Energy-Efficient Antenna Muting in Massive MIMO Networks Using Deep Reinforcement Learning

Alexandros Kouvatseas*, Anh-Khoa Dang†‡
*Université Paris Dauphine-PSL, Paris, France {alexandros.kouvatseas}@dauphine.eu
†Cnam, Paris, France {anh-khoa.dang}@cnam.fr
‡Ericsson, Massy, France {anh.khoa.dang}@ericsson.com

***Abstract*—Massive multiple-input multiple output (MIMO) has become a key enabler of 5G capacity, but the large number of active antenna elements significantly increases the Radio Access Network (RAN) energy consumption. Antenna muting dynamically reduces the number of active Transmit-Receive (TR) chains and offers a promising path to energy savings, provided that user Quality of Service (QoS) is preserved. In this article, we present a simulation framework that integrates live network measurements from a commercial 5G deployment with the high-fidelity Sionna RT ray-tracing simulator to study and optimize Massive MIMO muting policies under realistic simulation conditions. We formulate the muting problem as a Markov Decision Process (MDP) in which the agent selects an antenna configuration based on traffic load, user distribution across distance bins, and temporal context. Deep Reinforcement Learning (DRL) models are utilized and outperform standard threshold-based algorithms, which either achieve significantly lower energy savings or result in unacceptable QoS loss.**

***Index Terms*—Massive MIMO, antenna muting, energy efficiency, deep reinforcement learning, 5G**

I. Introduction

Massive MIMO arrays are central to the performance of 5G systems, enabling high spectral efficiency via directional multi-user beamforming. However, operating these large antenna panels with many RF chains and power amplifiers comes at a substantial energy cost. As networks densify and traffic patterns become more variable, reducing the energy footprint of Massive MIMO sites without compromising the QoS has become a key objective [1].

Antenna muting, also known as Massive MIMO sleep mode or TR de-activation, addresses this challenge by dynamically reducing the number of active antenna elements when the traffic demand is low. Practical base stations typically support only a small set of allowed configurations (e.g., 64, 32, 16, 8 TR) in order to preserve the structure of the underlying uniform planar array [2], [3]. While fewer active elements directly reduce energy consumption they also reduce beamforming gain leading to a potential throughput degradation if the muting is too aggressive.

This research work was supported by the French government, in the framework of France 2030 program (INTENTION-6G project).

ISBN 978-3-903176-82-9 © 2026 IFIP

The goal of this work is to design and evaluate intelligent muting strategies that balance energy savings and QoS in a realistic 5G setting. We focus on a single-cell scenario where muting decisions are taken at the timescale of Key Performance Indicators (KPI) measurements (15-minute intervals), following 3GPP specifications [4], using aggregated information about load and user distribution.

The simulation framework is built on the following format:

- **Real KPI traces** from a commercial (anonymous) 5G deployment collected over one month, providing realistic time-varying traffic and user distributions.
- **A Sionna RT ray tracing model** of an urban scenario to provide physically accurate channels and throughput under different TR configurations.

On top of this environment we formulate an antenna muting algorithm as a MDP and train several DRL agents, which we compare against classical threshold-based muting schemes.

A. Contributions

The main contributions of this work are:

- **Simulation for Massive MIMO muting.** We introduce a simulation pipeline in which real 5G KPI data drives a Sionna RT-based physical layer simulator, yielding realistic channels and throughput for different antenna configurations.
- **MDP formulation with QoS aware reward.** We formulate the muting decision as an MDP, and design a reward that trades off normalized energy consumption against relative throughput loss with respect to a 64-TR baseline, with explicit handling of high-use regimes.
- **DRL architectures for muting.** We utilize Proximal Policy Optimization (PPO) and Deep Q-Network (DQN), highlighting the impact of state representation and reward shaping on performance.
- **Benchmark against threshold baselines.** We compare DRL policies to simple utilization thresholds and an adapted dynamic muting scheme inspired by prior Massive MIMO sleep mode work, showing that DRL can achieve higher energy savings at comparable or better QoS levels.

