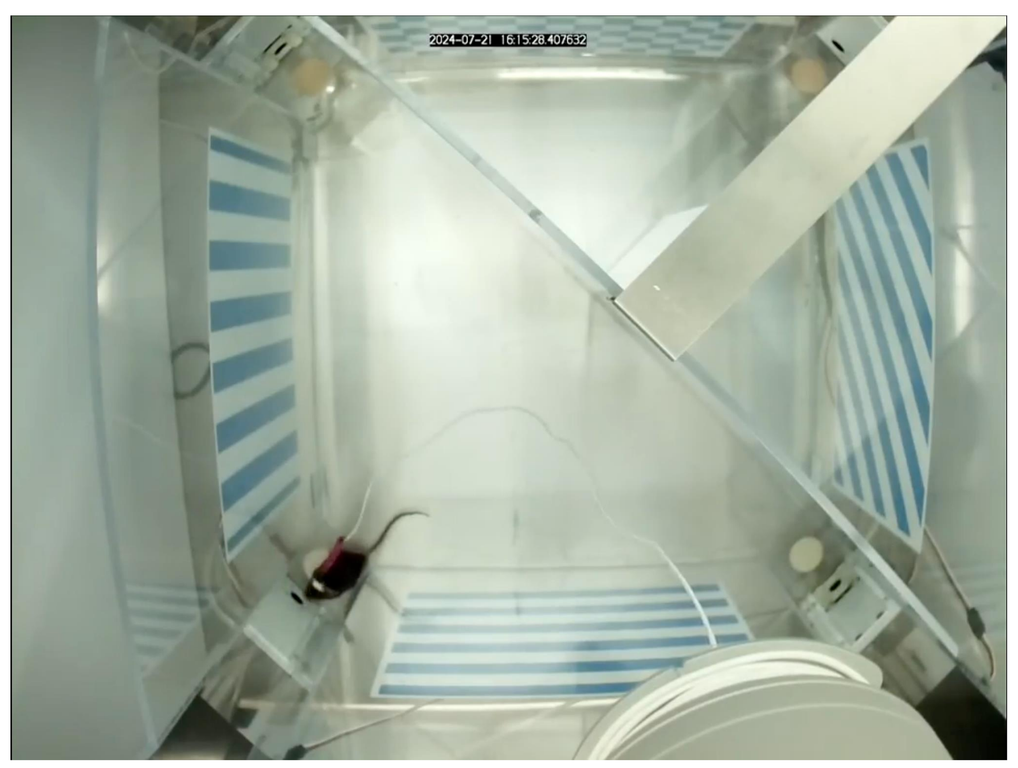
Interpretability Analyses of a Deep Reinforcement Learning Model of Sensory-Place Association

Juan Mendez[1], Alexander Fleischmann[2], Jason Ritt[3], Andrea Pierré[4]

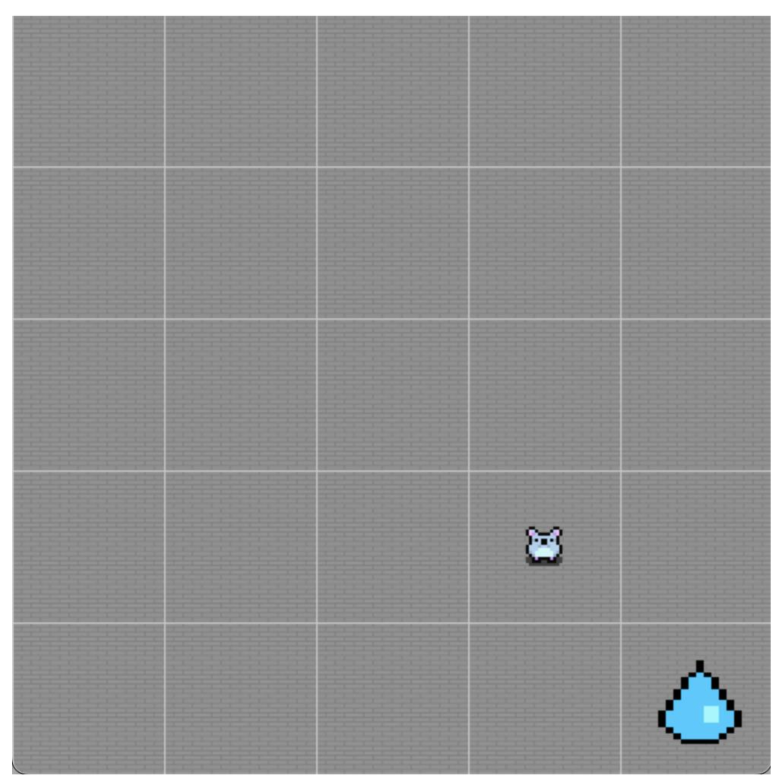
1. Williams College, 2. Dept. of Neuroscience, Brown University, 3. Carney Institute for Brain Science, Brown University, 4. University of Massachusetts Lowell

