Investigating Biologically-Inspired Approaches for Continual Reinforcement Learning

Olya Mastikhina
mastikhi@ualberta.ca
University of Alberta

Golnaz Mesbahi
mesbahi@ualberta.ca
University of Alberta

Martha White
whitem@ualberta.ca
University of Alberta

Abstract

Despite the brain's natural ability to continuously learn, biological insights are rarely leveraged in continual reinforcement learning (RL). In this paper, we aim to help bridge this gap by briefly examining four under-investigated biologically-motivated modifications within the context of continual RL: energy minimization, wire length constraints, sparse distributed memory multilayer perceptrons, and fuzzy tiling activations. We show that some of these modifications help increase plasticity and decrease catastrophic forgetting, and we provide an analysis of the learned representations.

1 Introduction

In order to adapt to changing environments in the real world, reinforcement learning (RL) agents need to be able to continually learn by sequentially integrating new information. However, despite advances in the field of deep RL, RL agents are still not great continual learners, limited by losses in plasticity (Lyle et al., 2024) and by catastrophic forgetting (Khetarpal et al., 2022).

In this paper, we propose to draw more inspiration from the remarkable ability of biological brains to continuously learn. We examine four modifications, only one of which has been applied to RL, that can be categorized into two overarching themes inspired by brain functionality: energy constraints and memory indexing. Both themes are related as energy constraints have been suggested to lead to the separation of neural populations representing individual concepts in the brain (Whittington et al., 2023).

Within machine learning itself, energy constraints have previously been shown to lead to disentanglement (Whittington et al., 2023) and to the organization of neurons into clusters that correspond to functional areas within brains (Margalit et al., 2023). Indexing mechanisms, mainly through sparse activations, have been shown to improve transfer learning (Wang et al., 2024) and decrease catastrophic forgetting (Bricken et al., 2023).

For energy constraints, we look at non-negativity with weight and activation minimization (Whittington et al., 2023) and wiring-length constraints (Margalit et al., 2023). For indexing, we examine fuzzy tiling activations (Pan et al., 2021) and a neural network variant of sparse distributed memory (Bricken et al., 2023).

Despite not exhaustively optimizing the implementations, we show that when applied to a Soft Actor Critic algorithm (Haarnoja et al., 2019), a few of the modifications already show promising benefits for continual reinforcement learning through increasing plasticity, reducing overfitting, or decreasing forgetting.

2 Background

2.1 Reinforcement learning

Reinforcement learning is goal-directed learning where an agent interacts with and learns from its environment (Sutton and Barto, 2018). the Markov Decision Process (MDP) framework (Puterman, 2014) formalizes these interactions by defining them in terms of states, actions, and rewards. For each interaction step, the agent is in state s out of a state space $\mathcal{S}$ and selects an action a from an action space $\mathcal{A}$ according to policy π, receives reward r, and transitions to s', another state from the state space $\mathcal{S}$. The agent's goal is to learn a policy π, which maps the agent's actions to states, that maximizes its expected sum of rewards. Past interactions between the agent and the environment are commonly saved in a replay buffer and are revisited as learning progresses (Mnih et al., 2015).

3 Main Experiments

3.1 General Experimental Setup

We evaluate the effects of different modifications on the Soft Actor Critic (SAC) reinforcement learning algorithm (Haarnoja et al., 2019) on sequences of environments. Three of the modifications have previously been implemented in supervised learning (Whittington et al., 2023; Bricken et al., 2023; Margalit et al., 2023), and one in reinforcement learning on DQN (Pan et al., 2021; Mnih et al., 2015). We use the default hyperparameter values for the base agent (Haarnoja et al., 2019), and tune additional modification-specific hyperparameters on the first environment within a sequence.

3.1.1 Environments

To evaluate the effects on negative and forward transfer, we use environments from the DeepMind Control Suite (Tassa et al., 2018). We evaluate negative transfer on `quadruped-run` following pre-training on `quadruped-walk` as SAC shows decreased performance on the second environment following the pre-training. For effects on forward transfer, we look at `humanoid-run` following pre-training on `humanoid-walk`, where unmodified SAC shows conversely higher performance on the second environment.

