Biological Neurons vs Deep Reinforcement Learning: Sample efficiency in a simulated game-world

Anonymous Author(s)
Affiliation
Address
email

Abstract

How do synthetic biological systems and artificial neural networks compete in their performance in a game environment? Reinforcement learning has undergone significant advances, however remains behind biological neural intelligence in terms of sample efficiency. Yet most biological systems are significantly more complicated than most algorithms. Here we compare the inherent intelligence of in vitro biological neuronal networks to state-of-the-art deep reinforcement learning algorithms in the arcade game 'pong'. We employed DishBrain, a system that embodies *in vitro* neural networks with in silico computation using a high-density multielectrode array. We compared the learning curve and the performance of these biological systems against time-matched learning from DQN, A2C, and PPO algorithms. Agents were implemented in a reward-based environment of the 'Pong' game. Key learning characteristics of the deep reinforcement learning agents were tested with those of the biological neuronal cultures in the same game environment. We find that even these very simple biological cultures typically outperform deep reinforcement learning systems in terms of various game performance characteristics, such as the average rally length implying a higher sample efficiency. Furthermore, the human cell cultures proved to have the overall highest relative improvement in the average number of hits in a rally when comparing the initial 5 minutes and the last 15 minutes of each designed gameplay session.

1 Introduction

The concept of reinforcement learning dates back to the early days of cybernetics and has been studied in statistics, psychology, neuroscience, and computer science. In the past decade, its use has become increasingly popular in the fields of machine learning and artificial intelligence. Its promise is highly convincing - a way of programming agents by rewarding and punishing them without having to specify how the task is to be accomplished. However, to deliver on this promise, formidable computational obstacles must be overcome. Reinforcement learning (RL) implies learning the best policy to maximize an expected cumulative long-term reward throughout many steps in order to achieve complex objectives (goals) [1]. A deep reinforcement learning (deep RL) approach integrates artificial neural networks with a reinforcement learning framework that helps the system to achieve its goals [2]. That is, it maps states and actions to the rewards they bring, combining function approximation and target optimization. Reinforcement algorithms that incorporate deep neural networks have been developed to beat human experts playing numerous Atari video games [3], poker [4], multiplayer contests [5], and complex board games, including go and chess [6, 7, 8]. Nevertheless, reinforcement learning still faces real challenges including but not limited to complexities in the selection of reward structure, sample inefficiency [9, 10], reproduciblity issues [11], as well as requiring high levels of computing power [12]. All of these suggest that deep RL algorithms may differ fundamentally from the underlying mechanisms of human learning while also being too

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

inefficient to be accepted as plausible models of human learning [10].
It was recently demonstrated that by using electrophysiological stimulation and recording in a real-time closed-loop system with a monolayer of living biological neurons, these cells could be trained to significantly improve performance in the simulated 'pong' gameworld [13]. The question arises as to whether this observed performance is notable in comparison to that of reinforcement learning at the same task. To examine this proposition and compare the performance and efficiency of such a biological neuronal network (BNN) to that of deep RL, we compare recent data gathered using the *DishBrain* system against time-matched learning from DQN, A2C & PPO algorithms. *DishBrain* is a novel system shown to display biological intelligence by harnessing the inherent adaptive computation of neurons. In this system, *in vitro* neuronal networks are integrated with *in silico* computing via high-density multi-electrode arrays (HD-MEAs). These cultured neuronal networks showcase biologically-based adaptive intelligence within a simulated gameplay environment in real time through closed-loop stimulation and recordings [13]. These cultures exhibit learning behaviour and possess an innate ability to self-organise activity and act intelligently in response to limited structured external information. Data was generated from cortical cells from embryonic rodent and human induced pluripotent stem cell (hiPSC) sources. We investigate whether these elementary learning systems achieve performance levels which can compete with state-of-the-art deep RL algorithms while varying the input information density required for training the RL algorithms to also determine the impact of information sparsity and ensure suitable comparisons to the biological system were made. This is the first comparison between a synthetic biological intelligence system and state-of-the-art RL algorithms.

