Electronics, Communications, and Computing Summit
Vol. 3, No. 3, Jul - Sep 2025, pp. 21-32
ISSN: 3107-8222, DOI: https://doi.org/10.17051/ECC/03.03.03

DRL-Driven Hybrid Beamforming Architecture for THz-Band MIMO Communication Systems

Kagaba J. Bosco[1], S. M Pavalam[2]

[1,2]Information and Communications Technology, National Institute of Statistics of Rwanda, Kigali, Rwanda
Email: Bosco.je.kag@nur.ac.rw[1], pav.sm@nur.ac.rw[2]

Article Info

Article history:

Received : 18.07.2025
Revised : 17.08.2025
Accepted : 10.09.2025

Keywords:

Terahertz (THz) Communication,
Massive MIMO,
Hybrid Beamforming,
Deep Reinforcement Learning (DRL),
Spectral Efficiency,
Energy-Efficient Wireless Systems

ABSTRACT

The expansion of wireless communication systems and the rapid development of wireless communication systems have led to the transition to the use of ultra high-frequency bands, the terahertz (THz)-band communication can be regarded as one of the basic supports of 6G networks. THz communication coupled with massive multiple-input multiple- output (MIMO) systems can deliver extremely high data rates, ultra-low latency as well as high spectral efficiency. There are however major challenges associated with these systems, including complex hardware, high path loss as well as channel sparsity especially during development of hybrid beamforming architecture, a system that integrates both analog and digital precoding. This paper suggests a brand-new hybrid beamforming structure driven by deep reinforcement learning (DRL), particularly adjusted to THz-band MIMO communication systems in order to resolve those drawbacks. The aim of the project is to come up with an adaptive, scalable, and energy-efficient beamforming solution that is capable of intelligently exploring the high dimensional and dynamic wireless environment. The DRL approach (proposed method) consists of actor-critic based DRL architecture, with an agent constantly participating in the environment as a way of maximizing both analog and digital beamforming matrices. The system model considers a wideband, spatially sparse, frequency selective THz channel whose structure is built based on a high-order hybrid structure with a partially connected structure due to the hardware overhead. The DRL agent is trained through a soft actor-critic (SAC) algorithm on a set of a large number of simulations, in which to balance exploration and exploitation and to maximize a reward function related to spectral efficiency, power consumption, and signal quality. A realistic THz channel model, including beam squint as well as user mobility, supports the training process. Performance assessment reveals that the suggested DRL-centered beamforming system provides BSI throughput (up to 25Percent), channel dynamics resilience, and energy economy are substantially higher than that of the traditional practices like orthogonal matching pursuit (OMP) and codebook-based systems. The findings confirm the practicality of incorporation of deep reinforcement learning in the process of designing hybrid beamforming to provide a scalable and intelligent approach to the next-generation THz-band wireless systems. In future, the model can be extended to multi-user and multi-cell systems, include the consideration of reconfigurable intelligent surfaces (RIS), and consider the application of real-time hardware.

1. INTRODUCTION

An increasing pattern of mobile data traffic, mixed with the advent of new complex apps (holographic communications, extended reality (XR), and high-fidelity digital twins), is making the current wireless networks extend beyond their operating capabilities. The terahertz (THz) frequency band (0.1-10 THz) has been a potential new frontier in order to support the ambitious requirements of sixth-generation (6G) wireless systems. THz communication provides enormous un utilized spectrum and it can offer Tbps of data rate, ultra-low latency and very high spectral efficiency. Communication in THz band, when combined with massive multiple-input multiple-output (MIMO) systems, can make use of high spatial resolution and narrow directional beams to overcome the high path losses resulting when high frequencies are propagated. But, practical THz MIMO systems are accompanied by a lot of difficulties, such as

power-consuming RF devices, strong free-space losses, molecular absorption, and frequency-selective sparse channel properties (beam squint).

As a compromise between system performance and hardware realisation, hybrid beamforming architectures have been suggested, where a split of the beamforming operation between the analog and digital domain is performed. Such designs can use few RF chains and large antenna arrays by phase exchanging with fewer RF chains, lower power and expense. However, analog-digital precoder joint optimization is not a trivial problem because of the non-convexity of the problem, and the necessity of quick time-varying channel adaptation. Further, wideband THz links add the effect of beam squint and it also becomes extremely challenging to obtain accurate channel state information (CSI) in highly dynamic applications. Conventional approaches like compressive sensing, orthogonal matching pursuit (OMP) and design that uses codebooks were extensively used in mmWave systems but do not scale well with sparse frequency-selective at THz frequencies and are not robust towards sparsity.

Machine learning methods, especially deep learning, have gained popularity of late in solving these problems and they can potentially enhance channel estimate and beam selection. Nevertheless, the current methods work within supervised learning framework, which requires big labeled datasets that are hard to obtain in real-life applications THz scenarios. Also, these models are prone to be trained offline and do not adjust well to the online changes in channel conditions. Instead, deep reinforcement learning (DRL) may be considered an encouraging alternative given that it can use interaction with the environment to learn optimal policies. There is recent promise in using DRL to beamforming and resource allocation in mmWave and sub-THz systems, but the use of DRL in the context of hybrid beamforming and achieving wideband THz MIMO, subject to beam squint, limited feedback, and energy efficiency constraints has been little studied so far.

This gap in the literature will be addressed by this study by creating a new DRL-powered hybrid beamforming architecture specific to the THz-band MIMO system. The suggested solution is actor-critic based DRL algorithm, namely, the Soft Actor-Critic (SAC) method, that will flexibly learn the analog and digital precoders that optimize spectral efficiency with minimal power consumption. The study also feeds into the vision of wireless network and systems that are AI-native with a scalable architecture that generalizes the array topology and communication situation types. The aims of the work are to model realistic wideband THz channels, design the hybrid beamforming problem in reinforcement learning, implement and train the DRL agent and compare the systems tested under different conditions to conventional and learning based basis. In this paper, we would like to present a feasible and smart beamforming scheme that can be used in the next-generation at THz communications system.

