CHALMERS
UNIVERSITY OF TECHNOLOGY

DRL-Assisted QoT-Aware Service Provisioning in Multi-Band Elastic Optical Networks

Citation for the original published paper (version of record):
Teng, Y., Natalino Da Silva, C., Arpanaei, F. et al (2025). DRL-Assisted QoT-Aware Service Provisioning in Multi-Band Elastic Optical Networks. Journal of Lightwave Technology, 43(19): 9090-9101. http://dx.doi.org/10.1109/JLT.2025.3601402

DRL-assisted QoT-aware Service Provisioning in Multi-band Elastic Optical Networks

Yiran Teng, Carlos Natalino, Farhad Arpanaei, Haiyuan Li,
Alfonso Sánchez-Macián, Paolo Monti, Shuangyi Yan*, Dimitra Simeonidou

***Abstract*—Multi-band (MB) optical transmission is a promising solution to support the ever-increasing network capacity demand of 5G/6G applications. By exploiting extra optical spectrum beyond the C- and L-bands, such as the L+C+S-band, the network can use up to 20 THz, quadrupling the original capacity of the C-band. The extensive spectrum resources and complex physical layer interactions in MB systems present challenges for traditional resource management solutions that are evaluated only for the C-band. Effective algorithms tailored for MB optical networks are needed to enable optical networks to provision services efficiently, thereby reducing service blocking and improving network throughput. In this study, we propose a deep reinforcement learning (DRL)-assisted framework for dynamic service provisioning in MB elastic optical networks. The proposed DRL framework aims to minimize long-term bit-rate blocking and includes several innovations. First, an accurate quality of transmission estimation model is employed to profile the performance of the supported modulation formats for each channel on pre-computed routes. Within the DRL agent design, a novel state representation incorporating both route-level and band-level features is designed to enhance the DRL agent's ability to perceive the network conditions. Moreover, a new reward function has been developed to enhance performance and accelerate convergence. Simulations are performed using a number of L+C+S MB systems with and without traffic grooming support. The results indicate that the proposed DRL-assisted framework can reduce bit rate blocking by an average of 35% to 85% compared to the existing heuristic methods from the literature while maintaining an appropriate inference time.**

***Index Terms*—deep reinforcement learning, quality of transmission, multi-band network, elastic optical networks, Gaussian noise model**

This is the authors' version of the paper. The final published version is available at 10.1109/JLT.2025.3601402.

Y. Teng, H. Li, S. Yan, and D. Simeonidou are with High Performance Networks Group, Smart Internet Lab, University of Bristol, Bristol, United Kingdom (e-mail: {ab20471, ocean.h.li, Shuangyi.Yan, Dimitra.Simeonidou}@bristol.ac.uk).

C. Natalino and P. Monti are with Dept. of Electrical Engineering, Chalmers University of Technology, 412 96 Gothenburg, Sweden (e-mail: {carlos.natalino, mpaolo}@chalmers.se).

F. Arpanaei and A. Sánchez-Macián are with Dept. of Telematic Engineering, Universidad Carlos III de Madrid, 28911 Leganes, Madrid, Spain (e-mail: farhad.arpanaei@uc3m.es, alfonsan@it.uc3m.es).

The authors acknowledge the support of EU-funded projects – Allegro (No. 101092766) and ECO-eNET (No. 10113933.) – and the UK EPSRC project HASC (No. EP/X040569/1). The UC3M authors would like to acknowledge the support of the Spanish funded Fun4date-Redes project (grant No.PID2022-136684OBC21). Y. Teng acknowledges the support from the China Scholarship Council (CSC) /University of Bristol joint-funded scholarships program (grant No. 202208410050).

I. Introduction

With the rapid development of emerging applications such as ultrahigh-definition video streaming and augmented/virtual reality (AR/VR), the amount of data traffic injected into transport networks is growing drastically [1]. Traditional single-band wavelength division multiplexing (WDM) optical networks operating primarily in the C- and L-bands can no longer meet such high data rate requirements. In recent years, elastic optical networks (EONs) [2] equipped with the bandwidth/bit-rate variable transceiver were introduced to enhance spectrum efficiency and allow multiple modulation formats (MFs) for each channel. However, the resources of the 12 THz C + L band spectrum are still insufficient to accommodate the network capacity demands driven by next-generation 6G applications [3]. In this context, multi-band (MB) transmission has emerged as a cost-effective solution, offering substantial capacity by leveraging multiple wavelength bands (e.g., L+C+S-band) to expand available spectrum resources [3]. In particular, recent advances in amplifiers and transceivers have enabled signal transmission throughout a wider spectral range [4]. Upgrading current core networks to be compatible with MB systems requires the expansion and utilization of multiple bands in existing single-core single-mode fibers, along with the installation of new amplifiers corresponding to the extended band [3], without the need to deploy additional fibers [5], [6]. Therefore, MB-EONs can serve as an effective and promising solution for addressing medium-term capacity limitations [7].

Designing effective resource management solutions for MB-EONs (e.g., for dynamic service provisioning), which offer a vast amount of optical spectrum resources, is essential to reduce the service blocking and enhance the network throughput. The provisioning of services in MB-EONs presents significant challenges, including routing, modulation format selection, band allocation, and spectrum assignment (RMBSA) issues. These complexities surpass those encountered in the traditional routing, modulation format, and spectrum allocation (RMSA) problem that is limited to the C-band [8]. First, physical layer impairments in the MB scenario are complex due to the wide active spectrum range, introducing additional non-linear interference (NLI) noise such as the inter-channel stimulated Raman scattering (ISRS) and the Kerr effect [9]. In addition, the different configurations of amplifiers for each band must be considered. Erbium-doped fiber amplifiers (EDFAs) are used for the C- and L-bands. In contrast, Thulium-doped fiber amplifiers (TDFAs) are used for the S-band [3], leading

to variations in amplified spontaneous emission (ASE) noise levels across different bands. These noises result in a degraded signal-to-noise ratio (SNR) of the lightpaths. To guarantee the quality of transmission (QoT) required by the services, we must ensure that the SNR of all established lightpaths remains within the receiver's sensitivity and the QoT threshold for the selected MF. Therefore, a precise QoT/SNR estimation model is necessary. Second, effective spectrum resource management needs to account for several aspects. The extensive spectrum resource across multiple bands can result in severe channel fragmentation, which reduces network utilization and overall throughput, requiring RMBSA algorithms to be upgraded with fragmentation concerns. Additionally, distinct NLI and ASE noise levels on each band cause channel SNR levels to vary significantly across bands [10]. This results in different available MFs of the channels across different bands for a certain path (e.g., channels in C- and L-band can support higher-level MF than in S-band) [3]. Therefore, the designed RMBSA algorithm needs to allocate services to bands according to the specific performance of each band. Finally, taking decisions on service provisioning affects the network's long-term resource availability, influencing future requests. Recognizing this impact helps the RMBSA algorithm develop an optimized provisioning strategy, enhancing overall network performance in the long run.

From the literature, most dynamic service provisioning solutions in MB-EONs focus on reducing the long-term network blocking probability (BP) while ensuring QoT for services by utilizing heuristic algorithms [11]–[15]. Some heuristics achieved satisfactory performance with moderate computational complexity. However, these fixed, human-designed rules (e.g., shortest path routing, L-band preference, maximum generalized signal-to-noise ratio (GSNR), and lowest maximum frequency slot index) lack the flexibility to adapt to varying network conditions and are not able to accurately assess the long-term impact of current provisioning decisions on future requests and network state. In recent years, the outstanding performance of deep reinforcement learning (DRL) in complex control tasks [16] has inspired growing interest in its application in optical networks. DRL has shown great potential in service provisioning and spectrum management within single-band EONs [17]–[21]. DRL employs deep neural networks (DNNs) to extract essential features from complex environments, effectively managing continuous observation spaces and adapting well across various network conditions. Additionally, the target of DRL algorithms is the maximization of a cumulative reward, making it suitable for solving long-term optimization problems. These properties make DRL intuitively more suitable than heuristics for dynamic service provisioning. However, the effectiveness of DRL in the MB scenario remains unproven. For example, DRL has yet to outperform basic heuristics like the k-shortest-path first-band first-fit (KSP-FB-FF) [8], [22], [23], or only slightly outperform KSP-FB-FF in the C+L-band scenario [24]. These suboptimal results come from a DRL design that simply extends the DRL-based model designed for single-band scenarios [17] without considering some critical features in MB scenarios such as fragmentation and band utilization [25]. Additionally, as the number of bands increases, achieving a high-performance DRL model becomes more challenging [8]. For instance, the expansion of the observation/action space significantly affects the capability of the DRL agent to learn effective policies. Finally, state-of-the-art solutions utilize a distance-adaptive approach to calculate channel impairments, assuming that all channels within a band have the same available MF. This method fails to accurately consider the effects of NLI and ASE in MB systems and oversimplifies the RMBSA problem.

In this paper, we extend the preliminary work [26], introducing a novel DRL-assisted QoT-aware RMBSA framework for EONs operating in the L+C+S band. This framework includes several innovative features. First, it adopts a QoT estimator based on the enhanced generalized Gaussian noise (EGGN) model to derive the channel-connection profile for all route/band/channel alternatives with optimum launch power for each span. Consequently, the bit rate/MF for each lightpath is contingent upon the corresponding channel QoT value. Second, the design of the DRL agent leverages a novel fragmentation-aware and band-usage-aware state representation, enhancing the agent's ability to effectively capture essential MB-EONs information. Third, action masking is applied to filter actions lacking sufficient lightpath capacity, ensuring that the DRL agent decision space only contains valid actions. Moreover, a new reward function is proposed, and shows advantages in guiding the agent towards efficient exploration. The performance of the proposed DRL-based framework is evaluated across various topologies of different scales, considering both grooming-available and non-grooming scenarios. Simulation results demonstrate that our solution outperforms simple heuristics in all scenarios with around 85% blocking reduction and surpasses specific advanced heuristics in most cases with around 35% blocking reduction. To the best of our knowledge, this is the first DRL-assisted RMBSA approach to utilize the EGGN model for accurate QoT estimation in MB-EONs with optimum power profile, and the first DRL-based solution to outperform advanced heuristics.