B. Organization of the Paper

Section II reviews related work on energy-aware RAN control, simulations, and DRL for network optimization. Section III describes the KPI dataset, the Sionna RT-based simulation, and how ray-tracing outputs are converted into RL-ready metrics. Section IV formulates the antenna muting problem, defines the energy/QoS objectives and presents the baseline schemes and the RL/DRL models including the reward design. Section V reports the experimental results and comparisons. Finally, Section VI concludes and outlines future directions.

II. Related Work

A. Energy-aware RAN control

Energy efficient operation of cellular networks has been widely studied at the levels of turning symbols on/off, carrier shutdown [5] and Massive MIMO adaptation [2]. Prior work on antenna muting and sleep modes has demonstrated that deactivating subsets of antenna elements can provide substantial energy savings, but often under idealized traffic or channel models and without advanced learning based control.

In particular, dynamic muting schemes have been proposed in which the number of active TR elements is adapted based on cell utilization thresholds and a multi-level sleep mode structure [3]. These algorithms have reported energy savings on the order of 30% in realistic scenarios, but could be considered heuristic as they do not explicitly leverage spatial UE information or temporal patterns in traffic.

Parallel efforts on RAN energy saving via carrier on/off control have explored heuristic rules, Bayesian decision models [6] and DRL methods [5], [7]–[11]. DRL based controllers outperform static thresholds by learning context dependent policies that adapt to time-of-day patterns and user mobility. However, many of these works operate at the cell or carrier level or do not directly address the multi-level antenna configuration in Massive MIMO panels.

B. Digital twins and DRL for RAN

Digital twin approaches, in which a simulation environment is driven by real network measurements, are gaining interest as a way to safely train and evaluate learning based controllers [12]–[14]. Recent works have combined realistic channels, traffic models and reinforcement learning to design energy efficient RAN control strategies. A good example is the use of PPO based agents to control antenna or cell activity under QoS constraints [5].

Graph based deep learning has also been applied to network control problems where the underlying structure is inherently spatial, such as inter-cell interference coordination or routing in core networks. In our setting, the UE distance distribution can be interpreted as a small graph structure (bins connected along the radial dimension), which motivated the use of GNNs to encode spatial patterns for muting decisions [9].

III. Data and Simulation Environment

A. KPI dataset

The dataset that was used during the research consisted of one month of measurements from a live 5G deployment. The KPIs are aggregated in 15-minute intervals and exported per cell [4]. For each interval, a compact set of indicators was extracted that are directly relevant for antenna muting:

- **RRC connected UEs:** The number of UEs with an active Radio Resource Control (RRC) connection, used as a proxy for instantaneous demand and multiplexing potential.
- **PRB utilization:** A utilization metric describing the fraction of Physical Resource Blocks (PRBs) used in the downlink; this is our main notion of cell load.
- **TA distribution:** A histogram of Timing Advance (TA) indexes, which we map to approximate distances from the base station. Grouping TA indexes into a small number of distance bins yields an estimate of how far each UE is from the cell.

Raw KPI time series are aligned on the 15-minute grid, and intervals with missing or inconsistent values are discarded. PRB utilization is transformed into a "load factor" in [0,1] that represents the demand, by dividing it by the number of UEs. We consider, for simplicity, that the demand is uniform towards all UEs. Finally, TA histograms are mapped to distance bins using the standard relation between TA and propagation delay.

The UE count defines how many users are present in the scenario and the PRB utilization captures their aggregate demand. High utilization with few UEs indicates heavy usage per UE; low utilization with many UEs indicates mostly idle users. For our RL environment, the total UE count and the load factor were used as scalar features, while the vector of UE counts per distance bin was used as a structured feature. The TA-based distribution does not give exact UE positions but it provides a sufficiently accurate radial representation of the user locations. For this reason, each bin is interpreted as an annular region around the base station and, for simulation purposes, the users are placed uniformly at random within that annulus and within a 120° sector matching the antenna orientation. This preserves realistic angular spread and sectorization while respecting the TA statistics.

