is important to note that this model does not implement a critic (commonly used in actor-critic frameworks [52]), so there is no reward prediction error signal.

2.2. Experimental design

We first validated the proposed learning mechanisms with a simpler network model of only one neuron and a easier experimental task, as can be seen in Supplementary Methods 5.1 and Supplementary Results 5.2.

The action-selection task used to test the model (Fig. 3B) works as follows: the agent has two possible actions to choose, A or B. An action is selected if the activity balance of its D1 and D2 neurons is biased to D1 in two spikes at least in one cycle (making the corresponding action neuron spike). The agent can do none, both, or any of them at a time. The input stream contains five different non-overlapping input patterns, each one presented 16% of the time (80% in total). The policy used to give rewards (excitation) and punishments (inhibition) to the agent (dopaminergic neuron) is the following. When pattern 1 or 2 is present, the agent is rewarded if action A is selected (action A neuron fires during the pattern presentation and action B neuron does not fire) but punished if action B is selected. When pattern 3 or 4 is present, the agent is rewarded if action B is selected but punished if action A is selected. When pattern 5 is present, the agent is punished if it selects action A or B. This policy applies no punishment or reward to the agent during noisy inputs, whatever the action taken is. In case of spiking both action neurons during a reinforced input, the network is punished.

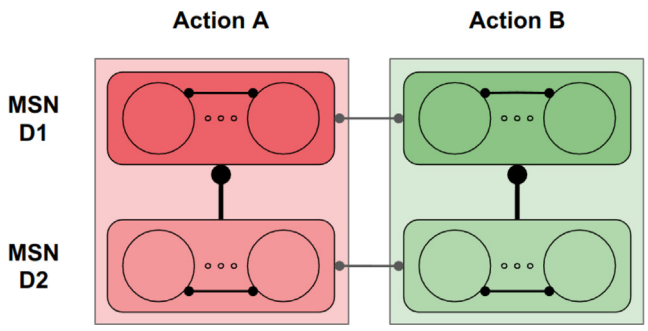

Fig. 2. Connectivity pattern used for the lateral inhibition, inspired on [6]. Two channels (action A and action B) are shown, each with two populations of D1 and D2 MSN.

2.3. Performance measurement

In the action-selection task we measure the performance of the models by calculating the percentage of correct action choices (i.e. the learning accuracy). This measure is widely used in classification problems when the objective is to describe the accuracy of a final map process [50]. To do so, for each pattern presentation we store the rewarded (expected) action in response to the presented pattern, and the finally selected (chosen) one during that pattern presentation. We only consider in the calculation those trials in which some reward or punishment can be delivered, ignoring those intervals with no repeating patterns conveyed to the inputs (only noisy inputs). We consider that an action has been taken if the corresponding action neuron has spiked at least once during the pattern presentation. Conversely, we consider that no action has been taken if none of the action neurons spikes during the same duration. In order to obtain an estimation of the temporal evolution of the accuracy we use a rolling mean of the last 100 values.

3. Results and discussion

We did extensive testing of the learning mechanisms we proposed. Some of these results demonstrate that the combination of STDE learning rule and homeostatic mechanisms allow learning (and re-learning) of rewarded patterns, or that there is no effect of the reward delay and the frequency of the input pattern on the learning process, among others. However, as they are not the main concern for this article, they are placed in the Supplementary Results 5.2 section for further examination.

The main results and discussion are structured as follows: we first show the general behavior of the network. Then we study the effect of the lateral connectivity pattern on the performance and the way neurons are processing information. Finally, we put our results in context by comparing our model with previously proposed models in the literature.

3.1. General network behavior

During the simulation of the action-selection task, each action group neuron becomes overall active in response to the presentation of the associated patterns as shown in the raster plots (Fig. 3C and D) and the activity balance for the action neuron groups (Fig. 3E), producing mainly dopaminergic rewarding (Fig. 3F). The action accuracy reveals steady-state performance after 200 s of simulations (Fig. 3G). According to these results, our combination of STDE learning rule [24] with homeostatic