The rest of the paper is organized as follows. Section II summarizes the current dynamic service provisioning solutions in the literature. Section III illustrates the physical layer model and RMBSA formulation in MB-EONs. Section IV introduces the proposed DRL-assisted RMBSA framework. The evaluation results of the proposed solution are presented and analyzed in Section V. Finally, Section VI concludes the paper.

II. Related Work

The dynamic RMBSA problem is usually divided into several sub-problems, which are addressed primarily through heuristics [11]–[15] and/or machine learning (ML)-based methods [8], [22], [23], [27]–[29].

A. Heuristics

Heuristics are the most investigated RMBSA solutions in the literature. Sambo *et al.* adopted the generalized Gaussian noise (GGN) model using GNPy [30] to evaluate the GSNR of lightpath requests, and investigated the impact of different

band allocation schemes (e.g., L band preferred) on the overall BP in multiple MB scenarios [11]. The authors in [12] proposed an ISRS-aware RMBSA in C+L-band that considers the time dimension to mitigate the fragmentation while satisfying the SNR of each established lightpath. Yao *et al.* presented a SNR re-verification-based RMBSA, which checks the impact of the new requests on the SNR of the existing lightpaths [13]. This method always tries to select lower-order MF with higher SNR threshold to satisfy the SNR verification. In [14], Mehrabi *et al.* proposed a low-margin QoT-aware RMBSA solution to achieve higher spectral efficiency. This solution firstly determines the highest-order MF under the SNR constraint and chooses the path with the lowest maximum FS index. The authors in [15] classified the request into multiple categories based on their transmission distance and bit rate requirement and set different band allocation priorities for each category. The extensive resources with diverse MFs make the network environment increasingly complex, posing challenges for heuristics to optimize across all variables. Moreover, rule-based heuristics lack a long-term feedback that may limit its adaptability to changing network conditions.

B. ML-based RMBSA

Among the AI/ML-assisted solutions, reinforcement learning (RL) has been reported as a viable technique for reducing BP by using the Q-table-based methods for service provisioning [27], [28]. However, Q-table suffers from scalability issues in complex and/or continuous observation-action spaces such as those experienced in the RMBSA problem, and lacks generalization abilities [31]. To this end, DRL-assisted RMBSA algorithms that use DNNs to effectively handle continuous input and enhance generalization abilities have been investigated [8], [22], [23], [29]. Sheikh *et al.* adopted the deep Q-network (DQN) model on a DRL agent trained with path-level state representation and simple reward function (+1/-1) [22]. Gonzalez *et al.* further improved the DRL model by introducing multiple advanced reward functions that account for spectrum compactness and fragmentation [23]. However, the resulting DRL agents did not outperform KSP-FB-FF, highlighting the need to further improve the DRL agents [8]. Dan *et al.* extended the DRL model proposed in [17] to the C+L band scenario, achieving 11% gains in reducing the request BP with compared to the KSP-FB-FF [24]. The authors in [29] proposed the closest work so far to our proposal. Their DRL-assisted algorithm adopts a GNPy-based QoT estimation that inserts the path and band available resources into the observation. For the reward, [29] considers the number of hops and the total number of assigned channels in the selected path, i.e., solutions with lower number of hops and occupying less spectrum will receive higher reward. The evaluated scenario uses the Japanese topology with 14 nodes. The services always request 400 Gbps that can be accommodated using two MFs, i.e., DP-16QAM and DP-QPSK. This method represented a state-of-the-art DRL-based algorithm that demonstrates superior performance when compared with the KSP-FB-FF, with 80% average gains in terms of request BP. However, current DRL-assisted service provisioning solutions in MB-EONs are largely inspired by the DeepRMSA framework which was designed for single-band EONs [17]. A critical aspect is the state representation that lacks band-level features which are essential in MB-EONs. Moreover, previous works fail to address the impact of the increased number of actions compared to the single-band scenario, and the fact that some of these actions may become infeasible. Compared to the closest work in the literature [29], our work includes multiple crucial band features in the observation space and path usage information in the reward function, which further improve the performance of the system. Moreover, we provide an in-depth performance analysis over three topologies. A recent study [32] evaluates various DRL-based resource management solutions for optical networks from the literature and concludes that DRL can provide greater benefit than heuristics in more complex network scenarios. By extrapolating their conclusions from a single-band environment, we believe there is potential in applying DRL in MB networks.

III. Problem Statement and Network Model

In this study, the RMBSA solution focuses on minimizing the long-term BP of MB-EONs while guaranteeing the QoT for each service. In this section, the RMBSA problem is introduced and formulated as a partially-observable Markov decision process (POMDP), and the physical layer model for evaluating the GSNR/QoT in MB-EONs is elaborated.

A. Dynamic RMBSA Problem in MB-EONs

Let $G(V, E)$ denote the topology of a MB-EON, where V, E represent the set of nodes and bidirectional links, respectively. The optical spectrum on each link is divided into multiple bands, and each band contains multiple channels with fixed bandwidth. We consider a set of services $\mathcal{R}$ composed of dynamic service requests $R_t(s, d, b)$ which arrive one at a time. Upon receiving a service request $R_t(s, d, b)$, the dynamic RMBSA problem needs to be solved by a policy. The policy needs to select one of a path among K pre-computed candidate paths between source node s and destination node d, and select a band on the path to allocate enough channels to satisfy the bit rate requirement b. In MB-EONs, the spectral continuity constraint must be considered, requiring that the channels assigned to a lightpath use the same spectrum across all links along the path.

The capacity provided by each channel depends on the adopted MF. Six MFs from DP-BPSK to DP-64QAM at 64 Gbaud are considered. The modulation level m for each MF is from 1 to 6, corresponding to bit rates from 100 to 600 Gbps, respectively. The available MFs for each channel on each candidate path are pre-computed based on the GSNR tolerance described in detail in Sec. III-C. Then, the highest-order available MF is assigned, provided that it does not result in capacity waste (e.g., using DP-64QAM with a 600 Gbps capacity to provision a 100 Gbps request); otherwise, the lowest-order MF that supports the requested bit rate is chosen. We assume that a lightpath can be supported by multiple channels/line card interfaces using different MFs [33]. In this study, two different spectrum assignment scenarios

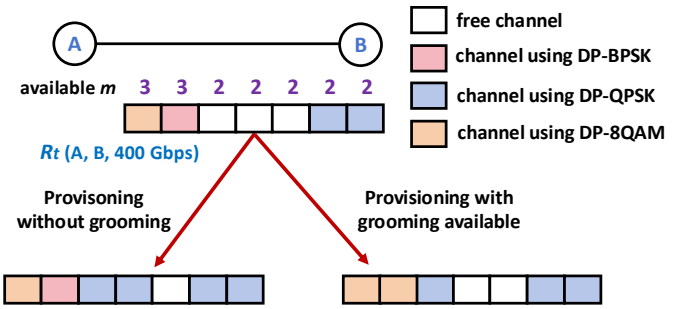


Fig. 1: RMBSA in MB-EONs with/without grooming.

are discussed based on whether the support for spectrum grooming in each channel is available. As shown in Fig. 1, the channel with unused capacity (e.g., second channel using DP-BPSK) cannot be used by other services when grooming is not allowed. In contrast, when grooming is available, the MF of channel with unused capacity can be upgraded to a higher order to accommodate other services using the same routing path as the current channels. For instance, in Fig. 1, the DP-BPSK used for the second channel is upgraded to DP-8QAM to provide additional 200 Gbps for the R_t so that one new channel is saved compared to the case where grooming is not available. Moreover, the grooming capability can reduce the number of line card interfaces and spectrum usage, thereby decreasing overall bandwidth utilization. Both the two scenarios are considered when we design the algorithms.