The resulting per-timestamp descriptor consists of:

- UE histogram over B distance bins,
- Total number of RRC UEs,
- Downlink load factor,
- Time-of-day and day-of-week encodings (sine/cosine)

This compact representation becomes the basis of the state vector for the RL agents.

B. Sionna RT simulation

To map KPIs and UE distributions to physical-layer performance we use Sionna, NVIDIA's GPU-accelerated link-level simulator, and its Sionna RT module built on top of Mitsuba 3. Sionna RT performs deterministic ray tracing in a 3D environment, capturing line-of-sight, reflections, refractions and scattering.

We proceeded with the simulations on built-in city scenes (Munich and Florence) as a representative dense urban environment. A 5G base station with a Massive MIMO panel is placed on a rooftop. The UEs are placed at ground level according to the distance-bin distribution. For each snapshot:

1) **UE placement.** The number of UEs per bin is drawn from the KPI histogram. Within each bin, UEs are uniformly distributed in angle within the sector and in radius within the bin limits.
2) **Channel computation.** Sionna RT computes the complex frequency response between the base station and each UE across all subcarriers, accounting for site-specific geometry and material properties.
3) **Antenna configuration.** The panel is configured with a given number of active transceiver elements $TR \in \{8, 16, 32, 64\}$, while preserving a uniform planar array geometry consistent with 3GPP recommendations.
4) **Beamforming and post-processing.** Conjugate (channel-based) beamforming is applied per UE and the effective channels are used to compute the post-equalization SINR and achievable rates.

Fig. 1a provides a top-down perspective of the urban layout, highlighting the transmitter (rooftop base station), the receiver locations at street level, and an example spatial map of the simulated radio conditions. Fig. 1b then depicts the corresponding deterministic ray-tracing propagation mechanisms between the base station and the UEs, including the line-of-sight component and dominant multi-path interactions (reflections/refraction and diffuse scattering) induced by the surrounding buildings.

The simulation was configured at 3.5 GHz with an OFDM grid (128 subcarriers) representative of a 5G deployment. For each snapshot and TR configuration, we obtain a per-UE and cell-level throughput under realistic propagation conditions. The 64 TR case serves as a full power reference and the reduced TR configurations are evaluated under the exact same UE placement and channel realization to isolate the effect of muting.

C. From ray tracing to RL-ready performance metrics

The ray-tracing outputs are converted into quantities usable by the RL environment by implementing a utility module that takes as input the complex channel frequency responses from Sionna RT, applies channel-based beamforming, and then computes the effective channel gains and SINR per UE stream (with LMMSE post equalization and uniform power allocation). Per-UE throughput is then evaluated via a Shannon-type capacity formula based on the SINR and noise assumptions.

This module is used offline to precompute, for each KPI snapshot, the throughput $R(\alpha)$ associated with each configuration α. The RL environment can then use these precomputed values at training time, avoiding expensive and computationally heavy ray tracing in the inner loop while preserving the realism of the original channels.

IV. Modeling

We consider a single Massive MIMO panel whose active TR configuration at time t can be chosen from the set of configurations $\{8, 16, 32, 64\}$. For each KPI snapshot and chosen configuration α, the simulator returns a cell throughput $R(\alpha)$. The full configuration throughput is denoted as $R_{\text{full}} = R(64)$.

