Check for updates

International Journal of Satellite Communications and Networking

WILEY

SPECIAL ISSUE PAPER OPEN ACCESS

Path-Based Deep Reinforcement Learning for On-Board Routing in Satellite Constellation Networks

Manuel M. H. Roth[1] | Thomas Jerkovits[1] | Anupama Hegde[1] | Thomas Delamotte[2] | Andreas Knopp[2]

[1]Institute of Communications and Navigation, German Aerospace Center (DLR), Cologne, Germany | [2]Chair of Signal Processing, University of the Bundeswehr Munich, Munich, Germany

Correspondence: Manuel M. H. Roth (manuel.roth@dlr.de)

Received: 31 August 2025 | **Revised:** 3 January 2026 | **Accepted:** 3 February 2026

Keywords: deep reinforcement learning | low earth orbit | on-board network management | routing | satellite constellation networks | traffic engineering

ABSTRACT

Efficient usage of available network resources is a crucial factor for broadband services in interconnected satellite constellations. To meet required quality of service standards under heavy network loads, it is essential to optimize traffic distribution among the intersatellite links. To address this challenge, we propose an adaptive traffic engineering framework based on deep reinforcement learning. Our approach employs a path-based decision-making strategy, using a centralized agent to distribute incoming flow requests on a set of candidate paths. This method approximates optimal solutions to the multicommodity flow problem with relatively low computational complexity, making it suitable for in-space network control despite on-board processing limitations. The performance of the proposed scheme is evaluated against state-of-the-art rule-based benchmarks in various scenarios. We quantify the impact on performance of different candidate path sets and traffic patterns. Overall, the proposed solution presents a viable approach for optimizing flow distribution in satellite constellation networks, suitable for the integration into the controller logic of software-defined networks.

1 | Introduction

In recent years, low Earth orbit (LEO) satellite constellation networks (SCNs) have experienced increasing interest, in particular in the context of broadband traffic services [1]. SCN designs typically rely on intersatellite links (ISLs) to create mesh networks, reducing the required number of ground stations and increasing service coverage [1, 2]. Satellites thus act as routing devices, which require configuration for efficient data transmission through the resulting space-borne network. However, depending on the routing approach, the traffic patterns resulting from the User Terminals (UTs) and Gateway stations (GWs) distributions on ground can cause bottlenecks. So, even if high-bandwidth ISLs are considered, congestion can decrease Quality of Experience (QoE) of users [2].

Therefore, various routing protocols tailored to the unique characteristics of nonterrestrial networks (NTNs) have been proposed. For instance, 6G standardization aims to seamlessly integrate the space segment with terrestrial services [3]. To handle the increasing number of users and broadband traffic volume, novel routing approaches specifically tailored to SCNs become thus more relevant. Given that 6G aims to be an AI-native network of networks, suitable routing frameworks are of interest.

A key paradigm in this context is software defined networking (SDN). In short, it offers improved flexibility through software and virtualization, scalability, fast reconfiguration, increased resiliency, and service-aware quality of service (QoS) handling [4, 5]. However, due to the physical size

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *International Journal of Satellite Communications and Networking* published by John Wiley & Sons Ltd.

of the constellation networks, significant end-to-end propagation delays of more than 100 ms are possible. Thus, distributed SDN-based approaches relying on in-space control have been proposed to enable higher reactivity and localized decision-making [6].

In order to maximize throughput by finding optimal routing configurations, the underlying multicommodity flow problem (MCFP) has to be solved. It formalizes the distribution of routes for a set of commodities with varied source-destination pairs and associated demands. The corresponding computation is complex, the integer MCFP is NP-hard in general [7, 8]. Given the limited on-board processing, finding optimal solutions for in-space routing is thus impractical. As flows typically arrive concurrently and subsequently, making meaningful step-wise choices is advantageous.

We have previously proposed a deep reinforcement learning (DRL)-based approach to approximate solutions to the MCFP with reduced complexity [9]. The approach is based on a locally centralized control unit which distributes incoming flows on sets of candidate paths. By maximizing expected future rewards, DRL can effectively handle the overlapping flow allocation subproblems. Therefore, it represents a well-suited method to enable a proactive traffic distribution that anticipates potential bottlenecks. Unlike greedy approaches or heuristics, DRL can learn from experience, adapt to changing conditions, and discover nonobvious traffic patterns.

In this work, we extend and refine our previous architecture, demonstrating its enhanced performance across diverse scenarios. Key improvements include a novel method for gathering candidate paths and a more compact observation space representation, resulting in improved outcomes with smaller neural networks. Furthermore, we investigate the suitability of this approach to various heat maps and larger network topologies. The main contributions of this work are as follows:

- We extend our previous path-based DRL approach [9] by introducing a new, simplified observation space based purely on aggregated path characteristics (e.g., minimum, maximum, average, variance, normalized path length), as well as global edge statistics (see Section 5.4). This observation space is entirely independent of the underlying graph topology, reducing complexity and improving generalization.
- We propose a new strategy for candidate path selection using four fixed path types: (i) shortest path (hop count), (ii) least-congested path (minimizing total utilization), and (iii–iv) two softmin-based selections with different variance scaling factors (see Section 5.3). This method outperforms the adaptive candidate selection used in our previous work [9].
- We conduct a comprehensive evaluation of the proposed approach across diverse network scenarios, including different traffic heat maps and larger network topologies. The results show that our method is not only competitive with state-of-the-art DRL and heuristic baselines but also resilient to shifts in traffic distribution and scalable to larger graphs.

