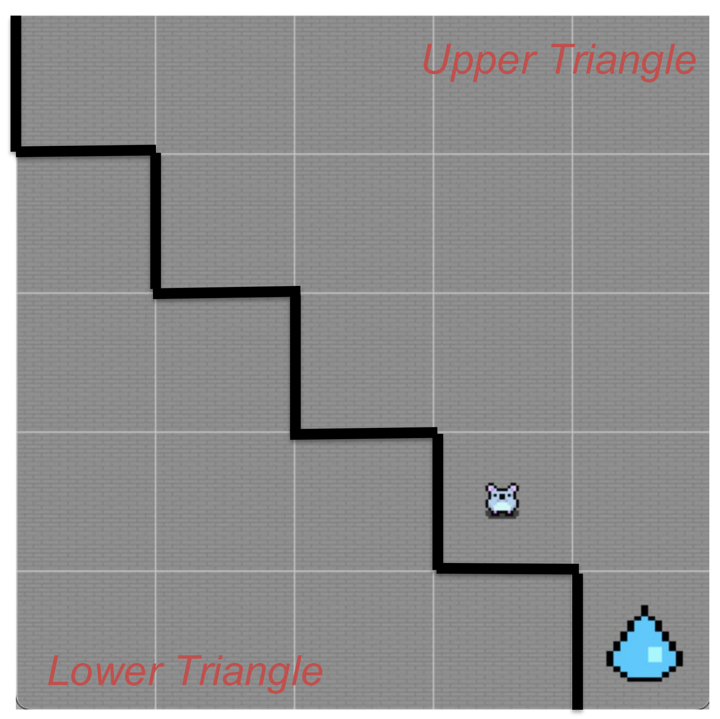
Interpretability Analyses of a Deep Reinforcement Learning Model of Sensory-Place Association

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

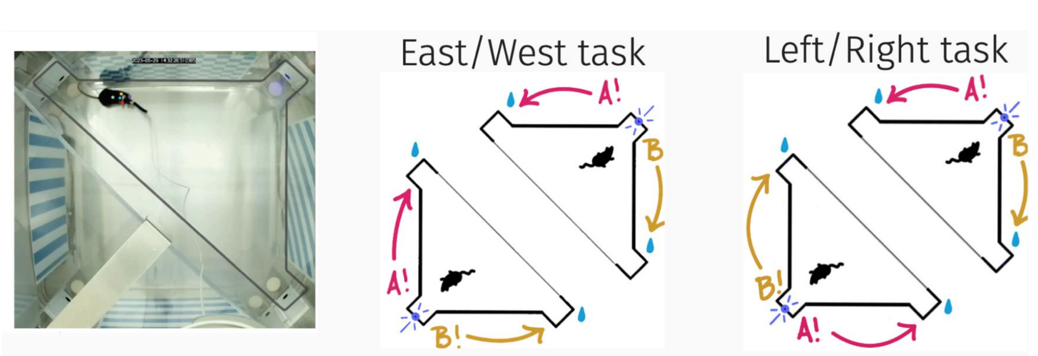
1. Williams College, 2. University of Massachusetts Lowell, 3. Carney Institute for Brain Science, Brown University, 4. Dept. of Neuroscience, Brown University

Introduction

- **We aimed to explore the underlying computational principles of both artificial and biological learning, in particular "sensory-place association" learning[1]. As a first step, we sought to gain interpretability into the learning process of a reinforcement learning (RL) digital agent.**
- **The *experimental* approach involves a mouse learning to associate an odor cue with a position on a triangular arena to receive a reward.**
- **The *computational* approach involves using an RL agent implemented with a Deep Q-Network (DQN)[2] to solve an analogous task in a digital arena.**