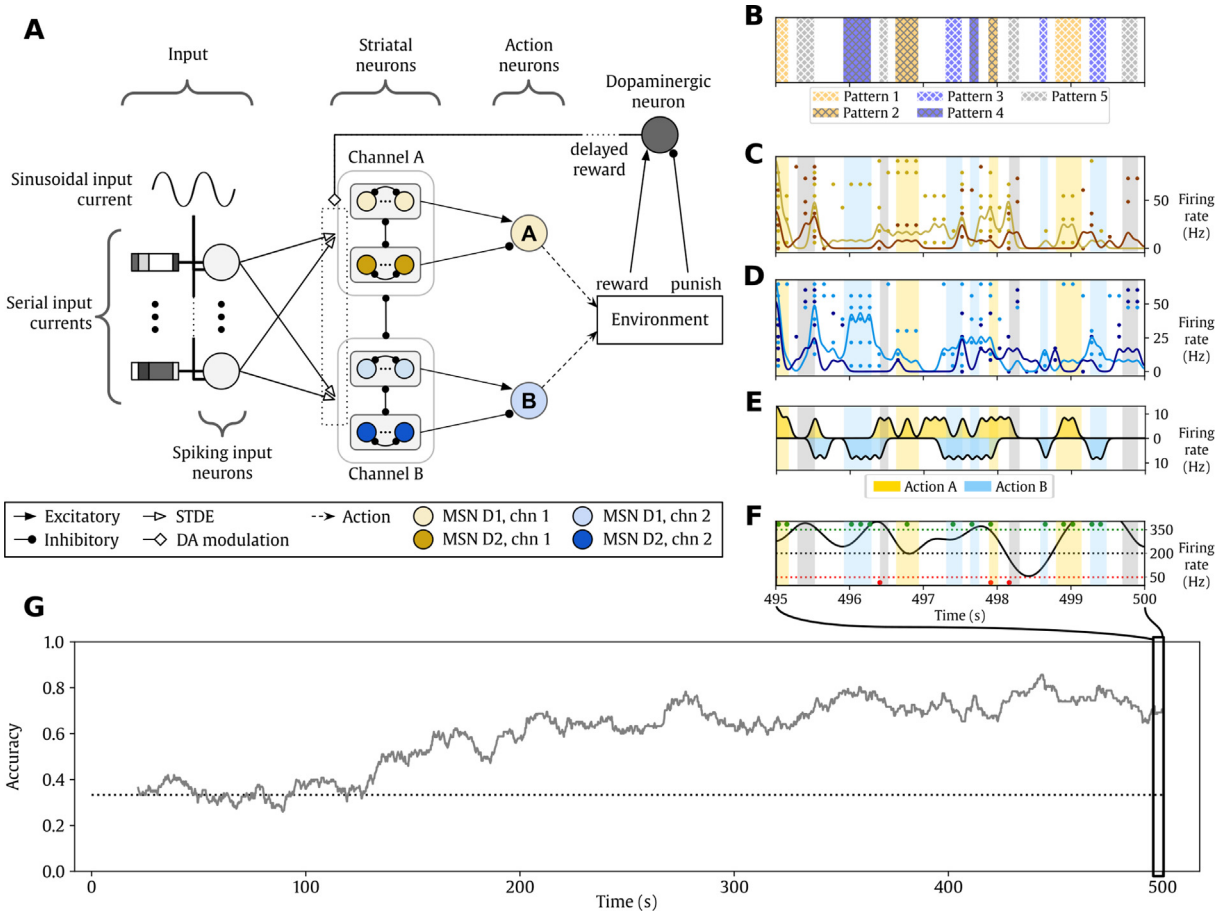

Fig. 3. Cortico-striatal network solving a RL task. **A**. Structure of the network. See Section 2.1.5 for a detailed explanation. **B-F**. The activity of the network during the last 5 s of simulation. Background color indicates the reward policy (yellowish colors, action A is rewarded and B is punished; bluish colors, action B is rewarded and A is punished; grey, any action is punished). **B**. Input pattern conveyed to the cortical layer. **C**. Raster plot of the channel-A action neurons. Yellow dots represent MSN D1 spikes, and orange dots are MSN D2 spikes. **D**. Raster plot of channel B. Cyan dots represent MSN D1 spikes, and dark blue dots are MSN D2 spikes. **E**. Action neuron firing rates. The middle horizontal line represents 0 Hz. Action A and B activity are represented in opposites directions for clarity. Action A neuronal activity increases in yellow zones while action B neuronal activity in cyan intervals. **F**. Firing rate of the dopaminergic neuron (black line). Dotted horizontal lines indicate the range of DA activity considered: black is the baseline, green is the maximum reward, and red represents the maximum punishment. Dots indicate rewards (green) and punishment (red) events delivered to the agent. **G**. Evolution of the learning accuracy of the agent, see Section 2.3 for further details. The dotted line marks the accuracy level by chance.

mechanisms and an oscillatory input signal in a cortico-striatal model learns to accurately select the most rewarding action.

The way our network learns to associate the corresponding input stimulus with sub-populations of D1 and D2 neurons in channel A or channel B is the following: If the agent takes the right action for a specific input pattern, the environment delivers a reward with some delay (high DA level in Fig. 3F). This reward potentiates the synapses between the cortical layer and the action-associated D1 sub-population, resulting in more frequent firing. On the other hand, if the agent takes a wrong action, then it receives a punishment sometime later (low DA level in Fig. 3F). This punishment weakens the synapses from the cortical layer to the action-associated D1 sub-population while strengthening the corresponding synapses to the D2 (inhibitory) sub-population of the same channel. This learning process makes the agent stick to the rewarded action and switch to a different one when punished. For the specific case when the environment punishes any action during a stimulus presentation, both D2 sub-populations increase their activity, and both action neurons remain silent.

The proposed model shows how combining two complementary dopamine-based STDE learning rules (Fig. 1) can facilitate the association between sensorial cortical inputs and rewarded actions with arbitrary rewarding policies. Previously, the STDE rule had been shown to be capable of learning to select an action channel driven by stronger cortical input [24], and here we show that this rule can also be used to learn inputs defined by the specific timing of their spikes (as all the inputs have the same average firing rate). This represents a higher complexity task and illustrates how STDE can be efficiently used for spike time pattern representation.

The model also is completely bioplausible, as all the mechanisms used have been described in biological systems: DA induces bidirectional, timing-dependent plasticity at MSNs glutamatergic synapses [49], *in vitro* pyramidal neural recordings are consistent with simulations of adaptive spike threshold neurons, and they lead to better stimulus discrimination than would be achieved otherwise [27], and rat hippocampal pyramidal neurons *in vitro* can use rate-to-phase transform [38]. Detailed discussion on the role of the homeostatic mechanisms can be found in Supplementary Materials.