For the QoS term we want to retain, we consider that the best way to describe it is by calculating the deviation of the average throughput with the configuration TR chosen by the agent, and the full configuration throughput:

$$\Delta(\alpha) = \frac{R_{\text{full}} - R(\alpha)}{R_{\text{full}}} \tag{1}$$

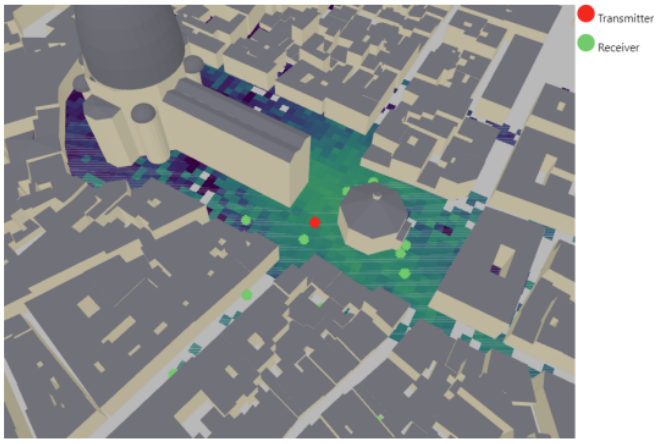
In high-use situations (i.e., when R_{full} exceeds a given threshold), we require $\Delta(\alpha)$ to remain below a small tolerance, reflecting that only limited QoS degradation is acceptable. In very low-use situations, larger deviations may be tolerated because the cell operates far from capacity. To model the energy cost, we use a normalized proxy based on the number of active TR chains:

$$E(\alpha) = \left(\frac{\alpha}{64}\right)^q \tag{2}$$

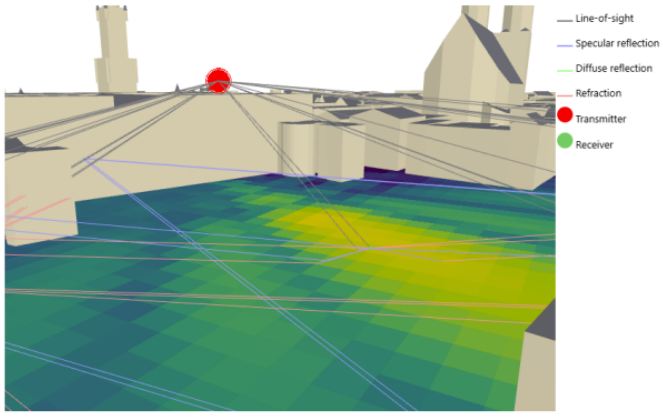
with $q > 1$ reflecting the super-linear growth of power consumption with the number of active RF chains. In all experiments, we set $q = 2$.

Finally, we distinguish high-use from very low-use conditions using the full-configuration throughput R_{full}. When R_{full} falls below a small activity threshold (here 0.05 bps/Hz), the cell operates far from capacity and users are effectively idle; in this regime, moderate throughput deviations have negligible practical impact, and QoS constraints can be relaxed compared to high-use intervals.

The objective is to design a muting policy π that, over a long horizon of KPI-driven time-slots, minimizes the expected energy cost while maintaining acceptable QoS, particularly under high-use conditions.



(a) Top view of the Florence scene with transmitter (Tx), receivers (Rx), and an example coverage/received-power map.



(b) Example ray-tracing paths showing LoS and multipath components (specular/diffuse reflections and refraction).

Fig. 1: Sionna RT urban simulation illustrations used in this work.

A. Baseline formulation

We evaluate two threshold-based muting schemes: (i) a simple two-level baseline that switches between 64 and 32 TR, and (ii) a four-level dynamic muting scheme (64/32/16/8 TR) inspired by Massive MIMO sleep-mode strategies [3]. Both baselines are implemented as *greedy QoS-constrained selectors*: at each time step t, we compute the full-configuration reference throughput $R_{\text{full},t} = R_t(64)$ and then choose the *smallest* TR configuration that respects a QoS tolerance $\epsilon = 5\%$ in high-use conditions. Concretely, for a candidate action $a \in \{32\}$ (two-level) or $a \in \{8, 16, 32\}$ (four-level, tested from smallest to largest), we evaluate the relative throughput loss, defined as:

$$\Delta_t(a) = \frac{R_{\text{full},t} - R_t(a)}{R_{\text{full},t}}. \tag{3}$$

We accept the first candidate satisfying $\Delta_t(a) \leq \epsilon$; if none satisfies the constraint, we fall back to $a = 64$. Performance is summarized using the same metrics as for the DRL agents: average energy savings relative to always using 64 TRs (with energy assumed proportional to the fraction of active TRs), and QoS retainment, measured as the mean ratio over high-use timesteps (cell-level throughput), i.e. $\frac{R_t(a_t)}{R_{\text{full},t}}$.

