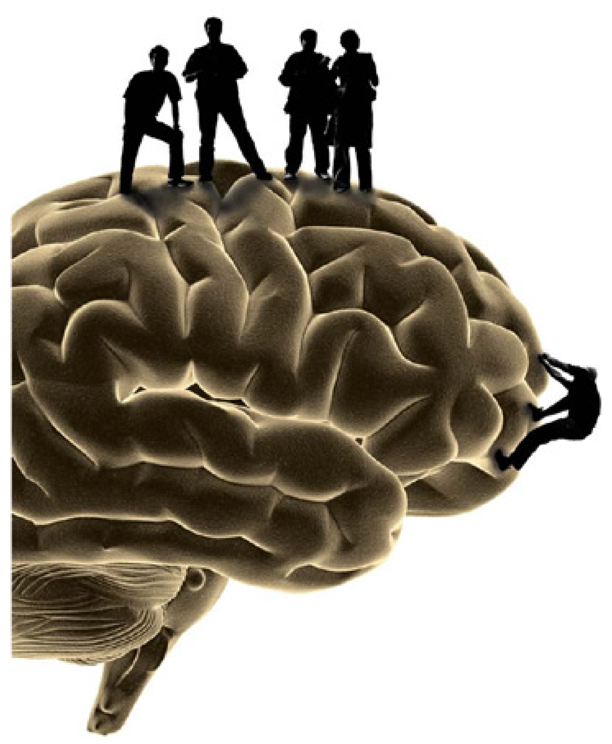


A general error-modulated STDP learning rule applied to reinforcement learning in the basal ganglia

Trevor Bekolay, Chris Eliasmith {tbekolay,celiasmith}@uwaterloo.ca
Centre for Theoretical Neuroscience, University of Waterloo <http://ctn.uwaterloo.ca>

University of Waterloo
CTN Waterloo
CNRG computational neuroscience research group
compneuro.uwaterloo.ca

Introduction

- Error-modulated rules have typically assumed simple scalar representations of error
- To learn complex tasks, models have employed attention or sparse local connectivity to do RL in high-dimensional spaces
- We propose an STDP rule that exploits multidimensional error signals to learn complex tasks without additional constraints

Methods

- Simulations use the Neural Engineering Framework[1]

Encoding: $a_i(\mathbf{x}) = G_i\left[\alpha_i \langle \mathbf{e}_i \mathbf{x} \rangle + J_i^{bias}\right]$

Decoding: $\hat{\mathbf{x}}(t) = \sum_{i,n} h(t - t_{i,n}) \mathbf{d}_i$

- Reinforcement learning: value function and reward prediction error

$$\hat{V}(s_t) = \hat{V}(s_t) + \alpha\delta$$
$$\delta = R_t + \gamma \hat{V}(s_{t+1}) - \hat{V}(s_t)$$