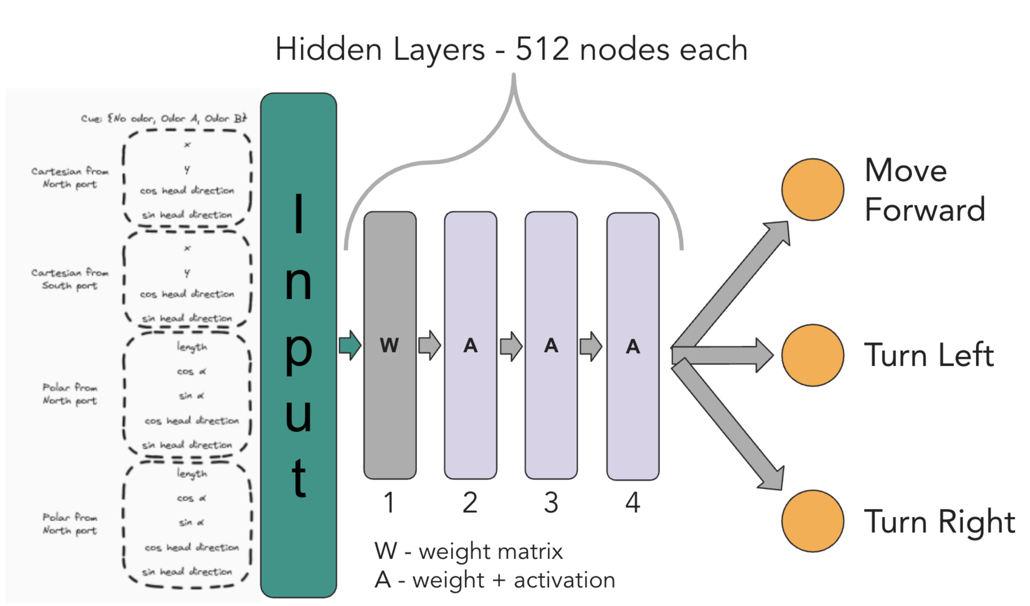
Artificial RL Task



- The *East/West* task emphasizes the learning of allocentric spatial representations
- The *Left/Right* task emphasizes the learning of egocentric spatial representations

How can we gain interpretability into RL agent learning?

Methods



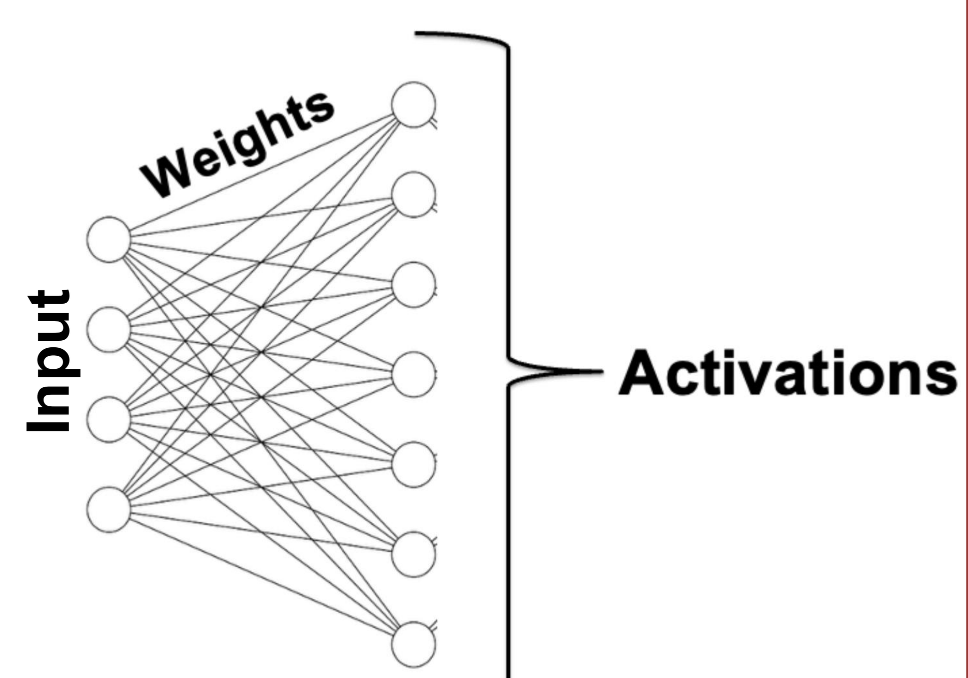
- The network takes in Odor Cue, and four different encodings of Position and Head Direction
- An *episode* is a complete sequence of initialization to reward/failure. The agent is trained over 350 episodes

Weights → Activations

- **Hierarchical clustering** and **strongest weight paths** can help us see what inputs the network pays attention to