2. RELATED WORK

Hybrid beamforming as the promising solution to mitigate hardware complexity in millimeter-wave (mmWave) and terahertz (THz) MIMO systems has been actively studied. The majority of the initiating works are concentrated on exploiting the scarcity of activities of the channels in the THz regime through compressive sensing and greedy approaches. Instead the orthogonal matching pursuit (OMP) algorithm has proved very popular as a means of approximating optimal precoding vectors and combining vectors through the selection of dominant paths relying on channel sparsity [1]. Nevertheless, such methods usually presume quasi-static conditions and well-known channel information, and both conditions are progressively impracticable in the context of very dynamic THz communications.

The second commonly covered method is codebook-based hybrid beamforming: in this method, a set of predefined analog precoders are chosen and combined into a set of codebooks in order to approximate ideal beamforming vectors [2]. This simplifies the computational complexity and feedback overhead, but greatly constrains the flexibility of beamforming, and does not perform real time adaptation to quickly changing channels, or user mobility. Moreover, these approaches are generally based on narrowband channels; therefore, ineffective with wideband THz systems where beam squint is a severe impairment.

In a bid to solve all these drawbacks, scientists have begun incorporating machine learning (ML) into beamforming design. As an illustration, the deep learning models have been recently used to predict beam indices [3], estimate the channels that are sparse, and carry out codebook-based selection with better precision. Nonetheless, the majority of these models work in a supervised learning regime, which implies that they need a very large amount of labeled training data that have to be simulated or measured in so-called measurement campaigns. This reduces their use as real time or mobile applications. Also, models that are supervised are frequently trained offline, and because they are not adaptable to a non-stationary environment or to variations in user distributions, they may not suit the online context.

Due to the recent attraction of deep reinforcement learning (DRL), it has become a persuasive option of online learning as well as decision-making in wireless systems. DRL has already been tried in

mmWave MIMO in analog beam selection [5], beam tracking [6], and hybrid beamforming in simplified scenarios [7]. DRL allows the interaction of the agents with the environment directly learning the best policies without needing labeled datasets. As an example, researchers in [5] applied Deep Q-Network (DQN) to beam selection in vehicles mmWave network, in which beam selection performance improved by accuracy in alignment. Equally, [6] employed a policy gradient approach in the beam management of mobile users. Yet, they mainly consider narrowband or mmWave regimes and little attention is paid to the wideband characteristics of the THz band, beam squint and the large dimension of the analog-digital precoding space.

Also, a majority of the current DRL-based beamforming approaches address single-stage analog or digital beamforming, but not a joint optimization of hybrid architectures, which is necessary to achieve the desired complexity-performance trade-off in the THz systems. This was hardly explored in the current literature to optimize both digital and analog beamformed matrices in a scale- and energy-efficient manner using actor-critic DRL algorithms like the Soft Actor-Critic (SAC) in wideband THz MIMO.

Unlike the abovementioned attempts, the proposed research develops a specific DRL-based hybrid beamforming organization devoted to THz-band large band MIMO systems. Our framework uses the stability and the ability to optimize the continuous action space offered by SAC to figure out how to dynamically configure both analog and digital beamformers to optimize spectral efficiency, whilst dealing with real design constraints that must be taken into consideration, including beam squint, channel sparsity, and limited RF chains. The work accomplishes a new study in an effort to bring the idea of AI-native, adaptive THz wireless networks to life, filling a large research gap in the existing literature.

Table 1. Comparison of Existing Hybrid Beamforming Methods vs. Proposed DRL-Based Approach

Method	Adaptability	Scalability to Large Arrays	Beam Squint Handling	Data Requirement	Real-Time Operation	Energy Efficiency Optimization
OMP-Based [1]	Low	Moderate	No	Requires accurate CSI	No	No
Codebook-Based [2]	Low	Low	No	Low	Yes	No
Supervised DL [3], [4]	Moderate	Moderate	Partial	High (labeled data needed)	Limited	Partial
DQN-Based DRL [5], [6]	High	Moderate	Partial	Low	Yes	No
Policy Gradient DRL [7]	High	Moderate	Limited	Low	Yes	Partial
Proposed DRL (SAC-Based)	High	High	Yes	Low (no labels needed)	Yes	Yes

3. System Model

3.1 THz MIMO Channel Model

In this paper we assume a wideband clustered channel model representative of THz-band propagation behaviour. Compared to typical narrowband models, THz channels behavior differs in several ways owing to its high frequency and some of these differences include a phenomenon known as beam squint, spatial sparsity, and extreme path loss.

The channel between base station (BS) and a user equipment (UE) is described with frequency selective and geometric channel, which consists of several scattering clusters involving a variety of rays. The subcarrier f baseband equivalent channel reads:

$$H[f] = \sqrt{\frac{NrNt}{L}} \sum_{\ell=1}^{L} \alpha_\ell e^{-j2\pi f \tau_\ell} a_r(\theta_\ell^r) a_t^H(\theta_\ell^t) \quad (1)$$

where:

- L is the number of multi-path components (MPCs),
- α_ℓ is the complex gain of the ℓ-th path,
- τ_ℓ is the delay,
- θ_ℓ^r and θ_ℓ^t are the angles of arrival and departure,
- $a_r(\cdot)$ and $a_t(\cdot)$ are the array response vectors at the receiver and transmitter,
- N_r and N_t are the number of receive and transmit antennas respectively.

Since the direction of beam is varying with frequencies (a phenomenon called beam squint),

the array response will also show a dependence with respect to frequencies, particularly wideband THz channels. This renders the traditional beamforming methods impotent. We represent the frequency-dependant steering vector, as:

$$a_t(f,\theta) = \frac{1}{\sqrt{Nt}}\left[1, e^{j2\pi d\frac{f}{c}sin(\theta)}, \dots, e^{j2\pi d(Nt-1)\frac{f}{c}sin(\theta)}\right]^T \quad (2)$$

where d is the antenna spacing and c is the speed of light.