B. POMDP Formulation of the Dynamic RMBSA Problem

Under the traditional RL/DRL framework, the dynamic RMBSA process can be modeled as a partially-observable Markov Decision Process (POMDP) [34], characterized by the tuple $\langle s_t, a_t, T, r_t, \gamma \rangle$ at each step t. Here, the state s_t represents the observation of the RMBSA environment, and the action a_t denotes the RMBSA decision. The transition function $T(s_{t+1}|s_t, a_t)$ describes how the environment evolves after executing action a_t. The reward r_t provides feedback based on the outcome of the action a_t, and γ is the discount factor, with a value in the range $[0, 1)$. The goal of the DRL-based RMBSA algorithm within the POMDP framework is to find a policy $\pi(a_t|s_t)$ – the probability distribution of action a_t for each state s_t –that maximizes the cumulative reward (i.e., return) G_t, which is calculated as:

$$G_t = \sum_{i=t}^{\infty} \gamma^{i-t} r_i \tag{1}$$

C. Physical Layer Model

According to the incoherent Gaussian noise (GN) model for uncompensated optical transmission links [35], the GSNR for a lightpath on channel i can be derived from (2) as follows:

$$GSNR^{i}_{\text{LP}}|_{\text{dB}} = 10 \cdot \log_{10} \left[\left(\sigma_{\text{ASE}} + \sigma_{\text{NLI}} + \sigma^{-1}_{\text{TRx}} \right)^{-1} \right] - \sigma_{\text{Flt}}|_{\text{dB}} - \sigma_{\text{Ag}}|_{\text{dB}}, \tag{2}$$

where σ_{ASE} and σ_{NLI} can be obtained by:

$$\sigma_{\text{ASE}} = \sum_{s \in S} \frac{P^{s,i}_{\text{ASE}}}{P^{s+1,i}_{\text{tx}}}, \tag{3}$$

$$\sigma_{\text{NLI}} = \sum_{s \in S} \frac{P^{s,i}_{\text{NLI}}}{P^{s+1,i}_{\text{tx}}}. \tag{4}$$

Moreover, $P^{s+1,i}_{\text{tx}}$ is the launch power at the beginning of span $s+1$, NLI noise power ($P^{s,i}_{\text{NLI}}$) is calculated from equation (2) in the reference [36], and $P^{s,i}_{\text{ASE}}$ is the noise power caused by the doped fiber amplifier (DFA) equipped with the dynamic gain equalizer, which is computed as follow:

$$P^{s,i}_{\text{ASE}} = n_{\text{F}} \cdot h \cdot f^{i} \cdot (G^{s,i} - 1) \cdot R_{\text{ch}}, \tag{5}$$

where n_{F}, h, f^{i}, $G^{s,i}$, S, and R_{ch} are the DFA's noise figure, Plank's coefficient, channel frequency, DFA's gain, set of spans, and channel symbol rate, respectively. The $G^{s,i}$ can be expressed as:

$$G^{s,i} = \frac{P^{s+1,i}_{\text{tx}}}{P^{s,i}_{\text{rx}}}. \tag{6}$$

$P^{s,i}_{\text{rx}}$ is the received power at the end of span s. σ_{TRx}, σ_{Flt}, σ_{Ag} are the transceiver SNR, SNR penalty due to wavelength selective switches filtering, and SNR margin due to the aging. Moreover, the line card interfaces in the MB-EON are equipped with state-of-the-art flexible bit rate/MF transceivers (TRxs). The MF of TRxs can be arbitrarily changed based on the need as long as the GSNR value meets the MF requirement.

This study examines the performance of L+C+S-band MB-EONs, deploying a combined bandwidth of 20 THz (6 + 6 + 8 THz), which is divided into 268 channels of 75 GHz each with a 400 GHz spacing between adjacent bands. The channel spacing of 75 (6×12.5) GHz was chosen to match current commercial fixed-grid systems using 64 GBaud coherent transceivers. Instead of investigating flex-grid WDM systems with 12.5 GHz base granularity, our work focuses on fixed-grid flexible optical networks where MFs and bit rates (ranging from 100 to 600 Gbps) adapt based on GSNR. This spacing provides sufficient guard bands and aligns with state-of-the-art deployments. In this context, this work employs an EGGN semi-closed form model for estimating NLI noise, incorporating effects such as the Kerr nonlinearity and ISRS [36]. This model adjusts for frequency-specific physical parameters, including attenuation, dispersion, and nonlinear coefficients. The actual Raman gain profile is calculated by solving coupled Raman differential equations based on pump frequency and offset values. Further, to enhance model precision, adjustments for MF and dispersion are implemented, validated through field experiments [37]. Unlike prior studies [8], [22], [23], [27], [28], this approach considers a fully-loaded spectrum scenario that utilizes a hyper-accelerated flat received power optimization technique that integrates amplified spontaneous emission (ASE) noise loading in unused channels [33], which is in line with real-world practice for multi-band systems.

For each source-destination pair, the first K = 5 shortest paths are pre-computed. The MF assignment is based on the estimated GSNR at the destination node, which incorporates

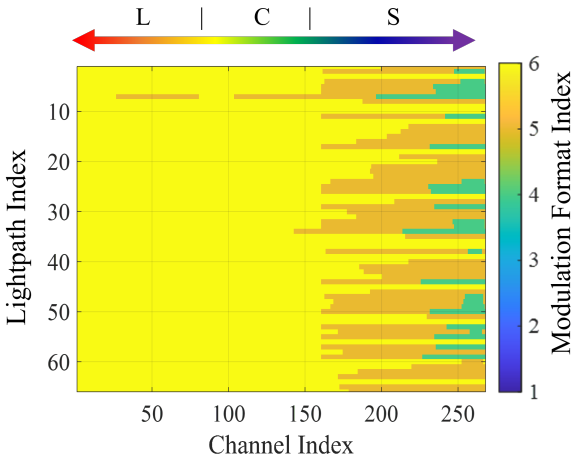


Fig. 2: Modulation format index profile for JPN12.

ASE noise, NLI, and ISRS effects. We assign available MFs that meet the pre-FEC BER threshold. Specifically, the highest-order available MF is assigned if it does not lead to capacity waste; otherwise, the lowest-order MF that satisfies the requested bit rate is selected. Since we assume that fully loaded links where unoccupied channels are filled with ASE-shaped noise, the optimal power profile, as well as the GSNR and available MF of each channel on all candidate paths can be pre-computed. We also assume a sufficient aging margin, ensuring that the selected MF for each established lightpath remains unaffected by subsequently established lightpaths or component aging. Under these assumptions, the GSNR of each lightpath is considered constant and does not vary with the establishment or teardown of other lightpaths. Therefore, although the traffic is dynamic—i.e., demands with different bit rates arrive and depart over time—the MF per channel is pre-determined, and only the bit rate assignment is carried out during provisioning, using both grooming and non-grooming strategies. This approach reduces online computational complexity while preserving physical-layer accuracy. For example, Fig. 2 illustrates the MF level index profile (i.e., the channel MF database), which indicates the highest available MF for each channel along the first shortest path across all connections in the JPN12 network. This calculation has been performed for all $K = 5$ paths per source-destination pair.

IV. DRL-Assisted QoT-Aware RMBSA Framework

The proposed framework is depicted in Fig. 3. The environment and the DRL agent are the two main components within the framework. In the following, we illustrate the design of these components and explain how they interact to form the complete RMBSA solution.

A. RMBSA Environment

The lower part of Fig. 3 illustrates the components of the RMBSA environment. In our proposed framework, the RMBSA environment includes *(1)* a simulated MB-EON constrained by the physical layer formulation described in Sec. III-C; *(2)* a database recording channels and the respective available MF of each channel on all candidate paths; *(3)* a feature extractor to generate the state s_t for the DRL agent based on the current state of the MB-EON; and *(4)* an evaluator to compute the reward r_t, returned to the DRL agent as the feedback of the action a_t.

B. DRL Agent Design

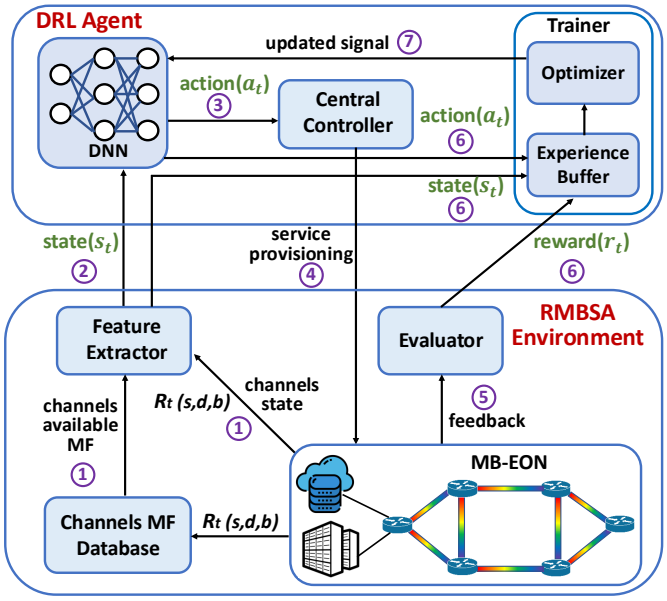


Fig. 3: Framework and workflow of the DRL-assisted QoT-aware service provisioning solution in multi-band elastic optical networks.

The upper part of Fig. 3 illustrates the components of the DRL agent. The DRL agent is responsible for formulating the RMBSA policy for each service, and optimizing its policy through consecutive interactions with the environment. Below, we elaborate on the design of the DRL agent, including the specific configurations of the state (s_t), action (a_t), and the reward (r_t).

1) State Representation (s_t)

The state s_t represents the observation of the MB-EON environment and should encompass features that effectively describe the current network condition, while also being relevant to the optimization objective (i.e., BP). In our case, the s_t is a vector with $K \times (H_{\max} + 5 \times B)$ elements, where K, B, and $H_{\max}$ are the number of candidate paths between each node pair, the number of bands, and the highest number of hops among all candidate paths, respectively. Out of these elements, $K \times H_{max}$ are related to route features, while $K \times B \times 5$ represent the band features.

The route features include, for each of the K paths between s and d, the indices of the links within the path. An equal number of elements (i.e., H_{max}) is assigned to each path, with unused elements padded with -1. This design uses fewer elements to effectively indicate the routing information compared to the one-hot [17] and binary [19] route representations used in the previous studies.

