Approaches to dynamic provisioning in multiband elastic optical networks

A. Beghelli*, P. Morales†, E. Viera†, N. Jara†, D. Bórquez-Paredes‡, A. Leiva §, G. Saavedra¶
* *Optical Networks Group, Department of Electronic and Electrical Engineering, University College London, WC1E 7JE, UK*
† *Department of Electronic Engineering, Universidad Técnica Federico Santa María, Av. España 1680, Valparaíso 2390123, Chile*
‡ *Faculty of Engineering and Sciences, Universidad Adolfo Ibáñez, Av. Padre Hurtado 750, Viña del Mar, 2520001, Chile*
§ *School of Electrical Engineering, Pontificia Universidad Católica de Valparaíso, Av. Brasil 2950, Valparaíso, 2362804, Chile*
¶ *Electrical Engineering Department, Universidad de Concepción, Víctor Lamas 1290, Concepción 4070409, Chile*

Abstract—**Adopting multiband transmission in optical networks can cost-effectively increase network capacity without deploying new fibre. In this paper, we focus on the solutions explored by the research community to address the problem of resource allocation in dynamic multiband elastic optical networks. We start by summarising the main challenges and contributions of the design of ad-hoc heuristics. Next, we review the few recent approaches based on deep reinforcement learning and evaluate the efficacy of different techniques to improve their performance. We also discuss possible future directions for research in the area.**

Index Terms—**Multiband optical networks, elastic optical networks, Heuristics, Reinforcement Learning.**

I. Introduction

Dynamic multiband elastic optical networks (MB-EON) have the potential to make efficient use of network capacity by flexibly allocating spectrum only where and when required. In doing so, it might become a promising, cost-effective technology to increase the capacity of already deployed fibre [1]. Among the many technical challenges to overcome to make dynamic MB-EON a reality, resource provisioning deals with the problem of determining - on demand - a route (R), a band (B), a modulation format (M), and a block of contiguous spectrum slots (S) for arriving connection requests. This is known as the dynamic RBMSA problem. Dynamic RBMSA solvers must achieve a good trade-off between computational simplicity (for fast resource allocation in a dynamic environment) and performance (blocking ratio). This paper briefly reviews previous work on different approaches to solve the dynamic RBMSA problem and discusses ways forward for this research area.

II. Dynamic RBMSA approaches

To date, the dynamic RBMSA problem has been addressed by using ad-hoc heuristic methods [2]–[10] and deep reinforcement learning approaches [11]–[13]. To the best of our knowledge, the research community has not yet explored metaheuristics (such as bio-inspired algorithms) to solve the dynamic RBMSA problem.

A. Ad-Hoc Heuristics

To decrease computational complexity, heuristic-based dynamic RBMSA approaches address the R, B, M, and SA sub-problems separately. In principle, almost any order might be used to solve the sub-problems. However, in practice, the SA problem is the last to be tackled since the number of spectrum slots required by a connection can only be determined after selecting a modulation format. Additionally, it is computationally simpler to find a block of available slots for a specific route and band than for all possible combinations of them. Also for computational complexity reasons, the M problem is usually solved after the R problem: the most efficient modulation format given the length of the selected route is chosen. However, recent works have also explored M before R [7]. As a result of these considerations, the orders used to date have been reduced to RMBSA [2], RBMSA [4], [6], MBRSA [7] and BRMSA [5], [8], [9]. Irrespective of the order used, we will term them RBMSA solvers.

Dynamic RBMSA solvers can be classified along three dimensions affecting computational complexity or performance:

- **Route pre-processing**: To decrease computational complexity, most proposals pre-compute a set of K routes between each source and destination node [2], [4]–[7], [10]. Typically, K does not exceed 5. In this way, the cost of running Dijkstra's algorithm for each request (with a computational complexity ranging from $O(N^2)$ to $O(L + NlogN)$ depending on implementation [14], where N and L denote the number of network nodes and links, respectively) is avoided.
- **Quality of transmission (QoT) evaluation**: Noise figure variation of optical amplifiers in different bands and the impact of Inter-channel Stimulated Raman Scattering (ISRS) become relevant in multiband environments, affecting the QoT of signals. Heuristics have resorted to evaluating QoT in an online or offline manner. In an online manner, the QoT of the new lightpath [2] or the QoT of the new lightpath and the already established lightpaths [7], [9] is evaluated for each connection request according to the network state at the moment the new request is processed. Online QoT evaluation can be

Financial support from EPSRC Programme Grant TRANSNET (EP/R035342/1), ANID FOVI210082, ANID Fondecyt Iniciación 11220650 and 11201024 is gratefully acknowledged.