We evaluate catastrophic forgetting and overfitting on robot arm tasks from Metaworld Yu et al. (2021). For catastrophic forgetting, we use separate output heads and task IDs, and train on a thematically-related sequence: `faucet-close-v2` → `window-close-v2`→ `faucet-close-v2`. We look at forgetting of the previous environments, and on how quickly the third environment is remembered. For overfitting, we use one output head and train on a sequence of three `hammer-v2` tasks with separate one-hot vector input IDs for each. We reset the replay buffers between environments for all cases.

3.2 Modifications Details

3.2.1 Energy Constraints

In addition to broader regions of functional organization, brains contain cells that code for individual factors of variation within a task space, such as object vector cells (Høydal et al., 2019), or border cells (Solstad et al., 2008). How these localized representations form is not fully understood, but energy constraints have been hypothesised to play a key role (Whittington et al., 2023; Margalit et al., 2023).

As this form of disentanglement may aid with continual learning, we evaluated two different types of energy constraints on SAC: biological constraints of non-negativity and energy minimisation (Whittington et al., 2023), which we refer to as "small-bio", and wiring length constraints (Margalit et al., 2023).

We found that energy constraints worked best when applied to the actor as well as the critics (not shown).

Non-negativity and Energy Minimization - "small-bio" Biological constraints of non-negativity and energy minimization in conjunction, or "small-bio", have been shown to promote disentanglement in neural networks (Whittington et al., 2023), with disentanglement defined here as more individual neurons coding for individual factors of variation in the input data.

For `small-bio`, energy minimization is simply imposed through l2 regularization of activations as well as weights, and non-negativity can be imposed with the ReLU activation function (Nair and Hinton, 2010; Whittington et al., 2023), which is what we do in this paper. Consistently with the small-bio paper, we found l2 regularization of both activation and weights to have better effects than l2 regularization applied to only the weights or only the activations (not shown).

Wiring Length Constraints - "wire" Wiring length constraints have been shown to recreate representations created by the brain by introducing a loss to encourage nearby neurons to have similar representations (Margalit et al., 2023). Wiring length constraints have not originally been shown to have a clear performance benefit, but to increase interpretability; we nonetheless saw improvement with `wire` in certain cases (shown in the Preliminary Results section). We applied wiring length constraints on the second hidden layer in the actor and critic networks.

3.2.2 Indexing

We use indexing as an umbrella term for theories and findings behind the brain developing neurons that index into a broader concept or memory (Teyler and DiScenna, 1986; Teyler and Rudy, 2007; McClelland et al., 1995; O'Reilly et al., 2014), which may be beneficial for creating orthogonal representations that would reduce catastrophic forgetting in neural networks.

Below, SDMLP and FTA are sparse activation approaches that we see as consistent with the complementary learning systems theory, in which the hippocampus formation learns sparse, non-overlapping representations that then index the overlapping and distributed representations in the neocortex (McClelland et al., 1995; O'Reilly et al., 2014). Both modifications also have ties to circuits in the cerebellum (Xie et al., 2023; Sutton, 1995; Albus, 1971).

We found that both indexing constraints worked best when applied only to the actor, and not to the critics (not shown).

Fuzzy Tiling Activations (FTA) FTA is an activation function that induces sparsity in neural networks by binning inputs with a fuzzy indicator function (Pan et al., 2021). It has previously been shown to increase transfer learning in Deep Q-Networks (Mnih et al., 2015; Wang et al., 2024). We apply FTA to the actor's second hidden layer.

Sparse Distributed Memory Multilayer Perceptron (SDMLP) Closely related to Hopfield networks, Sparse Distributed Memory (SDM) is a mathematical associative memory model of how concepts, or patterns, are stored and retrieved in the brain (Kanerva, 1988; 1992).

SDMLP is a one hidden layer neural network implementation of SDM that treats input weights into the hidden layer as addresses, and output weights as patterns (Bricken et al., 2023). To make this work for continual learning, modifications include employing a top-k activation function, eliminating neural network bias terms, and enforcing l2 normalization and non-negativity constraints on the weights and data (Bricken et al., 2023).