For each candidate path, band features are extracted for each band based on the candidate channel(s) determined using

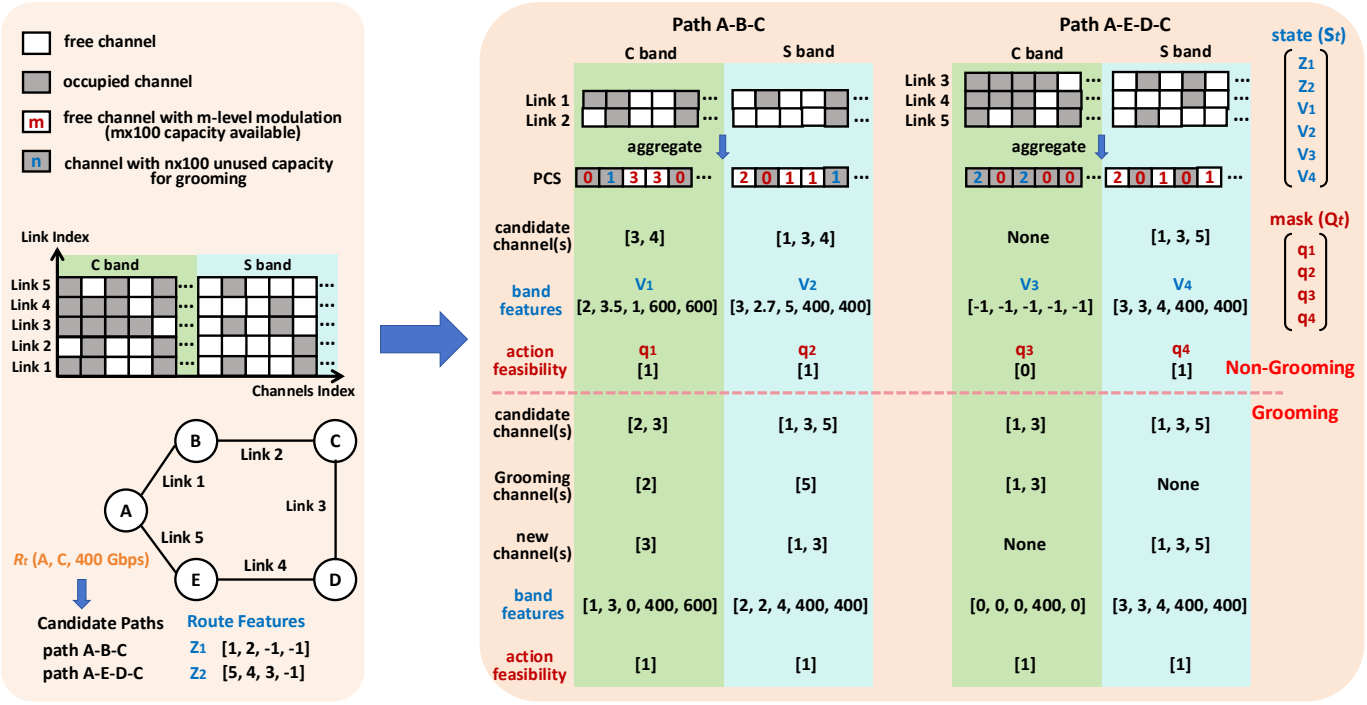


Fig. 4: State representation and action mask used in the DRL-RMBSA algorithm. Only the C+S-band scenario is shown in this figure but results for the L+C+S-band scenario are presented in Sec. V.

the first-fit scheme. In the case where the grooming is not allowed, four features are extracted from candidate channel(s) in each band: *(i)* the number of the channel(s) needed to fulfill the bit rate requirements; *(ii)* the average index of the channel(s) (related to channel frequency); *(iii)* the number of free channels on the adjacent links that occupy the same spectrum as these channel(s) (i.e., misaligned channels that may result in fragmentation) [38]; and *(iv)* the maximum bit rate supported by the channel(s) when using their highest MF. The last feature *(v)* is the total bit rate supported by the free channels of this band (related to band capacity utilization). For the bands without sufficient capacity to accommodate the bit rate requested by R_t, all the related band features are set to -1. When grooming is available, the candidate channels can be classified into new channel(s) that are entirely free and the grooming channel(s) with unused capacity. In this scenario, the features *(i)* to *(iii)* are extracted based on the new channel(s) instead of all the candidate channel(s), while the feature *(iv)* is still generated based on the candidate channel(s). Moreover, the feature *(v)* is obtained following the same procedure as the non-grooming scenario.

Fig. 4 illustrates an example where request R_t arrives requesting a service from node A to node C with 400 Gbps. Let us use the C-band of the path A-B-C as an example. The route features for this path are $[1, 2, -1, -1]$ (i.e., Z_1), as it includes link 1 (A-B) and link 2 (B-C). The last two elements are padded with -1 to match the $H_{\max} = 4$, which corresponds to the maximum hop count in this network (e.g., the path A–E–D–C–B). For band features (i.e., V_1), we firstly aggregate the channel state of each link across the path A-B-C to obtain the path channel state (PCS). The PCS denotes the channel availability of the path under the spectrum continuity constraint. The channel MF database is then traversed to identify the highest available MF level for each free channel along this path (from those identified in the pre-computation based on the GSNR tolerance), indicated by the red number (e.g., MF level 3, DP-8QAM, for channels 3 and 4 providing 300 Gbps). The basic channel capacity is 100 Gbps, so channels [3, 4] with 600 Gbps available capacity in total is a suitable candidate to meet the 400 Gbps requirement of R_t. Therefore, in Fig. 4 the number (feature *(i)*), average index (feature *(ii)*), and max supported bit rate (feature *(iv)*) of the candidate channels are 2, 3.5, and 600, respectively. The misaligned channels (feature *(iii)*) are the free channels on the adjacent links (i.e., links 3 and 5) that occupy the same spectrum as the candidate channels (i.e., third and fourth channels). In our case, only the fourth channel on link 5 is free, so the value is 1. Additionally, the total available bit rate of the band is 600 Gbps at free channels [3, 4], which is the value of the feature *(v)*. When grooming is available, the candidate channels include the grooming channel (i.e., second channel) with 100 Gbps of unused capacity (as indicated by the blue number) and the new channel (i.e., third channel) with full capacity. The features *(i)* to *(iii)* are extracted only based on the third channel, which have values 1, 3, and 0, respectively. The feature *(iv)* corresponds to the total bit rate supported by all candidate channels (i.e., second channel and third channel), which is 400. The feature *(v)* has the same value as non-grooming scenario, which is 600.

2) Action Space (a_t)

The a_t denotes a specific RMBSA action selected by the agent to be adopted for the provisioning of service request R_t. The

action encompasses choosing one of the B bands on one of the K candidate paths to assign the channels using the first-fit spectrum allocation scheme. An action masking scheme [39] is employed to prevent the selection of invalid actions (i.e., candidates without sufficient channels to accommodate R_t). As illustrated in Fig. 4, the mask $Q_t = [q_1, \ldots, q_{K\times B}]$ is generated at each step, where $q_i \in \{0, 1\}$ denotes the action feasibility of each band on the candidate path. Specifically, $q_i = 1$ indicates that the band has sufficient channels to accommodate R_t and thus can be selected; otherwise $q_i = 0$. After applying the mask Q_t, the original policy $\pi(a_t|s_t) = [p_1, \ldots, p_{k\times B}]$ is adjusted to the masked policy $\tilde{\pi}(a_t|s_t)$ as:

$$\tilde{\pi}(a_t|s_t) = \frac{Q_t \odot \pi(a_t|s_t)}{\sum_{i=1}^{K\times B} q_i p_i} \tag{7}$$

In $\tilde{\pi}(a_t|s_t)$, the probability corresponding to each invalid action with $q_i = 0$ (e.g., the C-band of path A–E–D–C in Fig. 4) is set to 0. The action a_t is then obtained by sampling from $\tilde{\pi}(a_t|s_t)$, ensuring that only valid actions can be chosen. This approach has shown that the agent can enhance performance [40] and focus on more important features [41].

3) Reward Function (r_t)

The reward r_t represents the feedback from taking action a_t. The agent's objective is to maximize the long-term return G_t, as expressed in (1), which is obtained by collecting rewards over multiple steps. Therefore, an effective reward function must be closely aligned with the optimization objective (i.e., minimization of long-term BP). In this study, we investigate two reward functions: the *simple reward function* and the *path-capacity-aware reward function*.

The *simple reward function* works as follows. If the request R_t is successfully provisioned, the simple reward function assigns $r_t = 1$, otherwise it assigns $r_t = -1$. This approach is commonly used in the previous DRL-based solutions [8], [17], [22]. Although intuitive, this reward function has shown to inefficiently explore the action space [20], which can be quite detrimental in MB-EONs where a large action space is observed.

In this work, we propose a *path-capacity-aware reward function*. This reward function design is inspired by the approach proposed in [20] but adapted to the MB-EON environment. The proposed reward function for a given action a_t is calculated as follows:

$$r_t = \begin{cases} 1 & \text{if } R_t \text{ is provisioned and } C(a_t) = C_{\max} \\ 0.9, & \text{if } R_t \text{ is provisioned and } C(a_t) \neq C_{\max} \\ -1, & \text{if } R_t \text{ is rejected} \end{cases} \tag{8}$$

Here, $C(a_t)$ represents the total available capacity of the free channels on the routing path selected by a_t, while $C_{\max}$ denotes the maximum available capacity of the free channels among all candidate paths. This reward function introduces capacity utilization information to the agent, encouraging it to select paths with higher capacity to mitigate the overuse of bottleneck links.