978-3-903176-54-6 © 2023 IFIP

TABLE I
SUMMARY OF HEURISTIC-BASED DYNAMIC RBMSA APPROACHES

Solver	Sub-problem order	Routes	QoT/VEL	Serv. Diff.	Computational Complexity
[2]	RMBSA	Precomputed (K=1)	Online/No	No	$O(M \times B \times G \times S \times L)$
[3]	RMBSA	Dijkstra (online)	Offline/No	No	$O(N \times log(N) + M \times L \times S)$
[4], [6]	RBMSA	Precomputed (K=1)	Offline/Yes	Yes	$O(M \times B \times S \times L)$
[5]	BRMSA	Precomputed (K=10)	Online/Yes	No	$O(M \times B \times K \times G \times S \times L)$
[7]	MBRSA	Precomputed (K=3)	Online/Yes	Yes	$O(M \times B \times K \times L \times S)$
[8]	BRMSA	Dijkstra (online)	Online/No	No	$O(N \times log(N) + M \times L^2 \times S^2)$
[9]	BRMSA	Dijkstra (online)	Online/Yes	No	$O(N \times log(N) + M \times L^2 \times S)$
[10]	RM(B)SA*	Precomputed (K=3)	Offline/No	Yes	$O(K \times M \times S \times L)$

performed either by using a numerical module based on the GNPy library [2], [7] or ad-hoc numerical evaluators [8], [9]. To avoid the high computational complexity of online QoT calculation - quadratic or linear with the number of channels, depending on the computational implementation [2], offline approaches have resorted to deep learning-based QoT estimation [10] or worst-case scenarios where all channels are assumed to be active and the optical reach of different modulation formats and bands, defined by the spectrum slot with the lowest QoT, is pre-computed and stored in a table [4], [6]. If the QoT is not good enough, another solution is explored (e.g., a different modulation format or transmission band) or the connection request is rejected. Online evaluation leads to increased computational complexity, but it might result in a lower blocking probability by considering the exact network state rather than the worst-case scenario or the estimation of the offline evaluation [10].

- **Service differentiation**: Some heuristics apply different band allocation strategies depending on the characteristics of the request being served [4], [6], [7], [10]. For example, higher bitrate requests on short routes are first attempted in higher-capacity bands with lower QoT performance, such as the E band [4]. This service differentiation idea has also been proposed as a network management option where each band can be considered a different virtual network slice for different traffic categories and services [15]. Others treat all connection requests in the same way, exploring the bands in the same order every time a new connection must be processed [2], [3], [5], [8], [9]. Service differentiation has the potential to outperform strategies without this feature [4], [6], [10].

Table I summarises the heuristics proposed to date to solve the dynamic RBMSA problem (considering fully dynamic or incremental traffic). For each, we describe: a) the order used to solve the sub-problems, b) whether they use pre-computed routes or not, c) the type of QoT evaluation used (offline or online), and whether the QoT of already established lightpath (VEL: Verification of Established Lightpaths) is carried out, d) whether service differentiation is offered and e) their computational complexity. For the computational complexity, K, M, B, S, and L represent the number of alternative paths, modulation formats, bands, spectrum slots, and links. G represents the computational complexity of evaluating the QoT of signals online ($O(S)$ or $O(S^2)$, depending on the computational implementation [2]). Algorithms with an asterisk in the second column only work for C+L scenarios. The static RBMSA problem (all demands are known beforehand by the RMSA solver) is out of the scope of this paper. An example of a static RMSA solver can be found in [16].

Currently, heuristics offer the best performance to solve the dynamic RBMSA problem. Further work is required to benchmark different heuristics to identify the best-performing ones or those with a good trade-off between computational complexity and performance. Until now, only an evaluation of different spectrum allocation approaches in the C+L scenario has been carried out in [17]. For dynamic RMSA solvers, the different assumptions and evaluation scenarios considered in the reported simulation experiments hamper a comprehensive comparison of heuristics in terms of performance and computational complexity. Additionally, although some initial work has focused on evaluating the energy consumption of progressive band exploitation [18], further work is still needed in this area as well as migration strategies beyond the C+L band scenario.

