Biologically Inspired Reinforcement Learning for Locomotion: A Central-Pattern Generator Approach

1st Adan Domínguez-Ruiz
Institute for the Future of Education
Instituto Tecnológico de Monterrey
Escuela de Ingeniería y Ciencias
Ave. Eugenio Garza Sada 2501, Monterrey 64849, NL, Mexico
0000-0002-0721-2853

2nd Edgar Omar López-Caudana
Institute for the Future of Education
Escuela de Ingeniería y Ciencias
Instituto Tecnológico de Monterrey
Monterrey, NL, Mexico
0000-0002-1216-4219

3rd Oscar Loyola
Facultad de Ingeniería
Universidad Autónoma de Chile
Santiago, Chile
0000-0001-9355-2346

4th Pedro Ponce-Cruz
Institute of Advanced Materials for Sustainable Manufacturing
Instituto Tecnológico de Monterrey
Mexico City, Mexico

***Abstract*—Bipedal locomotion, such as walking or running, involves complex coordination of rhythmic and cyclic movements that must be stable, smooth, and adaptable to varying terrains, which is challenging to achieve in robotic and simulated environments. Current reinforcement learning (RL) approaches often fail to generate stable and natural locomotion patterns due to a lack of inherent rhythmic control, resulting in jerky, unstable, and inefficient gaits. These methods typically do not incorporate the biological principles of rhythmicity and adaptability, which are crucial for achieving natural bipedal locomotion. This study presents a novel approach integrating Central Pattern Generators (CPG) with multiple RL algorithms, including Maximum a Posteriori Policy Optimization, Deep Deterministic Policy Gradient, and Soft Actor-Critic (SAC), using Matsuoka Oscillators to generate rhythmic patterns. By comparing these RL-CPG hybrid methods, the research demonstrates improvements in energy efficiency and synchronization for SAC+CPG in controlled environments. While other algorithms may have advantages in different conditions, SAC+CPG showed the most stable, and rhythmic gait, while minimising energy usage under the tested parameters. This study highlights the first multi-algorithm application of RL combined with CPGs for rhythmic control in bipedal locomotion, contributing to the future of robotics and cyber-physical systems.**

***Index Terms*—locomotion, reinforcement learning, central-pattern-generators, bioinspired-ai, educational-innovation, higher-education.**

I. Introduction

Bipedal locomotion is a fundamental and complex challenge in robotics and biomechanics, requiring the coordination of multiple joints, muscles, and sensory feedback to maintain balance, stability, and forward motion[1]. Achieving stable, smooth, and adaptable locomotion across varying terrains is particularly difficult due to the high degrees of freedom and non-linear dynamics involved in human-like movement. Conventional control methods rely heavily on precise calculations of the center of mass (CoM) and constant adjustments to maintain balance, which often leads to non-natural, jerky movements and energy inefficiencies [2]. Such methods struggle to replicate the smooth, cyclic motions observed in biological systems[3], which achieve efficiency and stability through complex neural and muscular coordination.

In recent years, reinforcement learning (RL) has emerged as a powerful tool for enabling autonomous agents to learn and optimize control policies through interaction with the environment. RL-based algorithms such as Maximum a Posteriori Policy Optimization (MPO), Deep Deterministic Policy Gradient (DDPG), and Soft Actor-Critic (SAC) have demonstrated success in motor control tasks, particularly in robotic locomotion [4]–[6]. These algorithms allow agents to improve their gait performance over time by maximising cumulative rewards associated with balance, stability, speed, and energy efficiency. However, many of these approaches struggle with generating stable and natural locomotion patterns due to a lack of rhythmic control mechanisms [2].

Central Pattern Generators (CPGs), inspired by biological neural circuits, have been explored as an alternative for generating cyclic locomotion in robots. CPGs are neural networks that generate rhythmic motor patterns, such as walking and running, without the need for continuous sensory input [7], [8]. These models inherently produce stable and smooth rhythmic patterns but lack the adaptability needed to adjust gait dynamically in response to environmental changes. To address this, reinforcement learning (RL) has been proposed to optimize gait control. However, RL alone often fails to generate consistent cyclic motion without an explicit rhythmic structure. CPGs are particularly well-suited for generating smooth, cyclic movements that are adaptable to various conditions, making them an ideal complement to RL approaches [9].

The integration of RL with CPGs offers a novel approach to achieving more natural and efficient bipedal locomotion. CPGs can generate rhythmic patterns for leg movements, while RL algorithms optimise the control policies based on

environmental feedback [10]. This hybrid method leverages the adaptability of RL with the smooth, energy-efficient movements generated by CPGs, offering significant potential for improvements in robotic and biomechanical applications [11]. This study integrates CPGs with RL, leveraging their rhythmic stability while allowing adaptive gait learning. Unlike previous works that treat CPGs as isolated controllers, our approach actively optimizes CPG dynamics using RL, resulting in energy-efficient and adaptive bipedal locomotion

Three prominent RL algorithms are also compared: MPO, DDPG, and SAC, each integrated with CPGs using Matsuoka Oscillators to control bipedal locomotion. We implement these RL-CPG approaches in progressively complex simulated environments, starting from the "Walker2d-v4" environment and advancing to the "Humanoid-v4" environment from MuJoCo physics simulator [12], with plans for future application in musculoskeletal models.