Introduction

- **Reinforcement learning (RL) is a branch of artificial intelligence where agents learn to make decisions through interaction with an environment.**
- **We studied a learning task in which mice associate an odor cue with a reward at a different spatial location—"sensory-place association" learning[1].**
- **We developed a virtual analog of this task using an RL agent implemented with a Deep Q-Network (DQN)[2]. We aimed to explore the underlying computational principles of both artificial and biological learning.**



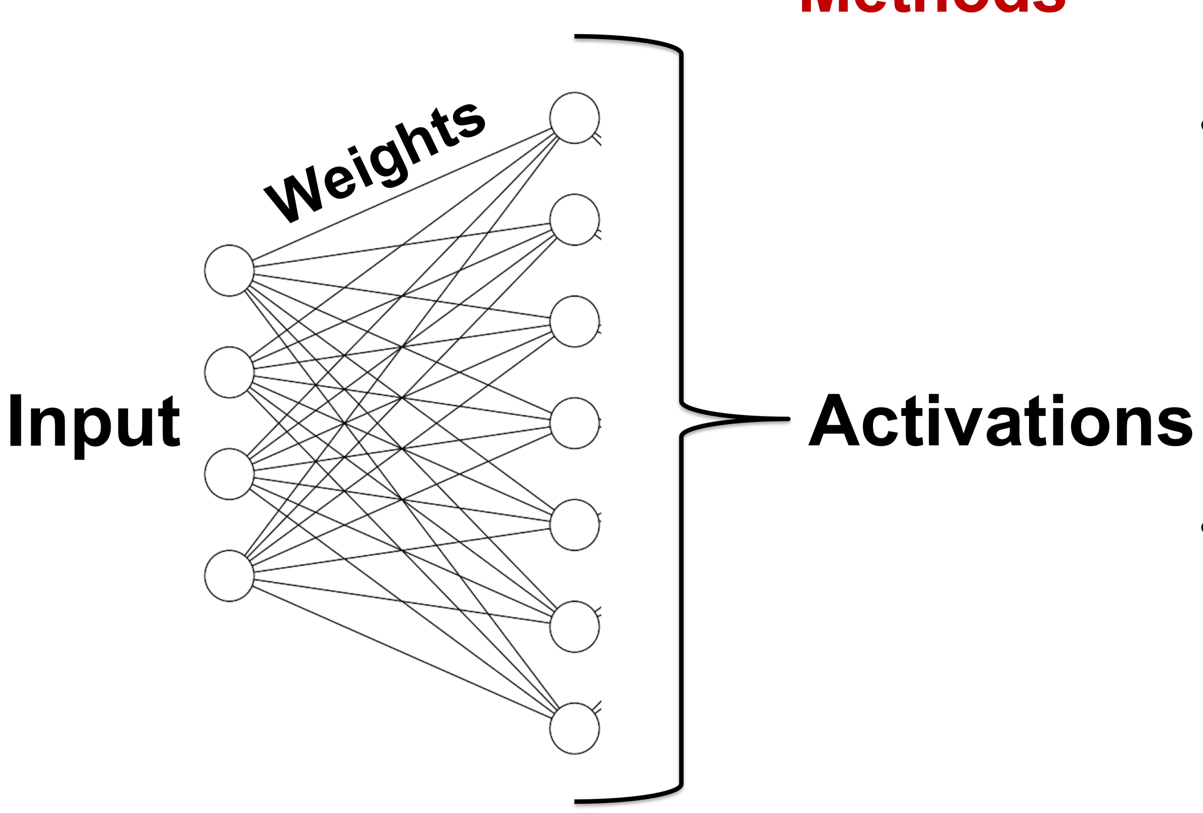
Biological Mouse Task



Artificial RL Task Analog