3.2. Effect of lateral inhibition patterns and task complexity

Once we have demonstrated how the striatal network can support RL, we wondered to what extent the connectivity pattern of the lateral inhibition in the striatum could impact the learning capabilities. So that we extensively explored different versions of connectivity.

We first study if there is any relationship between the connectivity pattern and difficulty of the task. We organized the lateral inhibitory connections in two groups: intra-channel (inhibitory connections from D2 MSNs to D1 MSNs within the same channel) and inter-channel (inhibitory connections between D1 MSNs of different channels, and between D2 MSNs of different channels). We obtained four possible subsets of connectivity patterns by keeping or removing each connection type (Fig. 2). We used three difficulty levels for the task: easy, normal and hard. The easy task uses only one stimulus associated with each action (stimulus 1 to action A, stimulus 2 to action B, stimulus 3 to no action). The normal task uses two stimulus per action, and one no-go stimulus. The hard task uses four stimuli per action, and two no-go stimuli.

The results of the easy version of the experiments are shown in the Fig. 4. The models without inter-channel inhibition work worse, as they stabilize with lower values of accuracy. The models with inter-channel inhibition seem to reach a similar level of accuracy but the intra-channel inhibition seems to reduce the learning rate.

In the normal version of the task, we again obtained the best learning performance when using the inter-channel lateral inhibition with asymmetrical structured connection pattern, and the difference increased. In this case, there is no apparent effect in of the intra-channel lateral inhibition in this task (Fig. 5). According to our simulations, lateral inter-channel inhibition facilitates the emergence of one action-related channel over the other one in a winner-take-all manner, as expected.

We saw in previous experiments that the inter-channel lateral inhibition is always increases accuracy, so we will use it always in the following tests. In the hard task we obtained small but significant differences: The accuracy of the network improves faster with the intra-channel lateral inhibition (see Fig. 6). Also, apparently the network with the intra-channel inhibition settled in a more stable regime as it maintains its performance, compared with the network without this intra-channel inhibition which slowly degrades (Supplementary Fig. 13). The results so far suggest that both connectivity patterns contribute to a reliable action-selection paradigm.

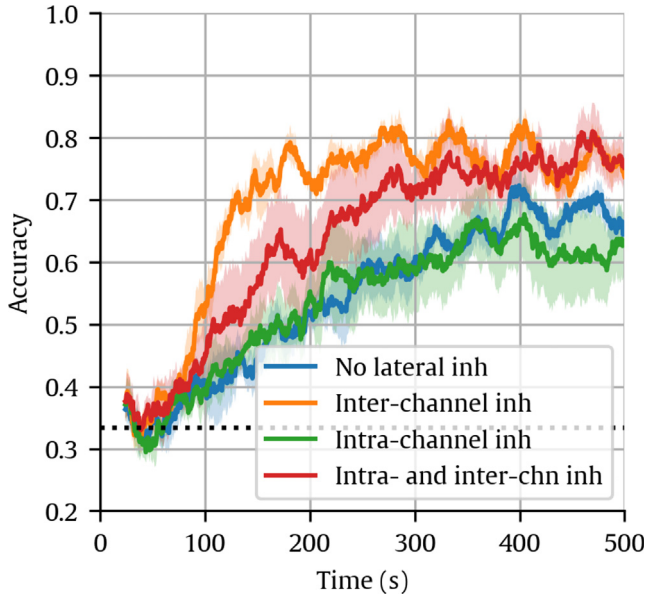

Fig. 4. Effect of the lateral inhibitory connectivity on the performance during a simpler version of the RL task. The horizontal dotted line represents the accuracy obtained by a random agent. The curves represent the mean and the standard error of the mean of the evolution of each agent during the task (n = 5).

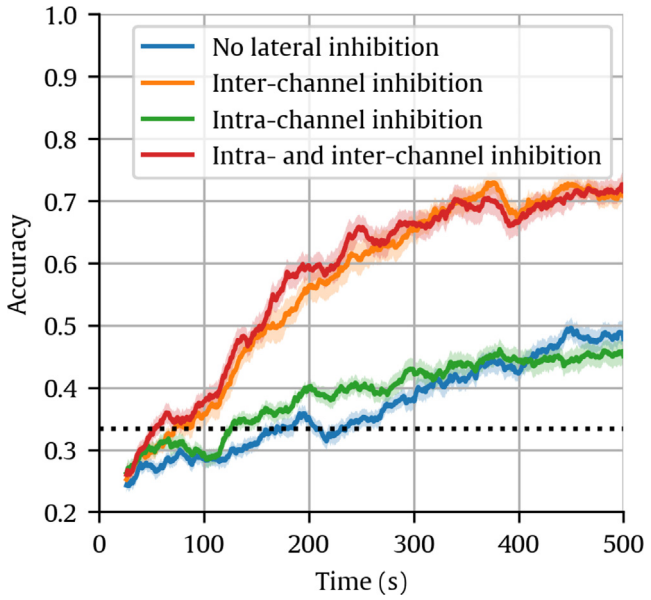

Fig. 5. Effect of the lateral inhibitory connectivity on the performance during the normal RL task. The curves represent the mean accuracy and the shaded areas represent the standard error (n = 30). Four different configurations are tested, depending on the presence of two types of lateral connectivity: intra- and inter-channel inhibition. The horizontal dotted line represents the accuracy obtained by a random agent with no learning mechanisms.