The key contributions of this work are:

1) The first demonstration of using multiple RL algorithms (MPO, DDPG, SAC) combined with CPGs for rhythmic pattern generation in bipedal locomotion.
2) Comparative analysis against state-of-the-art locomotion generation algorithms, demonstrating improvements in cyclic motion consistency, energy expenditure, and distance travelled.
3) Exploration of the scalability of these methods from simple to complex models, with potential real-world applications in robotics and prosthetic device control [3], [13]

This study represents a significant step forward in biologically inspired locomotion control, providing a scalable framework that can be adapted for a wide range of robotic and biomechanical applications. The combination of RL and CPGs creates a powerful and adaptable system for generating smooth and efficient gait patterns in both simulated and real-world environments.

This system architecture illustrates the integration of MPO, DDPG, and SAC with CPG-based oscillators to achieve rhythmic and adaptive bipedal locomotion. Sensory input (e.g., terrain data) feeds into the RL algorithms, which optimize gait control through action selection and policy updates. The CPG generates cyclic patterns for leg movements, and the oscillators for each leg are phase-synchronized for coordinated locomotion. Evaluation metrics (smoothness, energy, stability) provide feedback, allowing the RL algorithms to refine their policies over time for improved locomotion performance.

II. Methodology

This section outlines the experimental methodology designed to evaluate the integration of Central Pattern Generators (CPGs) with various reinforcement learning (RL) algorithms, specifically Maximum a Posteriori Policy Optimization (MPO), Proximal Policy Optimization (PPO), and Soft Actor-Critic (SAC). The goal is to compare the performance of these hybrid RL-CPG systems for generating rhythmic, stable, and efficient bipedal locomotion patterns across different environments, with an emphasis on smoothness, stability, and energy efficiency.

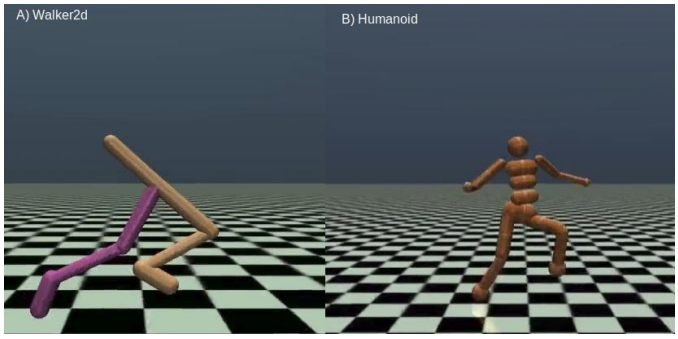

Fig. 1. Mujoco Bipedal Enviroments.

A. Experimental Setup:

The experiments are conducted in progressively complex environments using the Gymnasium library, the goal is to simulate bipedal locomotion under varying conditions to evaluate the effectiveness of RL-CPG integration. The experiments are performed in the MuJoCo physics simulator using 2 different bipedal environments:

Walker2d-v4: A 2D bipedal robot is tasked with walking in a horizontal plane. This environment is designed to evaluate basic locomotion and rhythmic movement generation (Figure1.A).

Humanoid-v4: A 3D humanoid robot is tasked with walking and balancing in a more complex environment with higher degrees of freedom. This environment introduces greater challenges in maintaining stability and coordination due to the increased complexity of the model and environment dynamics (Figure 1.B).

Future Musculoskeletal Model: Testing will later be expanded to a detailed musculoskeletal model, which includes elastic tendons and actuators, simulating real-world human locomotion with a more biologically accurate representation.

B. Reinforcement Learning Algorithms:

The following RL algorithms are compared in this study, each integrated with CPGs to provide rhythmic control:

MPO (Maximum a Posteriori Policy Optimization): MPO seeks to optimize the policy by limiting divergence between successive policies, ensuring stable learning [4]. MPO is expected to deliver stable control with relatively high-speed gait generation but may require more computational resources.

DDPG (Deep Deterministic Policy Gradient): An off-policy reinforcement learning algorithm designed for continuous action spaces, combining elements from Deep Q-Networks (DQN) and Actor-Critic architectures [14]. It employs an actor network to learn a deterministic policy and a critic network to estimate the Q-value function. DDPG applies an Ornstein-Uhlenbeck noise process for exploration, making it particularly effective in high-dimensional, continuous control tasks. DDPG

23rd LACCEI International Multi-Conference for Engineering, Education, and Technology: "Engineering, Artificial Intelligence, and Sustainable Technologies in service of society". Hybrid Event, Mexico City, July 16 - 18, 2025

is widely used in robotic locomotion, autonomous vehicles, and prosthetic control due to its ability to learn smooth, adaptive policies without predefined motion trajectories.

SAC (Soft Actor-Critic): SAC aims to maximize entropy, encouraging exploration while optimizing policies [5]. SAC may show advantages in adaptability to changing conditions, though it may produce less smooth movement than MPO.

The training configurations used in this study are based on predefined configurations from the Zoo-RL library [15], a widely adopted framework that provides optimized and validated hyperparameter settings for standard OpenAI Gym and MuJoCo environments. Hyperparameters were not tuned further, allowing us to isolate the effects of the CPG integration on locomotion quality across agents.

Tables I and II summarize the hyperparameter configurations used for the SAC, MPO, and DDPG algorithms in the Walker2d-v4 and Humanoid-v4 environments, respectively. Parameters include batch size, replay buffer size, learning rate, learning starts, noise values, and model architecture.

The reward function used in both Walker2d-v4 and Humanoid-v4 environments is the default one provided by Gymnasium/MuJoCo. It is composed of a weighted sum of forward velocity, control cost (energy usage), and posture penalties. No reward shaping or modifications were introduced, which ensures that the comparison across RL and RL-CPG variants is consistent and unbiased.