The two-level baseline is a coarse, operationally simple rule that only considers a single downshift option ($64 \rightarrow 32$ TR): it mutes to 32 TR when the QoS-loss constraint $\Delta_t(32) \leq \epsilon$ is met, and remains at 64 TR otherwise, yielding limited granularity in the energy–QoS trade-off. The dynamic threshold baseline introduces finer muting granularity by searching over multiple progressively smaller configurations ($8 \rightarrow 16 \rightarrow 32$ TR): it selects the smallest feasible TR that satisfies $\Delta_t(a) \leq \epsilon$, enabling larger energy savings in low-load regimes while still reverting to 64 TR when none of the reduced configurations meets the tolerance. This better approximates stepwise sleep-mode strategies such as those in [3].

B. Reinforcement Learning formulation

We model the muting problem as an MDP defined by a state space (variables describing the environment), an action space (the discrete set of choices available to the agent), and a reward function the agent seeks to maximize. In our case, the MDP follows the structure below:

- **State space** s_t. At each time step t each s_t consists of:
  - The UE histogram over distance bins
  - The total number of RRC connected UEs
  - The cell utilization (PRB utilization)
  - Temporal features such as hour-of-day and day-of-week
- **Action Space** α_t. The α_t is the chosen TR configuration for the next interval: $\alpha_t \in \{8, 16, 32, 64\}$.
- **Transition.** Given s_t and α_t, the environment queries precomputed throughput values obtained offline from Sionna RT. The next state s_{t+1} is derived from the next KPI snapshot in the time series. In this work, traffic dynamics are treated as exogenous: the muting decision does not feed back into the KPI evolution.

In this paper, we adopt Deep Q-Network (DQN) [15], an off-policy value-based reinforcement learning method well suited to discrete control and known for its sample efficiency. As a comparison baseline, we also consider Proximal Policy Optimization (PPO) [16], an on-policy reinforcement learning method trained with the same reward function, allowing us to compare off-policy and on-policy learning under the same discrete action space.

C. Reward Design

The reward design needs to balance out the relative throughput loss Δ_t that we defined (where $\Delta_t = 0$ corresponds

to matching the full configuration, while larger Δ_t indicates stronger QoS degradation) and the energy cost. In addition, we distinguish high-use and low-use regimes via an indicator H_t, in which:

$$H_t = \begin{cases} 1, & R_{\text{full},t} \geq R_{\text{high}}, \ 0, \\ \text{otherwise}, \end{cases} \tag{4}$$

where R_{high} is a throughput threshold above which QoS loss is considered critical. In low-use regimes the reward can focus more on energy savings while in high-use the QoS degradation is strongly penalized.

Initially, we used a simple linear form to show the trade off between the relative throughput and energy cost:

$$r_t^{\text{lin}} = H_t\big(1 - \Delta_t\big) - \lambda_{\text{lin}}, E(a_t) \tag{5}$$

When the cell is in high-use, the first term rewards configurations that keep throughput close to the 64-TR reference; the second term penalizes energy consumption in all regimes. In practice, this formulation turned out to be too conservative, and the corresponding DQN tended to stay close to 64 TR, yielding limited energy savings.