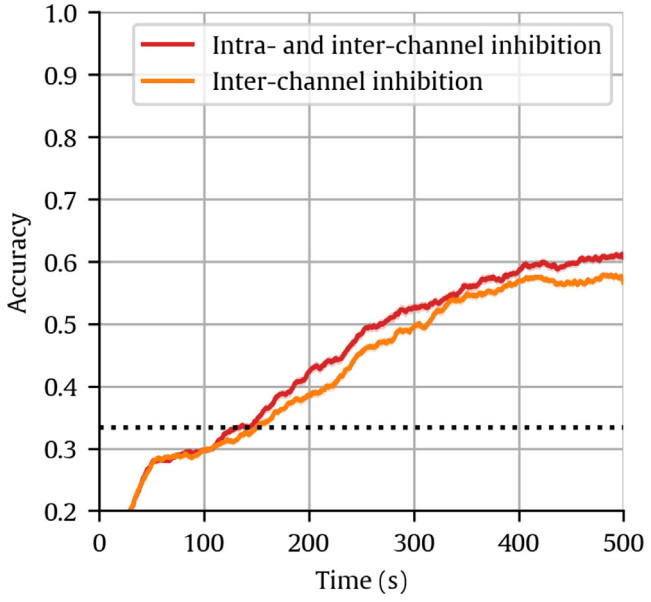

Fig. 6. Effect of the intra-channel lateral inhibitory connectivity on the performance during a harder version of the RL task. The horizontal dotted line represents the accuracy obtained by a random agent. The curves and the filling color represent the mean, the standard error of the mean, respectively, of the evolution of each agent during the task (n = 150), simulated for 500 s.

Taking these results together, it seems that when we use several stimuli associated with each action, intra-channel inhibition improves the RL action selection task. However, when only one stimulus is associated with each action, this intra-channel inhibition does not impact learning performance. When compared with the results in Fig. 5 and 6, it seems that the intra-channel lateral inhibition improves the learning capabilities only with a harder task,

but when the task is too simple then the intra-channel connection increases the learning time.

We also explored the effect of connectivity patterns of lateral inhibition different from the proposed by [6], by adding or removing lateral connections within a channel, within each subpopulation, and between subpopulations of the same channel. All variations from the original resulted in reduced learning performance (Supplementary Fig. 15). In this Figure, the curve #5 represents the network with both lateral inhibition in D1 layer and D2 layer, as well as intra- and inter-channel lateral inhibition. This structure (similar to the one proposed by [6] obtains the best accuracy.

3.3. Effect of intra-channel lateral inhibition on neuronal specialization

Intra-channel inhibition seems to facilitate learning in more complex tasks, possibly because it enhances neuron specialization. We saw a strong reduction of correlation at time difference $\delta t = 0$ between action A and B D1 sub-populations caused by intra-channel inhibition (data not shown), but this does not seem to justify the improved accuracy for more complex tasks.

Then, we hypothesized that intra-channel inhibition could encourage neuron specialization to specific cortical patterns. We tested this idea by analyzing the preferred stimuli for each neuron after the learning process (Fig. 7), and obtained the opposite result: the intra-channel lateral inhibition affects D1 neurons by forcing them to share more evenly their activity over several stimuli, in addition to reducing their average activity. This is in contrast with the network without intra-channel lateral inhibition, where the activity is more focused on the favorite stimuli and has higher mean activity.

According to these results, although individual neurons of the network with intra-channel inhibition have less precise representation of individual sensorial stimuli, these models have higher precision to associate rewarding actions. This can be explained assuming some sparse representation of the stimuli, where the simultaneous firing of several (but not many) neurons are needed to indicate the presence of an input stimuli. This more sparse representation emerges due to the combination of stronger inhibition and the homeostatic mechanisms: a neuron avoids firing when it is inhibited, so the homeostatic mechanisms tend to compensate for this activity reduction by increasing its chances to fire in response to several stimuli. This sparse representation has been suggested to facilitate sensorial pattern recognition in other brain areas, such as the cerebellar cortex, the mushroom body, and the dentate gyrus of the hippocampus [7].

In the context of our model, the sparse representation due to intra-channel inhibition plays a role in the action selection process, which can be seen as a form of classification. Here, the goal is not to classify stimuli per se, but to assign stimuli to appropriate actions. The sparse coding helps to achieve more efficient and robust action selection by reducing the overlapping between representations of different sensorial states, minimizing interference, and enabling more reliable decision-making.

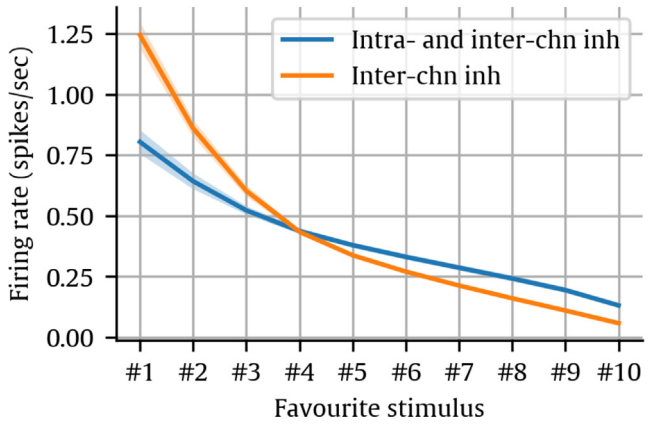

Fig. 7. Effect of the intra-channel lateral inhibitory connectivity on the firing rate pattern on their preferred stimuli of D1 neurons. Higher and more specialized firing patterns occur in networks without intra-channel lateral inhibition, while more sparse representations occur in networks with it. Lines and shaded areas represent mean and 95% confidence intervals of the mean ($n = 150$), respectively.