TABLE I
HYPERPARAMETERS FOR WALKER2D-V4 EXPERIMENTS

Algo	Batch	LR	Buffer	γ	Dual LR	Layers
SAC	256	0.001	4096	0.99	x	[400, 300]
DDPG	265	0.001	4096	0.99	x	[400, 300]
MPO	265	3e-4	8192	0.95	1.57e-4	[400, 300]

TABLE II
HYPERPARAMETERS FOR HUMANOID-V4 EXPERIMENTS

Algo	Batch	LR	Buffer	γ	Dual LR	Layers
SAC	265	0.00034	4000	0.99	x	[400, 300]
DDPG	265	0.00034	8192	0.97	x	[400, 300]
MPO	256	3.57e-5	16384	0.95	3.57e-4	[256, 256]

The integration of CPGs with these RL algorithms introduces an additional layer of rhythmic control, where Matsuoka Oscillators are used to generate cyclic patterns for the bipedal leg movements. The phase synchronization between the legs is managed by the CPGs, while the RL algorithm optimizes the overall gait policy.

C. Central Pattern Generators (CPGs)

For the purpose of generating cyclic motion, Matsuoka Oscillators are employed as CPGs for each leg. The Matsuoka Oscillator is a biologically inspired model capable of generating rhythmic output without continuous sensory feedback, mimicking natural motor control in biological organisms [7].

In the case of each of the oscillators, the generation of movement was defined as one oscillator per joint section, to keep in synchrony the movement of both legs. The architecture was chosen to keep one neuron to control either left or right joint, following the update function in Eq. 1 originally proposed by Matsuoka and updated for robotics motion [10], [16] .

$$dx = (-x - W_{in} * y_{prev} + A * S_{cpg} - \beta * Z) * \frac{dt}{\tau_r} \quad (1)$$

$$dz = (y - Z) * \frac{dt}{\tau_a} \quad (2)$$

$$y = relu(x) \quad (3)$$

The output value is gathered from the Eq. 3, where the relu function is chosen as excitation signal. To calculate x value from Eq. 1 we require the weights W_{in} which are provided by a DRL algorithm to adapt to the environment. y_{prev} is a rolled value from the output to conect each neuron with the output of the other one. A simbolize amplitude, chosen as the maximum value of the action space from the environment, S_{cpg} is the firing output for the Matsuoka neurons, and $\beta * Z$ is the decay factor. τ_r and τ_a are time learning factors taken by the original Matsuoka paper [10].

Following the control system architecture defined in Figure 2 The CPG system provides rhythmic activation for each leg, and the RL algorithms optimize the movement by adjusting CPG parameters (frequency, amplitude, and phase synchronization between the legs) based on environmental feedback. This allows the system to adapt to terrain changes and maintain a consistent gait cycle.

D. Steps of the Experiment

The experimental procedure involves the following steps for each RL algorithm (MPO, DDPG, SAC):

Initial Training: For both environments, the goal is to produce an stable, rhytmic and synchronized bipedal movement, while increasing the maximum distance covered in the minimum of time, for a maximum of 1500 simulation steps or 15 seconds, with no external disturbances.

The CPG is set to generate a basic rhythmic gait, with the RL algorithm adjusting the policy to optimize gait based on rewards for smoothness, stability, and energy efficiency. Training continues for a fixed number of episodes until the gait stabilizes.

Testing and Comparison:

After training, the agent's performance is measured for key variables (see below).

The process is repeated for each algorithm (MPO, DDPG, SAC), and the results are compared.

Scaling to Complex Environment:

The best algorithms working in simplier environments, are used and retrained to adapt to higher degrees of freedom and planes of movement. From Walker-2d, with only 6 degrees of freedom, jumping to Humanoid with 18 DoF, and making a test run on a more extensive musculoskeletal model with 36 DoF.

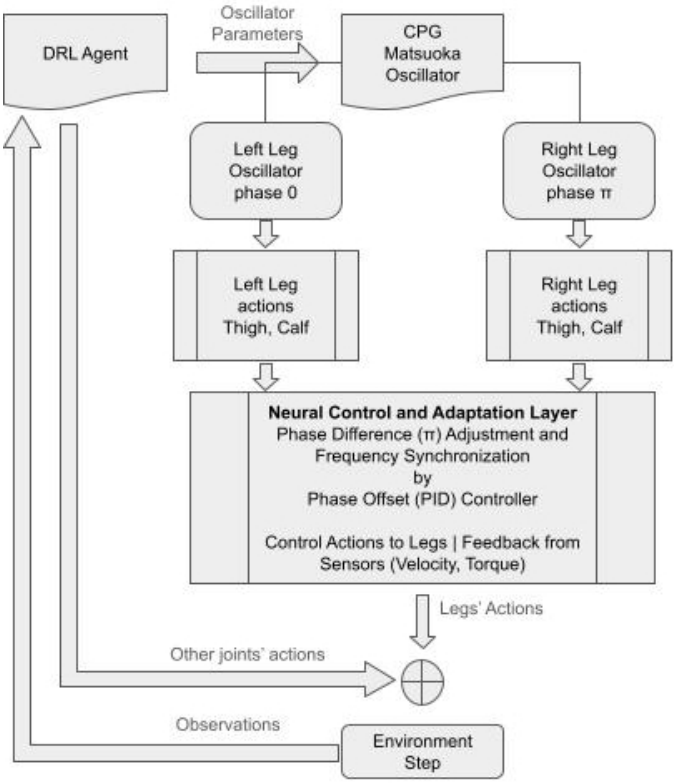

Fig. 2. Control Architecture describing how DRL agent provides the parameters for CPG and other joints for the action space.

Future Testing:

Once performance is validated in the Humanoid-v4 environment, testing will extend to a musculoskeletal model simulating real-world bipedal locomotion.