Incorporating reward prediction error in a synaptic learning rule

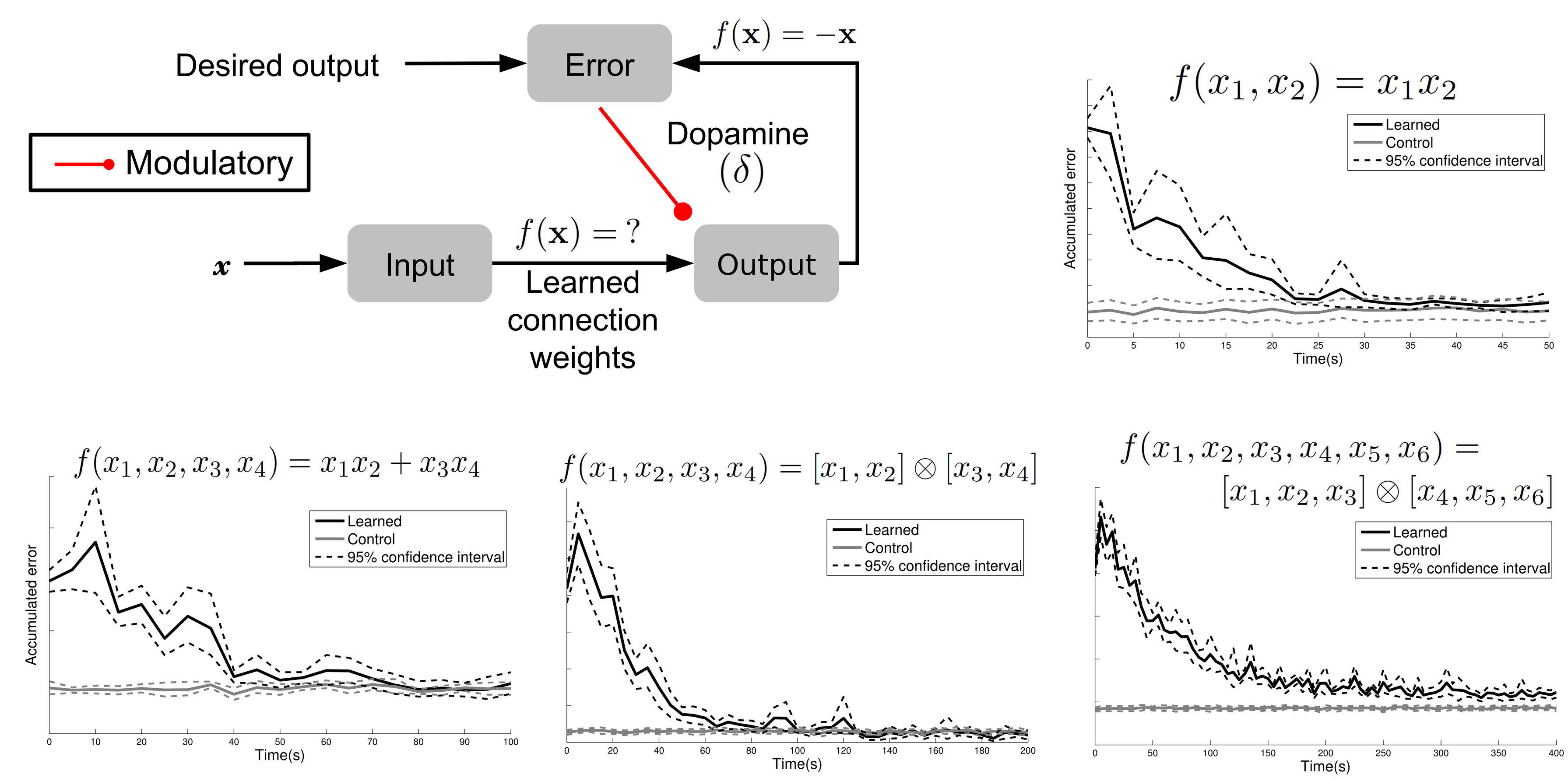
- Dopamine release has been shown to closely resemble reward prediction error (δ)
- Applied to NEF: $\Delta\omega_{ij} = \kappa\alpha_j \langle \mathbf{e}_i \delta \rangle a_{ij}$ [2]
- A triplet based STDP rule:[3]

$$\dot{r}_1(t) = -\frac{r_1(t)}{\tau_+} \quad \text{if } t = t^{pre}, \text{ then } r_1 = r_1 + 1$$
$$\Delta\omega_{ij}(t^{post}) = r_1(t)\left[A_2^+ + A_3^+ o_2(t - dt)\right]$$

- Combined:

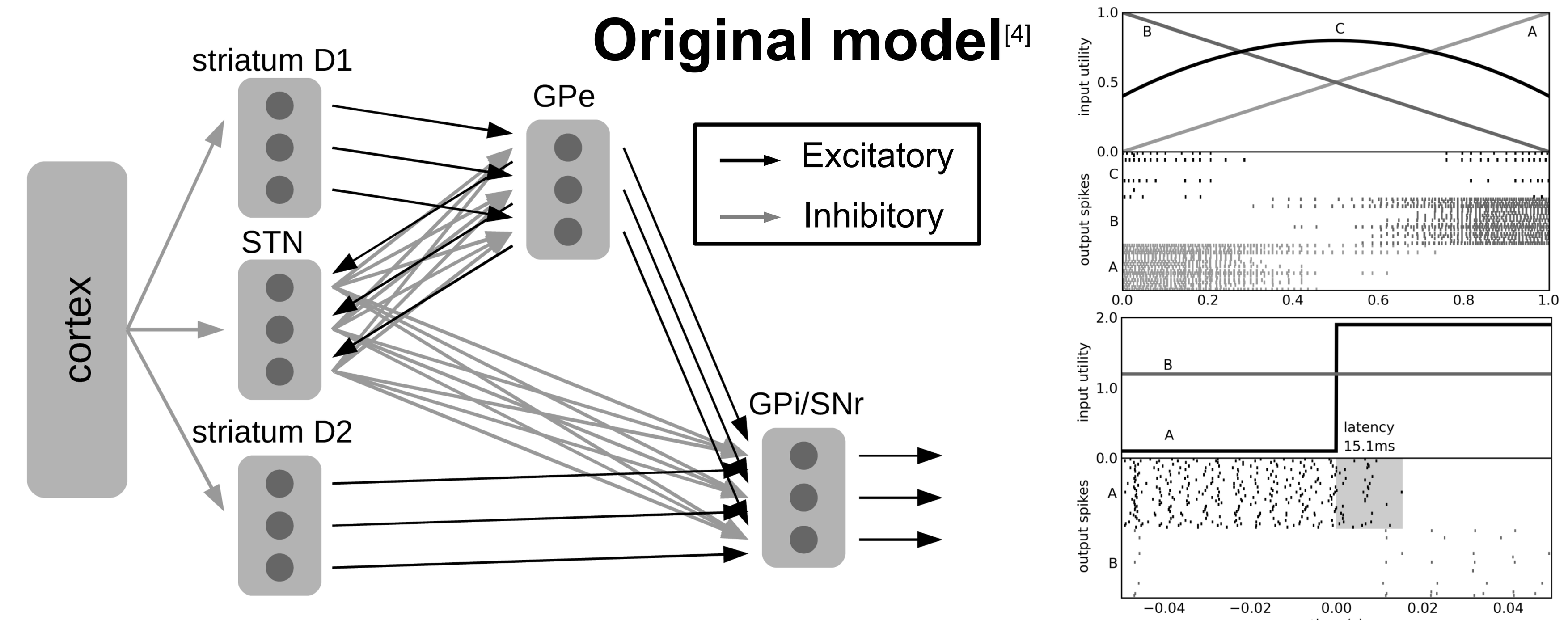
$$\Delta\omega_{ij}(t^{post}) = \kappa\alpha_j \langle \mathbf{e}_i \delta \rangle r_1(t)\left[A_2^+ + A_3^+ o_2(t - dt)\right]$$