3.4. Comparison with previous models of reinforcement learning and basal ganglia

We presented a point-neuron model of the BG that can solve complex action-selection tasks using a RL paradigm. We do so by using multiple mechanisms proposed in the literature: the STDE learning rule that implements synaptic modification in cortex-MSN connections [24], combined with homeostatic mechanisms [15] and an oscillatory input signal [37,16] in a network with asymmetrical structured lateral inhibition [6] can rapidly and consistently learn to detect the presence of rewarded input patterns. These processes have been described in biological systems and here proved to be robust.

Simpler STDP-like rules have been used for RL tasks [31,32], but they were employed in simpler networks, single neurons, and simple tasks. Beyond the state-action mapping role proposed in this article for the striatum, other theories exist about the action decision process. However, computational models of BG in the literature have considerably evolved during the last two decades [46], and there is still no consensus about how to achieve goal-oriented learning in a BG model. Previous models ranged from those with action-selection features but no learning [2,18,29,35,3,23,48,13,45,4] (but see [12]) to simple forms of learning, with RL [5], rate-based learning rules [26], or based on modulated STDP with eligibility traces [28,24,1]. These models considered direct and indirect pathways (as "selection" and "control" routes, respectively), composed of MSN D1 and D2 striatal neurons controlling GPe and SNr. Many models assume that the BG work as an actor-critic model [5,41], and actor-critic frameworks have been used for RL tasks like maze navigation [14,43,55] and cartpole [14]. More biologically-constrained models of the BG have been proposed to explain the origin of diseases like Parkinson's disease [34] and the role of specific interneurons [20] or pathways [19] during action-selection. Recent accumulation-to-bound models describe the decision process as an accumulation of evidence for each alternative action until a decision threshold is exceeded in one of these actions [40]. It would be interesting to explore how these models could be incorporated with the proposed model, potentially requiring additional brain areas. In this regard, some models incorporate recurrent activity loops with the cortex through the thalamus [36].

Moreover, we acknowledge that similar models can already deal with more complex action-selection tasks than the one used in this work, such as cart-pole, inverted pendulum, or simple mazes [14]. However, there exist some important differences between their model and the one proposed in this article. First, our network does not include a critic. Second, their learning rule requires a temporal difference (TD) signal that would need additional circuitry. Third, their model requires an additional place-cell layer with unsupervised learning to represent complex input patterns. However, it remains as a future work to embed the network model into a closed-loop experimental setup requiring continuously graded output (instead of selecting an action in a discrete set of possibilities). This way, the model could deal with a larger set of RL

tasks. In our case, we have integrated a spiking neural network with spike-time pattern representation that scales well with different patterns complexity at the pattern classification layer. Future work will explore how our model could be extended for such complex action control frameworks.

4. Conclusion

In this article we tested the respective roles in learning of the different mechanisms used during our simulations: homeostatic mechanisms make the neurons change their response to compensate for long-lasting changes in the input level, making learning faster and more robust to the configuration. The asymmetrical lateral inhibition consistently outperformed other connectivity configurations. By adding intra-channel lateral inhibition to the network model, we induced the channels to generate a sparse representation of each stimulus relevant for the task. This made the network less prone to errors as the model had to recruit more neurons to take an action. Lastly, by segregating striatal and action neurons in independent channels for each action and incorporating MSN D1 (Go neurons) and MSN D2 (No-Go) sub-populations with different learning kernels, the model effectively learned arbitrary mappings from sensorial input states to action output in a two-choice action-selection task. MSN D1 neurons and MSN D2 neurons cooperatively facilitated action selection with contrary effects; MSN D1 neurons learned to potentiate preferred actions while MSN D2 neurons learned to inhibit non-preferred actions.

CRediT authorship contribution statement

Álvaro González-Redondo: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Visualization. **Jesús Garrido:** Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing - review & editing. **Francisco Naveros Arrabal:** Software. **Jeanette Hellgren Kotaleski:** Conceptualization, Supervision. **Sten Grillner:** Conceptualization, Supervision. **Eduardo Ros:** Conceptualization, Methodology, Investigation, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Data availability

Data will be made available on request.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research is supported by the Spanish Grant INTSENSO (MICINN-FEDER-PID2019-109991 GB-I00), Regional grants Junta Andalucía-FEDER (CEREBIO P18-FR-2378 and A-TIC-276-UGR18). This research has also received funding from the EU Horizon 2020 Framework Program under the Specific Grant Agreement No. 945539 (Human Brain Project SGA3) and the EU Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 891774 (NEUSEQBOT). Additionally, the main author has been funded with a national research training grant (FPU17/04432). Finally, this research was also supported by the Vetenskapsrådet (VR-M-2017–02806, VR-M-2020–01652); Swedish e-science Research Center (SeRC); KTH Digital Futures. Funding for open access charge: Universidad de Granada / CBUA.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.neucom.2023.126377.