E. Evaluation Metrics

The following metrics are used to evaluate the performance of each RL-CPG model:

Cyclic Stability: Measured by the consistency of gait patterns over time. Stable movement involves maintaining consistent cyclic motions without interruptions or falls.

Energy Efficiency: Calculated as the energy consumed per unit distance traveled (measured in terms of joint torque). Energy efficiency is crucial for practical applications in robotics and prosthetics.

Cross-Correlation Between Legs: Evaluates the synchronization between left and right legs, ensuring that the phases of leg oscillators are correctly aligned for balanced movement.

Cumulative Reward: Tracks the accumulated reward over training episodes to assess the agent's learning progress.

F. Models to Be Compared

Three distinct combinations of RL and CPG models are compared in the experiments. Each of these models will be compared across the Walker2d-v4 and Humanoid-v4 environments, with future testing extending to the musculoskeletal model.

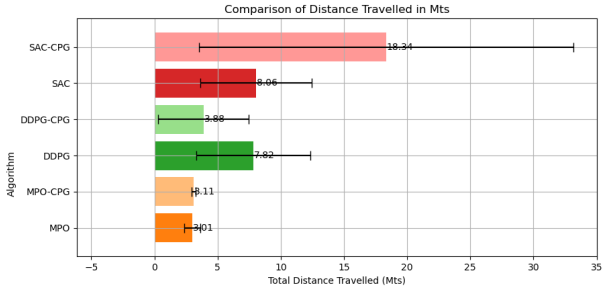

Fig. 3. Distance travelled from the Walker2d Mujoco environment

- MPO + CPG: MPO optimizes the gait policy while CPGs handle the cyclic generation of leg movements.
- DDPG + CPG: DDPG provides the policy optimization, integrated with CPGs for rhythmic control.
- SAC + CPG: SAC maximizes entropy to explore optimal policies, integrated with CPGs to generate cyclic movement.

III. Results

In this section, we present the outcomes of our experiments using the DRL-enhanced Central Pattern Generator (CPG) model applied to locomotion tasks in both Humanoid-v4 and Walker2d-v4 environments. The results highlight the efficacy of the DRL model in achieving stable and synchronized movements through controlled oscillations, optimized for each environment. We analyze various metrics to assess the stability, energy efficiency, and coordination achieved by the trained agent.

To evaluate the effectiveness of the trained DRL-CPG model, we tracked multiple metrics over a total of 100 episodes, including:

- Stability: The agent's ability to maintain balance over extended episodes, measured with the distance travelled without falling down.
- Energy Efficiency: Energy expenditure and efficiency in movement generation, measured in energy consumed per second.
- Coordination: The phase offset consistency between leg oscillators, contributing to synchronized walking patterns.

Each metric was assessed across multiple training sessions and was compared to a baseline model without CPG integration, providing insight into the improvements introduced by the CPG.

A. Experiment 1: Baseline vs. CPG-Enhanced DRL

We first compared the performance of a baseline DRL agent against the CPG-enhanced DRL agent in the Walker2d-v4 environment. The CPG model, with its oscillatory signals driving opposing leg movements, aimed to enhance stability and coordination over the baseline model. The following observations were noted:

23rd LACCEI International Multi-Conference for Engineering, Education, and Technology: "Engineering, Artificial Intelligence, and Sustainable Technologies in service of society". Hybrid Event, Mexico City, July 16 - 18, 2025

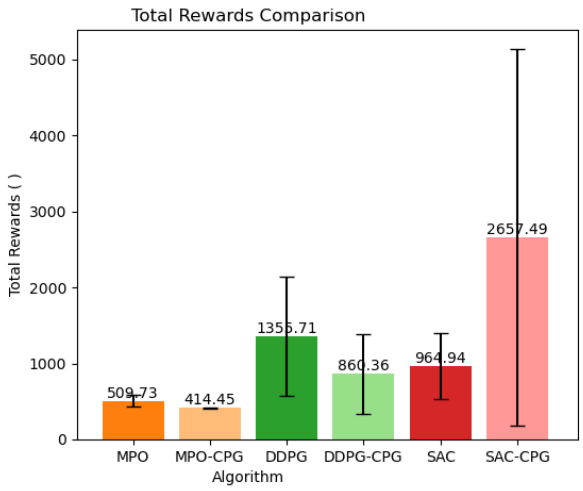

Fig. 4. Mean of rewards of the episodes from the agents performance in the Walker-2d environment.

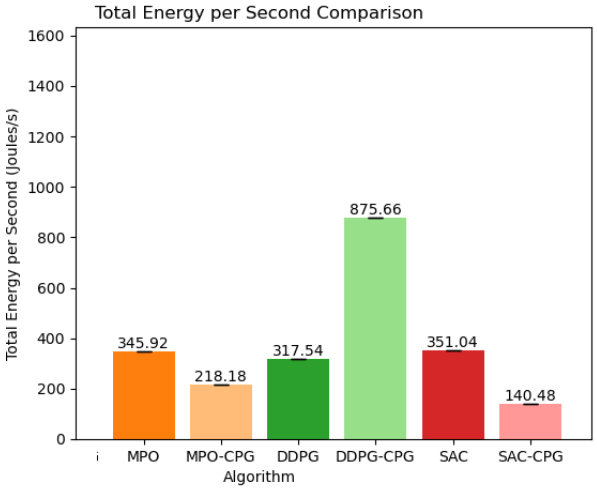

Fig. 5. Total energy expended per second from the Walker environment

In terms of distance travelled over the environemnt, Figure 3 shows how the best performance was the agent learned with the CPG Wrapper with Matsuoka Oscillators and trained with SAC algorithm, showing **18.34 mts** travelled, meaning more than the double of distance from the other methods. In the case for MPO and DDPG, we can see, that the variation with the CPG wrapper was insignificant.