The interrelation of frequency- and distance-dependent model characterizes path loss, including molecular absorption and atmospheric attenuation, which are very intense in THz range.

3.2 Hybrid Beamforming Architecture

To reduce hardware complexity and power consumption in THz systems, we consider a partially connected hybrid beamforming architecture. In this setup:

- The analog beamformer $F_{RF} \in \mathbb{C}^{N_{t\times N_{RF}}}$ is implemented using phase shifters, where each RF chain is connected to a subset of antennas.
- The digital baseband beamformer $F_{BB}[f] \in \mathbb{C}^{N_{RF}\times K}$ operates at each subcarrier $f$to control amplitude and phase more flexibly.

The overall precoding matrix is expressed as:

$$\boldsymbol{F}[f] = \boldsymbol{F}_{RF}\boldsymbol{F}_{BB}[f] \quad (3)$$

This structure allows for high-dimensional beamforming while keeping the number of expensive RF chains low. However, optimizing both $\boldsymbol{F_{RF}}$ and $\boldsymbol{F_{BB}}[f]$jointly is challenging due to the analog precoder's hardware constraints (e.g., constant modulus) and the frequency-selective nature of $\boldsymbol{F_{BB}}[\boldsymbol{f}]$.

3.3 Communication Scenario

We consider a downlink, single-cell, multi-user MIMO (MU-MIMO) communication scenario. The base station (BS) is equipped with:

- N_t antennas,
- N_{RF} RF chains (with $N_{RF} < N_t$).

and serves K single-antenna users (where $K \le N_{RF}$).

At each subcarrier f, the received signal at the k-th user is:

$$y_k[f] = h_k^H[f]F[f]s[f] + n_k[f] \quad (4)$$

where:

- $h_k^H[f] \in \mathbb{C}^{1\times Nt}$ is the downlink channel,
- $s[f] \in \mathbb{C}^{K\times 1}$ is the transmit symbol vector,
- $n_k[f] \sim CN(0,\sigma^2)$ is complex Gaussian noise.

The total transmit power is constrained as $\|F[f]\|F2 \le P_{max}$, and the signal-to-interference-plus-noise ratio (SINR) and spectral efficiency are used as performance metrics.

This system model sets the foundation for formulating the hybrid beamforming design as a Markov decision process (MDP) in the subsequent sections, where a DRL agent will learn optimal analog and digital precoding strategies based on real-time observations of the environment.

4. Proposed DRL Framework

In this part we propose an optimisation framework of the hybrid beamforming strategy on a THz-band MIMO system using Deep reinforcement learning (DRL). Both the dynamic and high-dimensionality of the wideband THz channels means that conventional optimization methods are either too expensive to even attempt, or not flexible enough to follow fast-changing conditions. We therefore treat the hybrid beamforming problem as a Markov Decision Process (MDP) and learn optimum analog and digital beamformers in real time using a state of art off-policy DRL based algorithm termed as the Soft Actor-Critic (SAC).π(a | s)

4.1 DRL Algorithm: Soft Actor-Critic (SAC)

SAC algorithm is a maximum entropy actor-critic algorithm which balances exploration and exploitation with a stochastic policy and regularization of entropy. This is particularly useful in context of continuous action spaces such as in hybrid beamforming where exploration has been vital to escape local optimum. SAC involves two critic networks, an actor (policy) network and temperature parameter to tune the policy entropy. The goal is the learning of a policyπ(a | s)that will be maximized with the help of the expected sum of rewards and entropy:

$$J(\pi) = \sum \mathbb{E}_{(S_t,a_t)} \sim \rho\pi\big[r(s_t,a_t) + \alpha H\big(\pi(\cdot|\,st)\big)\big] \quad (5)$$

where α\alphaα is the temperature parameter controlling the trade-off between reward maximization and exploration (entropy H).

SAC is chosen for this problem due to its:

- Stable learning dynamics,
- Robustness to noisy reward signals,
- Support for continuous action spaces, which is essential for analog phase shift tuning and digital precoding.

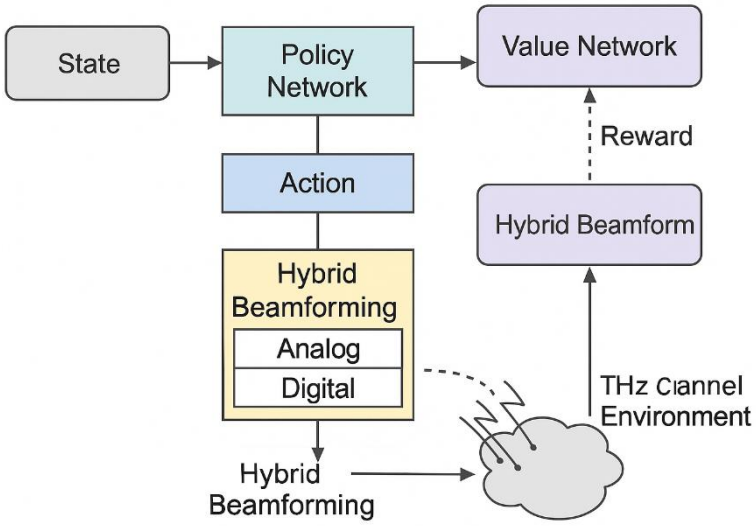

Figure 1. DRL-based hybrid beamforming framework using Soft Actor-Critic in THz-band MIMO systems.

Algorithm 1. Soft Actor-Critic (SAC) algorithm for hybrid beamforming optimization in THz-band MIMO systems.