The work is structured as follows. In Section 2, we summarize related works in the field of routing in satellite networks. The considered reference system and its relevant characteristics are provided in Section 3. The DRL background and the theoretical framework are presented in Section 4. Subsequently, we present the proposed architecture in Section 5. In Section 6, the simulator environment, comparative quantitative analyses, as well as comparisons to benchmark solutions are provided. The strengths and limitations of the approach as well as the implications for further developments are discussed in Section 7. Finally, the paper is concluded in Section 8.

2 | Related Works

For routing in wireless communications, various machine learning (ML)-based approaches have been explored [10], often facilitated by SDN frameworks [11]. In the context of SCNs, a complete centralization of network functionality is difficult, so adapted SDN-based schemes are being investigated [12]. For reconfigurable in-space network control in LEO constellations, distributed SDN architectures with flat hierarchies have been proposed [6]. This framework has also been investigated in the context of load-balancing with a focus on rule-based path selection making use of the inherent network structure [2]. In this work, we extend this architecture by enabling the in-space controller to approximate solutions to the MCFP via ML techniques.

In general, there is a wide array of research focusing on the MCFP [13]. Most commonly heuristic approaches are applied for feasible computations. This is also the case for SCNs, for which a suitable linear programming (LP) scheme has been investigated [14]. A dynamic programming (DP) formulation of intended MCFP optimization is provided in the derivation of our approach.

Recently, instead of explicitly defining heuristics, ML has been applied to directly learn characteristics of the problem in the given scenario. Various DRL-based routing schemes have been proposed, including hop-based deep Q-network (DQN) methods [15, 16]. However, the need to learn stable, loop-free paths can pose problems in terms of scalability and robustness. These issues have been addressed by constraints [16] or by hierarchical architectures [17]. Moreover, a hop-based DRL approach has been proposed for the MCFP [18]. We adopt a similar concept but avoid hop-based decision-making, using paths as actions.

In this context, the size of the set of paths can become large. To minimize overhead, K-means and cosine similarity approaches have been proposed for creating a set of probable paths, from which the least congested one is chosen [19]. This effective reduction of considered paths is applied in the proposed approach to enable fixed-size action spaces.

A path-based approach for load-aware routing in SCNs has also been proposed [20]. While it also utilizes clustering and path sets within a double deep Q-network (DDQN) architecture, their focus is primarily on QoS metrics (latency and bottleneck avoidance), not the MCFP directly.

15420981, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sat.70043 by Universitat der Bundeswehr Munchen, Wiley Online Library on [24/06/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 | Comparison of related architectures and routing approaches including their design characteristics for LEO SCNs.

Approach category	Decision	Core technique	Demand	On-board cost	References
Distributed SDN for SCNs	Path	Dijkstra's, k-best paths	Online	Low	Papa et al. [6], Roth et al. [2]
Heuristic MCFP solutions	Path	LP, DP	Offline	High	Grislain et al. [14]
Hop-based DQN	Hop	DQN (value-based)	Online	Low-medium	Zuo et al. [15], Hegde et al. [16]
Hop-based DRL for MCFP	Hop	Actor–critic	Online	Medium	Tsai et al. [18]
Least-congested from set	Path	k-means, cosine similarity	Online	Low	Kumar et al. [19]
Path-based QoS-aware DRL	Path	DDQN	Online	Medium	Cigliano et al. [20]
GNN-enhanced DRL	Hop/path	GNN and DRL	Online	Medium-high	Geyer et al. [21], Wang et al. [22]
Path-based Soft Actor–C ritic (this work)	Path	SAC, topology-agnostic	Online	Low	Roth et al. [9]

Various preprocessing concepts have been investigated, including graph neural networks (GNNs), which aim to extract structural features of the network for improved decision-making [21]. These embeddings are typically passed to a DRL agent as part of its observation space [22]. However, such methods remain tightly coupled to the graph structure with significant compute overhead. They are primarily suited to static or fully known networks, aligning with the assumptions of the offline MCFP.

In this work, we target online or ad-hoc routing scenarios, where demands are not known in advance and decisions must be made with local or partial information. Here, the MCFP serves only as an upper bound, not as an achievable solution. Our approach avoids expensive GNN-based feature extraction reliant on the full graph. Instead, we utilize a simplified, topology-agnostic observation space focusing on path states and demand characteristics. By focusing on candidate paths and learning to allocate flows efficiently, we demonstrate a scalable, adaptable, and lightweight alternative well-suited for real-world space networks.

An overview over the related works is provided in Table 1. Here, offline means the requests are known a priori. While prior research focused on demonstrating the feasibility of the approach [9], this work concentrates on refining the design and evaluating its performance in diverse scenarios.

3 | Reference System

As a baseline for the considered ML-based method, we assume an SDN-based network control architecture for SCNs. The proposed approach is integrated within the routing logic of the controller, enabling informed decision-making. While this investigation focuses on subsequent flow allocations, the approach can also be modified to apply proactive routing configurations.