Theory: Learning mathematical functions

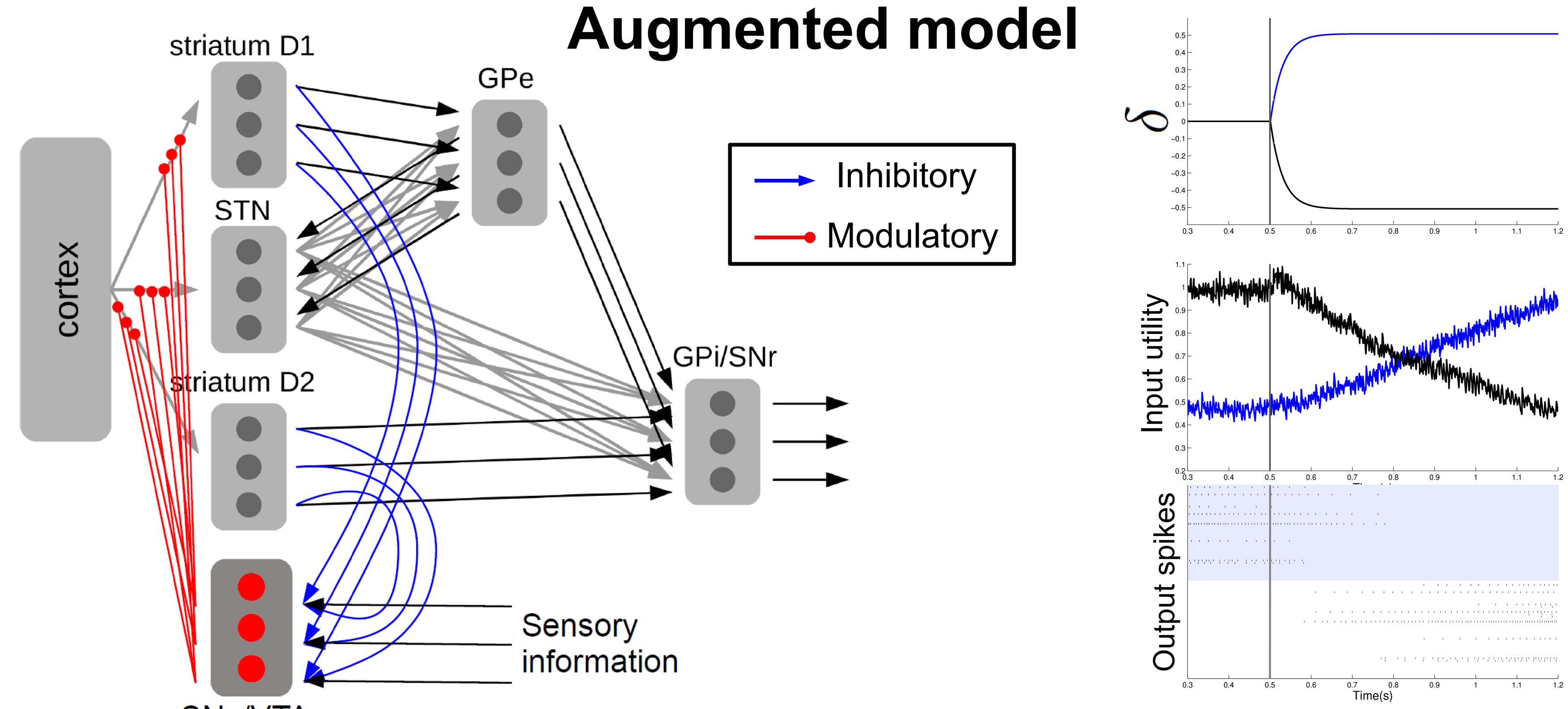


Application: Dynamic action selection in a biologically plausible basal ganglia model

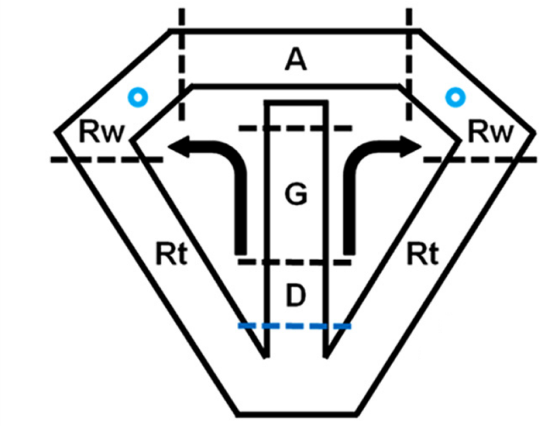
Original model[4]