We implement SDMLP on the actor network and handle data and output non-negativity by passing the inputs through a CReLU activation function (Shang et al., 2016), and by doubling the output layer and subtracting one half of the outputs from the other half. For top-k, we follow the method of top-k annealing, where the top-k number gradually decreased throughout training, instead an alternative GABA switch method also presented in the SDMLP paper (Bricken et al., 2023). However, from preliminary analysis, the GABA switch method appears to generate better performance in at least one set of the environments tested (not shown).

The SDMLP paper authors additionally identified a stale momentum problem with optimizers like Adam and RMSProp (Swerksy et al., 2012; Kingma and Ba, 2015), and found that using Stochastic Gradient Descent (SGD) without momentum (Rosenblatt, 1958) worked the best for SDMLP. Unfortunately, we were not yet able to achieve good results ourselves with SGD. Because of this, we expect that the GABA switch method, which reportedly suffers less from the stale momentum problem (Bricken et al., 2023), will likely end up being the most appropriate implementation in future work.

3.3 Preliminary Results

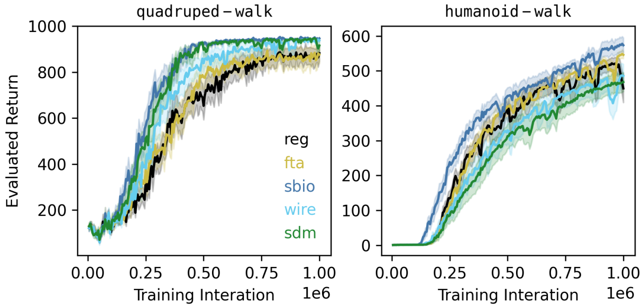

Figure 1: Performance on the first environment for quadruped and humanoid. sbio is small-bio, and sdm is SDMLP. The shading is SEM, and there are 20 seeds per run.

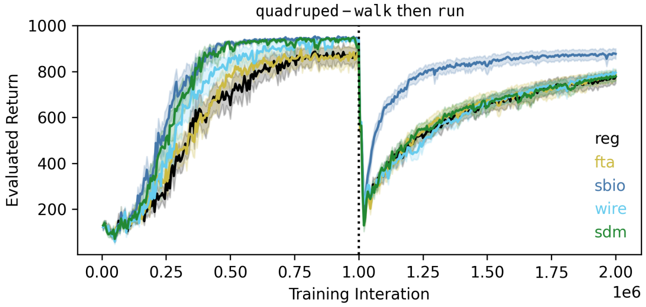

Figure 2: Performance on the second environment for quadruped and humanoid. The shading is SEM, and there are 20 seeds per run.

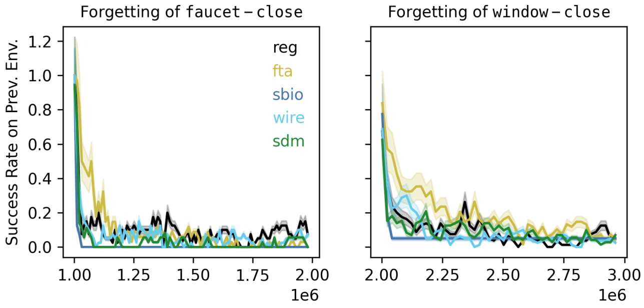

Figure 3: FTA slightly slows down forgetting. Forgetting of the first (left) and second (right) environment in the sequence of faucet-close -> window-close -> faucet-close Metaworld robot arm tasks. The shading is SEM, and there are 20 seeds per run.

Forward and Negative Transfer Figure 1 shows that `small-bio` in particular improves performance on individual environments. `small-bio`, `wire`, and SDMLP increase performance in `quadruped`, but not `humanoid`. However, Figure 2 only shows that only `small-bio` increases performance following pretraining in both. We suspect that the poorer results with SDMLP may be due to the stale momentum problem reported in the original SDMLP paper (Bricken et al., 2023) that we have not yet been able to resolve in this paper with SAC, and FTA and `wire` may benefit from more refined implementations in future work.

Overfitting Figure 4 shows that FTA and `small-bio` decrease overfitting on robot arm tasks by showing comparable performance across training as well as global success on one environment with different task IDs. SDMLP, however, fully collapses.