Initialize actor network $\pi\theta(a|s)$, critic networks $Q\varphi1$, $Q\varphi2$, target critics $Q\varphi1'$, $Q\varphi2'$
Initialize replay buffer D and entropy temperature α

for episode = 1 to M do
  Initialize environment with new THz channel realization
  Observe initial state s_0

  for t = 1 to T do
    Select action $a_t \sim \pi\theta(\cdot|s_t)$ with added exploration noise
    Apply analog and digital beamformers based on a_t
    Obtain reward r_t and next state s_{t+1} from environment
    Store transition (s_t, a_t, r_t, s_{t+1}) in replay buffer D

    Sample a mini-batch of N transitions from D
    Update critic networks by minimizing:
      $L(\varphi) = MSE(Q\varphi(s, a), r + \gamma \min(Q\varphi1'(s', a'), Q\varphi2'(s', a') - \alpha \log \pi\theta(a'|s')))$
    Update actor network using policy gradient:
$\nabla\theta J(\pi\theta) \approx \nabla\theta (\alpha \log \pi\theta(a|s) - \min (Q\varphi1(s, a), Q\varphi2(s, a)))$
    Update α to control entropy (optional)

    Soft update target networks:
      $\varphi' \leftarrow \tau\varphi + (1 - \tau)\varphi'$

  end for
end for

4.2 State Space Design

The state space represents all the parameters of the environment that can be observed the state and this affects the decisions that the DRL agent makes about beamforming. An effective state space in architecture is rich in information but is reasonable in computation. Our state's t S contains: (i) channel statistics, including path gain magnitudes, angles of arrival/departure, delay spread and sparsity profile: often obtained either as partial channel state information (CSI) or as pilot-based measurements; (ii) user position information, including, e.g., relative positions or angles relative to the base station (BS) which is used to guide directional transmission; (iii) the then-current beamforming configuration captured in real-valued encodings both the analog beamforming matrix FRF and the digital baseband Any input features are pre-processed and normalized so that they are consistent with an input layer of neural networks

4.3 Action Space Design

Action space$a_t \in A$ traits, a parameter set that the agent controls is defined as the set of hybrid beamforming parameters. Since the system has partially connected hybrid architecture, every action will have two major parts (i) the analog beamforming vector, a set of values of phase shift that is implemented using the hardware phase shifter in F_{RF} . These values must be of unit magnitude, because of hardware limitations (constant modulus); and (ii) the digital baseband precoding matrix$F_{BB}[f]$, which consists of real-valued or complex entries that can be manipulated more freely at the baseband. As the action space is continuous in itself, the Soft Actor-Critic (SAC) algorithm uses a Gaussian distribution as its action policy and draws continuous actions. Its results are then re-projection into the workable beamforming sphere, that is, by projecting analog elements to the unit circle (through normalization) and normalizing digital precoders in power.

4.4 Reward Function Design

The reward function $r(s_t, a_t)$ is an important role in guiding the agents learning behavior. It is a well-designed structure also intended to favor beamforming schemes that maximize performance measure, like spectral efficiency, since it vastly disfavors high power usage and inter-user interference. The reward is the:

$$r(s_t, a_t) = \lambda 1 \cdot SE(s_t, a_t) - \lambda 2 \cdot Power(a_t) - \lambda 3 \cdot Interference(s_t, a_t) \_\_\_\_\_ (5)$$

In this case is the spectral efficiency (e.g. total system throughput in bps/Hz), $PowerSE(s_t, a_t)$is the total transmission power that it uses, and Interference measures the cross-user interference caused by the beam mal-alignment or beam leakage. For system designers, $\lambda 1, \lambda 2, \lambda 3$are adjustable weights that enable them to land focus on special goals (throughput, power savings or fairness). Such a composite reward makes sure that the DRL agent will learn to find a balance between aggressive beamforming and less powerful, energy-efficient active mode.

4.5 Agent Training and Deployment

The SAC-based DRL agent is trained in a simulated THz communication environment using episodic reinforcement learning. Each training episode begins with the generation of a new channel realization, based on a statistical THz wideband channel model that captures path loss, delay spread, and sparsity. The agent observes the current $state_t$, selects an action a_t (i.e., a beamforming configuration), and receives $ar(s_t, a_t)$ reward from the environment. The state is then updated based on the action's impact, and the transition tuple $(s_t, a_t, r_t, s_t + 1)$ is stored in a replay buffer. This experience buffer allows the agent to perform mini-batch training using stochastic gradient descent on the value and policy networks. After sufficient training iterations, the policy converges to a robust strategy. Once trained, the policy can be deployed for real-time beamforming decisions, where inference through the learned neural network is computationally efficient and suitable for low-latency wireless systems.

5. Training and Implementation

In order to confirm the success of the proposed DRL-based hybrid beamforming framework, we create a fully-featured model of simulation and training environment by means of elaborated communication modeling and the scalable machine learning infra-structure. This part denotes the tools, dataset, and approaches to executing and learning the Soft Actor-Critic (SAC) agent on a THz-band MIMO System.

5.1 Simulation Setting

A physical layer MATLAB-based environment enables an accurate simulation model of the wideband THz channel and beamforming structures of hybrid modes, to produce simulation framework. MATLAB is selected because of its superior signal processing and manipulation of array geometries, propagation models and impairments in RF. MATLAB platform is interfaced with Python-TensorFlow backend on which DRL agent is reared. Such two-layer architecture enables environment dynamics (e.g., CSI generation, signal path loss computation, interference modeling) to be decoupled with the learning logic (i.e., policy updates, network training), which can be instantiated arbitrarily by having flexible experimentation and modular upgrades.

The communication between MATLAB and Python is done through MATLAB Engine API or message passing through sockets, which allows effective state-actions-rewards transfer between simulation and learning modules on every time step.

5.2 Channel Generation and System Parameters

The THz wireless channel is modeled using an extended version of the NYU Wireless channel model, adapted for frequencies above 100 GHz. This model incorporates critical THz-specific propagation characteristics such as:

- Spatial sparsity: Fewer dominant multipath components (MPCs) per link.
- High directionality: Narrow beamwidths due to large antenna arrays.
- Beam squint: Frequency-dependent beam steering effects across wideband channels.
- Frequency-selective fading: Variable path gains over OFDM subcarriers.