References

[1] J. Baladron, A. Nambu, F.H. Hamker, The subthalamic nucleus-external globus pallidus loop biases exploratory decisions towards known alternatives: a neuro-computational study, Eur. J. Neurosci. 49 (6) (2019) 754–767.
[2] D.G. Beiser, S.E. Hua, J.C. Houk, Network models of the basal ganglia, Curr. Opin. Neurobiol. 7 (2) (1997) 185–190.
[3] G.S. Berns, T.J. Sejnowski, A computational model of how the basal ganglia produce sequences, J. Cognit. Neurosci. 10 (1) (1998) 108–121.
[4] R. Bogacz, Optimal decision-making theories: linking neurobiology with behaviour, Trends Cognit. Sci. 11 (3) (2007) 118–125.
[5] R. Bogacz, T. Larsen, Integration of reinforcement learning and optimal decision-making theories of the basal ganglia, Neural Comput. 23 (4) (2011) 817–851.
[6] D.A. Burke, H.G. Rotstein, V.A. Alvarez, Striatal local circuitry: a new framework for lateral inhibition, Neuron 96 (2) (2017) 267–284.
[7] N.A. Cayco-Gajic, R.A. Silver, Re-evaluating circuit mechanisms underlying pattern separation, Neuron 101 (4) (2019) 584–602.
[8] J. Cox, I.B. Witten, Striatal circuits for reward learning and decision-making, Nat. Rev. Neurosci. 20 (8) (2019) 482–494, https://doi.org/10.1038/s41583-019-0189-2.
[9] M.R. DeLong, M.D. Crutcher, A.P. Georgopoulos, Primate globus pallidus and subthalamic nucleus: functional organization, J. Neurophysiol. 53 (2) (1985) 530–543.
[10] E. Fino, L. Venance, Spike-timing dependent plasticity in the striatum, Front. Synaptic Neurosci. 2 (2010) 6.
[11] N.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, C. Cao, S. Yamashita, A. Tugangui, A. Li, T. Jiang, X. Jia, Z. Feng, S. Aquino, H.S. Mun, M. Zhu, A. Santarelli, N.L. Benavidez, M. Song, G. Dan, M. Fayzullina, S. Ustrell, T. Boesen, D.L. Johnson, H. Xu, M.S. Bienkowski, X.W. Yang, H. Gong, M.S. Levine, I. Wickersham, Q. Luo, J.D. Hahn, B.K. Lim, L.I. Zhang, C. Cepeda, H. Hintiryan, H.W. Dong, The mouse cortico–basal ganglia–thalamic network, Nature 598 (78797879) (2021) 188–194, https://doi.org/10.1038/s41586-021-03993-3.
[12] M.J. Frank, Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive deficits in medicated and nonmedicated parkinsonism, J. Cognit. Neurosci. 17 (1) (2005) 51–72.
[13] M.J. Frank, Hold your horses: a dynamic computational role for the subthalamic nucleus in decision making, Neural Networks 19 (8) (2006) 1120–1136.
[14] N. Frémaux, H. Sprekeler, W. Gerstner, Reinforcement learning using a continuous time actor-critic framework with spiking neurons, PLoS Comput. Biol. 9 (4) (2013) e1003024.
[15] S.E. Galindo, P. Toharia, Ó.D. Robles, E. Ros, L. Pastor, J.A. Garrido, Simulation, visualization and analysis tools for pattern recognition assessment with spiking neuronal networks, Neurocomputing 400 (2020) 309–321, https://doi.org/10.1016/j.neucom.2020.02.114, arXiv:2003.06343.
[16] J.A. Garrido, N.R. Luque, S. Tolu, E. D'Angelo, Oscillation-Driven Spike-Timing Dependent Plasticity Allows Multiple Overlapping Pattern Recognition in Inhibitory Interneuron Networks, Int. J. Neural Syst. 26 (05) (2016) 1650020, https://doi.org/10.1142/S0129065716500209.
[17] W. Gerstner, W.M. Kistler, Spiking neuron models: Single neurons, populations, plasticity, Cambridge University Press, 2002.
[18] A. Gillies, G. Arbuthnott, Computational models of the basal ganglia, Mov. Disord. 15 (5) (2000) 762–770.
[19] B. Girard, J. Lienard, C.E. Gutierrez, B. Delord, K. Doya, A biologically constrained spiking neural network model of the primate basal ganglia with overlapping pathways exhibits action selection, Eur. J. Neurosci. 53 (7) (2021) 2254–2277.
[20] L. Goenner, O. Maith, I. Koulouri, J. Baladron, F.H. Hamker, A spiking model of basal ganglia dynamics in stopping behavior supported by arkypallidal neurons, Eur. J. Neurosci. 53 (7) (2021) 2296–2321.
[21] A.M. Graybiel, The basal ganglia and chunking of action repertoires, Neurobiol. Learn. Mem. 70 (1) (1998) 119–136, https://doi.org/10.1006/nlme.1998.3843.
[22] S. Grillner, J. Hellgren, A. Ménard, K. Saitoh, M.A. Wikström, Mechanisms for selection of basic motor programs – roles for the striatum and pallidum, Trends Neurosci. 28 (7) (2005) 364–370, https://doi.org/10.1016/j.tins.2005.05.004, URL: https://www.sciencedirect.com/science/article/pii/S0166223605001293.
[23] K. Gurney, T.J. Prescott, P. Redgrave, A computational model of action selection in the basal ganglia. i. a new functional anatomy, Biolog. Cybern. 84 (6) (2001) 401–410.