B. Deep Reinforcement Learning

Despite the increasing number of works reporting on applying deep reinforcement learning (DRL) algorithms to different dynamic provisioning problems in optical networks, there have been just a few works reporting results for dynamic MB-EONs [11]–[13]. In [12] a new multiband environment developed in the Optical RL-Gym [19] was reported without any comparison with heuristic approaches. In [11], the performance of a DQN agent in a multiband environment was reported, but it did not outperform a simple KSP-FF-FF heuristic. In both cases, the agent solved the R, B, and SA sub-problems by selecting a route (out of K precomputed routes), a band, and a spectrum block, respectively. The agent did not select the modulation format (the environment identified the most efficient modulation format for the selected route using the optical reach values available in [20]). Recently, a DRL-assisted solution, where the agent only solves the R sub-problem was proposed [13]. Unlike [11], [12], the agent in [13] does not select a route out of a set of pre-computed routes but selects a sequence of links. Simulation results show that the

DRL-assisted approach outperformed the KSP-FF-FF heuristic in the studied topology.

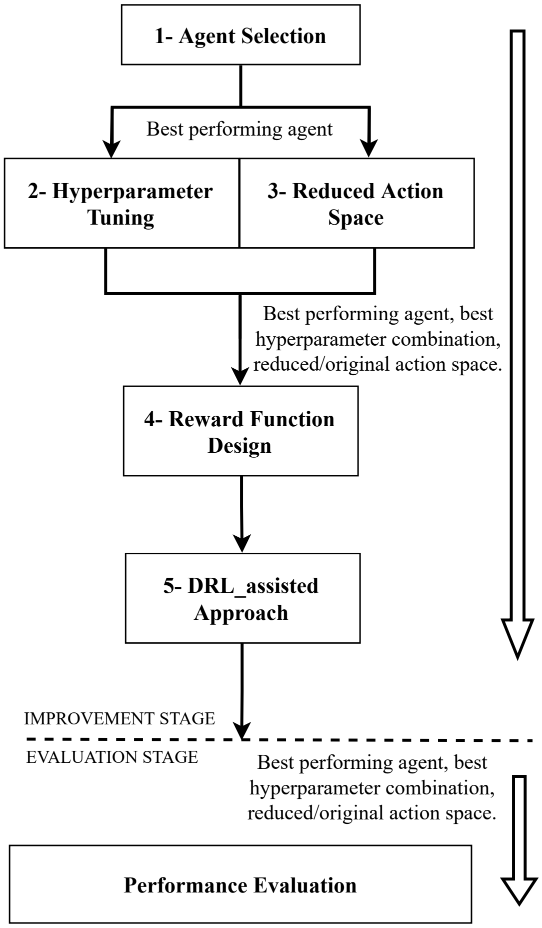

1- Agent Selection
Best performing agent
2- Hyperparameter Tuning
3- Reduced Action Space
Best performing agent, best hyperparameter combination, reduced/original action space.
4- Reward Function Design
5- DRL_assisted Approach
IMPROVEMENT STAGE
EVALUATION STAGE
Best performing agent, best hyperparameter combination, reduced/original action space.
Performance Evaluation

Fig. 1. Performance tuning process of a DRL framework for the dynamic RBMSA problem.

Given that no DRL agent solving the R, B, and SA subproblems has outperformed the KSP-FF-FF heuristic, we investigated what improvements to a DRL system could achieve better results. To do so, we applied agent, and environment-wise improvements to the original work reported in [12] for the C+L+S+E multiband scenario. Please, notice that we could not rely on automatic optimizers to improve the performance of the DRL system. On the agent side, automatic hyperparameter optimization was not possible (memory exhaustion), and automatic optimizers are unavailable on the environment side. Among the many approaches we could have taken, we have followed the one shown in Figure 1. We do not claim this approach to be near-optimal or even exhaustive (an exhaustive approach would have taken a prohibitively long time). It is just one out of several reasonable alternatives.