2 Methods

2.1 DishBrain System

To investigate whether cultured cortical networks learn efficiently in the task-present state, recordings from cultures integrated onto an MEA were used. The *DishBrain* environment is a low latency, real-time system which interacts with the MaxOne MEA (Maxwell Biosystems, AG, Switzerland) software to allow closed-loop stimulation and recording. In addition to the ability to record electrical activity in the neuronal cultures, this system delivers external electrical stimulation to the cells in a way that has been shown to be safe for the tissue with long-term chronic stimulation. Biphasic electrical stimulation was used to generate action potentials in the neurons [14]. The external electrical stimulation was arranged to transmit a variety of task-related information using appropriate coding schemes. Using this method, activity from a neuronal culture can be read, along with providing structured stimulation to the same culture in real-time. *DishBrain* was utilised to embody neural cultures in a virtual game-world, to simulate the classic arcade game 'Pong'. Stimulation was applied using a combination of rate coding (4Hz - 40Hz) electrical pulses to communicate position on the x-axis and place coding (on a given electrode that were arranged topographically from an egocentric representation for the culture) to communicate information on the y-axis into a predefined bounded two-dimensional sensory area consisting of 8 sensory electrodes to deliver this input information. The movement of the paddle was controlled by the level of electrophysiological activity measured in a predefined "motor area" of the cultured network, which was collected in real time. Cortical cells, either differentiated from human induced pluripotent stem cells (hiPSC) or derived from E15 mouse embryos, were subjected to the gameplay conditions in the *DishBrain* and their hit counts in each episode of the game before the ball was missed for the first time were compared with different deep RL baseline methods. Each recording session of the cultures during gameplay was 20 minutes. This equaled an average number of 70 episodes (games played until the ball was missed once) for each recording session. The measurements were carried out during gameplay, where cells adjusted paddle position through activity changes and received information about the position of the ball and the closed-loop response to their control of it. More details of this system are introduced in Supplementary Materials A.

2.2 Deep Reinforcement Learning Algorithms

In this work, we use three state-of-the-art deep reinforcement learning algorithms: Deep Q Network (DQN) [3], Advantage Actor-Critic (A2C) [15] and Proximal Policy Optimization (PPO) [16], established to have good performance in Atari games. Benefiting from deep learning advantages in automated feature extraction, specifically exploiting Convolutional Neural Networks (CNN) in their

structures, these methods are robust tools in reinforcement tasks, particularly in games where the system's input is an image. In this work, aiming to account for potential adversaries resulting from the increased dimensionality of the image input to the deep RL algorithms [17], we design two additional types of input information to the RL algorithms. We compare all three different designs with the performance of biological cultures. We attempt to study whether the curse of dimensionality and increased size of the feature vectors when directly utilizing image inputs affects the comparison between biological cultures and RL algorithms in terms of their sample efficiency. The three different input categories and RL algorithm designs are introduced below:

- **IMAGE INPUT:** All the algorithms follow a common strategy although different in structure. In this design, the current state is a tensor of the difference of pixel values from the two most recent frames (i.e. another 40×40 grayscale pixel image). This current state is then input into the CNN to obtain the selected action. Next, based on the action taken, a reward is received, and a new state is formed. The ultimate goal is to find a policy that indicates the best action in each state to maximise the reward function.
- **PADDLE&BALL POSITION INPUT:** In this case, instead of the grayscale image, we obtain a 4-dimensional vector encoding the x and y coordinates of the ball (distance to the paddle/wall and distance to the floor in pixels) and the y coordinates of the paddle's top and bottom, all being integer values in $[1, 40]$. The current state which is the input to each algorithm is then a tensor of the difference of values from the two most recent 4-dimensional location vectors. No additional CNN layer is utilized in this case.
- **BALL POSITION INPUT:** Finally, we aim to examine a design as similar to the DishBrain system's input structure as possible. For this case, we divide the y-axis of the gameplay environment to 8 equal segments each mimicking one of the sensory electrodes in the biological cultures and place coding the information about the ball's y-axis position as an integer in the $[1, 8]$ interval. Then, the ball's x-axis position is used as the second element of this input vector being an integer value in $[4, 40]$ similar to the rate coded component of the stimulation applied to the biological cultures. No additional CNN layer is utilized in this design.

The overview of the implemented DQN, A2C, and PPO algorithms are represented in Supplementary Materials A.4 (see Algorithms 1, 2, and 3).

All the deep RL implementations run on a 2.3 GHz Quad-Core Intel Core i5. We use PyTorch 1.8.1 to build neural network blocks and Open AI Gym environment to define our game environment represented by a 40×40 pixel grayscale image. In the training phase of all RL algorithms, we ran every algorithm for 40 random seeds and a total number of 70 episodes for each seed. These seeds imply 40 different neural networks trained separately, resembling 40 different recorded cultures. In this work, we report the average value of each metric among all seeds.

Figure 1 illustrates the comparison between the input information in the DishBrain system and the deep RL algorithms.

3 Results

We studied both human cortical cells (HCCs; 174 sessions) and mice cortical cells (MCCs; 110 sessions) and compared the game performance to the introduced RL baseline methods. To determine how the learning arises both in the cultures and the baseline methods, key gameplay characteristics were examined further. The hit counts in the gameplay in each episode before the ball was missed for the first time, the number of times the paddle failed to intercept the ball on the initial serve (aces), and the number of long rallies (> 3 consecutive hits) were calculated for this data.
For comparison purposes, we first mapped every 70-episode run of each RL algorithm to a real-time equivalent of 20 minutes by first normalizing to the actual total length of each run in minutes and then multiplying by 20 minutes. Figures 2, 3, and 4 represent the main findings for comparisons between the biological cultures and the the IMAGE INPUT, PADDLE&BALL POSITION INPUT, and BALL POSITION INPUT designs of the RL methods.
In all three designs, DQN is outperformed by all other groups in terms of the highest level of average hits per rally achieved, while the biological cultures (i.e. HCC and MCC) outperform all the RL baseline algorithms (see Subfigures 2.a, 3.a, and 4.a). This indicates the fact that the cultures