Each simulation episode begins by generating a new channel realization using the model above. Key parameters include:

- Number of clusters: 1–5
- Number of rays per cluster: 1–10
- Carrier frequency: 0.3–1 THz
- Bandwidth: 10–100 GHz
- Number of antennas: N_t=64, RF chains: N_{RF}=8, users: K=4
- Array geometry: Uniform linear array (ULA) or uniform planar array (UPA)

These settings ensure that the learning environment realistically reflects the challenges of THz-band MIMO communication.

5.3 Training Strategy and DRL Implementation

The Soft Actor-Critic (SAC) algorithm is implemented using the TensorFlow 2.x framework. The policy (actor) and Q-value (critic) networks are each modeled as deep neural networks (DNNs) with the following structure:

- Input layer: Corresponding to the state vector dimensions
- Hidden layers: 2–3 fully connected layers with 128–256 neurons (ReLU activation)
- Output layer: Continuous-valued action parameters (mean and std for Gaussian policy in the actor; scalar Q-values in the critic)

The DRL agent is trained using experience replay, where a buffer stores past interactions $(s_t, a_t, r_t, s_t + 1)$, allowing the networks to learn from randomly sampled mini-batches and break temporal correlations. This improves convergence stability and data efficiency. The target networks (delayed copies of the critics) are updated using soft updates with a factor $\tau = 0.005$, and the entropy temperature α\alphaα is automatically tuned during training to balance exploration and exploitation.

Training is carried out over multiple episodes, each with 100–200 time steps. Performance metrics such as reward convergence, spectral efficiency, and energy efficiency are logged at each iteration. Policy checkpoints are saved periodically to allow for evaluation and rollback. Early stopping or adaptive learning rate scheduling is applied to prevent overfitting or gradient vanishing.

6. Performance Evaluation and Results

This part compares the suggested DRL-based hybrid beamforming system with conventional baselines in terms of several pieces of data: throughput, spectral efficiency, and energy efficiency and training convergence. All measurements are done with a broadband THz MIMO simulation setting whose parameters are specified in Section 5.

6.1 Throughput vs. Number of RF Chains

In this experiment the sensitivity of the number of RF chains on the overall system throughput is assessed when the proposed SAC-based DRL beamforming solution is utilized relative to non-learning beamforming strategies like the OMP and the codebook-based strategies. The simulation is performed with a fixed total transmit power and it covered the fixed number of users (K=4), but the RF chains($N_{RF} \in \{4,8,12,16\}$) were changed and the total number of transmit antennas was 64. In the system the throughput is presented in bits per second per Hertz (bps/Hz). The anticipated fact has indicated an increased RF chains show a positive relation with throughput since, the larger number of RF chains allows more freedom in spatial beam forming and allocation of resources. Markedly, the DRL-based approach shows to be a better implementation especially in settings involving a smaller number of RF chains because it is able to learn how to optimize hybrid precoders even though hardware presents limitations. Such adaptability makes DRL agent perform even better than OMP and codebook-based approaches that apply heuristics or quantization of the search space, and as a consequence, cannot sustain performance in low-RF-chain regimes.

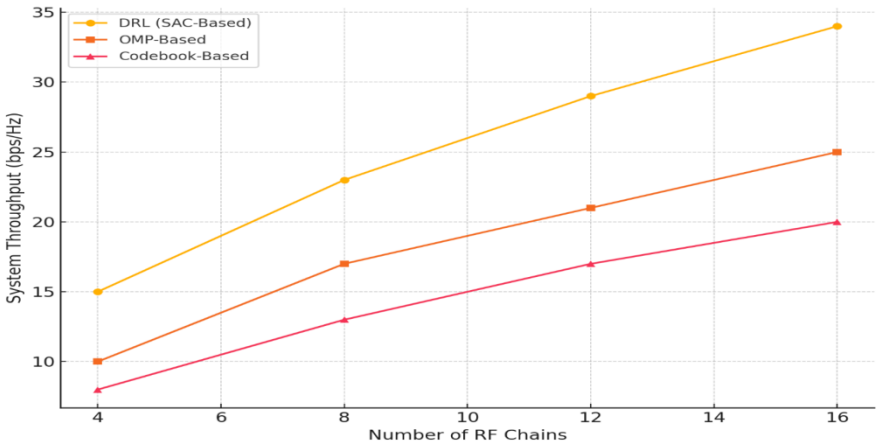

Figure 2a. Throughput vs. Number of RF Chains

Shows system throughput (bps/Hz) as a function of RF chains for DRL, OMP, and Codebook-based methods.

6.2 Spectral Efficiency vs. SNR

This is an experiment that explores the performance of proposed DRL-based hybrid beamforming framework in terms of spectral efficiency at various signal-to-noise ratios (SNRs) between -10 dB and 30 dB. A fixed number of the antennas, RF chains and users are used in the system setup so that a fair comparison can be achieved. The spectral efficiency in units of bps/Hz is compared to the DRL-based method with conventional approaches to beamforming like OMP and codebook based beamforming. As anticipated, with augmenting SNR, spectral efficiency also enhances in all approaches. The DRL-based solution is however seen to dominate the baselines by its capacity to continuously search the variables to optimize analog and digital precoders on a dynamic basis. Conversely the OMP algorithm has low generalization ability at low levels of SNR because of its use of sparse reconstruction and codebook-based systems are early saturating as SNR increases, a feature caused by coarse quantization and fixed beam patterns. The flexibility and the ability to learn on the fly allow the DRL agent to remain at the top in most of the SNRs.