C. Workflow and Interaction

Herein we recall Fig. 3, which illustrates the workflow of the proposed framework. Upon receiving a request $R_t(s, d, b)$, the available MF of the channels on each candidate path between the source s and destination d is obtained by querying the pre-computed channel MF database (step 1). The available MF for each channel, along with the channel states (i.e., free or occupied) and request information are sent to the Feature Extractor, which generates the s_t (step 2). In turn, s_t is fed to the DNN of the DRL agent which outputs the a_t (step 3). Next, the Central Controller provisions the service according to a_t (step 4). The result of the provisioning is sent to the Evaluator, which computes and returns r_t (step 5). Subsequently, the tuple (s_t, a_t, r_t) is stored in the Experience Buffer as a training sample (step 6). Once the Experience Buffer is full, a training epoch is triggered, during which the Optimizer adjusts the DNNs parameters (biases and weights) through gradient optimization (step 7). Through iterative cycles of these processes, i.e., episodes, the DRL agent progressively refines its provisioning policy to maximize the accumulative rewards G_t, which, in our case, minimizes the long-term BP.

V. Performance Assessment

A. Simulation Setup

The proposed framework was evaluated through a simulation environment. The simulations were carried out over three topologies with different scales, including the medium-scale NSFNET [42] with 14 nodes and 22 links, small-scale JPN12 [43] with 12 nodes and 17 links, and large-scale SPN30 with 30 nodes and 56 links [33]. The L+C+S-band system with 268 channels (80+80+128) with 75 GHz bandwidth per channel is considered. The bit rate requirement for each request is uniformly distributed among 100, 200, 400, and 1,000 Gbps. Six MFs from DP-BPSK to DP-64QAM at 64 Gbaud are considered. The channel MF database adopts the GSNR thresholds for each MF defined in the literature [6], corresponding to a pre-forward error correction bit error rate of 1.5×10^{-2}. The GSNR and MF profiles are pre-calculated based on (2) and the GSNR thresholds of the related MF [6], respectively, adopting an optimum power at each span [33]. Therefore, each channel can provide bit rates ranging from 100 to 600 Gbps when using DP-BPSK to DP-64QAM, respectively. We consider EDFAs with noise figures of 4.5 dB and 5 dB for the C- and L-band, respectively. For the S-band, we consider a TDFA with a noise figure of 6 dB. We assume a standard single mode fiber with a zero-water peak. The spectrum continuity constraint is considered for the channel assignment along a lightpath.

TABLE I: Hyperparameters of the DRL agent

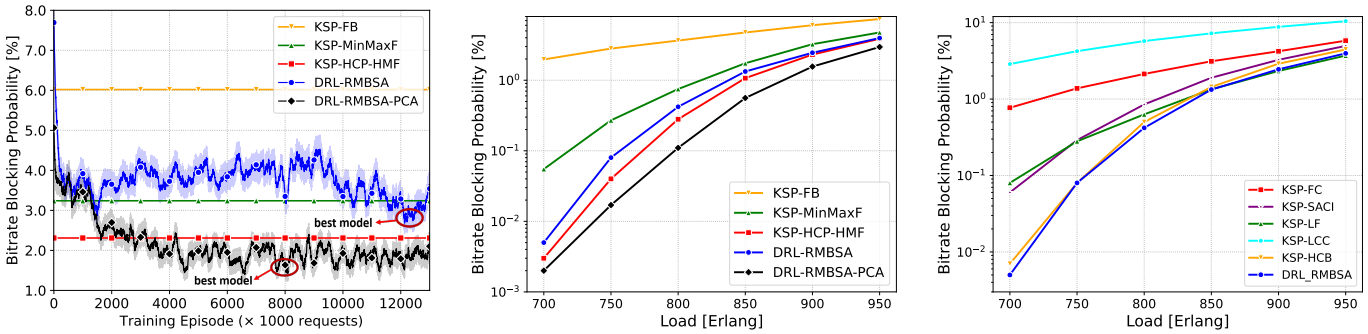
Hyperparameters	**Value**
Discount factor γ	0.95
Episode length L	1,000
Steps number N_s	200
Buffer size O	1,000
Batch size Z	500
Parallel environments N_e	5
Model record interval T	20,000
Learning rate	5e-5
Fully connected layers	5
Neurons per layer	128
Activation function	Relu

The MB-EON environment was established by extending the optical network environment in the Optical Networking Gym [44]. The advantage actor-critic (A2C) [45] algorithm was used for training. The A2C algorithm employs an actor (i.e., policy network) to select the action a_t, and a critic (i.e., value network) to estimate the value of the state s_t. The advantage function, which measures how much better a specific action is compared to the expected value of the state, is used to adjust the policy (actor). The advantage is typically calculated based on the difference between the cumulative rewards and the value estimated by the critic. A2C is commonly used to address problems with discrete action space [19]. The hyperparameters of the DRL agent are configured as shown in Table I. Specifically, the episode length L refers to the number of requests provisioned by the DRL agent in each episode, which is used to evaluate the agent's average blocking performance. The steps number N_s determines how many steps are accumulated to compute the discounted reward G_t. The buffer size O specifies the number of samples collected before each training epoch, meaning training is conducted after provisioning every O requests. The batch size Z defines the number of samples used in each update of the DNNs during training. Consequently, the DNNs are updated O/Z times in each training epoch. To accelerate training, N_e parallel environments are deployed, each hosting a DRL agent sharing the same DNNs. Each agent independently provisions requests in its own environment, enabling efficient and parallel collection of training samples. Additionally, the model record interval T specifies that after DRL agent provisions every T requests, its achieved average reward is evaluated, and the agent is updated as the best model if it outperforms the previous best.

Two DRL-based service provisioning algorithms were developed based on the proposed framework: *(1)* DRL-RMBSA, which adopts the simple reward function; and *(2)* DRL-RMBSA-PCA, which adopts the path-capacity-aware reward function. It is important to note that the available capacity of free channels on each candidate path (K features) is also integrated into the state of DRL-RMBSA-PCA. Apart from this, all other features of DRL-RMBSA-PCA are identical to those of DRL-RMBSA, as introduced earlier. The proposed DRL-based agents were compared with the following heuristics: *(1)* K-shortest-path first-band (KSP-FB) [11]; *(2)* KSP minimum-maximum frequency (KSP-MinMaxF) [46]; *(3)* KSP highest-capacity-path highest-MF (KSP-HCP-HMF); *(4)* KSP fewest-channels (KSP-FC); *(5)* KSP smallest-average-channels-index (KSP-SACI); *(6)* KSP least-fragmentation (KSP-LF); *(7)* KSP lowest-channels-capacity (KSP-LCC); and *(8)* KSP highest-capacity-band (KSP-HCB). KSP-FB prioritizes the use of lower frequency band (i.e., L band) on the shortest-available path. The KSP-MinMaxF and the KSP-HCP-HMF are more advanced heuristics. The KSP-MinMaxF always selects the action where maximum frequency candidate channel has the lowest frequency among all the feasible actions. The KSP-HCP-HMF is proposed in this paper. KSP-HCP-HMF adopts the same criteria as the reward function of the DRL-RMBSA-PCA suggests, characterizing a fair comparison with DRL. KSP-HCP-HMF attempts to use channels with the highest MF on the path with the highest available capacity (refer to the $C_{\max}$ used in the path-capacity-aware reward function). Moreover, if there are multiple actions with the same highest MF, KSP-HCP-HMF selects the action based on the MinMaxF criteria. Moreover, algorithms *(4)-(8)* are greedy heuristics designed based on the band features used in the state representation of the DRL agent. Specifically, KSP-FC selects the action with the fewest candidate channels; KSP-SACI selects the action whose candidate channels have the smallest average index; KSP-LF selects the action that results in the least fragmentation; KSP-LCC selects the action whose candidate channels have the lowest capacity; and KSP-HCB selects the action whose band offers the highest available capacity. All the DRL-based algorithms and heuristics employ the first-fit scheme for spectrum allocation. When grooming is available, the use of unused capacity on active (existing) channels is preferred over establishing new channels. The heuristics *(1)-(3)* are evaluated across all the scenarios, while the greedy heuristics *(4)-(8)* are only evaluated in NSFNET without grooming. We do not include previously-proposed DRL solutions [8], [22], [23] as they have been reported to not outperform KSP-FB. We indirectly compare the results with [29] by using the reported gains over KSP-FB-FF on the Japanese topology under similar conditions. The performance of the algorithms is evaluated based on their average achieved bit rate blocking probability (BRBP) over each episode with 1,000 dynamic requests.