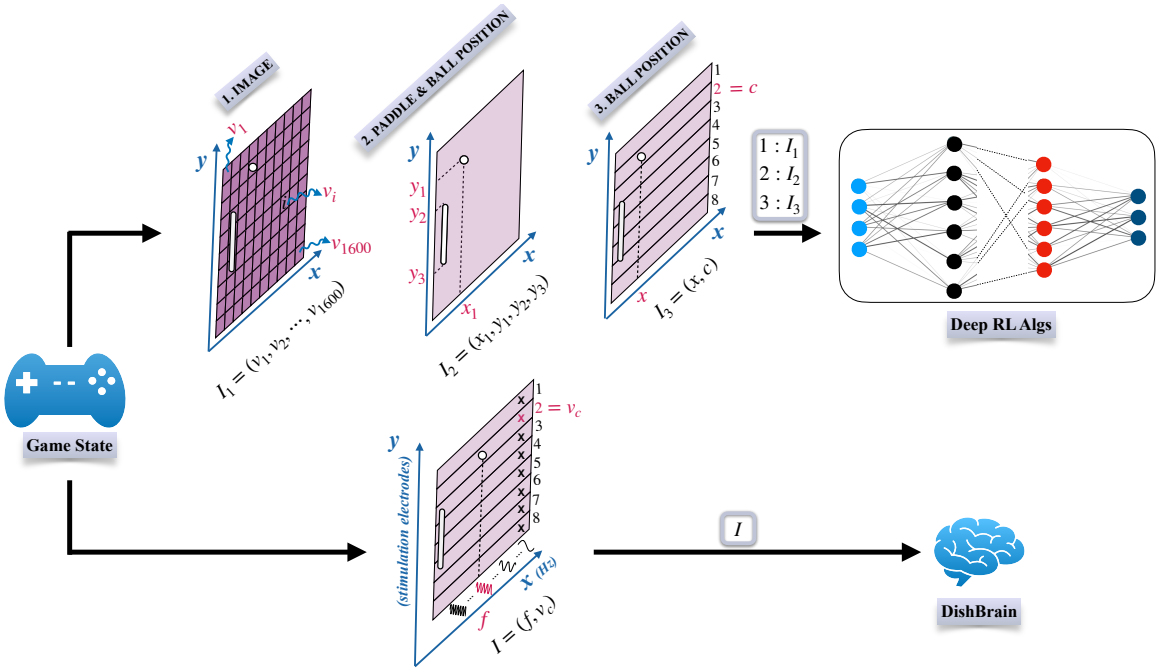

Figure 1: Schematic comparing the information feeding routes in the DishBrain system (bottom) and the three implementations of the deep RL algorithms (top). In each design, the input information to the computing module (deep RL algorithms or DishBrain) is denoted by a vector I.

represent faster growing learning rates in comparison. Subfigures 2.b, 3.b, and 4.b compare the % of missed balls on the initial serve, aces, among the biological cultures and the RL groups given the three different designs. HCC and MCC achieve the lowest percentage of aces compared to the deep RL algorithms in Subfigure 2.b and the other RL baseline designs in Subfigures 3.b, and 4.b. The increasing trend in the % of long rallies is observed in all groups and among all designs except PPO in the PADDLE&BALL POSITION INPUT design as illustrated in Subfigures 2.c, 3.c, and 4.c. Average % of long rallies reaches its highest levels for MCC, HCC compared to the RL baselines. Next, for all the groups, we compared the key activity metrics in the first 5 minutes versus the last 15 minutes in each session. Our aim was to identify any significant improvement occurring in the learning process within each group.

Panel (d) in Figures 2, 3, and 4 directly compares the average rally length between the two defined time intervals within all the groups. The results imply that the intra-group increasing trend in the length of rallies is significant only in the biological groups.

Panel (e) in Figures 2, 3, and 4 represents the change in the average percentage of aces over time in all groups. A significant decrease in the number of aces (where the ball was missed immediately in an episode with no accurate hits) implies an improved game performance. Only MCC and HCC groups had a significant decrease in the average ace percentage as opposed to the RL algorithms.

Panel (f) in Figures 2, 3, and 4, shows that the percentage of long rallies in the first 5 minutes versus the last 15 minutes only significantly increased for the biological cultures.

Pairwise inter-group comparison was also carried out for both time intervals (0-5 and 6-20 minutes) and all three metrics using Tukey's post hoc test as represented in panels (g), (h), and (i) in Figures 2, 3, and 4 for rally length (i.e. hit counts), % of aces, and % of long rallies respectively.

It should be noted that in the IMAGE INPUT design, where the performance of the deep RL methods comes closets to the biological cultures, the density of input information is starkly different between RL methods and the biological cultures. While RL agents receive pixel data with a density of 40 $\times$ 40 pixels, biological cultures only receive input from 8 stimulation points with a given integer rate code of 4Hz–40Hz, highlighting important efficiency differences in informational input between these learning systems. The possibility of the higher input information dimensionality having adverse effects on the overall sample efficiency of these RL algorithms is further nullified by evaluating the two alternative input structures (PADDLE&BALL POSITION INPUT and BALL POSITION INPUT designs).