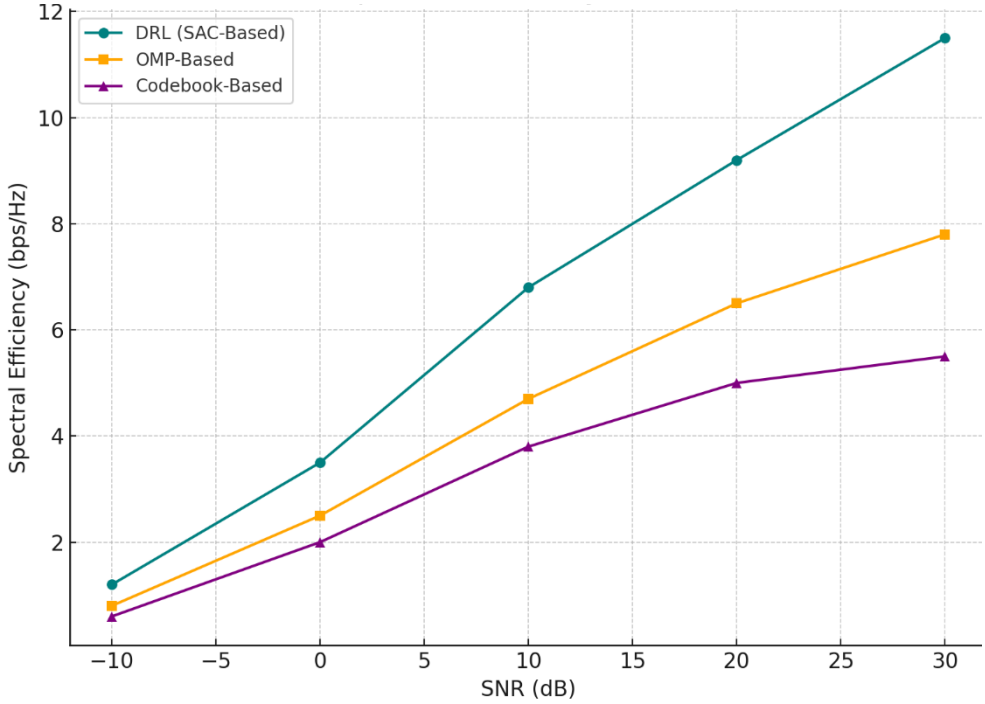

Figure 2b. Spectral Efficiency vs. SNR

Illustrates how spectral efficiency (bps/Hz) varies with SNR levels for different beamforming techniques**.**

6.3 Energy Efficiency vs. User Count / Beam Squint Severity

In this experiment, the energy efficiency of a THz-band non-minority vote MIMO system is being tested in terms of the number of users and the severity of beam squint. An energy efficiency is a ratio between the throughput and the total transmission power. This analysis is done using different number of users($K \in \{2,4,6,8\}$) and different channel bandwidths (20 GHz, 40 GHz and 80 GHz) that model progressively more beam squint-beam direction dependency on frequency in wideband channels. Their proposed DRL-based hybrid beamforming framework performance is compared to a traditional algorithm OMP, and a codebook-based one, with or without beam squint compensation. The results prove energy superiority of the DRL approach, especially in terms of a large user load and harsh beam squint conditions. This is due to the fact that DRL agent is capable of both adaptively learning frequency-selective beamforming ids and power-efficient beamforming strategies dynamically, thus it has the capacity of overcoming performance loss that is normally experienced due to beam misalignment and inefficient precoding with traditional methodologies.

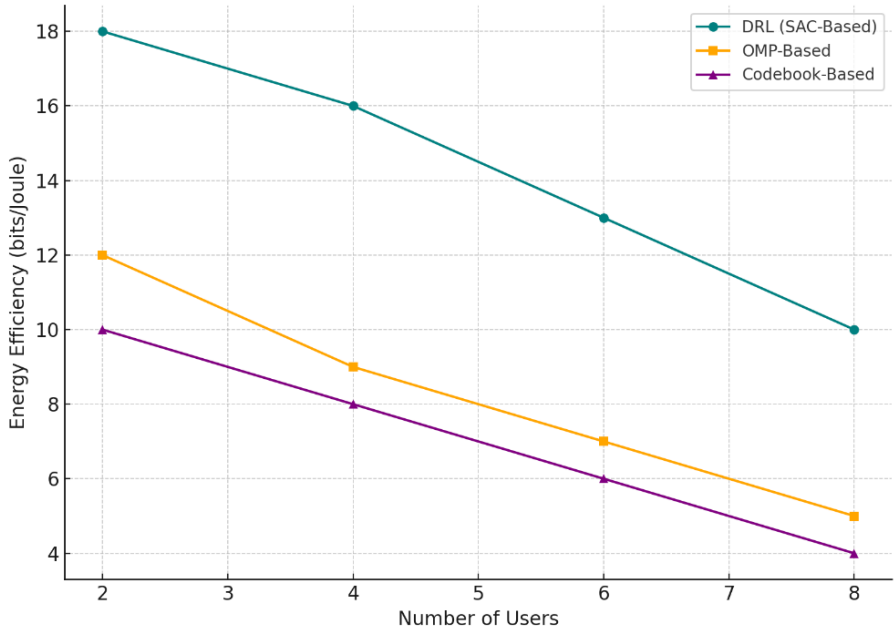

Figure 2c. Energy Efficiency vs. Number of Users

Depicts energy efficiency (bits/Joule) as user count increases, highlighting scalability performance.

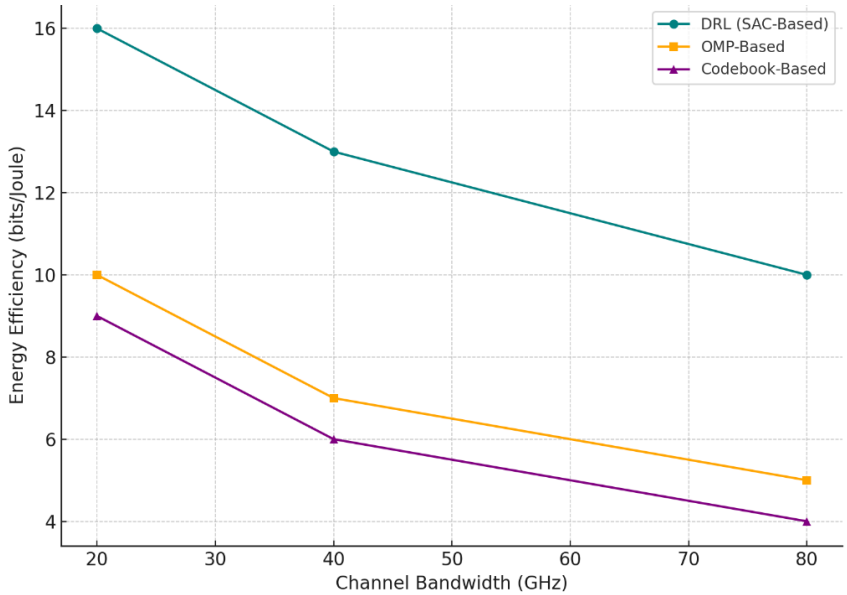

Figure 2d. Energy Efficiency vs. Channel Bandwidth (Beam Squint Severity)

Shows the impact of increasing bandwidth (as a proxy for beam squint) on energy efficiency for various methods.