[24] K.N. Gurney, M.D. Humphries, P. Redgrave, A New Framework for Cortico-Striatal Plasticity: Behavioural Theory Meets In Vitro Data at the Reinforcement-Action Interface, PLoS Biol. 13 (1) (2015) e1002034, https://doi.org/10.1371/journal.pbio.1002034.
[25] O. Hikosaka, Y. Takikawa, R. Kawagoe, Role of the basal ganglia in the control of purposive saccadic eye movements, Physiol. Rev. 80 (3) (2000) 953–978.
[26] S. Hong, O. Hikosaka, Dopamine-mediated learning and switching in cortico-striatal circuit explain behavioral changes in reinforcement learning, Front. Behav. Neurosci. 5 (2011) 15.
[27] C. Huang, A. Resnik, T. Celikel, B. Englitz, Adaptive spike threshold enables robust and temporally precise neuronal encoding, PLoS Comput. Biol. 12 (6) (2016) e1004984.
[28] M.D. Humphries, N. Lepora, R. Wood, K. Gurney, Capturing dopaminergic modulation and bimodal membrane behaviour of striatal medium spiny neurons in accurate, reduced models, Front. Comput. Neurosci. 3 (2009) 26.
[29] M.D. Humphries, R.D. Stewart, K.N. Gurney, A physiologically plausible model of action selection and oscillatory activity in the basal ganglia, J. Neurosci. 26 (50) (2006) 12921–12942.
[30] B.J. Hunnicutt, B.C. Jongbloets, W.T. Birdsong, K.J. Gertz, H. Zhong, T. Mao, A comprehensive excitatory input map of the striatum reveals novel functional organization, eLife 5 (2016) e19103, https://doi.org/10.7554/eLife.19103.
[31] E.M. Izhikevich, Solving the distal reward problem through linkage of STDP and dopamine signaling, Cereb. Cortex 17 (10) (2007) 2443–2452, https://doi.org/10.1093/cercor/bhl152.
[32] R. Legenstein, D. Pecevski, W. Maass, A learning theory for reward-modulated spike-timing-dependent plasticity with application to biofeedback, PLOS Computat. Biol. 4 (2008) 1–27, https://doi.org/10.1371/journal.pcbi.1000180, DOI: 10.1371/journal.pcbi.1000180.
[33] W. Levy, O. Steward, Temporal contiguity requirements for long-term associative potentiation/depression in the hippocampus, Neuroscience 8 (4) (1983) 791–797.
[34] M. Lindahl, J.H. Kotaleski, Untangling basal ganglia network dynamics and function: role of dopamine depletion and inhibition investigated in a spiking network model, eneuro 3 (6) (2016).
[35] C.C. Lo, X.J. Wang, Cortico–basal ganglia circuit mechanism for a decision threshold in reaction time tasks, Nature Neurosci. 9 (7) (2006) 956–963.
[36] C.C. Lo, X.J. Wang, Cortico–basal ganglia circuit mechanism for a decision threshold in reaction time tasks, Nature Neurosci. 9 (7) (2006) 956–963.
[37] T. Masquelier, E. Hugues, G. Deco, S.J. Thorpe, Oscillations, Phase-of-Firing Coding, and Spike Timing-Dependent Plasticity: An Efficient Learning Scheme, J. Neurosci. 29 (43) (2009) 13484–13493, https://doi.org/10.1523/JNEUROSCI.2207-09.2009.
[38] D. McLelland, O. Paulsen, Neuronal oscillations and the rate-to-phase transform: mechanism, model and mutual information, J. Physiol. 587 (4) (2009) 769–785.
[39] B.R. Miller, A.G. Walker, A.S. Shah, S.J. Barton, G.V. Rebec, Dysregulated information processing by medium spiny neurons in striatum of freely behaving mouse models of huntington's disease, J. Neurophysiol. 100 (4) (2008) 2205–2216.
[40] M.J. Mulder, The temporal dynamics of evidence accumulation in the brain, J. Neurosci. 34 (42) (2014) 13870–13871, https://doi.org/10.1523/JNEUROSCI.3251-14.2014.
[41] J. O'Doherty, P. Dayan, J. Schultz, R. Deichmann, K. Friston, R.J. Dolan, Dissociable roles of ventral and dorsal striatum in instrumental conditioning, Science 304 (5669) (2004) 452–454.
[42] A. Parent, L.N. Hazrati, Functional anatomy of the basal ganglia. ii. the place of subthalamic nucleus and external pallidium in basal ganglia circuitry, Brain Res. Rev. 20 (1) (1995) 128–154.
[43] W. Potjans, A. Morrison, M. Diesmann, A spiking neural network model of an actor-critic learning agent, Neural Comput. 21 (2) (2009) 301–339.
[44] B. Rajendran, A. Sebastian, M. Schmuker, N. Srinivasa, E. Eleftheriou, Low-power neuromorphic hardware for signal processing applications: A review of architectural and system-level design approaches, IEEE Signal Process. Mag. 36 (6) (2019) 97–110.
[45] R. Ratcliff, M.J. Frank, Reinforcement-based decision making in corticostriatal circuits: mutual constraints by neurocomputational and diffusion models, Neural Comput. 24 (5) (2012) 1186–1229.
[46] J.E. Rubin, C. Vich, M. Clapp, K. Noneman, T. Verstynen, The credit assignment problem in cortico-basal ganglia-thalamic networks: A review, a problem and a possible solution, Eur. J. Neurosci. 53 (7) (2021) 2234–2253.
[47] W. Schultz, Dopamine signals for reward value and risk: basic and recent data, Behav. Brain Functions 6 (1) (2010) 1–9.
[48] B. Sen-Bhattacharya, S. James, O. Rhodes, I. Sugiarto, A. Rowley, A.B. Stokes, K. Gurney, S.B. Furber, Building a spiking neural network model of the basal ganglia on spinnaker, IEEE Trans. Cognit. Develop. Syst. 10 (3) (2018) 823–836.
[49] W. Shen, M. Flajolet, P. Greengard, D.J. Surmeier, Dichotomous dopaminergic control of striatal synaptic plasticity, Science 321 (5890) (2008) 848–851.
[50] S.V. Stehman, Selecting and interpreting measures of thematic classification accuracy, Remote Sens. Environ. 62 (1) (1997) 77–89, https://doi.org/10.1016/S0034-4257(97)00083-7, URL: https://www.sciencedirect.com/science/article/pii/S0034425797000837.
[51] S.M. Suryanarayana, J.H. Kotaleski, S. Grillner, K.N. Gurney, Roles for globus pallidus externa revealed in a computational model of action selection in the basal ganglia, Neural Networks 109 (2019) 113–136, https://doi.org/10.1016/j.neunet.2018.10.003.
[52] R.S. Sutton, A.G. Barto, R.J. Williams, Reinforcement learning is direct adaptive optimal control, IEEE Control Syst. Mag. 12 (2) (1992) 19–22.
[53] A. Taherkhani, A. Belatreche, Y. Li, G. Cosma, L.P. Maguire, T.M. McGinnity, A review of learning in biologically plausible spiking neural networks, Neural Networks 122 (2020) 253–272.
[54] A. Tavanaei, M. Ghodrati, S.R. Kheradpisheh, T. Masquelier, A. Maida, Deep learning in spiking neural networks, Neural Networks 111 (2019) 47–63.
[55] E. Vasilaki, N. Frémaux, R. Urbanczik, W. Senn, W. Gerstner, Spike-based reinforcement learning in continuous state and action space: when policy gradient methods fail, PLoS Comput. Biol. 5 (12) (2009) e1000586.