TABLE 2 | Summary of an exemplary near-polar Walker star constellation consisting of 288 satellites [2, 23].

Space segment characteristic	Example constellation
Number of satellites	288
Number of planes	12
Number of satellites per plane	24
Satellite altitude (km)	780
Orbital inclination (°)	86.4
Cross-seam planes spacing (°)	30
Corotating planes spacing (°)	15
Phase offset corotating planes (°)	7.5
Interplane ISL shut-down latitude (°)	80
Number of ISLs per satellite	4

For SCNs in particular, given their physical size, distributed in-space control has been proposed to enhance reactivity and scalability [2, 5]. For this investigation, we adopt an architecture where the network is subdivided into domains (or clusters) of 48 satellites with dedicated controllers. We focus on intradomain routing, resulting in an environment consisting of 48 nodes. To assess performance in larger networks, we also consider a domain comprising 288 satellites. This is representative of a complete exemplary SCN. The considered Walker star constellation [23] is summarized in Table 2. A 48-node domain in this reference system is illustrated in Figure 1.

15420981, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sat.70043 by Universitat der Bundeswehr Munchen, Wiley Online Library on [24/06/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

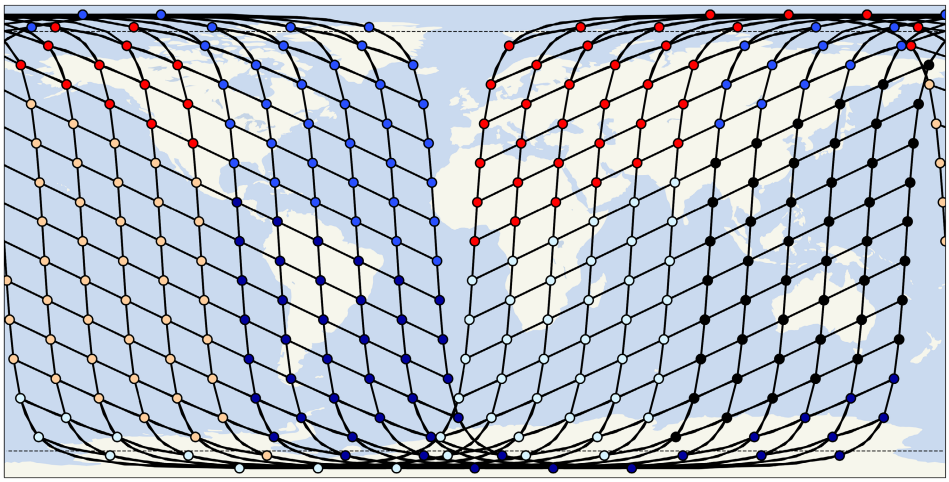

FIGURE 1 | Reference satellite constellation network. A polar constellation of 288 satellites is assumed. The red area highlights a potential SDN domain of the network in the context of a distributed in-space architecture. Each domain consists of 48 satellites [2, 24].

3.1 | Network Connectivity

We assume that each satellite of the constellation possesses four ISLs: two intraplane ISLs and two interplane ISLs. The bandwidth and availability of all ISLs are considered equal in this work. Based on this setup, a grid network emerges. We also omit potential seams due to counter-rotating planes and interplane ISL shutdown latitudes.

Notably, an inherent path diversity is characteristic for larger SCNs. For multihop transmissions, multiple viable paths of comparable end-to-end characteristics are oftentimes available. This enables an efficient distribution of incoming flows based on an understanding of underlying traffic patterns.

The ground segment impacts the intended performance evaluation only indirectly. Although the heat maps used in this investigation correspond to systems consisting of both UTs and GWs, we focus exclusively on the sampled flow requests.

3.2 | Alignment With 6G Functional Split

The deployment of 5G-RAN and 5G-core functions across ground and space segments enables flexible routing architectures for multi-orbit constellations. A critical design aspect is the functional split [25, 26], which governs the distribution of protocol stack functions between network entities. For 6G-NTN, this determines where routing computation occurs [3].

Of particular interest for in-space routing, and thus for the proposed approach, is functional split option 2. Within this standard-compliant option, the control layer (control plane) is generally centralized at the Central Unit (CU), while lower layers (data plane) remain distributed. This enables a centralized path computation with global network visibility as well as distributed path execution via programmable forwarding rules. Real-time adaption through control plane signaling, as in SDN, is possible.

However, the distributed units (DUs), which reside on satellites and handle data plane functionalities, can also take over some routing-related capabilities. Therefore, they can act as distributed SDN controllers as foreseen in our architecture. Overall, the functional split is generally suitable for ML-based routing using centralized learning and decentralized control, where the central entity has global knowledge of the SCN.

The proposed path-based SAC method directly integrates into this architecture:

- **Training:** SAC agent is trained centrally using global state data (topology, demands, link states)
- **Deployment:** Optimized policy can be deployed to on-board SDN controllers via 3GPP control plane
- **Execution:** Satellites apply path selections autonomously using:
  Topology-agnostic observations (path states, demand patterns)
  Potentially local telemetry for real-time adaptation

This setup enables efficient flow distribution while maintaining latency benefits of functional split Option 2: Ground round-trips are avoided, and control plane overhead is minimized.