1) Agent-wise improvements:

- **Learning algorithm**: As shown in Figure 1 (step 1), we trained 6 agents with different learning algorithms. They were: TRPO, A2C, PPO2, ACKTR, DQN, and ACER. To do so, the default configuration available in the `Stable Baselines` library for each agent was used. The training scenario is described in Table II. TRPO was the best-performing learning algorithm in this stage with an average reward of 40.7, followed by ACER, A2C, PPO2, DQN, and ACKTR with an average reward of 39.8, 37.2, 37.1, 35.9, and 22.7, respectively.

TABLE II
TRAINING PARAMETERS AND SCENARIO

Parameters	*Value*
AGENT	
Connection requests per episode	50 [21]
Simulated requests per training	100,000
Agent's parameters	By-default [19]
ENVIRONMENT	
Network Parameters	
Topology	*NSFNet*
Number of FSU per band (W_B)	344(C), 480(L), 760(S), 1136(E)
Modulation Formats	BPSK, QPSK, 8QAM, 16QAM 32QAM, 64QAM, 256QAM
Traffic Parameters	
Bit rates [Gb/s]	Randomly elected [10, 40, 100, 400, 1000] Gbps
Mean Holding $(1/\mu)$	200 requests per time unit
Mean Interarrival $(1/\lambda)$	1/5 time units
Traffic load	1000 Erlang
Reward function (RF0)	+1 (request accepted), -1 (request rejected)

- **Hyperparameter tuning (neural network)**: Although some authors recommend parameter adjustment using complex deep learning models [22], the computational effort of such approaches is high. Our experience using an optimizer (Optuna) led to memory exhaustion after 12 hours of computation. Thus, we manually selected a range for tuning the neurons and layers of the neural network that outputs the policy of TRPO (the best-performing agent in the previous step). We evaluated the accumulated reward for a number of layers equal to 2, 5, 10, 15, 25, and 50, and a number of neurons equal to 50, 128, 200, and 448. These values were selected taking the combination of neurons and layers used in [19], [21] (5 layers/28 neurons) as a baseline. The highest reward was achieved with 2 layers and 200 neurons.

2) Environment-wise improvements:

- **Reduced action space**: As shown in Figure 1, in parallel with hyperparameter tuning, we explored whether reducing the action space had a significant impact on the performance of the selected agent (TRPO), since previous work in single band networks worked with just 100 slots in the C-band [19], [21], [23]. In this case, we reduced the number of slots available at each band: the C-band capacity was reduced to 100 slots, as in [19], [21], and the capacity of the rest of the bands was decreased proportionally. This led to 140, 220, and 330 slots for the L, S, and E band, respectively. We kept the neural network parameters in 5 layers and 128 neurons as suggested in [12], [19], [21]. Results showed no performance improvement and hence, this technique was discarded from further experiments.
- **Reward function**: Previous work has shown the impact of different reward functions on the performance of DRL agents [23], [24]. Thus, we studied the impact of 5 reward functions - termed $RF0$ to $RF4$ - on the

performance of the DRL system. The baseline reward function ($RF0$) is the one described in the bottom row of Table II. The remaining functions were proposed in [24] and they consider band utilization ($RF1$), route/band slot availability ($RF2$), spectrum compactness ($RF3$) and mixed compactness-availability ($RF4$) to determine the reward sent to the agent. Due to space constraints we cannot ellaborate further on the reward functions (see [24] for further details).

$$RF1 = \begin{cases} 2 & \text{request accepted in most available band} \\ 1 & \text{request accepted in any other band} \\ -1 & \text{request rejected} \end{cases} \tag{1}$$

$$RF2 = \begin{cases} \widetilde{A}(B) & \text{, request accepted} \\ -1 & \text{, request rejected} \end{cases} \tag{2}$$

where $\widetilde{A}(B)$ is average percentage of frequency slots available in the selected band/route links.

$$RF3 = \begin{cases} 1 - \frac{\widetilde{C}}{W_B} & \text{, request accepted} \\ -1 & \text{, request rejected} \end{cases} \tag{3}$$

where $\widetilde{C}$ is the average compactness of the links of the selected route, which can be thought of as the chance of finding available contiguous frequency slots on it, and W_B is the number of slots of the selected band B shown in Table II [20].