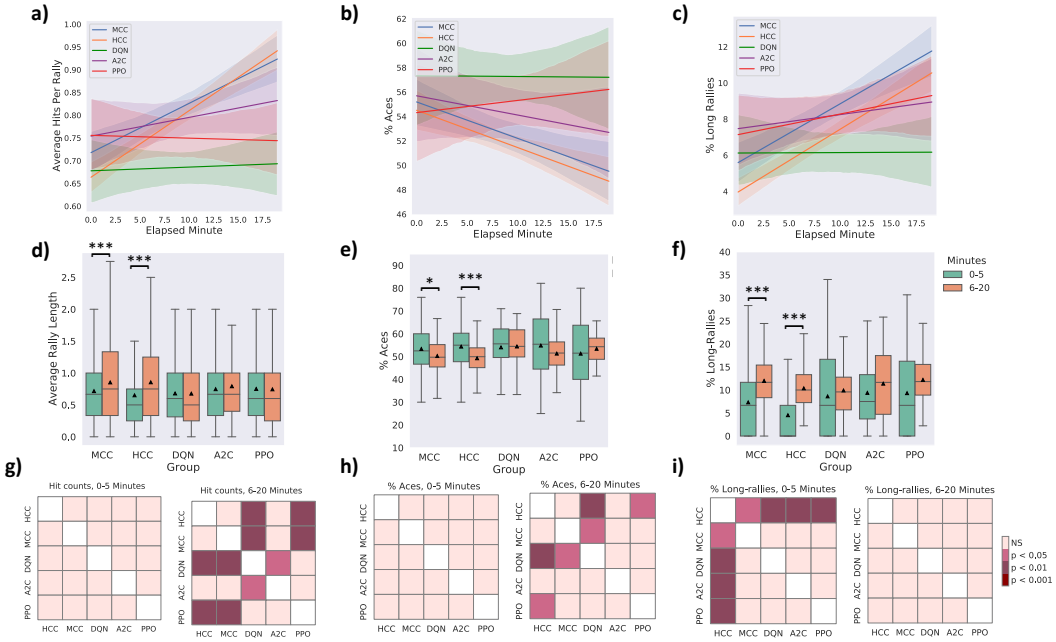

Figure 2: IMAGE INPUT to the deep RL algorithms: Average number of **a)** hits per rally, **b)** % of aces, and **c)** % of long rallies over 20 minutes real-time equivalent of training DQN, A2C, PPO, and MCC, HCC cultures. A regressor line on the mean values with a 95% confidence interval highlights the learning trends. Comparing the performance amongst all groups, the highest level of average hits per rally is achieved by the neuronal MCC and HCC cultures while DQN is outperformed by all the opponents. Average % of aces is lowest for the neuronal cultures compared to all deep RL baseline methods. Average % of long rallies reaches its highest levels for MCC and HCC. **d)** Average performance of the groups over time, where only the biological cultures have significant intra-group improvement and increase in their performance at the second time interval (One-way ANOVA test, p = 5.854e-6, p = 7.936e-17, for MCC and HCC respectively; p = 0.952, p = 0.354, and p = 0.905 for DQN, A2C, and PPO respectively). **e)** Average % of aces within groups and over time, only MCC and HCC (One-way ANOVA test, p= 0.014, p = 2.907e-08 respectively) differed significantly over time. No significant change was detected within the DQN, A2C, or PPO groups (One-way ANOVA test, p = 0.857, p = 0.212, and p = 0.448 respectively). **f)** Average % of long-rallies (>3) performed in a session. All the groups showed an increase in the average number of long rallies where this intra-group increase was significant only for MCC, HCC (One-way ANOVA test, p = 1.172e-7, p = 1.525e-24). **g)** Pairwise Tukey's post hoc test shows that HCC and MCC groups significantly outperform PPO and DQN only in the last 15 minutes interval with A2C also outperforming DQN. **h)** Using pairwise Tukey's post hoc test, HCC group significantly outperforms PPO and DQN in the last 15 minutes interval with a lower average of % Aces. DQN is also outperformed by the MCC group in this time interval. **i)** Pairwise comparison using Tukey's test only shows a significant difference in the percentage of long rallies between HCC and the rest of the groups in the first 5 minutes. However, this is later altered in the direction of all groups having an increased % of long rallies with no significant difference in the last 15 minutes of the game. Box plots show interquartile range, with bars demonstrating 1.5X interquartile range, the line marks the median and ▲ marks the mean. Error bands = 1 SE

To account for potential effects of paddle movement speed and whether it is playing an important role in determining the success rate of paddle control, we derived the average paddle movement (in pixels) for all groups. Subfigures 5.a,c, and e represent these results for the IMAGE INPUT, PADDLE&BALL POSITION INPUT, and BALL POSITION INPUT designs respectively. Using pairwise Tukey's post hoc tests it was found that a consistent significant difference is present between pairs of DQN and HCC in terms of the average paddle movement with DQN having the higher average. This is when all the RL algorithms with the PADDLE&BALL POSITION INPUT and BALL POSITION INPUT designs have significantly higher average paddle movement compared to both groups of biological cultures. Interestingly, the higher paddle movement speed of the RL algorithms is not reflected as better game performance according to previously discussed results.