Current 3GPP work on Option 2 variants [27] explicitly support containerized network functions for on-board ML as well as control plane extensions for policy distribution. Therefore, our approach complies with the direction of the standard while addressing throughput demands via optimized multicommodity flow allocation.

3.3 | Multicommodity Flow Problem

In order to propose paths resulting in routing configurations which enable the highest possible throughput, approaches which solve or approximate solutions to the MCFP have to be formulated. The MCFP provides a framework for modeling and

15420981, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sat.70043 by Universitat der Bundeswehr Munchen, Wiley Online Library on [24/06/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

solving the network flow distribution when multiple traffic demands (commodities) must be routed through a network with limited capacity.

If integer flows are considered, as in the integer multicommodity flow problem (I-MCFP), the problem is NP-complete [8]. The fractional multicommodity flow problem (F-MCFP), on the other hand, can be solved in polynomial time using LP [8]. Nevertheless, for increasingly large networks, the problem size grows rapidly. Therefore, approximate solutions may be considered for large networks since complete LP approaches are costly.

As the overall objective is to support as many incoming flows as possible without saturating links, we focus on a minimization of the maximum link utilization. The *utilization* $u(i,j)$ is the total flow on the edge (i,j) divided by its capacity $c(i,j)$:

$$u(i,j) = \frac{f(i,j)}{c(i,j)} \tag{1}$$

There are other objectives which can be considered, such as minimizing the average link utilization, but we will focus on avoiding saturation as packet drops are particularly costly in SCNs.

We omit a detailed description of the MCFP and its associated LP constraints for static networks, as these have been covered in our prior work [9]. Instead, we focus directly on dynamic networks, where the set of flows is subject to a temporal order. Accordingly, the graph is expanded over time steps t to model these dynamics. Thus, flows on edges depend on t:

$$f_t^k(i,j) \quad \text{for } t \in T \tag{2}$$

Similarly, the link capacity $u_t(i,j)$ is now time-dependent. For certain scenarios, the link capacities also change over time; hence, we use $c_t(i,j)$.

$$u_t(i,j) = \frac{f_t(i,j)}{c_t(i,j)} = \frac{\sum_{(i,j)\in E} f_t^k(i,j)}{c_t(i,j)} \tag{3}$$

There are two options to formulate the objective function of minimizing the maximum link utilization over time. First, by extending the problem over all time steps. This represents the general approach of minimizing in all time steps:

$$\min \max_t \max_{(i,j)\in\mathscr{E}} u_t(i,j) \tag{4}$$

Or, secondly, by minimizing the time-averaged maximum link utilization over the time horizon T:

$$\min \frac{1}{T} \sum_{t=1}^{T} \max_{(i,j)\in\mathscr{E}} u_t(i,j) \tag{5}$$

As the network states are directly impacted by the route selection, any path selection effects the subsequent routing decisions. So, if we want to optimize the overall performance over time, the sub-problems of individual routing decisions overlap. For such sequential decision-making, DP is a suitable approach. To find solutions, good knowledge of the network and its state evolution is required.

We denote the state variable s_t describing the network state at time t. In addition, we consider the matrix of flows $\overrightarrow{F_t}$ which contains all flows $f_t^k(i,j)$ at time t. As before, the decision variables are $f_t^k(i,j)$. Moreover, a transition function between time steps is required, which updates the current state based on the flows.

$$s_{t+1} = \Phi(s_t, f_t) \tag{6}$$

To formulate a suitable policy, the value function V_t is used to evaluate potential next states. This function considers the resulting cost, represented by a cost function C_t for the current moment in time, as well as the impact on the value of future states. So, the value function of the current state is based on the value function of subsequent states. Consequently, the computation relies on solving recursive subproblems. Leveraging the Bellman equation enables a formulation of an optimal policy:

$$V_t(s_t) = \min_{f_t} \left[C_t(s_t, f_t) + V_{t+1}(s_{t+1}) \right] \tag{7}$$

Given our objective function, the immediate cost function C_t is as follows:

$$C_t(s_t, f_t) = \max_{(i,j)\in\mathscr{E}} u_t(i,j) \tag{8}$$

So, for the complete time horizon from time t to T, the optimal value function representing the minimal total cost is as follows:

$$V_t(s_t) = \min_{f_t,\ldots,f_T} \left(\max_{t'\geq t} \left[\max_{(i,j)\in\mathscr{E}} u_t(i,j) \right] \right) \tag{9}$$

This DP formulation implicitly assumes several simplifications. It does not include delay functions, queuing models, and most importantly stochastic traffic models. Therefore, formulating and solving DP for the considered SCN scenario are difficult [2].

In addition, using DP results in high computational complexity due to the state of dimensionality [28]. The state space grows exponentially with the number of input parameters. So, for large state spaces, DP approaches may not be applicable in practice. Given the amount of parameters, heuristics are typically considered to make the problem manageable [14].

3.4 | Approximating Least Congested Path Computation

To approximate the least congested path, we reformulate the classical min-max path selection problem as a differentiable min-sum formulation. For each edge $(i,j) \in \mathscr{E}$ with current utilization $u(i,j) \in [0,1]$, we define a nonlinear edge weight:

$$w_\alpha(i,j) = \exp(\alpha \cdot u(i,j)),$$

15420981, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sat.70043 by Universitat der Bundeswehr Munchen, Wiley Online Library on [24/06/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

where $\alpha > 0$ is a scaling parameter that controls the sharpness of the approximation.