For this reason we proceeded with a quadratic dependence on Δ_t, where small throughput losses lead to a mild reduction of the reward, whereas larger losses are penalized much more severely. Our DQN is based on this reward and it produced stable learning and a good compromise between savings and QoS, and the same form was used for the PPO agent:

$$r_t^{\text{quad}} = H_t\big(1 - \Delta_t\big)^2 - \lambda_{\text{quad}}, E(a_t) \tag{6}$$

V. Experimental Results

Implementation details

Both agents use a multi-layer perceptron policy over the discrete action space $\alpha_t \in \{8, 16, 32, 64\}$, with a state vector of length $|B| + 6$ where $B = 10$ is the number of distance bins. DQN is trained with learning rate 10^{-3}, batch size 64, $\gamma = 0.95$, and $\lambda_{\text{quad}} = 0.15$. PPO uses learning rate 3×10^{-4}, batch size 64, $\gamma = 0.95$, rollout length 128, clip range 0.2, and $\lambda_{\text{lin}} = 0.22$. Both are trained for 3000 episodes. The high-use threshold is $R_{\text{high}} = 0.05$ bps/Hz, and the QoS tolerance for threshold baselines is $\epsilon = 5\%$. The one-month KPI trace is split chronologically into training and held-out test portions.

TABLE I: Energy Savings and QoS Retainment for Selected Policies for $R_{\text{full}} > 0.05$ bps/Hz

Method	Energy Savings [%]	QoS Retainment [%]
DQN	23.44	98.82
PPO	18.47	97.60
Simple threshold	8.40	97.18
Dynamic threshold	16.24	95.67

As shown in Table I, the DQN achieves the best overall trade-off, reaching 23.44% energy savings while retaining 98.82% of the full-power throughput in high-use intervals. The PPO agent is slightly more conservative, with 18.47% savings and 97.60% QoS retainment, which confirms that the quadratic reward drives both methods towards "safe" muting policies. In contrast, the simple threshold baseline preserves QoS at a similar level (97.18%) but only delivers 8.40% savings, illustrating the limitations of purely utilization-based rules. The dynamic threshold improves the savings to 16.24% with a modest QoS reduction to 95.67%, yet still remains below ours in both dimensions, indicating that learning-based policies can exploit richer state information to reach a more favorable energy–QoS operating point.

VI. Conclusion and Future Work

This work introduced a simulation framework for evaluating energy-efficient antenna muting strategies in Massive MIMO networks. KPI data from live networks were integrated with a Sionna RT-based ray-tracing simulator, and two reinforcement learning approaches (DQN and PPO) were benchmarked against utilization-based threshold policies. The results showed that learning-based muting can substantially increase energy savings while preserving QoS.

We acknowledge several limitations of the present study. The per-UE demand is assumed uniform across active users, which simplifies the load model but may not reflect heterogeneous traffic profiles. Traffic dynamics are treated as exogenous, so the muting decision does not feed back into KPI evolution. The evaluation is restricted to a single-cell scenario in two urban scenes (Munich and Florence), leaving inter-cell interference and multi-cell coordination out of scope. Finally, results are reported for a single representative training run; a full multi-seed sensitivity analysis is left for future work.

Future work will focus on incorporating forecasting capabilities and operating at finer time scales. In particular, integrating time-series load predictors would allow the algorithm to anticipate demand when real-time measurements are unavailable. In parallel, adapting the framework from 15-minute KPIs toward (sub-)second time granularity would enable more responsive control and a fairer comparison with dynamic threshold baselines. This finer granularity would also require a more efficient, GPU-friendly simulation pipeline and improved robustness across heterogeneous cells. Finally, extending the comparison to additional learning-based baselines (e.g., contextual bandits, multi-agent RL) would yield more comprehensive results.

References

[1] D. López-Pérez, A. D. Domenico, N. Piovesan, and M. Debbah, "Data-Driven Energy Efficiency Modeling in Large-Scale Networks: An Expert Knowledge and ML-Based Approach," *IEEE Transactions on Machine Learning in Communications and Networking*, vol. 2, pp. 780–795, 2024.