$$RF4 = \begin{cases} \left(1 - \frac{\widetilde{C}}{W_B}\right) + \widetilde{A} & \text{, request accepted} \\ -2 & \text{, request rejected} \end{cases} \tag{4}$$

where $\widetilde{A}$ is the availability for all bands on a given path. $RF4$ combines $RF2$ and $RF3$, rewarding higher those paths with low compactness and multiple available FSUs. The TRPO agent was trained with the different rewards using the original number of slots per band (since reducing the space action did not lead to better performance) and the best combination of hyperparameters found (2 layers, 200 neurons). For a fair comparison, reward values of the different functions were normalised using the *Standard Scaler* tool from the SkLearn Python library. As a result, values between -1 and 1 are obtained. The difference between different rewards was very small, but $RF1$ showed slightly better performance and thus, it was the reward function selected for the next steps.

- **DRL-assisted solver**: Finally, we tested the impact of decreasing the number of sub-problems solved by the agent. To do so, we made the agent to select one out of K pre-computed routes (R) and the band (B) whilst the spectrum allocation (SA) task was solved by the First-Fit heuristic. A TRPO agent was trained with the original number of slots per band, the reward function $RF1$ and the hyperparameters tuned for the original problem (2 layers, 200 neurons).

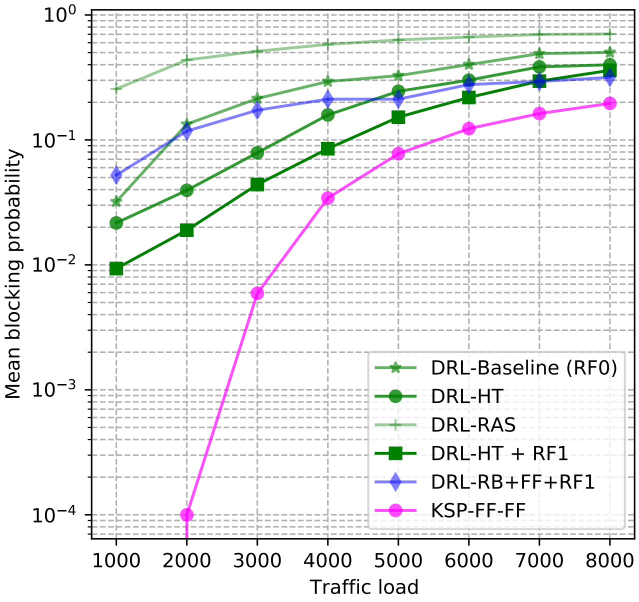
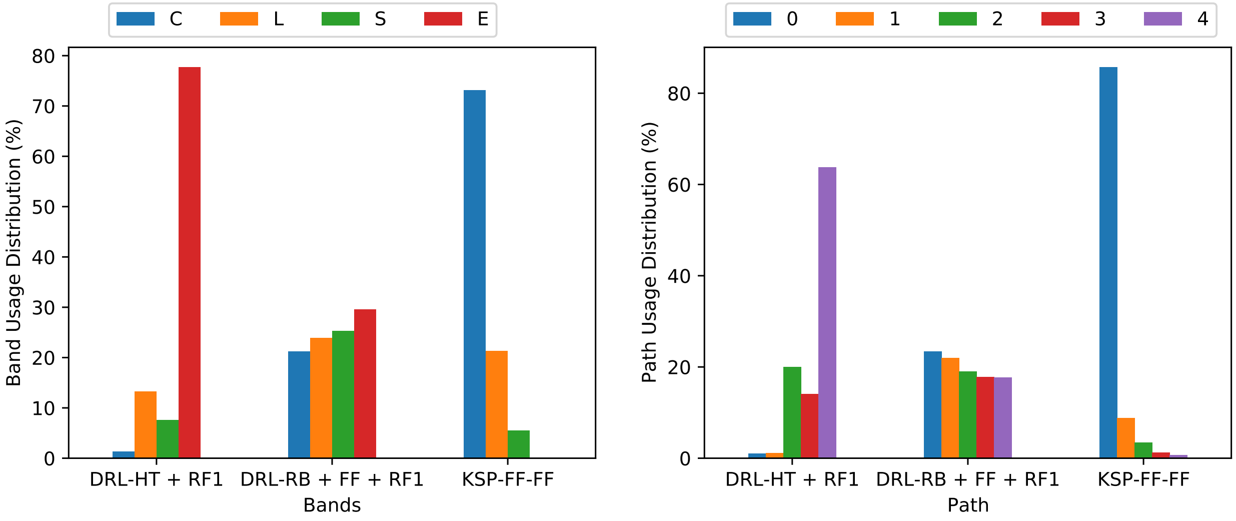


Fig. 2. Blocking probability versus traffic load for different DRL systems and KSP-FF-FF heuristic.