This formulation is inspired by the *log-sum-exp* identity, which provides a smooth approximation of the maximum:

$$\log\left(\sum_i \exp(\alpha \cdot x_i)\right) \approx \max_i x_i \quad \text{as } \alpha \to \infty \tag{10}$$

a well-known technique in optimization and machine learning [29].

Using this transformation, the cost of a path $p \subset \mathscr{E}$ becomes the following:

$$C(p) = \sum_{(i,j)\in p} \exp(\alpha \cdot u(i,j))$$

and minimizing this cost approximates minimizing the maximum link utilization along the path:

$$\log C(p) \approx \max_{(i,j)\in p} u(i,j)$$

We use this formulation to compute alternative candidate paths with varying sensitivity to congestion. These paths are included in the set of routing options provided to the DRL agent, as described in Section 5.3.

4 | DRL Framework

The application of DRL techniques has gained significant attention in recent years, particularly in solving complex problems that are challenging for traditional heuristic methods and DP approaches [10, 28]. For routing and network management in SCNs, DRL is a promising solution, as it enables low-complexity decision-making based on dynamically evolving network states.

However, the action space definition oftentimes poses a problem in the context of routing optimization. Given that variable action spaces are difficult to learn from, a common approach is to define the actions as the potential next hops of each node [30]. Such hop-based methods require an additional learning the pathfinding problem, which can introduce unnecessary complexity [16]. To circumvent this challenge, we formulate an approach which uses rule-based pathfinding algorithms to generate sets of candidate paths. This allows the DRL agent to focus on the primary objective of approximating the MCFP solution.

Our approach is based on the assumption that an optimal solution can be approximated by distributing incoming flows across a limited set of such pre-computed candidate paths. By learning the underlying traffic characteristics, the DRL agent is supposed to anticipate potential bottlenecks and make informed decisions, enabling the support of higher network loads and improved network performance. To this end, a suitable framework is required, which can effectively distribute flows on multiple candidate paths.

4.1 | Soft Actor–Critic

In order to comply with the path-based action space requirement and stochastic policies, we have proposed a soft actor critic (SAC) [31, 32] architecture [9]. SAC represents an off-policy actor–critic method which simultaneously maximizes the expected return and entropy. This promotes exploration and prevents premature convergence to sub-optimal policies. By learning a stochastic policy, value functions, and a temperature parameter, SAC balances the trade-off between exploration and exploitation. So, agents are able to navigate complex decision-making spaces. The approach is particularly useful in scenarios where convergence to suboptimal policies is a concern, as it encourages the agent to continue exploring and learning, rather than settling on a potentially inferior solution.

It consists of an actor (policy) network, parameterized by θ_a, and two critic (Q-value) networks, parameterized by ϕ_{Q1} and ϕ_{Q2}. While the original SAC formulation [31] applied a V-critic network ϕ_V, the utilized implementation relies on the now more commonly used twin Q-critic approach [33]. This approach can help reduce overestimation of Q-values, improving stability.

The temperature parameter $\alpha_{\mathscr{H}}$ adjusts the weight of the entropy term. It thus controls the trade-off between reward maximization (exploitation) and entropy maximization (exploration). When the policy entropy is below the target, $\alpha_{\mathscr{H}}$ increases to encourage exploration. Inversely, it can also decrease when entropy is too high. $\overline{\mathscr{H}}$ denotes the target entropy, which is set to $-|\mathscr{A}|$, so the negative cardinality of the action space. The value is tuned automatically using the following objective:

$$\mathscr{J}(\alpha_{\mathscr{H}}) = \mathbb{E}_{a\sim\pi_{\theta_a}(\cdot|s)}\left[-\alpha_{\mathscr{H}} \log \pi_{\theta_a}(a|s) - \alpha_{\mathscr{H}}\overline{\mathscr{H}}\right] \tag{11}$$

The objective functions $J_{Q1}(\phi_{Q1})$ and $J_{Q2}(\phi_{Q2})$ for updating the soft Q-function parameters correspond to the minimization of the soft Bellman residual. So, they are a measure of the difference between the current estimate of the action-value function and its updated value based on the Bellman equation:

$$\mathscr{J}_{Q1}(\phi_{Q1}) = \mathop{\mathbb{E}}_{(s_t,a_t,r_t,s_{t+1})\sim\mathscr{D}}\left[\frac{1}{2}\left(Q(s_t,a_t;\phi_{Q1}) - \hat{Q}(s_t,a_t;\phi_{Q1})\right)^2\right] \tag{12}$$

$$\mathscr{J}_{Q2}(\phi_{Q2}) = \mathop{\mathbb{E}}_{(s_t,a_t,r_t,s_{t+1})\sim\mathscr{D}}\left[\frac{1}{2}\left(Q(s_t,a_t;\phi_{Q2}) - \hat{Q}(s_t,a_t;\phi_{Q2})\right)^2\right] \tag{13}$$