[2] H. Asplund, D. Astely, M. Buchmayer, P. von Butovitsch, T. Chapman, S. Faxér, M. Frenne, C. Friberg, F. Ghasemzadeh, B. Göransson, M. Hagström, B. Hogan, Y. Jading, G. Jöngren, J. K. B, E. Larsson, M. Ljungberg, J. Rao, J. Rosenberg, and Y. Yang, *Massive MIMO Handbook, Third Edition: Extended Version*, 3rd ed., P. von Butovitsch, D. Astely, and E. Larsson, Eds. Ericsson AB, 2024. [Online]. Available: http://ericsson.com/massive-mimo

[3] P. Frenger and K. W. Helmersson, "Massive MIMO Muting using Dual-polarized and Array-size Invariant Beamforming," in *2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring)*. IEEE, 2021.

[4] 3GPP, "3GPP TS 28.552: Management and orchestration; 5g performance measurements," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 28.552, release 20; Version 20.0.0 (uploaded 2025-09-30). Accessed: 2026-01-07.

[5] A.-K. Dang, H. Khalifé, M. Sintorn, S. Rovedakis, and S. Secci, "Data-driven Energy Optimization in Mobile Networks with User Experience Guarantees," in *IEEE INFOCOM 2025 - IEEE Conference on Computer Communications*, London, United Kingdom, May 2025. [Online]. Available: https://hal.science/hal-05040259

[6] L. Maggi, C. Mihailescu, Q. Cao, A. Tetich, S. Khan, S. Aaltonen, R. Koblitz, M. Holma, S. Macchi, M. E. Ruggieri, I. Korenev, and B. Klausen, "Energy Savings under Performance Constraints via Carrier Shutdown with Bayesian Learning," in *2023 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*. IEEE, 2023, pp. 1–6.

[7] S. Bassoy, R. Behravesh, and J. Pujol-Roig, "SEEDRL: Smart Energy Efficiency using Deep Reinforcement Learning for 6G Networks," in *2023 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2023, pp. 732–737.

[8] J. S. Pujol-Roig, S. Wu, Y. Wang, M. Choi, and I. Park, "Deep Reinforcement Learning for Cell On/Off Energy Saving on Wireless Networks," in *2021 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2021, pp. 1–6.

[9] T. Cai, Q. Wang, S. Zhang, Özlem Tuğfe Demir, and C. Cavdar, "Multi-agent Reinforcement Learning for Energy Saving in Multi-Cell Massive MIMO Systems," *arXiv preprint arXiv:2402.03204*, 2 2024.

[10] M. Choi, K. Kim, H. Jang, H. Woo, J. S. Pujol-Roig, Y. Wang, H. Yeon, S. Choi, and S. Jang, "Cell On/Off Parameter Optimization for Saving Energy via Reinforcement Learning," in *2021 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2021, pp. 1–6.

[11] M. Hoffmann and M. Dryjański, "Energy efficiency in open ran: Rf channel reconfiguration use case," *IEEE Access*, vol. 12, pp. 118 493–118 501, 2024.

[12] R. Pegurri, F. Linsalata, E. Moro, J. Hoydis, and U. Spagnolini, "Toward digital network twins: Integrating sionna rt in ns-3 for 6g multi-rat networks simulations," 2025. [Online]. Available: https://arxiv.org/abs/2501.00372

[13] J. Hoydis, F. A. Aoudia, S. Cammerer, M. Nimier-David, N. Binder, G. Marcus, and A. Keller, "Sionna RT: Differentiable Ray Tracing for Radio Propagation Modeling," *arXiv preprint arXiv:2303.11103*, 7 2023. [Online]. Available: https://arxiv.org/abs/2303.11103

[14] Z. Yun and M. F. Iskander, "Ray Tracing for Radio Propagation Modeling: Principles and Applications," *IEEE Access*, vol. 3, pp. 1089–1100, 2015.

[15] V. Mnih, K. Kavukcuoglu, D. Silver *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[16] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.