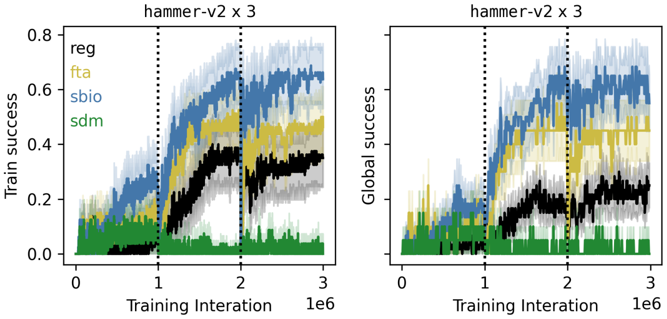

Figure 4: small-bio and FTA decrease overfitting on three hammer-v2 tasks with separate task IDs. Left is training success, right is overall success on three hammer-v2s. 20 seeds per run, and shading is SEM.

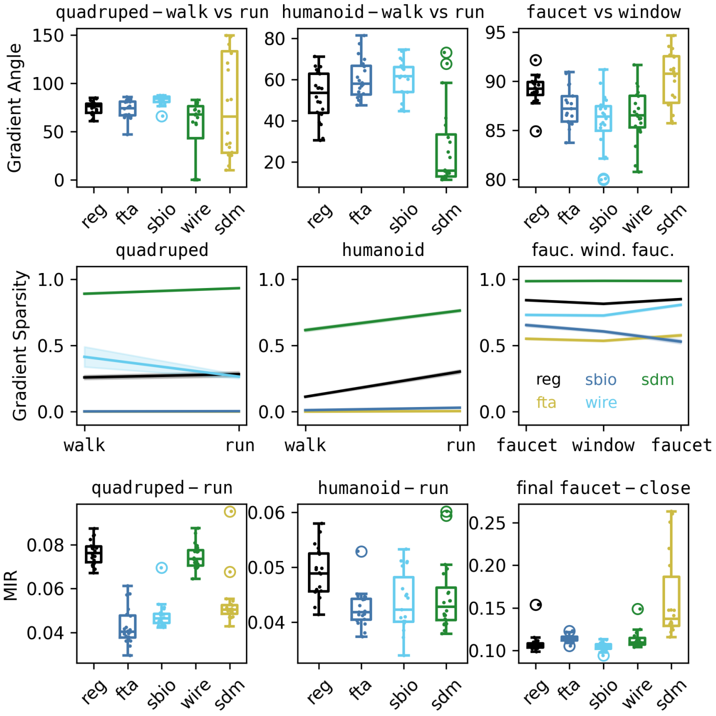

Figure 5: SDM maintains a high gradient sparsity, while small-bio and wire maintain a low one. The gradient orthogonality and disentanglement (MIR) measurements do not suggest a trend. The measurements are for the actor's final hidden layer.

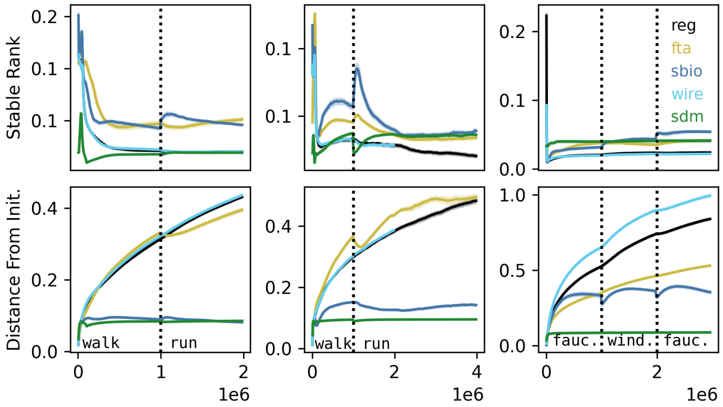

Figure 6: small-bio and FTA maintain a higher stable rank, and small-bio and SDMLP in particular maintain a low distance from initialization.