Here, $\mathscr{D}$ denotes the replay buffer. Lastly, the actor is updated to maximize the expected reward plus the entropy term. The objective of the actor network is to encourage the policy to take actions with high Q-values while maintaining diversity for exploration through entropy maximization:

$$\mathscr{J}_\pi(\theta_a) = \mathop{\mathbb{E}}_{(s_t,a_t,r_t,s_{t+1})\sim D}\Big[\mathbb{E}_{a_t\sim\pi_{\theta_a}(\cdot|s)} \left[\min\left(Q(s_t,a_t;\phi_{Q1}), Q(s_t,a_t;\phi_{Q2})\right) - \alpha_{\mathscr{H}} \log \pi(a_t|s_t;\theta_a)\right]\Big] \tag{14}$$

15420981, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sat.70043 by Universitat der Bundeswehr Munchen, Wiley Online Library on [24/06/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

These objective functions highlight the self-regulating design of SAC, in terms of its exploration-exploitation trade-off. This balancing of a primary objective, that is, reward maximization, with a regularization term which promotes diversity represents a formal relationship with the MCFP.

4.2 | Relationship to Multicommodity Flow Problem

Based on the presented objective functions, a formal relationship between the MCFP and the SAC framework can be established. This connection demonstrates the theoretical viability of SAC for the optimization problem at hand. While we derived why DRL-based approaches are well-suited for routing, this framework shares properties with throughput maximization under constraints, the QoS requirements.

The distinguishing feature of SAC is its entropy regularization term. It prevents the policy from becoming deterministic too quickly and encourages exploration of the state space. A similar entropy-regularization term can be introduced into the MCFP formulation to promote flow diversity.

In the context of throughput maximization in the MCFP, the objective function can be formulated as follows:

$$\mathcal{J}_{\text{MCFP}}(f^k) = \max \sum_{k \in K} \sum_{(i,j) \in \mathcal{E}} f^k(i,j) \tag{15}$$

Capacity and flow conservation constraints apply to this maximization. Using entropy regularization, a formal connection is visible. To control the importance of flow diversity, α_R is introduced, and we denote the flow-specific entropy term, $\mathcal{H}(f)$, by the following:

$$\mathcal{H}(f) = -\sum_{k \in K} \sum_{(i,j) \in \mathcal{E}} f^k(i,j) \log f^k(i,j) \tag{16}$$

Accordingly, the regularized MCFP objective function becomes the following:

$$\mathcal{J}_{\text{MCFP}}(f^k) = \max \sum_{k \in K} \sum_{(i,j) \in \mathcal{E}} f^k(i,j) - \alpha_R \mathcal{H}(f) \tag{17}$$

This formulation shares structural similarity with the SAC objective of the actor network, provided in Equation (14). In both cases, the primary objective, throughput or expected reward, is balanced with an entropy term that promotes diversity in solutions. The temperature parameter $\alpha_{\mathcal{H}}$ in SAC serves a similar role to α_R in the regularized MCFP, controlling the trade-off between the primary objective and solution diversity. As we have seen, the latter is an important aspect of proactive load balancing.

Overall, this formal relationship indicates that SAC has the potential to adequately approximate solutions to the MCFP. The automatic tuning of the exploration–exploitation trade-off is considered a significant advantage, as converging towards promising policies is nontrivial.

5 | Proposed Architecture

Based on this SAC framework, we propose a suitable DRL architecture for the reference scenario. The components and their interaction are summarized in the diagram shown in Figure 2.

5.1 | Environment

The environment is defined over a grid network graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each edge $(i,j) \in \mathcal{E}$ has a capacity $c(i,j) > 0$ and a dynamic utilization $u(i,j) \in [0, c(i,j)]$. At each time step t, a new flow request must be routed through the network. Each request is a tuple (x_t, y_t, d_t), where $x_t \in \mathcal{V}$ is the source