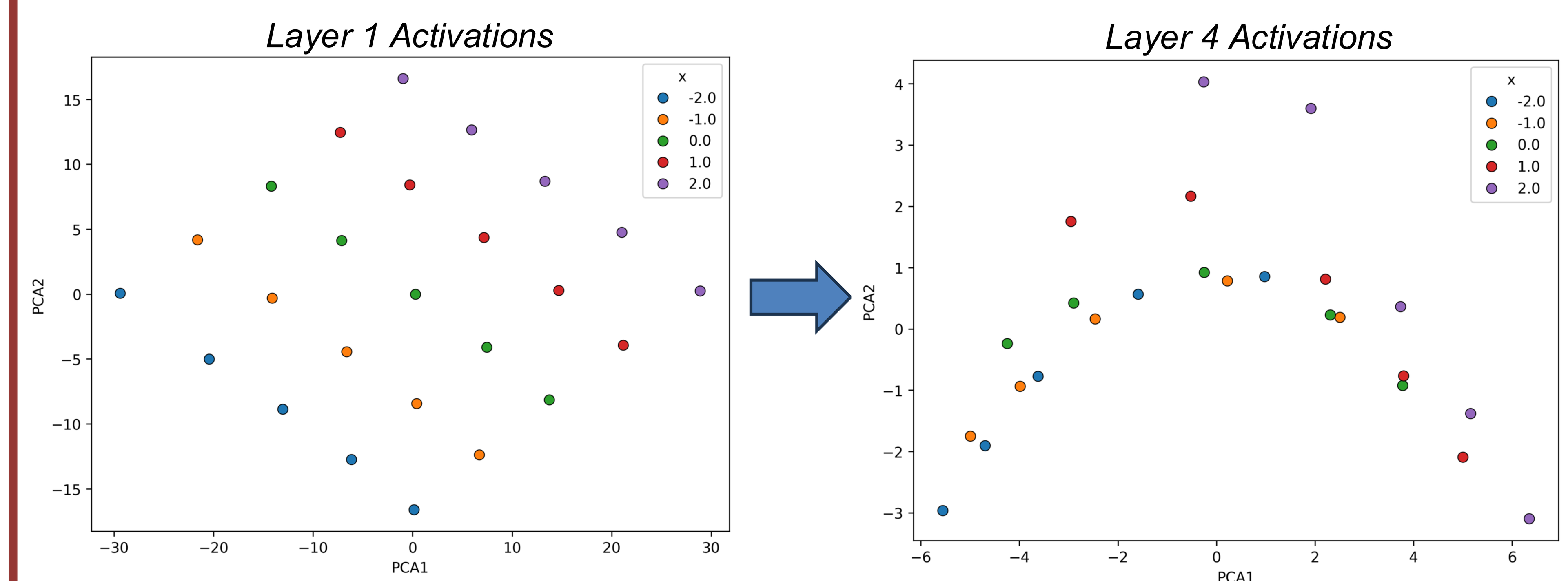
Activations → Behavior

- **PCA** can help us find patterns in the activations, which we can link to behavior using **policy maps**. **SHAP** helps us see feature importance