How can we gain interpretability into RL agent learning?

Methods



- We recorded activations of the network in response to different inputs to try to find any patterns.
- **PCA** is a dimensionality reduction technique that help us visualize high-dimensional activations.

Activations → Behavior

- We used **policy maps** and **3D scatter plots** to try to find some structure in the outputs/behavior of our network.

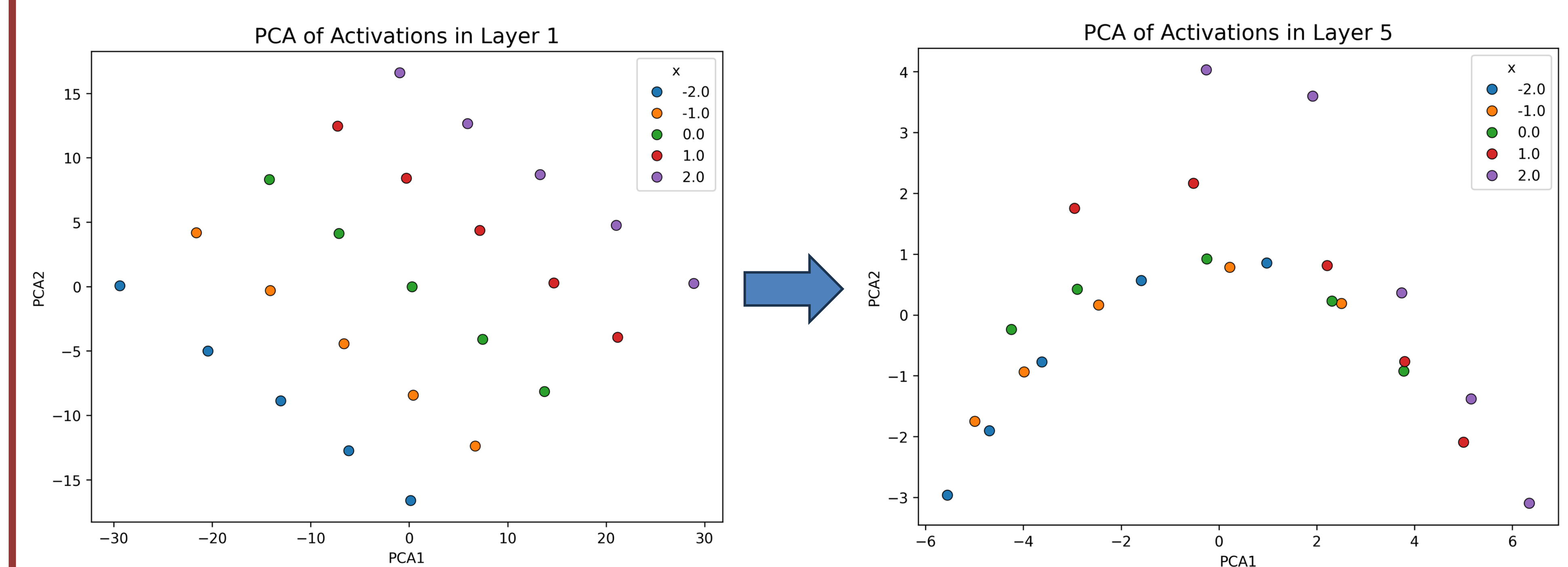
Weights → Activations

- **Hierarchical clustering** and **strongest weight paths** help us see any structure in the weights of our network, possibly explaining any activation patterns we see.