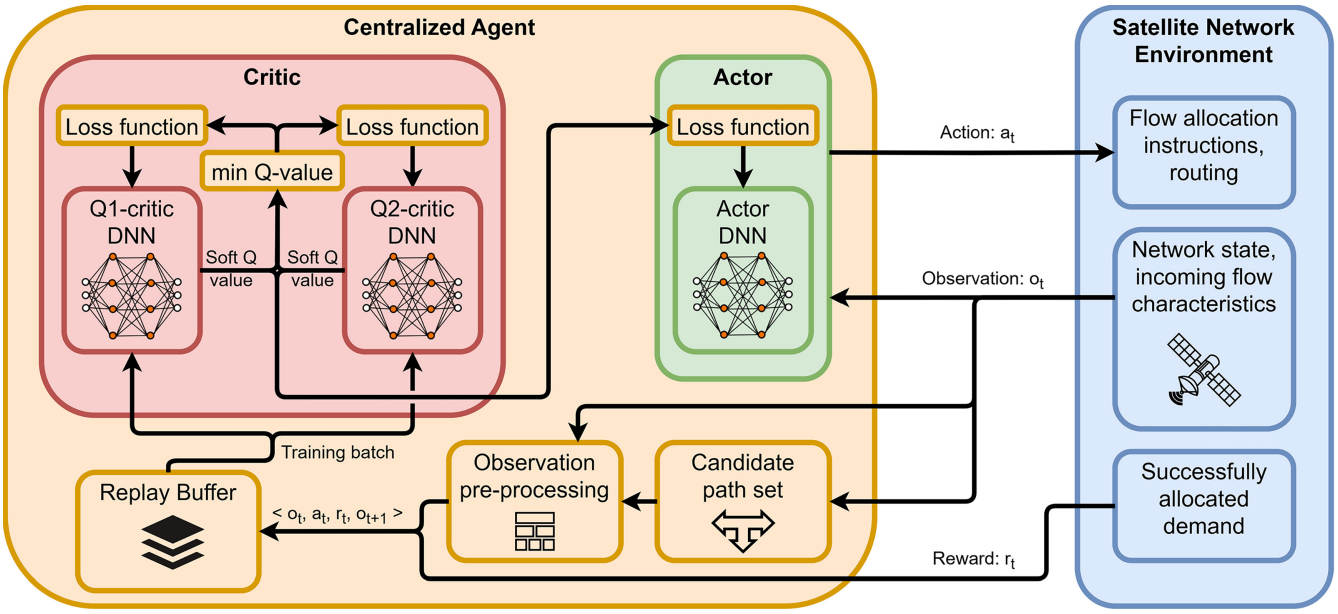

FIGURE 2 | Diagram of proposed path-based routing architecture. A centralized agent using a soft actor–critic framework interacts with the satellite network environment.

15420981, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sat.70043 by Universitat der Bundeswehr Munchen, Wiley Online Library on [24/06/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

node, $y_t \in \mathscr{V}$ is the destination node, and $d_t \in [0,1]$ is the flow demand.

The source and destination nodes are sampled according to a fixed heat map distribution, while the demand d_t is drawn from a uniform distribution. For each request, a set of k candidate paths from x_t to y_t is precomputed using efficient pathfinding algorithms (see Section 5.3). The agent then allocates the flow across these paths.

After allocation, link utilizations $u(i,j)$ are updated according to the assigned flow fractions. If the allocation would exceed the capacity on any edge, that portion of the flow is not routed.

Termination Condition: The episode terminates when the environment becomes saturated, that is, when an incoming flow request cannot be routed on any of the available paths without violating link capacity constraints. This reflects a realistic failure mode in network operation due to resource exhaustion. The overall objective of the agent is to route as much demand as possible before such a failure occurs.

5.2 | Action Space

The objective of the agent is to distribute the incoming flow request among a set of pre-generated candidate paths. To achieve this, we define the action space as a k-dimensional vector over $\mathbb{R}_{\geq 0}$. Here, k represents the number of computed candidate paths.

Notably, in some cases, the set of candidate paths may contain duplicates, for instance, when there are fewer than k viable paths available. However, the agent should be able to learn effective policies despite the presence of such overlapping actions. To ensure performance, a meaningful choice of k should be considered, taking into account the network topology and resulting sets of candidate paths. For the reference scenario, a constant value of $k = 4$ is assumed.

To guarantee that the agent's actions correspond to valid flow allocations, we normalize the action vector to ensure that its entries sum to 1. Such a normalization step enables the agent to output a probability distribution over the candidate paths, effectively dividing the flow proportionally among them based on the decision of the agent. For example, for $k = 4$, a potential normalized action vector of $\mathbf{a}' = [0.3, 0.0, 0.3, 0.4]$ is valid.

5.3 | Set of Candidate Paths

In order for the agent to learn effective routing policies, the design of the candidate path set is critical. These paths are precomputed before the agent selects any action and form the available routing options for a given flow request. While the action space itself consists of continuous allocations over this set (representing how much of the flow is routed along each path), the quality and diversity of the candidate paths directly impact the agent's ability to learn meaningful decisions and observe varied reward outcomes.

An effective set of candidate paths must strike a balance between viability and discriminability. On the one hand, it should include high-quality paths, such as those that minimize delay or congestion, to offer meaningful routing options. On the other hand, the paths should overlap minimally to ensure that the agent can distinguish their effects and learn from the resulting performance differences.

In our previous investigation [9], the candidate paths were selected using heuristics with varying degrees of overlap and cost formulations. This approach sometimes resulted in ambiguous behavior, as different paths could share similar characteristics or serve overlapping roles. Consequently, the agent could not reliably associate a specific allocation with a clearly defined routing objective.

In this work, we introduce a revised path selection method that improves both learning performance and interpretability. For every flow request, we construct four distinct paths based on the following criteria:

1. **Shortest Path**: path minimizing the number of hops, independent of link utilization.
2. **Least Congested Path**: path minimizing the sum of utilizations,
$$\sum_{(i,j)\in p} u(i,j).$$
3. **Softmin Path** ($\alpha = 0.5$): path minimizing the sum of exponential weights,
$$w(i,j)_{0.5} = \exp(0.5 \cdot u(i,j)).$$
4. **Softmin Path** ($\alpha = 10$): path minimizing the sum of sharper weights,
$$w(i,j)_{10} = \exp(10 \cdot u(i,j)).$$

Each path is uniquely associated with a specific routing principle, allowing the agent to interpret each action dimension as targeting a distinct traffic behavior. This structure simplifies the learning task and leads to improved performance, as the agent can learn to associate traffic conditions with appropriate allocation strategies more effectively than with the more ambiguous sets used previously.