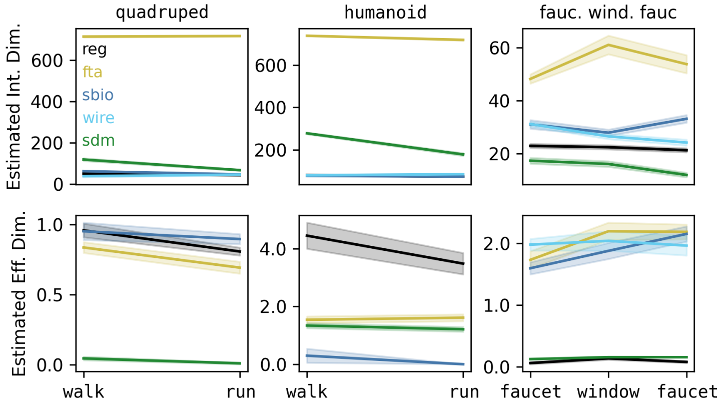

Figure 7: FTA greatly increases the intrinsic dimensionality of the data. The measurements are for the actor's final hidden layer.

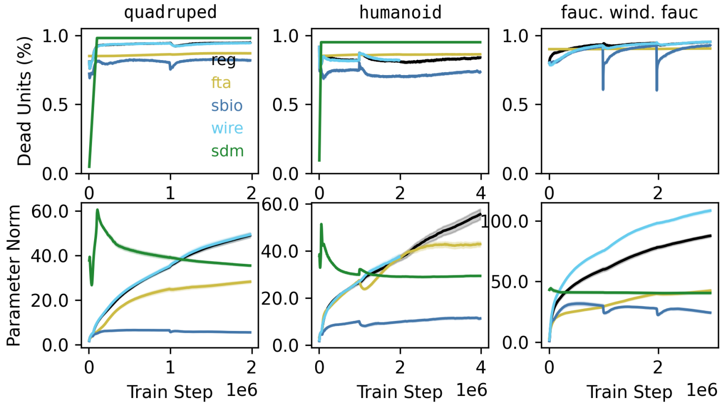

Figure 8: FTA and small-bio both decrease the percentage of dead units as well as the l2 norm of the weights. The measurements are for the actor's final hidden layer.

Preliminary Analysis Figures 5, and 6, 7 show an analysis of the representations produced by the actor networks. Despite previously demonstrated increases in disentanglement with `small-bio` using the mutual information ratio (MIR) metric (Whittington et al., 2023), we do not see this with `small-bio` applied to SAC. Figure 5 instead shows inconsistent results across environments for different modifications, which may indicate MIR being a poor metric for reinforcement learning data.

Figure 6 shows that the most successful modification, `small-bio`, maintains a high stable rank and a low distance from initialization, both beneficial for generalization (Sanyal et al., 2019; Nagarajan and Kolter, 2019). However, Figure 5 also shows it maintaining the lowest gradient sparsity, which may be an indication of why we see the fastest forgetting with `small-bio` in Figure 3.

Similarly, increases in dead units and in the norm of the parameters are potential mechanisms of plasticity loss (Lyle et al., 2024), and Figure 8 shows both `small-bio` and FTA, with `small-bio` in particular, maintaining the lowest percentage of dead units as well as parameters norms. Interestingly, `small-bio` also shows a decrease in dead units and in the parameter norm with task switches in the `faucet-window-faucet` set of environments.

4 Conclusion and Discussion

With this paper, we present a preliminary overview of biologically-inspired modifications that are understudied in continual reinforcement learning. While the other modifications may benefit from more thoroughly optimized implementations, energy constraints and non-negativity with `small-bio` have shown the most promise despite their simplicity. We hope that this work promotes further discussions and investigations into algorithms inspired by brain function in continual reinforcement learning.

References

James S. Albus. A theory of cerebellar function. *Mathematical Biosciences*, 10(1):25–61, February 1971. ISSN 0025-5564.

Trenton Bricken, Xander Davies, Deepak Singh, Dmitry Krotov, and Gabriel Kreiman. Sparse Distributed Memory is a Continual Learner. February 2023.

Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. Soft Actor-Critic Algorithms and Applications, January 2019.

Øyvind Arne Høydal, Emilie Ranheim Skytøen, Sebastian Ola Andersson, May-Britt Moser, and Edvard I. Moser. Object-vector coding in the medial entorhinal cortex. *Nature*, 568(7752), April 2019. ISSN 1476-4687.

Pentti Kanerva. *Sparse Distributed Memory.* MIT Press, 1988. ISBN 978-0-262-11132-4.