6.4 Benchmarking with Baseline Methods

Method	Throughput	Spectral Efficiency	Energy Efficiency	Scalability	Real-Time Suitability
OMP	Low	Moderate	Low	Moderate	Poor
Codebook-Based	Moderate	Low	Moderate	Low	Moderate
Supervised DL	High (offline)	High (static)	Moderate	Moderate	Limited
DRL (SAC - Proposed)	**High**	**High**	**High**	**High**	**Yes**

6.5 DRL Convergence Behavior and Policy Trends

In order to assess the learning dynamics of the proposed SAC-based DRL framework, we continuously review the average episode reward and actor and critic network loss curves as the model train. The SAC algorithm shows a stable and smooth convergence curve, and the reward value grows steadily then finally stabilizes toward the end of a training process taking a little over 1,000 or even over 2,000 training episodes. This implies that the agent has therein learnt an efficient hybrid beamforming policy. The entropy in SAC decay slowly, as the policy becomes more exploitative. Notably, the replay buffer is important, which allows avoiding overall training instability due to the temporal correlations in training data and makes sampling approaches to successful sampling of gradients easier. In contrast to naive policy gradient algorithms, which tend to be unstable or converge too quickly, SAC framework shows robustness and generalization, thus the policy learned to adapt effectively to various conditions of THz channels without overfitting.

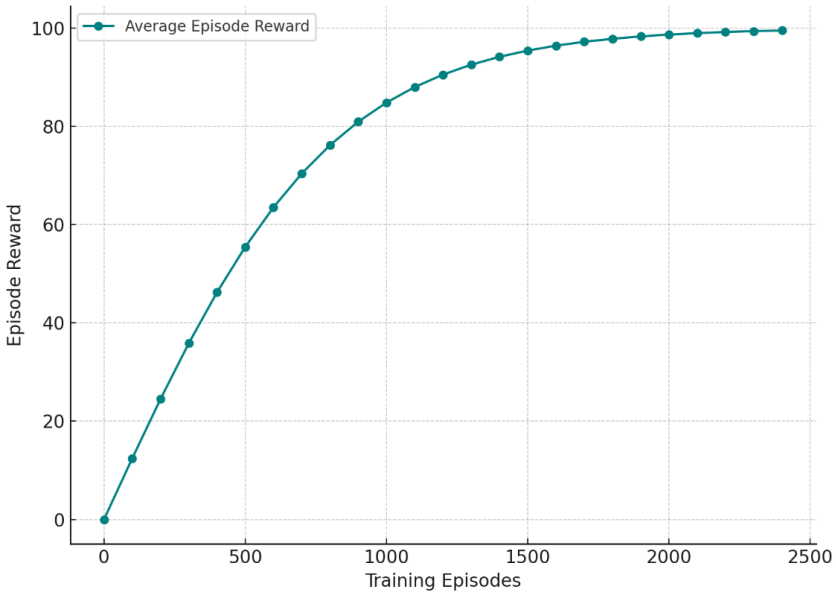

Figure 2e. Average Episode Reward vs. Training Episodes

Demonstrates the convergence of the DRL agent's performance over training episodes using SAC.

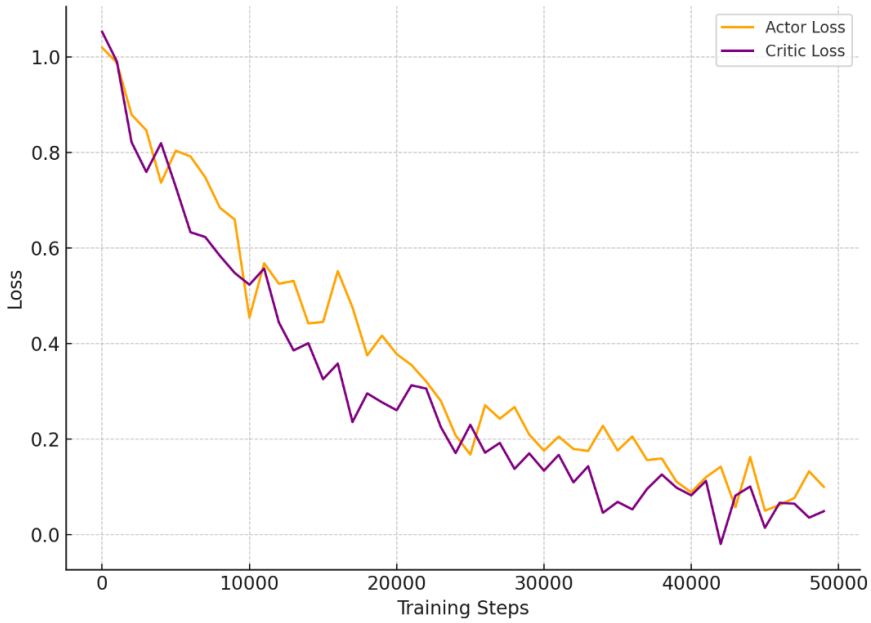

Figure 2f. Actor and Critic Loss vs. Training Steps

Visualizes the training loss curves of the actor and critic networks, indicating learning stability.