Figure 2 summarises the results obtained with the different improvements described above for a C+L+S+E scenario. The figure shows the network blocking performance for traffic loads ranging from 1000 to 8000 Erlang in the *NSFNet* topology. The colours of the curves represent 3 different approaches: heuristic (pink), DRL (shades of green) and DRL-assisted (blue). The heuristic used as a baseline for comparison was KSP-FF-FF: K-Shortest Path (using the same K routes made available to the DRL agents) for the routing problem, and the First Fit algorithm for both, the band and spectrum allocation problem. K was set 5, and the order of band selection was C, L, S, and E. The names of the DRL curves are: DRL-Baseline (TRPO agent with the original training parameters described in Table II), DRL-HT (TRPO agent with hyperparameter tuning: number of neurons/layers changed to 200/2); DRL-RAS (TRPO with reduced action space, remaining training parameters and scenario as described in Table II); DRL-HT+RF1 (TRPO with hyperparameters tuned and reward $RF1$, remaining training parameters and scenario as described in Table III) and DRL-RB+FF+RF1 (DRL assisted approach) where the agent solves the R and B problems only (TRPO with hyperparameters tuned and reward $RF1$, remaining training parameters and scenario as described in Table III), and FF is used for spectrum assignment. In all cases, an offline QoT computation was used. The QoT was obtained using the ISRS GN-model following the methodology from [4], [20]. Estimations were performed assuming a worst-case scenario: a fully loaded C+L+S+E transmission line and the channel with the lowest SNR in each band was used to represent the band QoT.



(a) Band usage distribution vs reward functions

(b) Path usage distribution vs reward functions

Fig. 3. Results obtained using different strategies: (a) Band usage distribution (b) Path usage distribution.

DRL-RAS exhibits a blocking probability even higher than DRL-Baseline, making this approach the worst performing one. It can also be seen that although hyperparameter tuning decreases the blocking probability (DRL-HT), the best results are achieved with the combination of hyperparameter tuning and the reward function $RF1$ (DRL-HT+RF1). However, none of the previous strategies outperform the heuristic, especially at low values of traffic loads (1000-3000 Erlangs). This is an unexpected result since previous works applying DRL in the more straightforward single-band problem outperformed the heuristic approach by either changing the reward function [23] or simplifying the problem solved by the agent [13]. However, in the latter case, this was done by making the agent solve a single problem (R) rather than two problems (R and B) as in our case.

To attempt an explanation for the poor performance of the agents, we studied how the selection of routes and bands was done by the best-performing agent (DRL-HT+RF1) and the agent that only solves the routing and band allocation problem (DRL-RB+FF+RF1). The heuristic was also added for comparison. Figure 3 shows the percentage of usage of the different bands by the different approaches (left) and paths (right), respectively. Very different strategies can be observed. In terms of the band, DRL-HT-RF1 has a clear preference for band E (the band with a higher number of slots). However, this band has the worst QoT. Instead, DRL-RB-FF achieves a well-balanced usage of bands as opposed to the heuristic that prioritizes the use of band C. In terms of routes, a similar pattern is observed: DRL-HT-RF1 prioritizes the longest routes, DRL-RB-FF makes balanced use of them, whiles KSP-FF-FF prioritizes the usage of the shortest routes. Clearly, the agents have failed to find novel and effective strategies to allocate resources.