Subfigures 5.b, d, and f compare the relative improvement in the performance of different groups over time when comparing the HCC and MCC groups to RL algorithms with IMAGE INPUT, PADDLE&BALL POSITION INPUT, and BALL POSITION INPUT respectively. This measure identifies

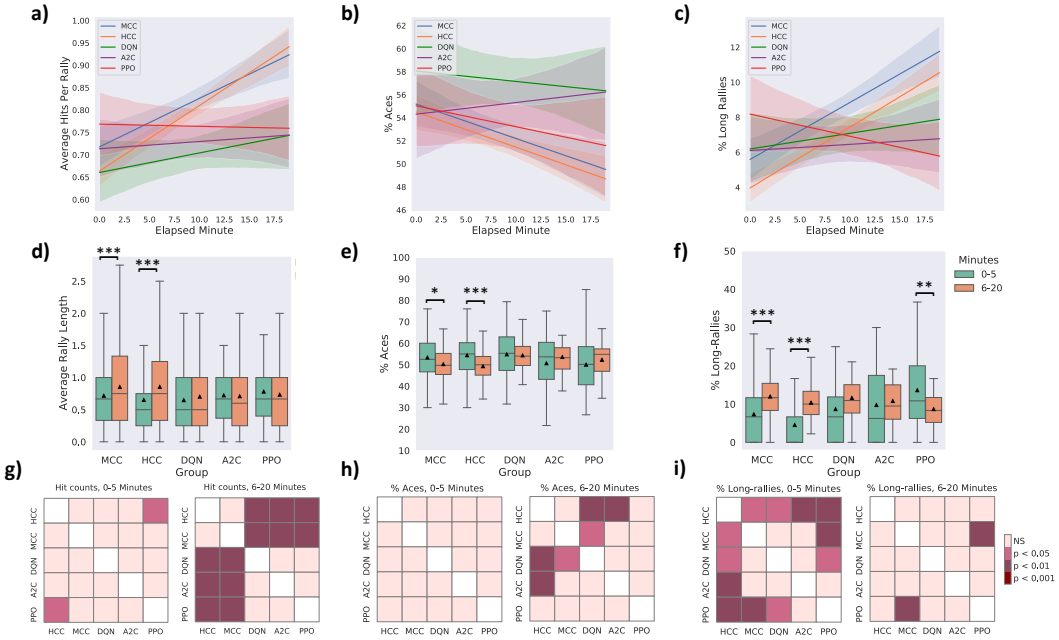

Figure 3: PADDLE&BALL POSITION INPUT to the deep RL algorithms: Average number of **a)** hits per rally, **b)** % of aces, and **c)** % of long rallies over 20 minutes real-time equivalent of training DQN, A2C, PPO, and MCC, HCC cultures. A regressor line on the mean values with a 95% confidence interval highlights the learning trends. The highest level of average hits per rally is achieved by the neuronal MCC and HCC cultures. Average % of aces is lowest for the neuronal cultures compared to all deep RL baseline methods. Average % of long rallies reaches its highest levels for MCC and HCC. Comparing to the same findings for the HCC and MCC groups, **d)** average rally length over time only showed a significant increase in the biological cultures between the two time intervals (One-way ANOVA test, p = 0.241, p = 0.756, and p = 0.315 for DQN, A2C, and PPO respectively). **e)** Average % of aces within groups and over time only showed a significant difference in the MCC and HCC groups. No significant change was detected within the DQN, A2C, or PPO groups (One-way ANOVA test, p = 0.858, p = 0.279, and p = 0.398 respectively). **f)** Average % of long-rallies (>3) performed in a session increased in the second time interval in all groups except the PPO group. This intra-group difference was significant for MCC, HCC, and PPO (One-way ANOVA test, p = 1.172e-7, p = 1.525e-24, p = 0.008 respectively). **g)** Pairwise Tukey's post hoc test shows that the HCC group significantly outperforms PPO in the first 5 minutes in terms of the hit counts or rally length. The biological cultures do significantly better compared to all deep RL opponents in the 15 minutes interval. **h)** Using pairwise Tukey's post hoc test, HCC group significantly outperforms A2C and DQN in the last 15 minutes interval with a lower average of % Aces. DQN is also outperformed by the MCC group in this time interval. **i)** Pairwise comparison using Tukey's test shows a significant difference in the percentage of long rallies between HCC and the rest of the groups in the first 5 minutes all outperforming the HCC. PPO also shows a significantly higher % of long rallies in the first time interval compared to MCC and DQN. However, this is later altered in the last 15 minutes with only MCC putperforming PPO significantly having an increased % of long rallies. Box plots show interquartile range, with bars demonstrating 1.5X interquartile range, the line marks the median and ▲ marks the mean. Error bands = 1 SE