In the case of the rewards metric (Figure 4), gathered by the environment in its default form, we found that the highest reward was on the CPG wrapped agent, performing with SAC algorithm, also with a value higher than double of the rest **(2,657)**. For DDPG, rewards were higher without CPG, and MPO results were insignificant.

At analysing energy expenditure from Figure 5, we are expecting to see less energy being used from either SAC **(140.48 J/s)** and MPO **(218.18 J/s)** algorithms, while DDPG with CPG used more than double from all the other algorithms.

The most important movement to take a look is the cross-correlation movement between right and left legs as seen in Figure 6, where is showing that the minimum error from each of the algorithms come from the CPG trained agent, with the biggest gap and more cyclic motion the SAC-CPG agent **(6.29%)**.

B. Experiment 2: Humanoid-v4 Environment Adaptation

In the Humanoid-v4 environment, we adapted the CPG to control the oscillatory patterns of a more complex bipedal structure. The DRL model demonstrated adaptability to the additional joints and complexities, which is critical for simulating human-like movement.

At increasing the DoF and plane complexity of the environment, the algorithms also gave more variation in the results. In all the trials, plain DRL outpuerformed CPG variant, as seen in Figure 7. Therefore, since distance is directly related to the rewards, in all agents, we saw similar results from Figure 8, with plain DRL outperforming CPG agents, with a smaller gap.

In Figure 9 the highest gap is between SAC and SAC-CPG, showing a higher energy expenditure on the CPG trained agent, while MPO with CPG, show a smaller amount, but only of 19.63%.

On the cross correlation analysis from Figure 10, all scenarios show a smaller error between the movement from each leg, being SAC-CPG the one showing the minimal value **(3.02%)**.

C. Experiment 3: Validation with Musculoskeletal Model

Following the performance analysis in Walker2d and Humanoid environments, the most promising algorithm, under the measured variables (stability, energy, and coordination), Soft Actor-Critic (SAC), was selected for further validation on a more complex musculoskeletal model with 36 DoF, as showcased in Figure 11. The results demonstrated a trade-off between energy efficiency and locomotion effectiveness.

The distance traveled as seen in Figure 12 was notably lower in the CPG-enhanced SAC model (11.88 meters) compared to plain SAC (15.92 meters), the model tending to fall. Figure 14, as expected, showcased the energy expenditure significantly lower with the CPG variant (256.89 Joules) versus plain SAC (344.60 Joules)a 25.4% reduction, reinforcing the energy efficiency of rhythmic control. However, the reward score accumulated, in Figure 13 followed a similar trend to the distance traveled, being lower in the CPG-enhanced model (295,566) compared to plain SAC (505,705).

An interesting result emerged in the correlation between both legs. The RMSE in leg synchronization was lower in SAC-CPG (7.24%) than in plain SAC (12.17%), indicating better coordination between limbs. This suggests that SAC-CPG produced a more synchronized gait pattern, whereas plain SAC exhibited greater variance in movement. Closer inspection of motion trajectories revealed that plain SAC caused both legs to move together in a jumping motion, as seen

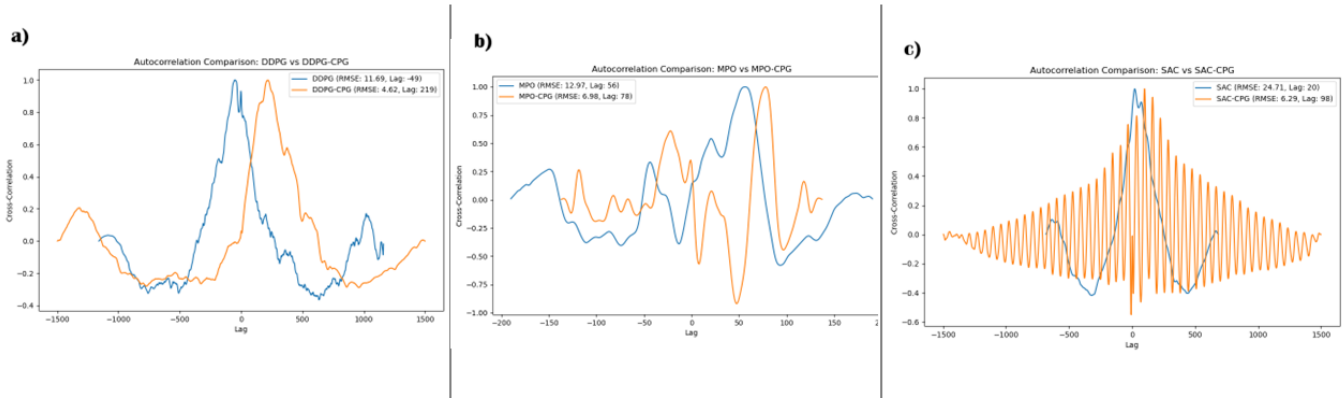

Fig. 6. Cross-Correlation for the left and right legs of the Walker-2d.