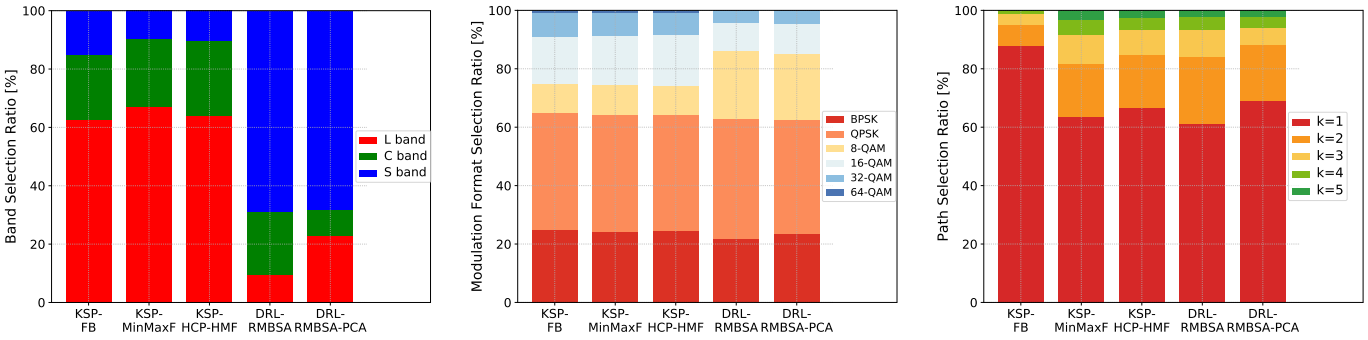
B. Results and Discussion

Fig. 5 shows the results for the NSFNET topology without grooming under a 900 Erlang traffic load. Fig. 5a demonstrates the evolution of the BRBP achieved by the DRL agents over the training episodes. The shadow around the curve represents the confidence interval with 95% confidence level. Since the performance of heuristics with fixed policies does not change significantly as training progresses, we show their average achieved BRBPs across all evaluated episodes as stationary horizontal lines. At the initial stage of the training, the DRL agents were initialized with random DNN parameters, thus achieving a poor BRBP. However, as training progresses, the DRL agents gradually optimized their policies, resulting in a sharp reduction in BRBP. During this period, 0-1,000 episodes, both DRL agents can outperform the KSP-FB. Specifically for DRL-RMBSA, its performance plateaued in a local minimum until around 9,000 training episodes, after which it further reduced the BRBP, and finally converged after approximately 10,000 training episodes. Its performance after convergence was slightly better than that of KSP-MinMaxF. Meanwhile, DRL-RMBSA-PCA was able to moderately improve its policy through exploration and eventually converged after around 4,500 training episodes. After convergence, DRL-RMBSA-PCA outperforms all the heuristics. Moreover, DRL-RMBSA-PCA achieves better BRBP performance and faster convergence compared to DRL-RMBSA. We speculate that this is due to the path-capacity-aware reward function used by DRL-RMBSA-PCA, which enables the agent to efficiently explore the environment and learn an effective routing policy in the



(a) Performance over training (b) Performance over different loads (c) DRL vs. greedy heuristics

Fig. 5: Bit rate blocking probability achieved by the DRL agents and the heuristics in the NSFNET topology without grooming under (a) a 900 Erlang traffic load, and (b)-(c) different traffic loads.



(a) Band selection (b) Modulation format selection (c) Path selection

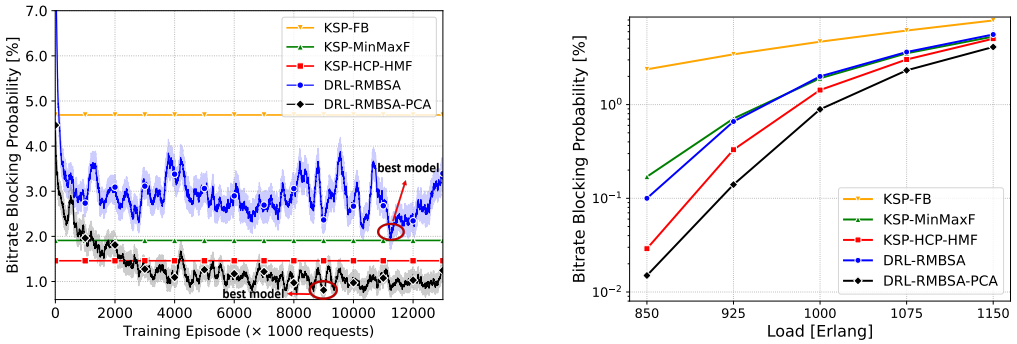
Fig. 6: Detailed band, modulation format, and path selected by the heuristics and DRL agents for the NSFNET topology without grooming under a 900 Erlang traffic load.

short term. Additionally, it progressively optimizes the band selection scheme by maximizing long-term returns, resulting in better performance than KSP-HCP-HMF, which is also path-capacity-aware.

To evaluate the best performance of DRL-RMBSA and DRL-RMBSA-PCA, the DNN parameters that achieved the lowest average BRBP over 20,000 requests (i.e., the red circle in Fig. 5a) were identified as the best model. Fig. 5b shows the results for the best models of the two DRL-based algorithms, along with the heuristics, evaluated with provisioning 200,000 randomly-generated (unseen) requests for each load. To fully exploit the effectiveness of the DRL agent, a greedy policy was employed during DNN inference, which always selects the action with the highest probability. Focusing on the load of 900 Erlang, Fig. 5b shows that the BRBP achieved by the DRL-RMBSA was under 2.45%, which represents a reduction by 59% and 24% compared to KSP-FB (6.02%) and KSP-MinMaxF (3.24%), respectively. However, DRL-RMBSA cannot outperform the more advanced KSP-HCP-HMF heuristic which achieves 2.31% BRBP. On the other hand, by adopting the path-capacity-aware reward, DRL-RMBSA-PCA, achieves a BRBP of 1.56%, which is 73%, 52%, and 32% lower than that achieved by the KSP-FB, KSP-MinMaxF, and KSP-HCP-HMF, respectively. Across all investigated loads, DRL-RMBSA-PCA can outperform all the heuristics and DRL-RMBSA also shows satisfactory performance (Fig. 5b). This validates the effectiveness and the generality of the proposed DRL-based framework.

In the following, to validate the efficient exploration of the DRL agent, we compare DRL-RMBSA to greedy heuristics that optimize each of the band features within the state representation. Fig. 5c shows that the DRL-RMBSA outperforms all the heuristics, confirming that the proposed solution does not overly rely on any single feature to determine the action (e.g., it does not always selecting the action with the highest band capacity or the lowest fragmentation). This indicates that the DRL agent avoids convergence to a local optimum.

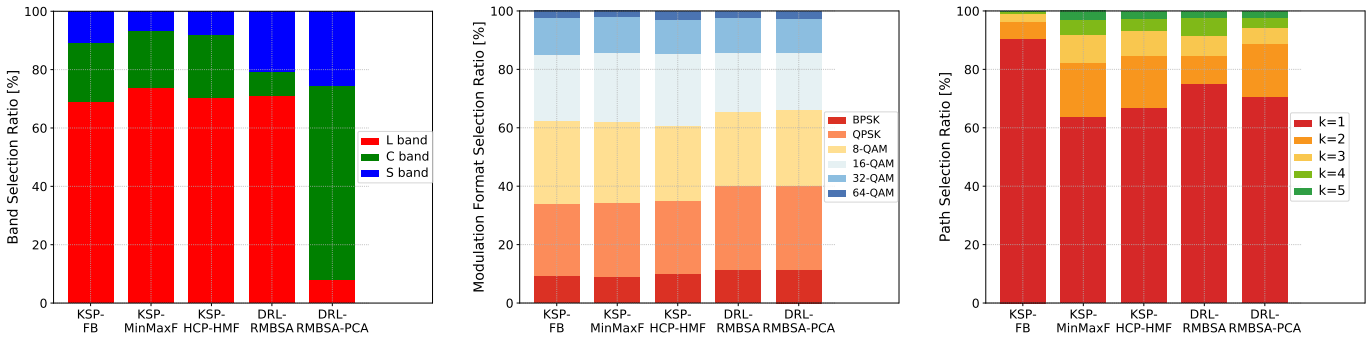
Fig. 6 shows the detailed policy of the heuristics and DRL agents in terms of selection of band, MF, and path. These metrics are calculated by dividing the total number of times each path, band, or MF is selected by the total number of accepted requests across the simulation. As shown in Fig. 6a and Fig. 6b, all the heuristics prefer to select the L-band with the higher modulation level, which aligns with their corresponding rules. When the grooming is not allowed, the use of channels with higher MF may lead to the waste of the capacity, particularly in provisioning the request with low bit rate requirement (i.e., 100 Gbps). In contrast, the DRL agent prefers the S-band (Fig. 6a) with lower MF to save the capacity and leave the channels with the high capacity on the bottleneck links/paths to provision the services with high bit rate requirement. As shown in Fig. 6c, the routing



(a) Performance over training

(b) Performance over different loads

Fig. 7: Bit rate blocking probability achieved by the DRL agents and the heuristics in the NSFNET topology with grooming available under (a) a 1,000 Erlang traffic load, and (b) different traffic loads.



(a) Band selection

(b) Modulation format selection

(c) Path selection

Fig. 8: Detailed band, modulation format, and path selected by the heuristics and DRL agents for the NSFNET topology with grooming available under a 1,000 Erlang traffic load.

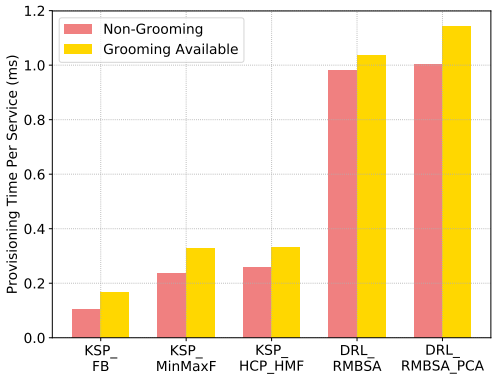
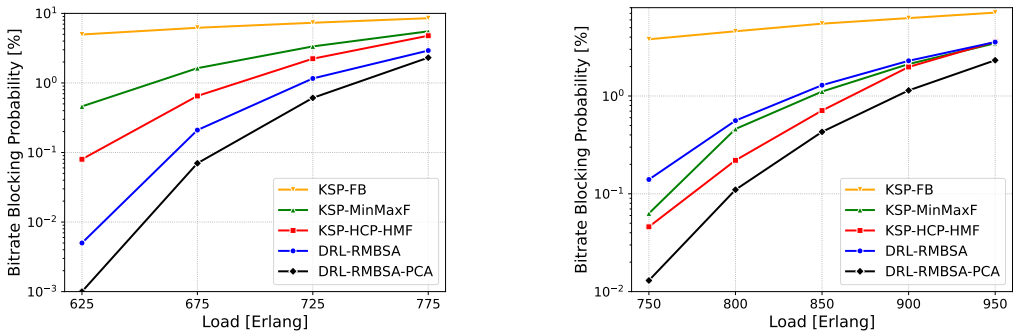


Fig. 9: Time complexity of each algorithm (including inference and environment transition) in NSFNET topology.

strategies applied by DRL-RMBSA and DRL-RMBSA-PCA tend to select the first and second shortest paths, unlike KSP-FB which relies mostly on the shortest path.