the relative increase in the average accurate hit counts in the second 15 minutes of the game compared to the first 5 minutes. The HCC group shows the highest improvement in time and performing Tukey's post hoc tests showed that the difference in this measure is significant between HCC and PPO, as well as HCC and DQN in the IMAGE INPUT case, between HCC and PPO and well as HCC and A2C in the PADDLE&BALL POSITION INPUT case, and between HCC and A2C in the BALL POSITION INPUT case. The MCC group also outperform DQN, PPO, or A2C groups in the IMAGE INPUT, PADDLE&BALL POSITION INPUT, and BALL POSITION INPUT designs respectively.

Eventually, Subfigures 5.g, h, i, and j compare the frequency tables for the distributions of mean summed hits per minute amongst groups for the IMAGE INPUT, PADDLE&BALL POSITION INPUT, and BALL POSITION INPUT designs respectively. These tables were found to be not significantly different (Two-sample *t*-test).

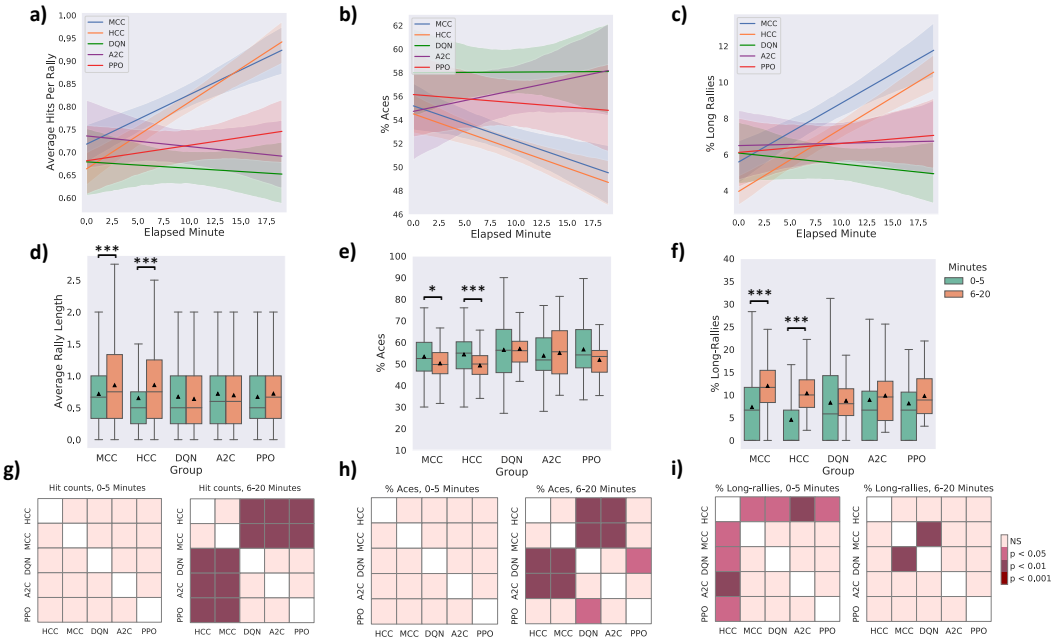

Figure 4: Ball Position Input to the deep RL algorithms: Average number of **a)** hits per rally, **b)** % of aces, and **c)** % of long rallies over 20 minutes real-time equivalent of training DQN, A2C, PPO, and MCC, HCC cultures. A regressor line on the mean values with a 95% confidence interval highlights the learning trends. The highest level of average hits per rally is achieved by the neuronal MCC and HCC cultures. Average % of aces is lowest for the neuronal cultures compared to all deep RL baseline methods. Average % of long rallies reaches its highest levels for MCC and HCC. Comparing to the same findings for the HCC and MCC groups, **d)** average rally length over time only showed a significant increase in the biological cultures between the two time intervals (One-way ANOVA test, p = 0.436, p = 0.612, and p = 0.240 for DQN, A2C, and PPO respectively). **e)** Average % of aces within groups and over time only showed a significant difference in the MCC and HCC groups. No significant change was detected within the DQN, A2C, or PPO groups (One-way ANOVA test, p = 0.858, p = 0.279, and p = 0.398 respectively). **f)** Average % of long-rallies (>3) performed in a session increased in the second time interval in all groups. This intra-group difference was only significant for MCC and HCC groups. **g)** Pairwise Tukey's post hoc test shows that biological cultures significantly outperform all deep RL groups in the last 15 minutes in terms of the hit counts or rally length. **h)** Using pairwise Tukey's post hoc test, HCC and MCC groups significantly outperform A2C and DQN in the last 15 minutes interval with a lower average of % Aces. DQN is also outperformed by the PPO group in this time interval. **i)** Pairwise comparison using Tukey's test shows a significant out-performance of all groups over HCC in the percentage of long rallies in the first 5 minutes. MCC also shows a significantly higher % of long rallies in the second time interval compared DQN. Box plots show interquartile range, with bars demonstrating 1.5X interquartile range, the line marks the median and ▲ marks the mean. Error bands = 1 SE