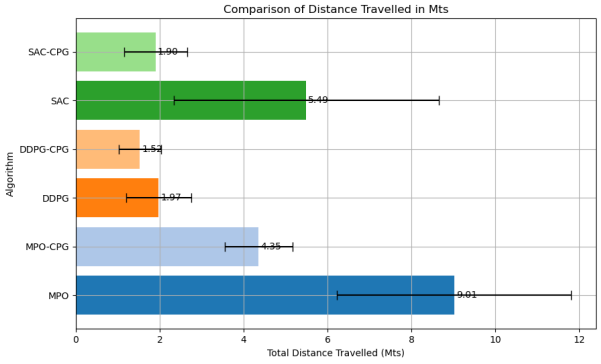

Fig. 7. Distance Travelled in Humanoid Mujoco environment

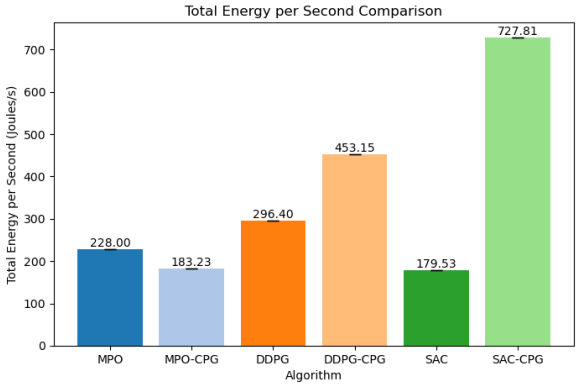

Fig. 9. Energy Expenditure from the humanoid agents

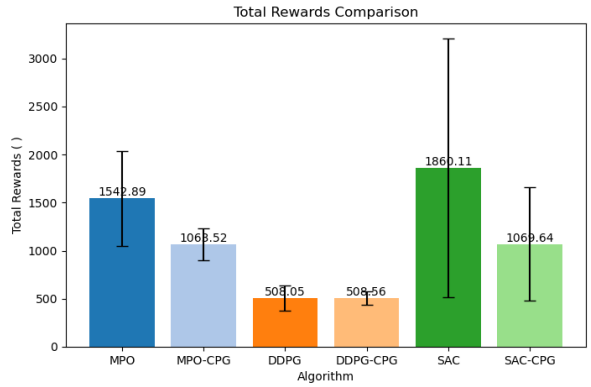

Fig. 8. Mean of rewards from the episodes of the Humanoid agent performance

in Figure 11, which artificially improved correlation scores despite being biomechanically unnatural. In contrast, SAC-CPG achieved a biologically inspired alternating gait, further supporting the role of CPGs in improving locomotion patterns.

IV. Discussions

The results from the three experiments highlight the strengths and trade-offs of integrating CPGs with DRL for bipedal locomotion. In Experiment 1 (Walker2d-v4), SAC-CPG showed better cyclic stability and energy efficiency, making it suitable for applications where rhythmic gait is prioritized. However, Experiment 2 (Humanoid-v4) revealed that plain DRL (SAC without CPG) performed better in distance traveled and reward accumulation, suggesting that raw reward maximization favors a more exploratory policy without predefined oscillations. These results are aligned with previously found experiments, where CPGs improved motion smoothness [8], and reduced energy cost in humanoid gait [11].

Given these trade-offs, we focused on SAC+CPG for musculoskeletal model testing, as it provided better synchronization and energy efficiency, key factors in real-world prosthetic applications. Nevertheless, alternative algorithms like MPO or DDPG could be explored for high-speed gait or terrain adaptation scenarios.

A key insight from Experiment 3 is that raw synchronization metrics like RMSE do not fully capture the qualitative differ-

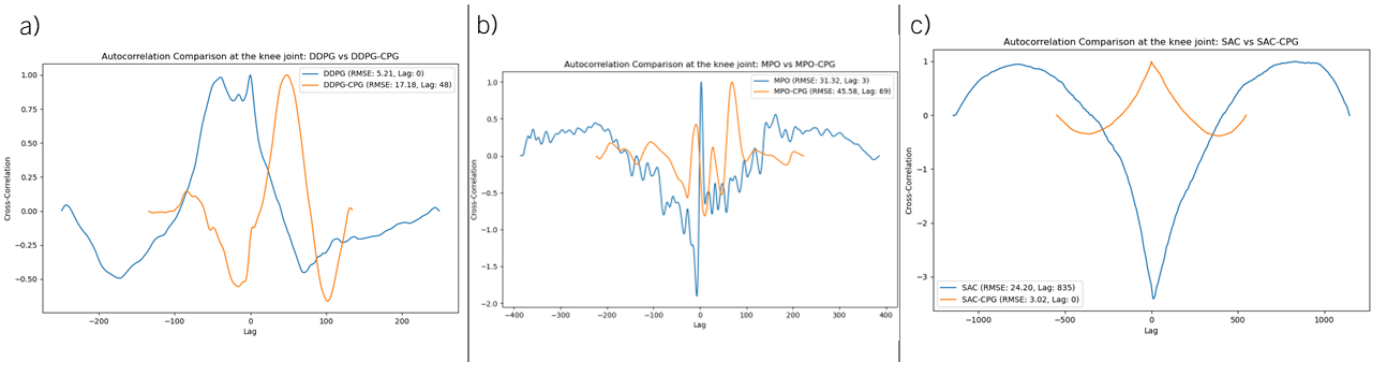

Fig. 10. Cross-Correlation for the left and right legs of the Humanoid-v4.