These results highlight two important facts. First, as the problem size increases, achieving a DRL agent that outperforms the best heuristics becomes more challenging. To date, there are still no reports of DRL agents solving the resource allocation problem in multiband and multicore elastic optical networks, where an extra dimension is added to action space (core selection) with respect to the problem solved here. Based on our results, achieving a satisfactory performing agent for that problem will be hard since reinforcement learning requires significant processing power and memory. Second, identifying the improvement strategies that most impact the DRL performance is essential. Since applying DRL to allocation problems in optical networks is a recent line of research, we still do not know what strategies guarantee improvement. Strategies that work in one environment have proved not to work as well in others. Further research on identifying generic winning strategies will shed light on how to design computationally simple optimization methods that guide the tuning process with some guarantee of success, a key step for the optical research community to benefit from the potential of DRL.

III. Conclusions and Future Remarks

This paper reviewed current strategies to address the resource allocation problem on dynamic multiband elastic optical networks (MB-EON): *ad-hoc algorithms (heuristics)* and reinforcement learning-based solutions, with the former (still) outperforming the latter.

Although heuristics proposed to date outperform current DRL approaches, they are difficult to rank due to the diversity of assumptions and evaluation scenarios. Further work is needed in terms of identifying the best-performing heuristic or developing tools that allow quick performance evaluation of different approaches for a given network scenario, considering computational complexity or execution times constraints. Additional work is also required in terms of power consumption and migration strategies for multiband networks.

The best-performing heuristics are usually the ones with higher computational complexity. Hence, DRL approaches are attractive since once trained, their execution time is extremely low (a requirement for dynamic environments). Additionally, DRL agents might find novel and effective strategies. However, to date, this has not been the case in MB-EONs. The very few current DRL approaches proposed so far still lag behind heuristics, with the exception of DRL-assisted approaches where the agent is in charge of solving the routing problem only. While this is a welcome addition to the pool of DRL approaches for the dynamic RBMSA problem, leaving the spectrum resource allocation to a heuristic might still be undesirable, since their execution time increases linearly with the number of frequency slots.

Further research on improved DRL approaches for the dynamic RBMSA problem is needed. Such research can be carried out along different lines: **a)** Performance tuning. The many tuning possibilities (agent selection, hyperparameter tuning, reduced action space, reward function design, among others) make this a time/processing-consuming endeavor. Therefore, identifying the set of effective improvement strategies is key to help researchers find competitive DRL solutions. **b)** DRL-assisted approaches. This has been little explored so far for the dynamic RBMSA problem. Performance evaluation of different DRL-assisted approaches (e.g. only R with precomputed routes; only R from scratch, only RB, etc.) would shed light on the level of complexity where a DRL agent can bring significant benefits with respect to heuristics. **c)** Multi DRL agents approaches. Having different DRL agents address the different sub-problems of the dynamic RBMSA might lead to a competitive performance. This research has not been reported yet.

References

[1] R. K. Jana, A. Mitra, A. Pradhan, K. Grattan, A. Srivastava, B. Mukherjee, and A. Lord, "When is operation over C+L bands more economical than multifiber for capacity upgrade of an optical backbone network?," in *2020 European Conference on Optical Communications (ECOC)*, pp. 1–4, 2020.

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

[3] M. Nakagawa, H. Kawahara, K. Masumoto, T. Matsuda, and K. Matsumura, "Performance evaluation of multi-band optical networks employing distance-adaptive resource allocation," in *2020 Opto-Electronics and Communications Conference (OECC)*, pp. 1–3, IEEE, 2020.

[4] 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 Communications Letters*, vol. 26, no. 2, pp. 379–383, 2022.

[5] D. Uzunidis, E. Kosmatos, C. Matrakidis, A. Stavdas, and A. Lord, "Strategies for upgrading an operator's backbone network beyond the c-band: Towards multi-band optical networks," *IEEE Photonics Journal*, vol. 13, no. 2, pp. 1–18, 2021.

[6] F. Calderón, D. Bórquez-Paredes, N. Jara, R. Olivares, A. Leiva, A. Beghelli, and G. Saavedra, "Dynamic resource allocation in different ultrawideband optical network topologies," in *2022 IEEE Photonics Society Summer Topicals Meeting Series (SUM)*, pp. 1–2, 2022.

[7] C. Wang, N. Yoshikane, and T. Tsuritani, "Towards more accurate and effective service provision in multiband transport networks," in *2022 European Conference on Optical Communication (ECOC)*, pp. 1–4, IEEE, 2022.