4 Discussion

In this work, we compared the performance of biological neuronal networks with that of state-of-the-art deep reinforcement algorithms (deep RL) in the game environment of *pong*. Human and mice cortical cells (HCC and MCC) as well as the three deep RL algorithms: DQN, A2C, and PPO were employed and compared in sessions with average episode number of 70 games played. We examined the overall performance of each group with respect to various gameplay characteristics. While the DQN algorithm was outperformed by all the other groups in all the tests, the biological cultures outperformed all RL baselines in terms of the ultimate average hit per rally count (Subfigure 2.a), ultimate % of aces or games lost in a single shot (Subfigure 2.b), and ultimate % of long rallies achieved (Subfigure 2.c). Moreover, the increase in the average rally length, the decrease in the number of aces, and the increase in the number of long rallies were only significant within the mice and human cell culture groups when comparing the first 5 minutes and the last 15 minutes of gameplay sessions (see Subfigures 2.d, e, and f). Additionally, we found that the HCC group had the highest relative improvement in the average number of hits between the first 5 minutes and last 15 minutes of the game as depicted in Subfigures 5.b, d, and f.

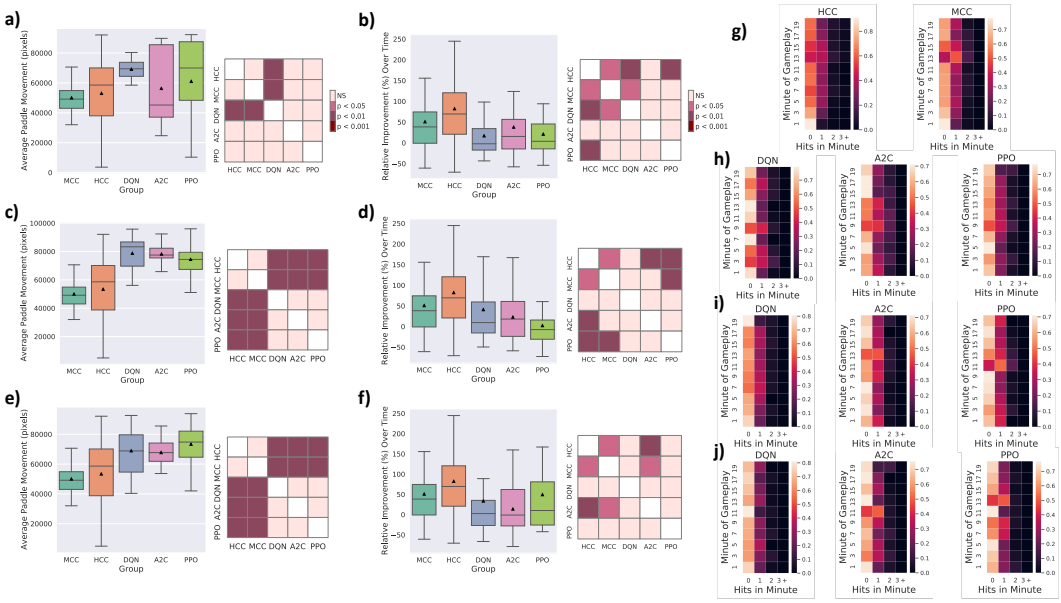

Figure 5: The average paddle movement in pixels in all the difference groups for the **a)** IMAGE INPUT, **c)** PADDLE&BALL POSITION INPUT, and **e)** BALL POSITION INPUT to the deep RL algorithms. Pairwise Tukey's post hoc test was conducted showing that DQN had a significantly higher average paddle movement compared to HCC and MCC in all scenarios. A2C and PPO also had a significantly higher average movement of the paddle in the case of Location Vector Input, and 2-dimensional Input. Relative improvement (%) in the average hit counts between the first 5 minutes and the last 15 minutes of all sessions in each separate group for the **b)** IMAGE INPUT, **d)** PADDLE&BALL POSITION INPUT, and **f)** BALL POSITION INPUT to the deep RL algorithms. The biological groups show higher improvements with HCC outperforming all. **b)** Using pairwise Tukey's post hoc test, the inter-group differences were significant for HCC and DQN, HCC and PPO, MCC and DQN, as well as MCC and HCC. **d)** HCC showed a significantly higher relative improvement compared to PPO and A2C while MCC also outperformed PPO in terms of relative improvement over time. **f)** Finally, MCC and HCC groups could significantly perform better than the A2C group with the 2-dimensional Input vector inputted to the deep RL algorithms. Distribution of frequency of mean summed hits per minute amongst groups for **g)** biological cultures and deep RL algorithms with **h)** IMAGE INPUT, **i)** PADDLE&BALL POSITION INPUT, and **j)** BALL POSITION INPUT.