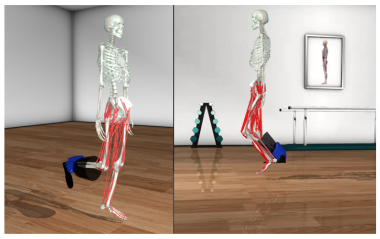

Fig. 11. Musculoskeletal Model, left showing the model walking with SAC-CPG algorithm, right showing the model locomotion, jumping with SAC algorithm

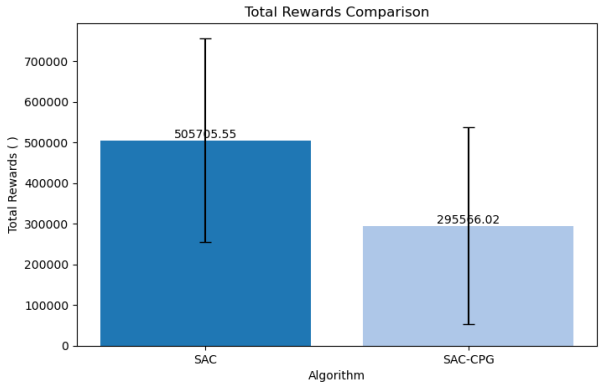

Fig. 13. Mean of rewards from the episodes of the ms-model agent performance

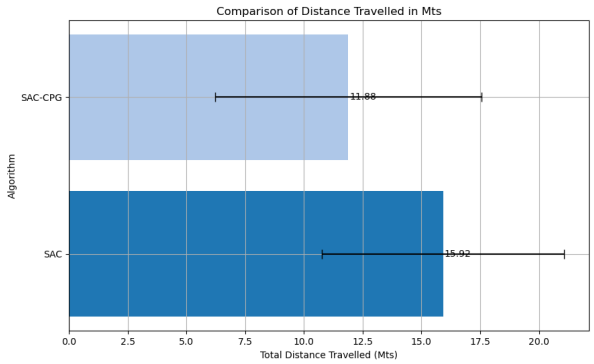

Fig. 12. Distance Travelled in the MS-model from MyoSuite

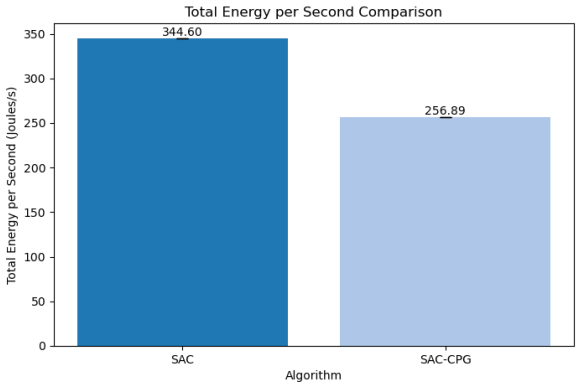

Fig. 14. Energy Expenditure from the humanoid agents

ences in gait. While plain SAC had a small RMSE, similar to CPG+SAC, this was due to unintended jumping motion rather than a biologically accurate gait cycle. The SAC-CPG model, achieved lower RMSE values and demonstrated a more natural and synchronized leg movement pattern, reinforcing the hypothesis that CPGs contribute to more stable and biomechanically realistic locomotion.

These findings suggest that while DRL alone can optimize reward-based locomotion, integrating CPGs leads to more energy-efficient and natural gait cycles. Future work should refine reward functions to better balance distance traveled and energy efficiency, as well as explore hybrid training approaches to leverage both trajectory-based learning and rhythmic control mechanisms.

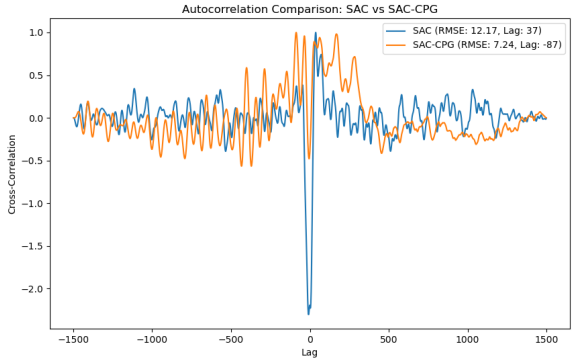

Fig. 15. Cross-Correlation for the left and right legs using SAC algorithm on the MS model.

V. Conclusions

The results from experiments confirmed SAC as the most effective RL algorithm for rhythmic locomotion under controlled conditions, particularly when combined with CPGs for energy-efficient gait synchronization. However, in cases where maximizing reward accumulation and distance traveled is a priority, plain SAC, or even MPO performed better, suggesting that different tasks may require different optimization objectives.

Comparative analysis demonstrated that SAC+CPG outperformed standalone SAC by producing a more energy-efficient and synchronized gait, highlighting its superiority over other RL approaches and simplifying future research directions. Additionally, the successful adaptation of SAC+CPG from simpler models from MuJoCo environments to complex musculoskeletal simulations in MyoSuite underscores its scalability for real-world robotic and prosthetic applications, reinforcing its potential for improving gait synchronization and reducing energy consumption in advanced prosthetic designs and human-robot interaction.

Acknowledgment

The authors would like to acknowledge the financial support of Writing Lab, Institute for the Future of Education, Tecnologico de Monterrey, Mexico, in the production of this work.

References

[1] J. Reher and A. D. Ames, "Dynamic walking: Toward agile and efficient bipedal robots," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 4, pp. 535–572, Volume 4, 2021 May 3, 2021, Publisher: Annual Reviews, ISSN: 2573-5144. DOI: 10.1146/annurev-control-071020-045021. [Online]. Available: https://www.annualreviews.org/content/journals/10.1146/annurev-control-071020-045021 (visited on 09/06/2024).

[2] C. Kaymak, A. Ucar, and C. Guzelis, "Development of a new robust stable walking algorithm for a humanoid robot using deep reinforcement learning with multi-sensor data fusion," *Electronics*, vol. 12, no. 3, p. 568, Jan. 2023, Number: 3 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2079-9292. DOI: 10.3390/electronics12030568. [Online]. Available: https://www.mdpi.com/2079-9292/12/3/568 (visited on 09/06/2024).