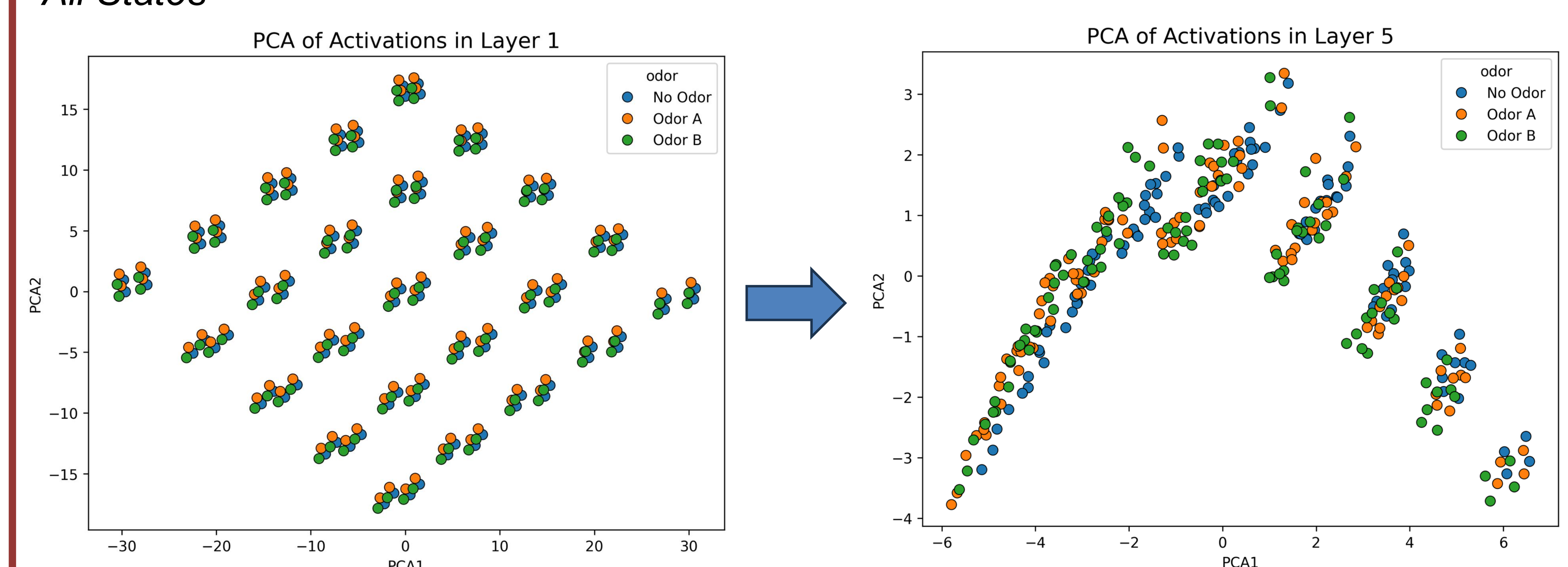
Results

Activations

Odor A, South States

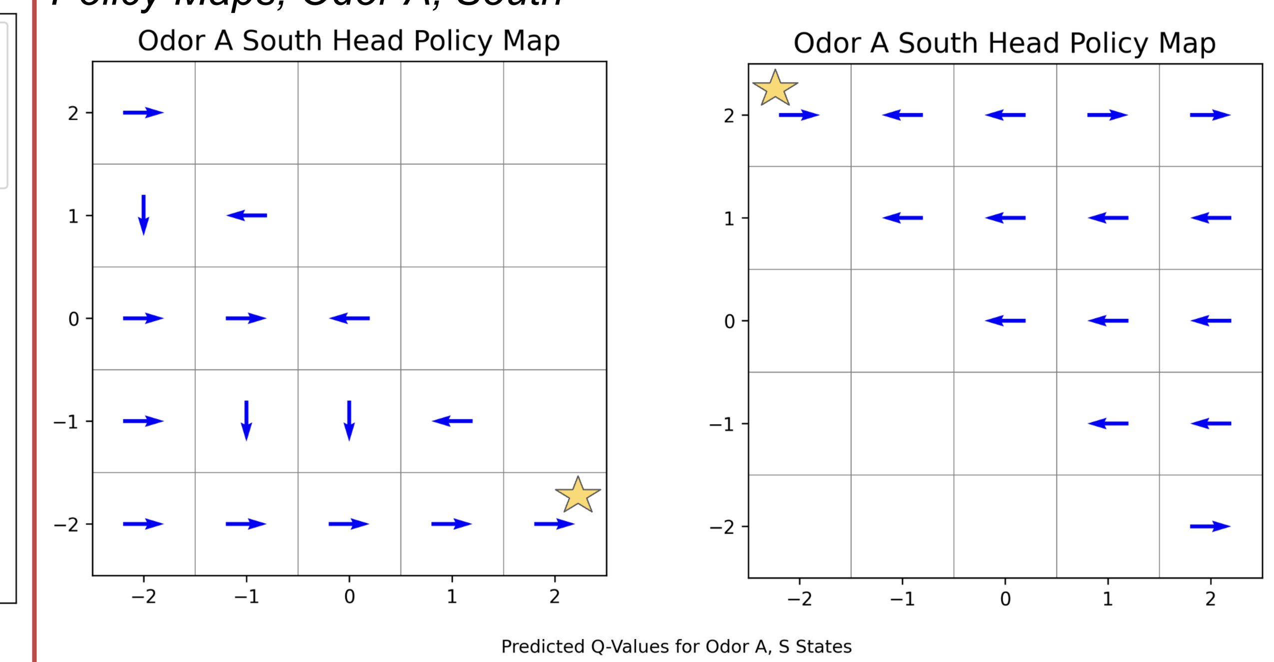


All States

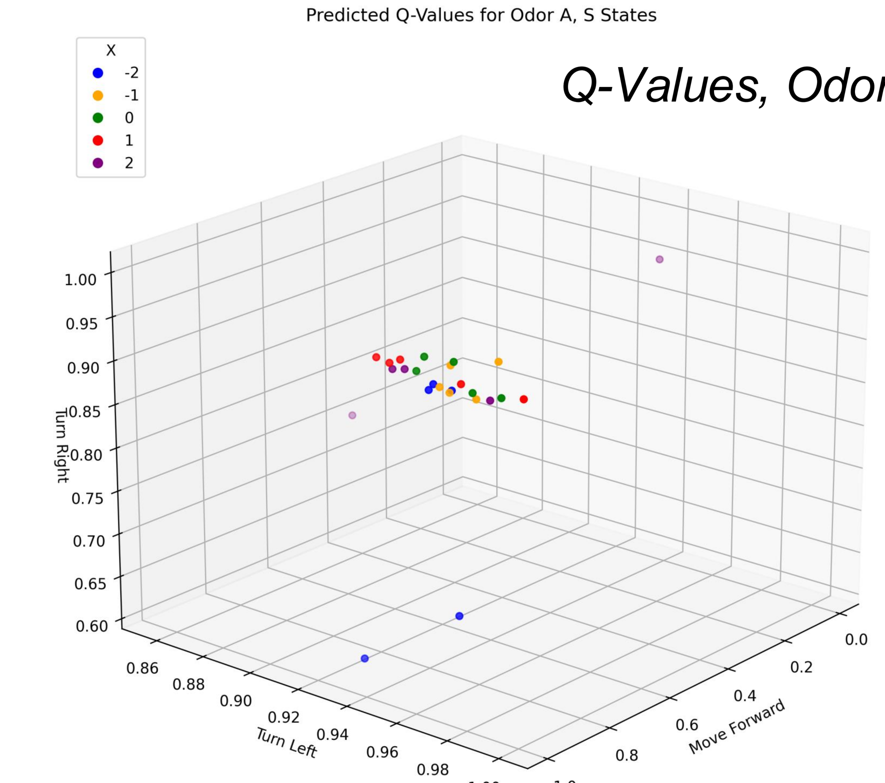


Behavior

Policy Maps, Odor A, South

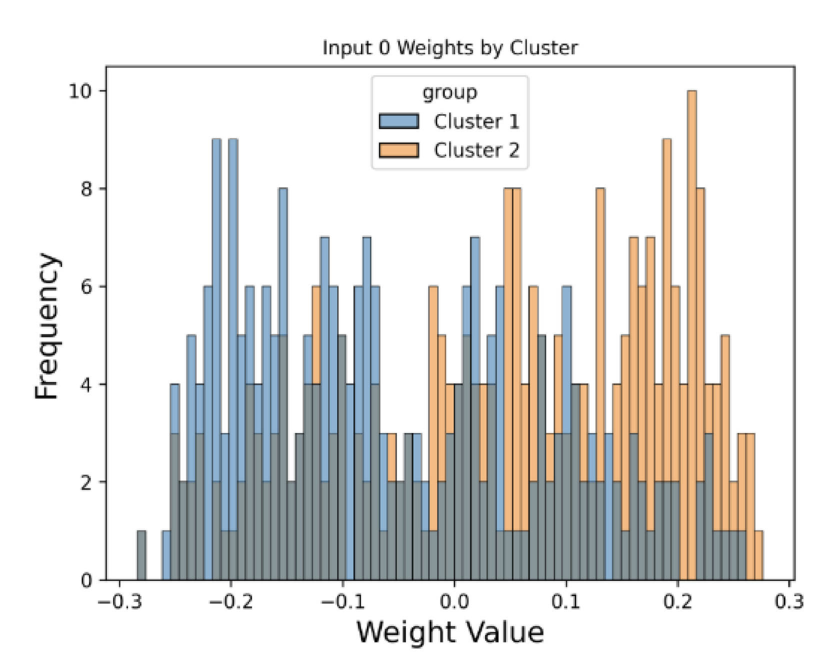


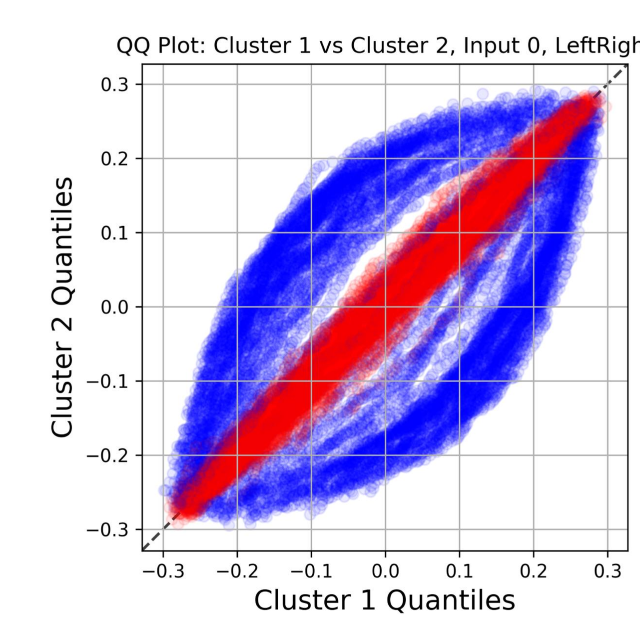
Q-Values, Odor A, South



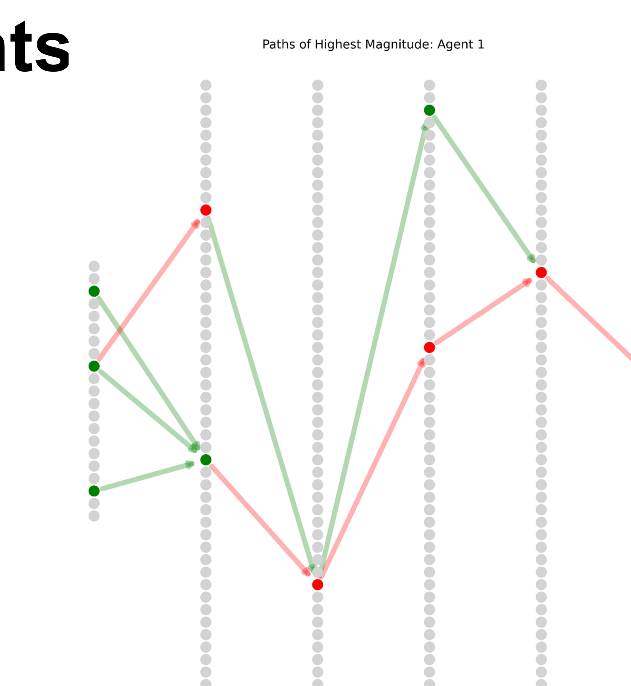
Weights

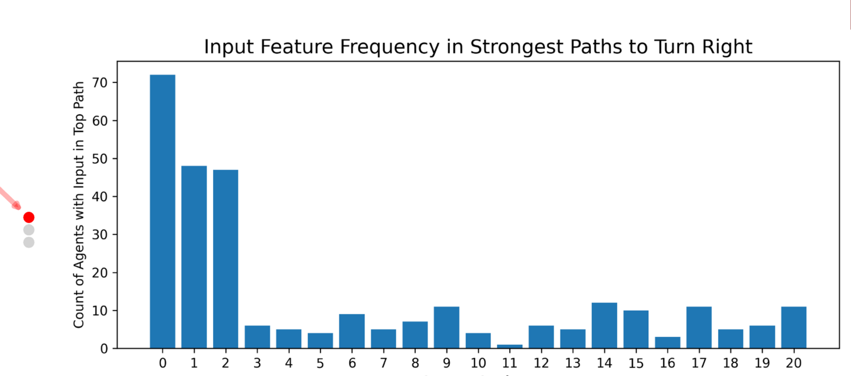
Hierarchical Clustering, NoOdor Weights





Strongest Weight Paths





Conclusions & Future Directions

Activations
- The agent encodes differences between the Upper and Lower triangle, and maintains a spatial map for each

Behavior
- The agent learns to perform the right egocentric action based on head direction, and differs in response between Upper/Lower triangle

Weights
- Weights may help modulate Odor activation differences

Future directions include further investigation into how the agent learns to obtain the Odor, different task comparisons, and whether the agent prefers particular spatial encoding formats (Cartesian coordinates vs. Polar coordinates)

References

(1) McKissick O, Klimpert N, Ritt JT, Fleischmann A. Odors in space. Front Neural Circuits. 2024 Jun 24;18:1414452. doi: 10.3389/fncir.2024.1414452. PMID: 38978957; PMCID: PMC11228174.

(2) Mnih, V., Kavukcuoglu, K., Silver, D. *et al.* Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015). https://doi.org/10.1038/nature14236

Acknowledgments

- Thanks to the Leadership Alliance and the Carney Institute for Brain Science at Brown University for offering the research program!
- Thanks to my advisors Dr. Fleischmann and Dr. Ritt for their careful guidance, as well as collaborator Andrea Pierré for building the RL model on which all these analyses are based on and providing advice.