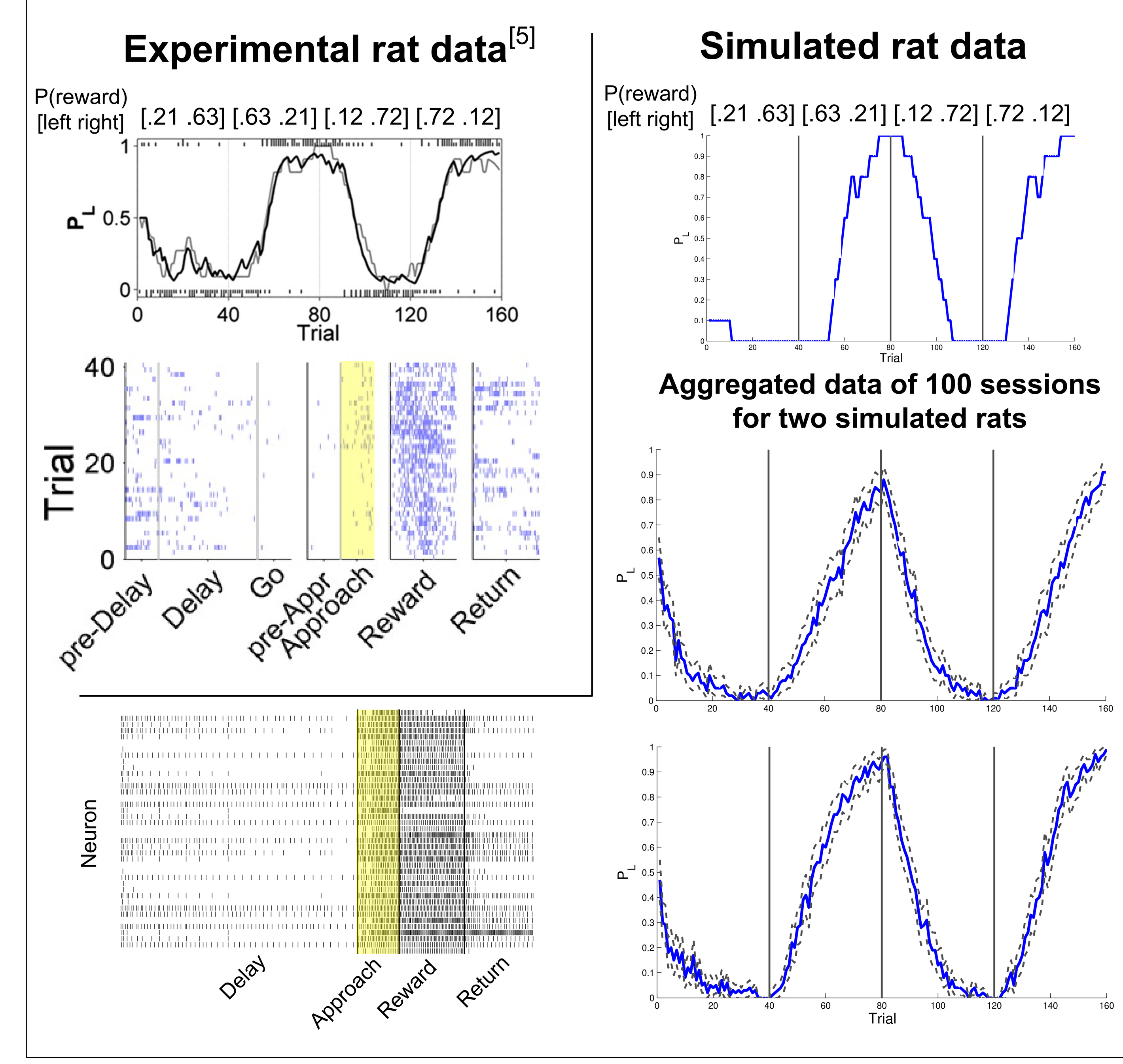
Augmented model



Simulation results from a 2-armed bandit task



1. Animal waits (**D**elay phase).
2. Bridge lowers, allowing animal to move (**G**o phase).
3. Animal reaches decision point, either turns left or right (**A**pproach phase).
4. Reward is stochastically delivered at reward site (**Rew**ard phase).
5. Animal returns to delay area (**Ret**urn phase).



Conclusions

- With a multidimensional error signal, the rule can learn arbitrary transformations in vector spaces
- The rule can be applied to a biologically plausible model of the basal ganglia without adding constraints
- The augmented basal ganglia can replicate behavioural and neural results from an experimental reinforcement learning task

[1] Eliasmith, Chris, and Charles H. Anderson. Neural engineering: computation, representation, and dynamics in neurobiological systems. MIT Press, 2003.
[2] Macneil, David, and Chris Eliasmith. "Fine-tuning and stability of recurrent neural networks." J.Neuroscience, 2010 (submitted).
[3] Pfister, Jean-Pascal, and Wulfram Gerstner. "Triplets of spikes in a model of spike timing-dependent plasticity." J.Neuroscience, 2006.
[4] Stewart, Terrence C., Xuan Choo, and Chris Eliasmith. "Dynamic behaviour of a spiking model of action selection in the basal ganglia." ICCM 2010.
[5] Kim, Hoseok, Jung Hoon Sul, Namjung Huh, Daeyeol Lee, and Min Whan Jung. "Role of striatum in updating values of chosen actions." J.Neuroscience, 2009.