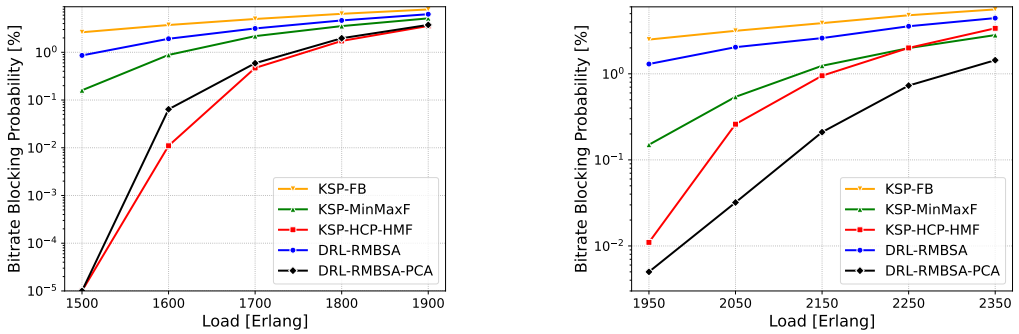
Fig. 7a illustrates the BRBP achieved by the DRL agent during training in the NSFNET topology with grooming available under a 1,000 Erlang traffic. DRL-RMBSA-PCA continues to outperform all heuristics after convergence. Meanwhile, DRL-RMBSA does not perform as well as it does in the non-grooming scenario. Nevertheless, DRL-RMBSA still surpasses KSP-FB during training, and its best model performs comparably to KSP-MinMaxF. Fig. 7b shows the BRBP achieved by all heuristic methods and the best model of the proposed DRL-based solutions, where DRL-RMBSA-PCA shows the best performance in all cases. Specifically, the DRL-RMBSA-PCA can reduce the BRBP by approximately 81%, 53%, and 38% compared with the KSP-FB, KSP-MinMaxF, and KSP-HCP-HMF, respectively.

Fig. 8a shows that the band allocation policies learned by the DRL agents focus on utilizing the bands that support higher MF levels. DRL-RMBSA prefers the L-band, while DRL-RMBSA-PCA prefers the C-band. This differs from the non-grooming scenario where both agents preferred the S-band. We speculate that this is because grooming reduces capacity waste on channels with higher MF, allowing the use of the C-band and L-band to yield effective performance. It is also worth noting that, since the cost per unit for C-band components is approximately 10%-30% lower than for L- and S-band components [3], the C-band-preferred DRL-RMBSA-PCA can offer a cost-effective solution. Fig. 8b shows that allowing grooming can improve the overall modulation level compared to the non-grooming scenario (Fig. 6b), thereby increasing network throughput. However, this can be attributed to the cases where without grooming there is no incentive to upgrade to higher-level MFs, since the extra capacity will



(a) Without grooming (b) With grooming available

Fig. 10: Bit rate blocking probability achieved by all algorithms under different traffic loads in the JPN12 topology.



(a) Without grooming (b) With grooming available

Fig. 11: Bit rate blocking probability achieved by all algorithms under different traffic loads in the SPN30 topology.

remain idle. Moreover, DRL-RMBSA and DRL-RMBSA-PCA do not overly rely on any single path but instead employ an effective routing policy to balance the capacity across all paths (Fig. 8c).

The time complexity of each algorithm was evaluated based on its average provisioning time per service. Our simulations were conducted on a 12th-generation Intel i7 CPU operating at 2.2GHz. Fig. 9 shows that the DRL-based algorithms take more time to provision requests compared to the heuristics. The provisioning time per service includes the inference time (i.e., the time for the RMBSA algorithm to find a solution) in addition to the environment transition time (i.e., the time to provision the service in our simulation tool). We can see that the DRL solutions take around $5\times$ the time taken by the heuristics. However, this difference is negligible considering the absolute time taken. In the worst-case scenario, our solution takes less than 1.2 ms, which is a negligible time considering the time necessary to setup a lightpath (in the order of several seconds to minutes).

The assessments were conducted on the JPN12 and SPN30 topologies, which have fewer and larger numbers of nodes and links, respectively. As shown in Fig. 10, DRL-RMBSA-PCA consistently outperforms all other benchmarks in both scenarios on the MB-EONs with the JPN12 topology, and DRL-RMBSA also demonstrates its effectiveness when grooming is not considered. Our proposed framework achieves a reduction of up to 91% in terms of BRBP compared to KSP-FB-FF. The BP gains reported in [29] reach 71% of the BP in the same blocking range, indicating that our proposed framework outperforms the one in [29]. In the more complex SPN30 topology (Fig. 11), DRL-RMBSA-PCA retains its advantage when grooming is available. However, the DRL-based solutions are unable to outperform KSP-HCP-HMF when grooming is not allowed. This indicates that for SPN30 the DRL agent may need to undergo a hyper-parameter tuning campaign, which is left for future work.

VI. Conclusions

In this paper, we proposed a DRL-assisted QoT-aware RMBSA framework for dynamic service provisioning in MB-EONs. Within the solution, an EGGN QoT estimator was employed to create the route-channel MF profiles for the DRL agent. A comprehensive state representation tailored for MB-EONs was developed, encompassing essential features of each path and band. To improve training efficiency, action masking was employed to filter out the large number of invalid actions in the decision space of the DRL agent. A path-capacity-aware reward function was designed to guide the agent in effectively exploring the MB-EON environment and accelerating training. The effectiveness of our solutions was validated across multiple network topologies and traffic loads, demonstrating significant BRBP reduction compared to the heuristics in the literature. In this study, we modeled the state transitions of the simulation environment as a function of several network behavior distributions (e.g., data rate, requested node pair). Our assumption is that network

operators can compute current or estimate expected network behavior distributions in practical networks. However, we acknowledge the offline simulations may not fully emulate the physical layer impairments or capture the dynamics of practical networks. As future work, an important aspect is to deploy the DRL agents trained in the simulated environment on real-world testbeds to evaluate the generalizability of their learned policies. Moreover, we plan to further unlock the potential of DRL in MB-EONs by leveraging graph neural networks with spatial feature extraction capabilities to capture hidden relationships between network links and formulate a network-level state representation.

Code and Data Availability

The source code used to generate the results presented in this study, and the generated results, are available at https://github.com/YiranTeng/DRL-RMBSA.

References

[1] E.-K. Hong, I. Lee, B. Shim, Y.-C. Ko, S.-H. Kim, S. Pack, K. Lee, S. Kim, J.-H. Kim, Y. Shin *et al.*, "6G R&D vision: Requirements and candidate technologies," *J. Commun. Netw.*, vol. 24, no. 2, pp. 232–245, Apr. 2022.

[2] O. Gerstel, M. Jinno, A. Lord, and S. B. Yoo, "Elastic optical networking: a new dawn for the optical layer?" *IEEE Commun. Mag.*, vol. 50, no. 2, pp. s12–s20, Feb. 2012.

[3] A. Souza, B. Correia, A. Napoli, V. Curri, N. Costa, J. Pedro, and J. Pires, "Cost analysis of ultrawideband transmission in optical networks," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 16, no. 2, pp. 81–93, Feb. 2024.

[4] L. Rapp and M. Eiselt, "Optical amplifiers for multi–band optical transmission systems," *J. Lightw. Technol.*, vol. 40, no. 6, pp. 1579–1589, Mar. 2022.

[5] F. Arpanaei, M. R. Zefreh, C. Natalino, P. Lechowicz, S. Yan, J. M. Rivas-Moscoso, O. Gonzalez de Dios, J. P. Fernandez-Palacios, H. Rabbani, M. Brandt-Pearce, A. Sanchez-Macian, J. A. Hernandez, D. Larrabeiti, and P. Monti, "Ultra-high-capacity band and space division multiplexing backbone EONs: multi-core versus multi-fiber," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 16, no. 12, pp. H66–H78, Dec. 2024.

[6] F. Arpanaei *et al.*, "Enabling seamless migration of optical metro-urban networks to the multi-band: unveiling a cutting-edge 6D planning tool for the 6G era," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 16, no. 4, pp. 463–480, Apr. 2024.

[7] A. Ferrari, E. Virgillito, and V. Curri, "Band-division vs. space-division multiplexing: A network performance statistical assessment," *J. Lightw. Technol.*, vol. 38, no. 5, pp. 1041–1049, Mar. 2020.

[8] A. Beghelli, P. Morales, E. Viera, N. Jara, D. Bórquez-Paredes, A. Leiva, and G. Saavedra, "Approaches to dynamic provisioning in multiband elastic optical networks," in *Proc. Int. Conf. Opt. Netw. Des. Model.*, 2023, pp. 1–6.

[9] T. Hoshida, V. Curri, L. Galdino, D. T. Neilson, W. Forysiak, J. K. Fischer, T. Kato, and P. Poggiolini, "Ultrawideband systems and networks: Beyond C + L-band," *Proc. IEEE*, vol. 110, no. 11, pp. 1725–1741, 2022.

[10] N. Guo, B. Correia, G. Shen, and V. Curri, "Performance analyses of wavelength assignment algorithms in wideband fixed-grid/flex-rate optical networks," in *Proc. Int. Conf. Transparent Opt. Netw.*, 2024, pp. 1–5.