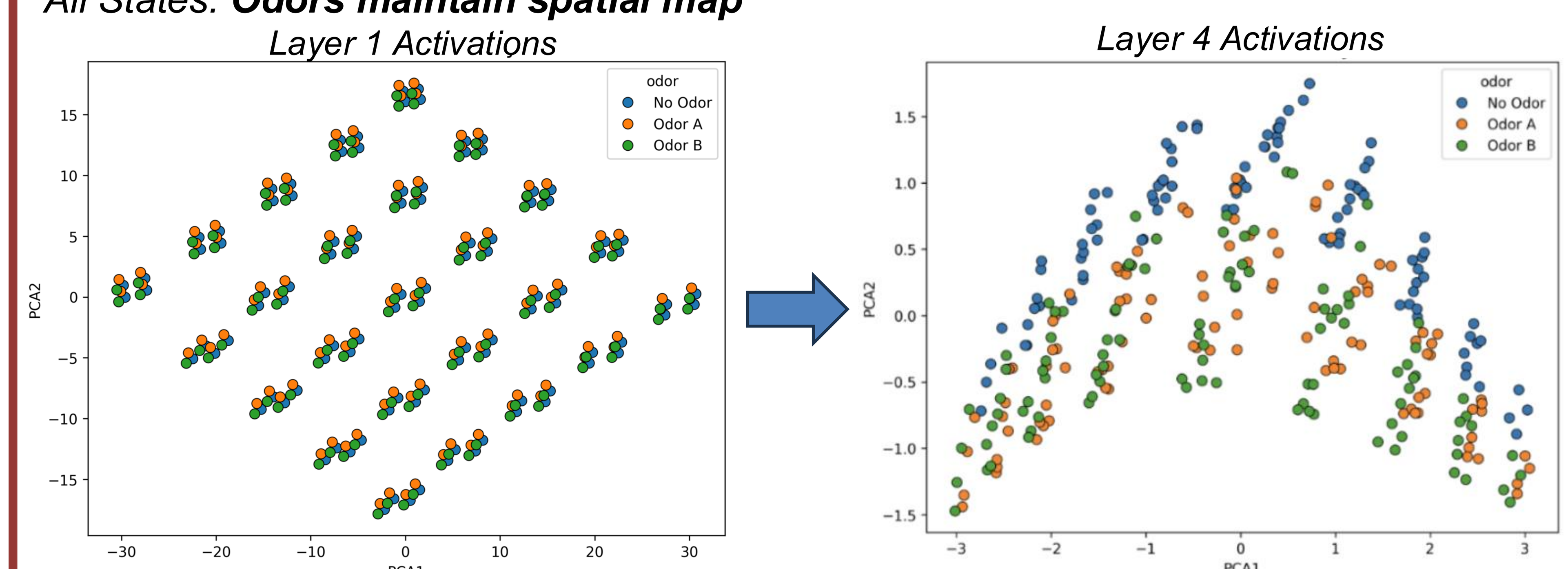


Results

Upper and lower triangle activates differently

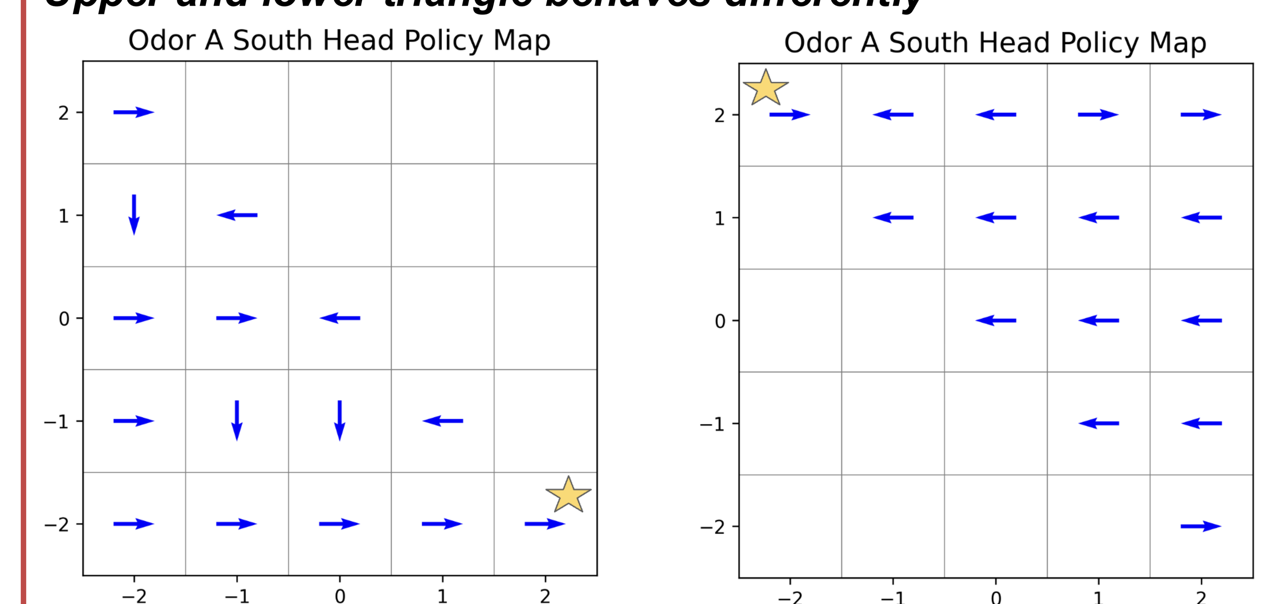


All States. ***Odors maintain spatial map***

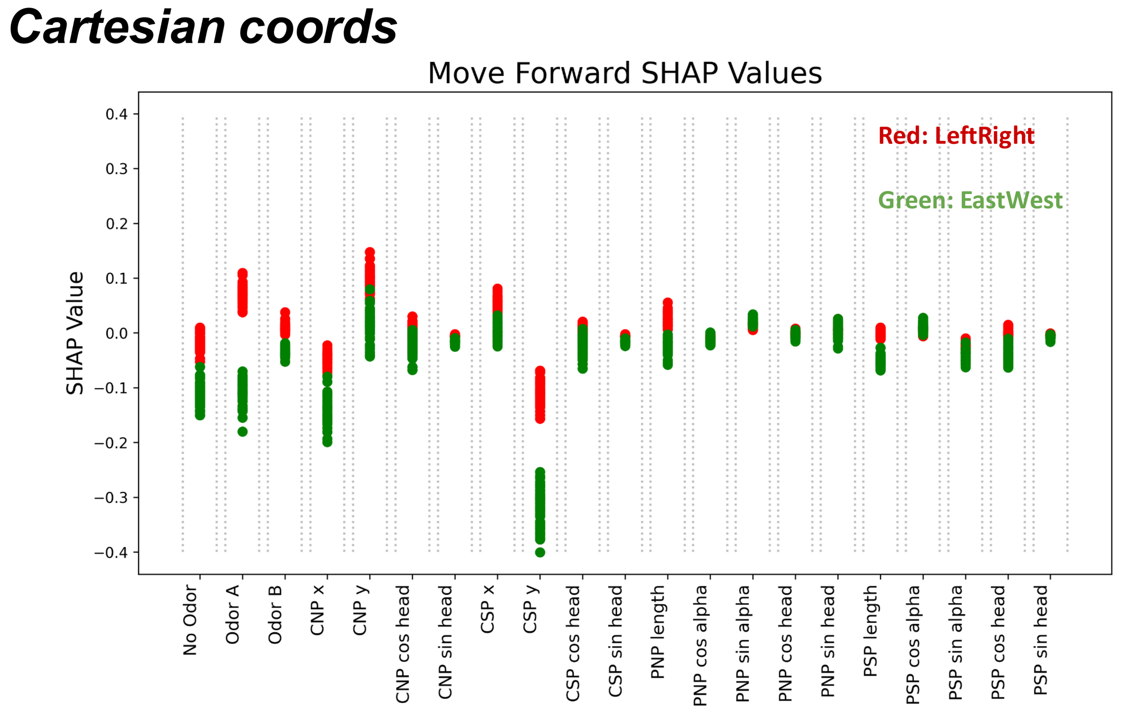


Behavior

Upper and lower triangle behaves differently

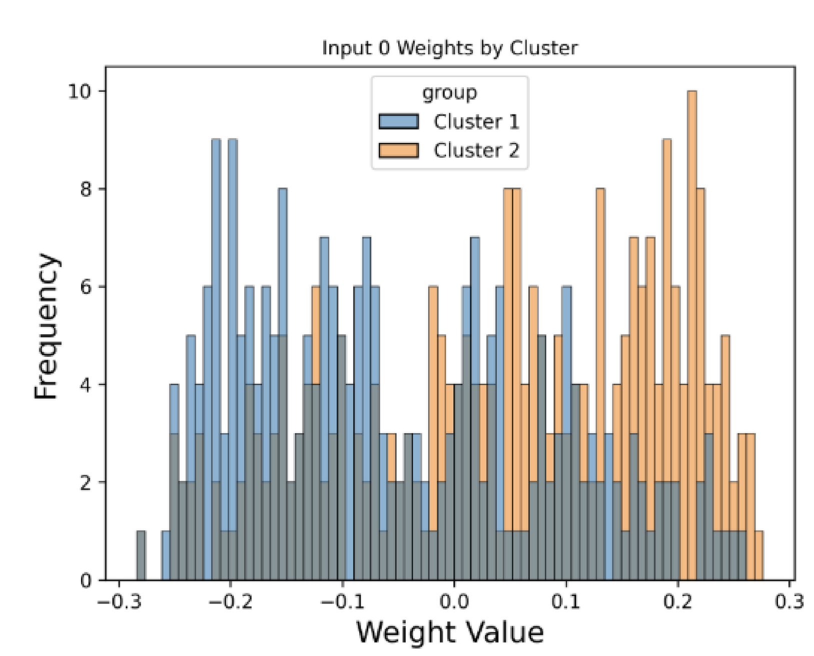


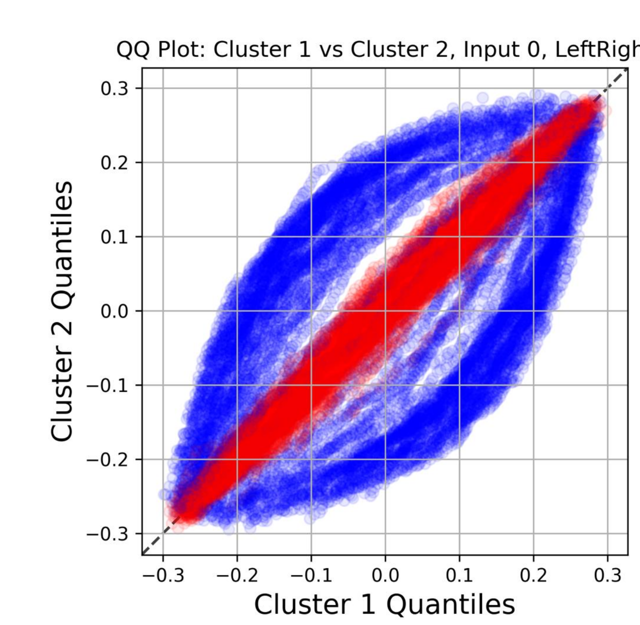