Álvaro González-Redondo is a PhD student at the Computational Neuroscience and Neurorobotics Lab of the University of Granada. He holds a degree in Psychology and a second degree in Computer Science. He is MSc in Neuroscience at the University of Granada. His main research interests are motor control and learning for robotics, and in particular the basal ganglia, a brain region involved in reward evaluation and action selection.

Jesús A. Garrido is Associate Professor at the School of Computer Science and Engineering from the University of Granada. He earned his M.Sc. degree in Computer Science in 2006 and his PhD degree in Computer Engineering and Networks in 2011 respectively, all from the University of Granada (Spain). From 2012 to 2015, he joined the Brain and Behavioral Science department at the University of Pavia (Italy) under supervision of Dr. Egidio D'Angelo. In 2015, he was awarded with a Young Researchers Fellowship by the University of Granada in the Applied Computational Neuroscience Research Group of the University of Granada (ACN-UGR). He is the author of more than 25 articles. His main research interests include cerebellar information processing and learning, motor control, neuromorphic engineering, spiking neural networks and neurorobotics.

Francisco Naveros received his M.Sc. and Ph.D degrees in Telecommunication and Computational Neuroscience from the University of Granada (Spain) in 2011 and 2017 respectively. In 2020 he obtained a position as Assistant Professor at the School of Computer Science and Engineering from the Polytechnic University of Madrid (Spain). In 2021 he was awarded with a Marie Curie Global Fellowship in the Baylor College of Medicine (Houston, Texas, USA) and the University of Granada (Granada, Spain). He is a member of the Applied Computational Neuroscience Research Group of the University of Granada (ACN-UGR). He is the author of 14 articles. His main research interests include experimental and computational neuroscience in the cerebellum, parallel and real-time spiking neural network simulations and neurorobotics.

Jeanette Hellgren Kotaleski has a M.Sc in Engineering Physics and a Ph.D. in Computer Science from KTH the Royal Institute of Technology (Sweden). She pursued postdoc studies at the George Mason University, Krasnow Institute (USA). Since 2007 she is Full Professor in Neuroinformatics at Dept. Computational Science and Technology at the School of Electrical Engineering and Computer Science, KTH. The main focus of her research is to use computational modeling to understand the neural mechanisms underlying information processing and learning in motor system such as the basal ganglia. The levels of investigation range from simulations of large-scale neural networks, using both biophysically detailed as well as abstract systems level models, down to kinetic models of subcellular processes.

Sten Grillner is professor in the Department of Neuroscience at the Karolinska Institute in Stockholm. His initial research focus was on the intrinsic function of the spinal networks that generate locomotion in vertebrates and subsequently on forebrain mechanisms underlying selection of behavior with a focus on the basal ganglia, combined with an interest in the evolution of the nervous system. He has used a variety of experimental techniques including data-driven modelling. He has been elected into the National Academy of Science (US), National Academy of Medicine (US), the Royal Swedish Academy of Science and other Academies, and has been awarded several prizes including the Kavli Prize in Neuroscience 2008.

Eduardo Ros received his M.Sc. and Ph.D. degrees in Physics and Computational Neuroscience from the University of Granada (Spain) in 1992 and 1997 respectively. He is currently Full Professor in the Department of Computer Architecture and Technology of the University of Granada. He is the head of the Applied Computational Neuroscience Research Group of the University of Granada (ACN-UGR). He is the author of more than 100 articles. His main research interests include bio-inspired processing and neuromorphic engineering, spiking neural networks, hardware processing architectures and computational neuroscience.