[11] N. Sambo, A. Ferrari, A. Napoli, N. Costa, J. Pedro, B. Sommerkorn-Krombholz, P. Castoldi, and V. Curri, "Provisioning in multi-band optical networks," *J. Lightw. Technol.*, vol. 38, no. 9, pp. 2598–2605, May. 2020.

[12] Y. Teng, H. Yang, Q. Yao, R. Gu, Z. Sun, A. Xu, F. Liu, and J. Zhang, "SRS-proactive-aware resource allocation based on all-optical wavelength converters in C+L band optical networks," *J. Lightw. Technol.*, pp. 1–15, Oct. 2024.

[13] Q. Yao, H. Yang, B. Bao, J. Zhang, H. Wang, D. Ge, S. Liu, D. Wang, Y. Li, D. Zhang, and H. Li, "SNR re-verification-based routing, band, modulation, and spectrum assignment in hybrid C-C+L optical networks," *J. Lightw. Technol.*, vol. 40, no. 11, pp. 3456–3469, Jun. 2022.

[14] M. Mehrabi, H. Beyranvand, M. J. Emadi, and F. Arpanaei, "Efficient statistical qot-aware resource allocation in eons over the c+l-band: a multi-period and low-margin perspective," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 16, no. 5, pp. 577–592, May. 2024.

[15] F. Calderón, A. Lozada, P. Morales, D. Bórquez-Paredes, N. Jara, R. Olivares, G. Saavedra, A. Beghelli, and A. Leiva, "Heuristic approaches for dynamic provisioning in multi-band elastic optical networks," *IEEE Commun. Lett.*, vol. 26, no. 2, pp. 379–383, Feb. 2022.

[16] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature.*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[17] X. Chen, B. Li, R. Proietti, H. Lu, Z. Zhu, and S. J. B. Yoo, "DeepRMSA: A deep reinforcement learning framework for routing, modulation and spectrum assignment in elastic optical networks," *J. Lightw. Technol.*, vol. 37, no. 16, pp. 4155–4163, Aug. 2019.

[18] T. Tanaka and M. Shimoda, "Pre-and post-processing techniques for reinforcement-learning-based routing and spectrum assignment in elastic optical networks," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 15, no. 12, pp. 1019–1029, Dec. 2023.

[19] L. Xu, Y.-C. Huang, Y. Xue, and X. Hu, "Deep reinforcement learning-based routing and spectrum assignment of EONs by exploiting GCN and RNN for feature extraction," *J. Lightw. Technol.*, vol. 40, no. 15, pp. 4945–4955, Aug. 2022.

[20] B. Tang, Y.-C. Huang, Y. Xue, and W. Zhou, "Heuristic reward design for deep reinforcement learning-based routing, modulation and spectrum assignment of elastic optical networks," *IEEE Commun. Lett.*, vol. 26, no. 11, pp. 2675–2679, Nov. 2022.

[21] E. Etezadi, C. Natalino, R. Diaz, A. Lindgren, S. Melin, L. Wosinska, P. Monti, and M. Furdek, "Deep reinforcement learning for proactive spectrum defragmentation in elastic optical networks," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 15, no. 10, pp. E86–E96, Oct. 2023.

[22] N. E. D. E. Sheikh, E. Paz, J. Pinto, and A. Beghelli, "Multi-band provisioning in dynamic elastic optical networks: a comparative study of a heuristic and a deep reinforcement learning approach," in *Proc. Int. Conf. Opt. Netw. Des. Model.*, 2021, pp. 1–3.

[23] M. Gonzalez, F. Condon, P. Morales, and N. Jara, "Improving multi-band elastic optical networks performance using behavior induction on deep reinforcement learning," in *Proc. IEEE Lat. Amer. Conf. Commun.*, 2022, pp. 1–6.

[24] D. Yan, N. Feng, Z. Gu, X. Zuo, S. Fan, and J. Zhao, "DRL-based impairment-aware resource allocation algorithm in C + L band elastic optical networks," in *Proc. Int. Conf. Opt. Commun. Netw.*, 2024, pp. 1–3.

[25] E. Etezadi, F. Arpanaei, C. Natalino, E. Agrell, L. Wosinska, P. Monti, D. Larrabeiti, and M. Furdek, "Joint fragmentation- and QoT-aware RBMSA in dynamic multi-band elastic optical networks," in *Proc. Int. Conf. Transparent Opt. Netw.*, 2024, pp. 1–5.

[26] Y. Teng, C. Natalino, F. Arpanaei, A. Sanchez-Macian, P. Monti, S. Yan, and D. Simeonidou, "DRL-assisted dynamic QoT-aware service provisioning in multi-band elastic optical networks," in *Proc. Eur. Conf. Opt. Commun.*, 2024, pp. 1912–1915.

[27] A. B. Terki, J. Pedro, A. Eira, A. Napoli, and N. Sambo, "Routing and spectrum assignment assisted by reinforcement learning in multi-band optical networks," in *Proc. Eur. Conf. Opt. Commun.*, 2022, pp. 1–4.

[28] ——, "Routing and spectrum assignment based on reinforcement learning in multi-band optical networks," in *Proc. Int. Conf. Photon. Switch. Comput.*, 2023, pp. 1–3.

[29] ——, "Deep reinforcement learning for resource allocation in multi-band optical networks," in *Proc. Int. Conf. Opt. Netw. Des. Model.*, 2024, pp. 1–4.

[30] V. Curri, "GNPy model of the physical layer for open and disaggregated optical networking [invited]," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 14, no. 6, pp. C92–C104, Jun. 2022.

[31] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26–38, 2017.

[32] M. Doherty, R. Matzner, R. Sadeghi, P. Bayvel, and A. Beghelli, "Reinforcement learning for dynamic resource allocation in optical networks: hype or hope?" *IEEE/OSA J. Opt. Commun. Netw.*, vol. 17, no. 9, pp. D1–D17, Jun. 2025.

[33] F. Arpanaei, M. Ranjbar Zefreh, Y. Jiang, P. Poggiolini, K. Ghodsifar, H. Beyranvand, C. Natalino, P. Monti, A. Napoli, J. M. Rivas-Moscoso, Ó. González de Dios, J. Pedro Fernández-Palacios, O. A. Dobre, J. Alberto Hernández, and D. Larrabeiti, "Synergizing hyper-accelerated

power optimization and wavelength-dependent QoT-aware cross-layer design in next-generation multi-band EONs," pp. 1840–1855, 2025.

[34] M. L. Puterman, "Markov decision processes," *in Handbooks in operations research and management science*, vol. 2, pp. 331–434, 1990.

[35] P. Poggiolini, G. Bosco, A. Carena, V. Curri, Y. Jiang, and F. Forghieri, "The GN-model of fiber non-linear propagation and its applications," *J. Lightw. Technol.*, vol. 32, no. 4, pp. 694–721, Feb. 2014.

[36] P. Poggiolini and M. Ranjbar-Zefreh, "Closed form expressions of the nonlinear interference for uwb systems," in *Proc. Eur. Conf. Opt. Commun.*, 2022, pp. 1–4.

[37] Y. Jiang, A. Nespola, S. Straullu, F. Forghieri, S. Piciaccia, A. Tanzi, M. R. Zefreh, G. Bosco, and P. Poggiolini, "Experimental test of a closed-form EGN model over C+L bands," *J. Lightw. Technol.*, vol. 43, no. 2, pp. 439–449, Jan. 2025.

[38] Y. Yin, H. Zhang, M. Zhang, M. Xia, Z. Zhu, S. Dahlfort, and S. J. B. Yoo, "Spectral and spatial 2D fragmentation-aware routing and spectrum assignment algorithms in elastic optical networks [invited]," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 5, no. 10, pp. A100–A106, Oct. 2013.

[39] S. Huang and S. Ontañón, "A closer look at invalid action masking in policy gradient algorithms," *The International FLAIRS Conference Proceedings*, vol. 35, May 2022.

[40] M. Shimoda and T. Tanaka, "Mask RSA: End-to-end reinforcement learning-based routing and spectrum assignment in elastic optical networks," in *Proc. Eur. Conf. Opt. Commun.*, 2021, pp. 1–4.

[41] O. Ayoub, C. Natalino, and P. Monti, "Towards explainable reinforcement learning in optical networks: The RMSA use case," in *Proc. IEEE Opt. Fiber Commun. Conf. Exhibit.*, 2024, pp. 1–3.

[42] D. L. Mills and H.-W. Braun, "The NSFNET backbone network," in *Proc. ACM Workshop Front. Comput. Commun. Technol.*, 1987, pp. 191–196.

[43] T. Sakano, Y. Tsukishima, H. Hasegawa, T. Tsuritani, Y. Hirota, S. Arakawa, and H. Tode, "A study on a photonic network model based on the regional characteristics of Japan," *IEICE Tech. Rep.*, vol. 113, no. 91, pp. 1–6, Jun. 2013.

[44] C. Natalino, T. Magalhaes, F. Arpanaei, F. R. L. Lobato, J. C. W. A. Costa, J. A. Hernandez, and P. Monti, "Optical Networking Gym: an open-source toolkit for resource assignment problems in optical networks," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 16, no. 12, pp. G40–G51, Dec. 2024.

[45] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1928–1937.

[46] F. Arpanaei, J. M. Rivas-Moscoso, M. R. Zefreh, J. A. Hernández, J. P. Fernández-Palacios, and D. Larrabeiti, "A comparative study on routing selection algorithms for dynamic planning of EONs over C+L bands," in *Proc. OSA Photon. Netw. Devices.*, 2023, pp. NeM3B–4.