[3] N. Heess, D. TB, S. Sriram, *et al.*, *Emergence of locomotion behaviours in rich environments*, Jul. 10, 2017. DOI: 10.48550/arXiv.1707.02286. arXiv: 1707.02286[cs]. [Online]. Available: http://arxiv.org/abs/1707.02286 (visited on 04/11/2025).

[4] A. Abdolmaleki, J. T. Springenberg, Y. Tassa, R. Munos, N. Heess, and M. Riedmiller, *Maximum a posteriori policy optimisation*, Jun. 14, 2018. arXiv: 1806.06920[cs, math, stat]. [Online]. Available: http://arxiv.org/abs/1806.06920 (visited on 08/30/2023).

[5] T. Haarnoja, S. Ha, A. Zhou, J. Tan, G. Tucker, and S. Levine, "Learning to walk via deep reinforcement learning," *arXiv:1812.11103 [cs, stat]*, Jun. 19, 2019. arXiv: 1812.11103. [Online]. Available: http://arxiv.org/abs/1812.11103 (visited on 11/30/2021).

[6] L. Carvalho Melo and M. R. Omena Albuquerque Máximo, "Learning humanoid robot running skills through proximal policy optimization," in *2019 Latin American Robotics Symposium (LARS), 2019 Brazilian Symposium on Robotics (SBR) and 2019 Workshop on Robotics in Education (WRE)*, ISSN: 2643-685X, Oct. 2019, pp. 37–42. DOI: 10.1109/LARS-SBR-WRE48964.2019.00015. [Online]. Available: https://ieeexplore.ieee.org/document/9018554 (visited on 10/03/2024).

[7] K. Matsuoka, "Mechanisms of frequency and pattern control in the neural rhythm generators," *Biological Cybernetics*, vol. 56, no. 5, pp. 345–353, Jul. 1987, ISSN: 0340-1200, 1432-0770. DOI: 10.1007/BF00319514. [Online]. Available: http://link.springer.com/10.1007/BF00319514 (visited on 08/15/2024).

[8] A. J. Ijspeert, "Central pattern generators for locomotion control in animals and robots: A review," *Neural Networks*, vol. 21, no. 4, pp. 642–653, May 2008, ISSN: 08936080. DOI: 10.1016/j.neunet.2008.03.014. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0893608008000804 (visited on 09/06/2024).

[9] D. Gough, S. Oliver, and J. Thomas, *An Introduction to Systematic Reviews*. SAGE Publications, 2017, ISBN: 978-1-4739-6822-6. [Online]. Available: https://books.google.com.mx/books?id=41sCDgAAQBAJ.

[10] Y. Wang, X. Xue, and B. Chen, "Matsuoka's CPG with desired rhythmic signals for adaptive walking of humanoid robots," *IEEE Transactions on Cybernetics*, vol. 50, no. 2, pp. 613–626, Feb. 2020, ISSN: 2168-2267, 2168-2275. DOI: 10.1109/TCYB.2018.2870145. [Online]. Available: https://ieeexplore.ieee.org/document/8488532/ (visited on 08/15/2024).

[11] G. Li, A. Ijspeert, and M. Hayashibe, "AI-CPG: Adaptive imitated central pattern generators for bipedal locomotion learned through reinforced reflex neural networks," *IEEE Robotics and Automation Letters*, vol. 9, no. 6, pp. 5190–5197, Jun. 2024, ISSN: 2377-3766, 2377-3774. DOI: 10.1109/LRA.2024.3388842. [Online]. Available: https://ieeexplore.ieee.org/document/10499824/ (visited on 08/15/2024).

[12] E. Todorov, T. Erez, and Y. Tassa, "MuJoCo: A physics engine for model-based control," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Vilamoura-Algarve, Portugal: IEEE, Oct. 2012, pp. 5026–5033, ISBN: 978-1-4673-1736-8 978-1-4673-1737-5 978-1-4673-1735-1. DOI: 10.1109/IROS.2012.6386109. [Online]. Available: http://ieeexplore.ieee.org/document/6386109/ (visited on 05/12/2023).

[13] L. De Vree and R. Carloni, "Deep reinforcement learning for physics-based musculoskeletal simulations of healthy subjects and transfemoral prostheses' users during normal walking," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 607–618, 2021, ISSN: 1534-4320, 1558-0210. DOI: 10.1109/TNSRE.2021.3063015. [Online]. Available: https://ieeexplore.ieee.org/document/9366532/ (visited on 05/07/2022).

[14] T. P. Lillicrap, J. J. Hunt, A. Pritzel, *et al.*, "Continuous control with deep reinforcement learning," *arXiv:1509.02971 [cs, stat]*, Jul. 5, 2019. arXiv: 1509.02971. [Online]. Available: http://arxiv.org/abs/1509.02971 (visited on 01/25/2022).

[15] A. Raffin, *RL baselines3 zoo*, original-date: 2020-05-05T05:53:27Z, May 2, 2025. [Online]. Available: https://github.com/DLR-RM/rl-baselines3-zoo (visited on 05/02/2025).

[16] K. Matsuoka, "Sustained oscillations generated by mutually inhibiting neurons with adaptation," *Biological Cybernetics*, vol. 52, no. 6, pp. 367–376, Oct. 1985, ISSN: 0340-1200, 1432-0770. DOI: 10.1007/BF00449593. [Online]. Available: http://link.springer.com/10.1007/BF00449593 (visited on 08/15/2024).