The results show that the game performance of the deep RL algorithms in terms of relative learning improvement in time and the ultimate number of average hits per rally is outperformed by biological cultures. Furthermore, their performance in the average rally length and percentage of aces only matches those of neuronal cultures at best. This is while to achieve this level of learning, the biological cell cultures require only a fraction of the input information density compared to their RL opponents (8 pixel combination of rate coded and place coded stimulation compared to 40×40 pixels of input image). The RL algorithms showed the lowest sample efficiency having the lowest improvement in learning given the 70 episode training duration provided for all the groups. To account for potential disadvantages occurring as a result of increased input dimensionality, we also examined two alternative designs in terms of the input structure to the RL algorithms (i.e. PADDLE&BALL POSITION INPUT and BALL POSITION INPUT designs). The in-depth comparison between the biological cultures' performance and these alternative RL algorithms did not provide any significantly different outcome in favour of the RL baselines' sample efficiency (see Figures 3 and 4).

This is the first comparison between a synthetic biological intelligence system and state-of-the-art RL algorithms. This early work establishes that even the most rudimentary SBI systems with limited informational input are a viable learning system that can compete and even defeat the established RL algorithms which receive significant more information input. Coupled with the promise of significant gains in power efficiencies, flexibility of tasks, and as data representation to the SBI system is improved, these biological based intelligence systems present a compelling pathway for realizing real-time learning unachievable by current silicon-based approaches.

Table 1 brings a summary of comparisons between different gameplay characteristics among groups

with the IMAGE INPUT to the RL algorithms and whether the differences were statistically significant.

Table 1: Summary of characteristic differences between the groups.

Mean Gameplay Characteristics	**0-5 minutes**			**6-20 minutes**		
	Min	Max	Significance	Min	Max	Significance
% Aces (immediately missed)	MCC	A2C	NO	HCC	DQN	YES
% Long-Rallies	HCC	PPO	YES	DQN	PPO	NO
Avg Rally Length	HCC	PPO	NO	DQN	HCC	YES

Mean Gameplay Characteristics	**Total Session Duration**		
	Min	Max	Significance
Relative Improvement (%)	DQN	HCC	YES
Avg Paddle Movement	MCC	DQN	YES
Training time (mins) per 70 episodes	MCC/HCC (20 mins)	DQN (131.4 mins)	YES

References

[1] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[2] Matteo Hessel et al. "Rainbow: Combining improvements in deep reinforcement learning". In: *Thirty-second AAAI conference on artificial intelligence*. 2018.

[3] Volodymyr Mnih et al. "Human-level control through deep reinforcement learning". In: *nature* 518.7540 (2015), pp. 529–533.

[4] Matej Moravčík et al. "Deepstack: Expert-level artificial intelligence in heads-up no-limit poker". In: *Science* 356.6337 (2017), pp. 508–513.

[5] M Jaderberg et al. "Human-level performance in first-person multiplayer games with population-based deep reinforcement learning. arXiv". In: *arXiv preprint arXiv:1807.01281* (2018).

[6] David Silver et al. "Mastering chess and shogi by self-play with a general reinforcement learning algorithm". In: *arXiv preprint arXiv:1712.01815* (2017).

[7] David Silver et al. "Mastering the game of go without human knowledge". In: *nature* 550.7676 (2017), pp. 354–359.

[8] David Silver et al. "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play". In: *Science* 362.6419 (2018), pp. 1140–1144.

[9] Pedro A Tsividis et al. "Human learning in Atari". In: *2017 AAAI spring symposium series*. 2017.

[10] Gary Marcus. "Deep learning: A critical appraisal". In: *arXiv preprint arXiv:1801.00631* (2018).

[11] Elizabeth Gibney et al. "This AI researcher is trying to ward off a reproducibility crisis". In: *Nature* 577.7788 (2020), pp. 14–14.

[12] Seyed Sajad Mousavi, Michael Schukat, and Enda Howley. "Deep reinforcement learning: an overview". In: *Proceedings of SAI Intelligent Systems Conference*. Springer. 2016, pp. 426–440.

[13] Brett J Kagan et al. "In vitro neurons learn and exhibit sentience when embodied in a simulated game-world". In: *Accepted In Neuron* (2022).

[14] Maria Elisabetta Ruaro, Paolo Bonifazi, and Vincent Torre. "Toward the neurocomputer: image processing and pattern recognition with neuronal cultures". In: *IEEE Transactions on Biomedical Engineering* 52.3 (2005), pp. 371–383.

[15] Kai Arulkumaran et al. "Deep reinforcement learning: A brief survey". In: *IEEE Signal Processing Magazine* 34.6 (2017), pp. 26–38.

[16] John Schulman et al. "Proximal policy optimization algorithms". In: *arXiv preprint arXiv:1707.06347* (2017).

[17] Richard Bellman and Robert Kalaba. "Dynamic programming and statistical communication theory". In: *Proceedings of the National Academy of Sciences* 43.8 (1957), pp. 749–751.