The network assigns importance to Odor and Cartesian coords

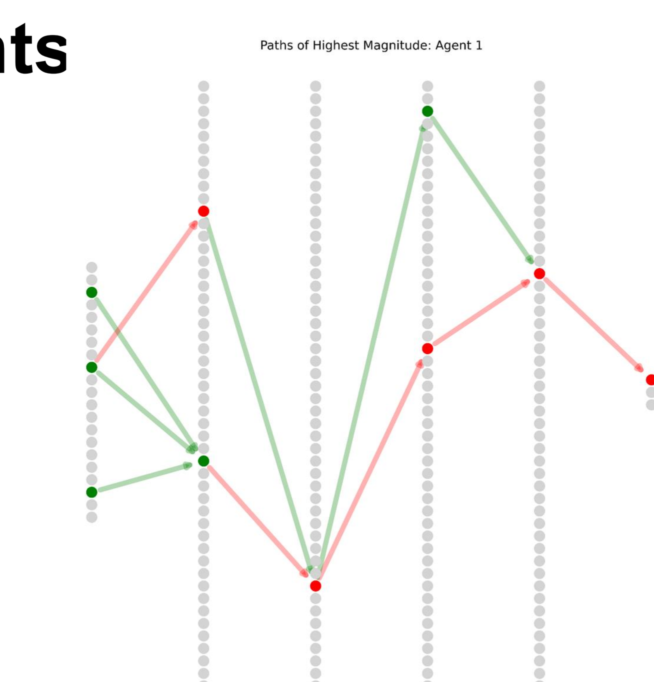


Weights

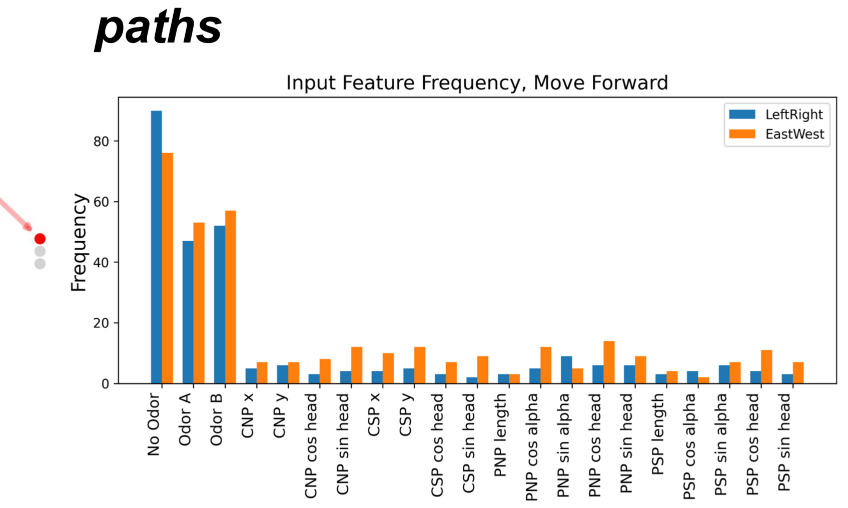
Hidden nodes in first layer split based on NoOdor weight







Odor inputs frequently implicated in 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 behaves differently in the Upper and Lower triangle (corresponding to different reward locations), and assigns more importance to Odor and Cartesian coordinates

Weights

- Weights may help modulate Odor activation differences

Future directions include compressing network size via an autoencoder to reduce the amount of complexity, as well as conduct input perturbation experiments

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!