Pentti Kanerva. Sparse distributed memory and related models. Technical Report NASA-CR-190553, April 1992.

Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup. Towards Continual Reinforcement Learning: A Review and Perspectives. *Journal of Artificial Intelligence Research*, 75: 1401–1476, December 2022. ISSN 1076-9757.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. International Conference on Learning Representations (ICLR), 2015.

Clare Lyle, Zeyu Zheng, Khimya Khetarpal, Hado van Hasselt, Razvan Pascanu, James Martens, and Will Dabney. Disentangling the Causes of Plasticity Loss in Neural Networks, February 2024.

Eshed Margalit, Hyodong Lee, Dawn Finzi, James J. DiCarlo, Kalanit Grill-Spector, and Daniel L. K. Yamins. A Unifying Principle for the Functional Organization of Visual Cortex, May 2023.

James L. McClelland, Bruce L. McNaughton, and Randall C. O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3), 1995. ISSN 1939-1471.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540), February 2015. ISSN 1476-4687.

Vaishnavh Nagarajan and J. Zico Kolter. Generalization in Deep Networks: The Role of Distance from Initialization, January 2019.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 807–814, Madison, WI, USA, June 2010. Omnipress. ISBN 978-1-60558-907-7.

Randall C. O'Reilly, Rajan Bhattacharyya, Michael D. Howard, and Nicholas Ketz. Complementary learning systems. *Cognitive Science*, 38(6):1229–1248, August 2014. ISSN 1551-6709.

Yangchen Pan, Kirby Banman, and Martha White. Fuzzy Tiling Activations: A Simple Approach to Learning Sparse Representations Online. January 2021.

Martin L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley and Sons, 2014.

F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 1958. ISSN 1939-1471.

Amartya Sanyal, Philip H. Torr, and Puneet K. Dokania. Stable Rank Normalization for Improved Generalization in Neural Networks and GANs. September 2019.

Wenling Shang, Kihyuk Sohn, Diogo Almeida, and Honglak Lee. Understanding and Improving Convolutional Neural Networks via Concatenated Rectified Linear Units. In *Proceedings of The 33rd International Conference on Machine Learning*. PMLR, June 2016.

Trygve Solstad, Charlotte N. Boccara, Emilio Kropff, May-Britt Moser, and Edvard I. Moser. Representation of Geometric Borders in the Entorhinal Cortex. *Science*, 322(5909), December 2008.

Richard S Sutton. Generalization in Reinforcement Learning: Successful Examples Using Sparse Coarse Coding. In *Advances in Neural Information Processing Systems*, volume 8. MIT Press, 1995.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: an introduction*. Adaptive computation and machine learning series. The MIT Press, Cambridge, Massachusetts, second edition edition, 2018. ISBN 978-0-262-03924-6.

Kevin Swerksy, Geoffrey Hinton, and Nitish Srivastava. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. In *Coursera Deep Learning Lecture Slides*, 2012.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy Lillicrap, and Martin Riedmiller. DeepMind Control Suite, January 2018.

Timothy J. Teyler and Pascal DiScenna. The hippocampal memory indexing theory. *Behavioral Neuroscience*, 100(2), 1986. ISSN 1939-0084.

Timothy J. Teyler and Jerry W. Rudy. The hippocampal indexing theory and episodic memory: Updating the index. *Hippocampus*, 17(12), 2007. ISSN 1098-1063.

Han Wang, Erfan Miahi, Martha White, Marlos C. Machado, Zaheer Abbas, Raksha Kumaraswamy, Vincent Liu, and Adam White. Investigating the properties of neural network representations in reinforcement learning. *Artificial Intelligence*, 330:104100, May 2024. ISSN 0004-3702.

James C. R. Whittington, Will Dorrell, Surya Ganguli, and Timothy Behrens. Disentanglement with Biological Constraints: A Theory of Functional Cell Types. February 2023.

Marjorie Xie, Samuel P Muscinelli, Kameron Decker Harris, and Ashok Litwin-Kumar. Task-dependent optimal representations for cerebellar learning. *eLife*, 12, September 2023. ISSN 2050-084X.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Avnish Narayan, Hayden Shively, Adithya Bellathur, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-World: A Benchmark and Evaluation for Multi-Task and Meta Reinforcement Learning, June 2021.