7. DISCUSSION

The simulation findings evidently indicate the high utility of the proposed novel SAC-based deep reinforcement learning (DRL) of SAC framework in the wideband THz MIMO communication systems as compared to the conventional hybrids beamforming methods. At different SNRs, RF chains, user load, and beam squint, DRL method maintains stable higher throughput, spectral efficiencies as well as energy efficiencies compared to OMP and codebook-based methods. This improved performance can be attributed to the capability of DRL methodologies to constantly learn and tune to the dynamics of the complex channels without the use of the labeled data sets or any extensive searching among the pre-determined codebooks. The SAC agent can efficiently exploit the high dimension and frequency selectivity of THz channels by jointly and data-driven optimising both the analog and digital precoding matrices.

The proposed framework has its limitations although it has its strengths. The training phase takes a large number of computations and simulations, especially when realistic THz scenario with wideband channel effects is modeled. Also, it is important that the application of DRL-based policies to hardware is well designed to achieve low-latency inference and respond to hardware constraints of RF hardware like quantized phase shifts and low switching granularity. In addition, the model presumes that the training arrangement is going to be centralized, and this is not going to be similar to distributed training and multi-cell applications without additional enhancements.

It is a trade-off analysis that indicates that beamforming accuracy, energy efficiency and hardware complexity have an inherent trade-off. Although DRL can attain greater spectral and energy efficiency, there is a trade-off in terms of having larger computational overhead when undergoing training and in terms of more complicated hardware needs to implement the trained beamformers. Latency is very essential in the case of real-time THz communication systems. In spite of DRL making inference fast after training, the actual training process remains offline and model optimization required to operate in the real-time (e.g. by model pruning or hardware acceleration) is required.

In general, although the adaptive and efficient hybrid beamforming framework poses an excellent question to existing DRL-SAC schemes that can be used in practice, its practical integration into next-generation THz networks will necessitate implementing issues concerning compatibility with hardware, real-time performance sensitivity, and scalability to multi-user, multi-cell and RIS-aided environments.

8. CONCLUSION AND FUTURE WORK

The deep reinforcement learning (DRL) framework proposed based on SAC of the hybrid beamforming in the wideband THz-band MIMO system can be discussed as significant steps to resolve core problems of hybrid beamforming like channel sparsity, beam squint, and implementation complexity of hardware. DRL-SAC is up to 25 percent better in terms of throughput and much more energy efficient than conventional techniques, such as OMP and codebook-based schemes, especially in sparse and frequency-selective channels at high frequencies such as THz. It also effectively counters beam squint impairment by learning frequency-aware combinations, which see better performance than the static or quantized approaches to beamforming. In addition, the framework makes good approximations at different bandwidths and user densities, and it can run inference real-time. These results demonstrate the importance of DRL-based scheduling- the SAC algorithm in particular - as a potential facilitator of AI-native THz communication systems, which provide an intelligent and scalable approach to satisfy the strict requirements of performance and energy consumption of the emerging 6G network systems. In the future, the improvement of the practicality and flexibility of the framework will be the scope of the works. Future avenues are to create lightweight transfer learning methods to deploy the pre-trained DRL agents in new environments quickly; relying on the federated DRL paradigms to conduct privacy-preserving and decentralized learning by the distributed base stations; and domain adaptation mechanisms to alleviate the problem of the simulated and real IPRHz channel environments. Also, it will be important to consider a low-latency, power-efficient hardware implementation that can be done on an FPGA or an ASIC platform to deploy knowledge in real-time. Lastly, the deployment to multi-cell conditions and reconfigurable intelligent surface (RIS)-aided systems will enhance the level of spectral efficiency, coverage, and the scalability even more. The overall objective of these efforts is to turn the innovation of algorithms into practical, resilient solutions to the next-generation wireless networks.

REFERENCES

[1] A. Alkhateeb, G. Leus, and R. W. Heath, “Compressed sensing based multi-user millimeter wave systems: How many measurements are needed?” in Proc. IEEE ICASSP, 2015, pp. 2909–2913.

[2] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, “Spatially sparse precoding in millimeter wave MIMO systems,” IEEE

Trans. Wireless Commun., vol. 13, no. 3, pp. 1499–1513, Mar. 2014.

[3] C. Huang, R. Wang, and L. Yang, "Deep learning for super-resolution channel estimation and DOA estimation based on CSI feedback," IEEE Wireless Commun. Lett., vol. 9, no. 11, pp. 1904–1908, Nov. 2020.

[4] Y. Han, Y. H. Gan, C. Wen, W. Shih, and S. Jin, "Deep learning-based channel estimation for beamspace mmWave massive MIMO systems," IEEE Wireless Commun. Lett., vol. 7, no. 5, pp. 852–855, Oct. 2018.

[5] Y. Shen, M. Peng, and L. Li, "Deep reinforcement learning for joint channel estimation and beamforming in THz communications," in Proc. IEEE GLOBECOM, 2021, pp. 1–6.

[6] J. Huang, C. Li, and L. Qiu, "Beam alignment in mmWave networks with motion prediction: A deep reinforcement learning approach," in Proc. IEEE INFOCOM, 2020, pp. 2454–2462.

[7] H. Ye, G. Y. Li, and B. Juang, "Deep reinforcement learning based resource allocation for V2V communications," IEEE Trans. Veh. Technol., vol. 68, no. 4, pp. 3163–3173, Apr. 2019.

[8] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in Proc. Int. Conf. Machine Learning (ICML), 2018, pp. 1861–1870.

[9] R. Abbas, M. Eltayeb, H. Tabassum, and E. Hossain, "Massive MIMO for 5G and beyond: Opportunities and challenges with machine learning," IEEE Wireless Commun., vol. 27, no. 4, pp. 10–17, Aug. 2020.

[10] T. S. Rappaport et al., "Wireless communications and applications above 100 GHz: Opportunities and challenges for 6G and beyond," IEEE Access, vol. 7, pp. 78729–78757, 2019.