[8] B. Bao, H. Yang, Q. Yao, C. Li, Z. Sun, J. Zhang, S. Liu, and Y. Li, "Lorb: Link-oriented resource balancing scheme for hybrid c/c+ l band elastic optical networks," *Optical Fiber Technology*, vol. 74, p. 103071, 2022.

[9] Q. Yao, H. Yang, B. Bao, J. Zhang, H. Wang, D. Ge, S. Liu, D. Wang, Y. Li, D. Zhang, *et al.*, "Snr re-verification-based routing, band, modulation, and spectrum assignment in hybrid c-c+ l optical networks," *Journal of Lightwave Technology*, vol. 40, no. 11, pp. 3456–3469, 2022.

[10] R. K. Jana, B. C. Chatterjee, A. P. Singh, A. Srivastava, B. Mukherjee, A. Lord, and A. Mitra, "Quality-aware resource provisioning for multiband elastic optical networks: a deep-learning-assisted approach," *Journal of Optical Communications and Networking*, vol. 14, no. 11, pp. 882–893, 2022.

[11] N. E. D. El 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 *2021 International Conference on Optical Network Design and Modeling (ONDM)*, pp. 1–3, IEEE, 2021.

[12] P. Morales, P. Franco, A. Lozada, N. Jara, F. Calderón, J. Pinto-Ríos, and A. Leiva, "Multi-band environments for optical reinforcement learning gym for resource allocation in elastic optical networks," in *2021 International Conference on Optical Network Design and Modeling (ONDM)*, pp. 1–6, IEEE, 2021.

[13] 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 *European Conference and Exhibition on Optical Communication*, pp. Tu5–63, Optica Publishing Group, 2022.

[14] U. Zwick, "Exact and approximate distances in graphs—a survey," in *European Symposium on Algorithms*, pp. 33–48, Springer, 2001.

[15] V. Curri, "Multiband optical transport: a cost-effective and seamless increase of network capacity," in *Photonic Networks and Devices*, pp. NeTu2C–3, Optical Society of America, 2021.

[16] M. Mehrabi, H. Beyranvand, and M. J. Emadi, "Multi-band elastic optical networks: inter-channel stimulated raman scattering-aware routing, modulation level and spectrum assignment," *Journal of Lightwave Technology*, vol. 39, no. 11, pp. 3360–3370, 2021.

[17] R. K. Jana, B. C. Chatterjee, A. P. Singh, A. Srivastava, B. Mukherjee, A. Lord, and A. Mitra, "Performance evaluation of conventional spectrum-allocation policies for c+ l band elastic optical networks," in *2021 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)*, pp. 348–353, IEEE, 2021.

[18] R. Sadeghi, B. Correia, A. Souza, N. Costa, J. Pedro, A. Napoli, and V. Curri, "Transparent vs translucent multi-band optical networking: Capacity and energy analyses," *Journal of Lightwave Technology*, vol. 40, no. 11, pp. 3486–3498, 2022.

[19] C. Natalino and P. Monti, "The optical rl-gym: An open-source toolkit for applying reinforcement learning in optical networks," in *2020 22nd International Conference on Transparent Optical Networks (ICTON)*, pp. 1–5, 2020.

[20] E. Paz and G. Saavedra, "Maximum transmission reach for optical signals in elastic optical networks employing band division multiplexing," 2021.

[21] X. Chen, B. Li, R. Proietti, H. Lu, Z. Zhu, and S. B. Yoo, "Deeprmsa: a deep reinforcement learning framework for routing, modulation and spectrum assignment in elastic optical networks," *Journal of Lightwave Technology*, vol. 37, no. 16, pp. 4155–4163, 2019.

[22] X. Liu, J. Wu, and S. Chen, "Efficient hyperparameters optimization through model-based reinforcement learning and meta-learning," in *2020 IEEE 22nd International Conference on High Performance Computing and Communications; IEEE 18th International Conference on Smart City; IEEE 6th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pp. 1036–1041, IEEE, 2020.

[23] 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 Communications Letters*, vol. 26, no. 11, pp. 2675–2679, 2022.

[24] M. Gonzalez, F. Condon, P. Morales, and N. Jara, "Improving multi-band elastic optical networks performance using behavior induction on deep reinforcement learning," in *2022 IEEE Latin-American Conference on Communications (LATINCOM)*, pp. 1–6, IEEE, 2022.