The candidate paths are recomputed at each step based on the current network state. The agent's action is a normalized vector over $[0,1]^4$, specifying how the current demand is distributed across the four paths. This compact formulation supports flexible, adaptive behavior while maintaining semantic clarity for each component of the decision.

In the experimental results, we refer to this new scheme as **P1**, reflecting its use of four fixed and semantically distinct path types. For comparison, the earlier version based on heuristically generated paths with varying overlap and less clearly defined roles is referred to as **P2**.

15420981, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sat.70043 by Universitat der Bundeswehr Munchen, Wiley Online Library on [24/06/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

5.4 | State Space

We use a compact, statistical representation of the network state, avoiding path-level structural encodings. For each of the k candidate paths, the following five features are included:

$$\min(u_p), \quad \max(u_p), \quad avg(u_p), \quad \mathrm{var}(u_p), \quad \ell_p$$

where u_p denotes the vector of link utilizations on the path p, and $\ell_p \in [0,1]$ is the normalized path length (relative to the maximum possible length).

In addition, four global features are included over all network links:

$$\min(u), \quad \max(u), \quad avg(u), \quad \mathrm{var}(u)$$

where $u \in [0,1]^{|\mathscr{E}|}$ is the vector of current link utilizations across the graph.

Finally, the current flow demand $d_t \in [0,1]$ is appended.

The total observation vector is real-valued, normalized, and of size

$$S = 5k + 4 + 1 = 5k + 5 \tag{18}$$

5.5 | Reward

The reward reflects the amount of demand successfully routed at each step. Let d_t be the demand of the current flow request. If the allocation is feasible (i.e., does not exceed any link capacities), the agent receives a reward equal to d_t. Otherwise, the reward is zero:

$$r_{t+1} = \begin{cases} d_t & \text{if the flow is successfully routed} \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

This formulation encourages the agent to maximize the total routed demand over the episode. If no valid allocation exists, meaning that routing the flow would exceed link capacity on any path, the episode terminates. This reflects the saturation of the network and enforces early stopping once no further traffic can be accommodated.

Due to the step-wise nature of the task and the absence of long-term shaping rewards, a high discount factor is beneficial. It allows the agent to value long-term throughput and learn to avoid early network saturation.

Alternative reward formulations, including penalties for exceeding link capacity and exponential shaping functions, were evaluated but did not yield improved performance. The simple, demand-based reward proved most effective for encouraging stable and efficient routing behavior.

5.6 | Complexity

The complexity of the proposed approach after training can be broken down into two components: the complexity of calculating candidate paths, denoted as $\mathscr{TC}_{\text{paths}}$ and the complexity of model inference, denoted as $\mathscr{TC}_{\text{inference}}$. This relationship can be represented by the following:

$$\mathscr{TC}_{\text{deploy}} = \mathscr{TC}_{\text{paths}} + \mathscr{TC}_{\text{inference}} \tag{20}$$

The time complexity of the k-shortest paths algorithm can be used as a baseline for computing the candidate paths:

$$\mathscr{TC}_{\text{paths}} = O(|\mathscr{E}| + |\mathscr{V}| \log |\mathscr{V}| + k) \tag{21}$$

Here, k denotes the number of paths. $\mathscr{V}$ and $\mathscr{E}$ denote the sets of vertices and edges respectively, as before. Naturally, the path selection does not directly follow this algorithm. Instead, it is based on a mix of pathfinding algorithms which typically results in slightly higher complexity. Nevertheless, it is assumed to remain in a similar domain.

On the other hand, the inference process involves a forward pass through the actor network. The process consists primarily of matrix multiplications representing the layers of the trained neural network. Its complexity in time can be represented by the following:

$$\mathscr{TC}_{\text{inference}} = O\left(\sum_{\iota \in \mathscr{L}} N_n^{\iota-1} N_n^{\iota}\right) \tag{22}$$

In this context, $\mathscr{L}$ denotes the set of layers in the neural network and N_n^{ι} represents the number of nodes or neurons in layer ι.

In contrast, the training complexity is significantly higher. It can be expressed as the product of the number of iterations, denoted as N_{its}, and the time complexity associated with forward and backward passes through both the actor network and two critic networks, denoted as $\mathscr{TC}_{\text{actor}}$ and $\mathscr{TC}_{\text{critic}}$, respectively,

$$\mathscr{TC}_{\text{training}} = O\left(N_{\text{its}}\left(\mathscr{TC}_{\text{actor}} + 2\mathscr{TC}_{\text{critic}}\right)\right) \tag{23}$$

It is important to note that increasingly large network sizes impact performance. A key issue is the more expensive computation of candidate paths ($\mathscr{TC}_{\text{paths}}$). But there are further aspects which have to be considered. These are discussed in Section 6.4.

6 | Performance Evaluation

To evaluate the effectiveness of the proposed DRL-based routing method, we conduct simulations on different network configurations under varying traffic loads. The agent's goal is to maximize the total amount of routed demand before the network reaches saturation. We compare its performance to baseline routing schemes and investigate the impact of different candidate path selection strategies.

6.1 | Simulator Environment

We implement the environment using the Python-based `Gymnasium` library [34]. The SAC algorithm is implemented

15420981, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sat.70043 by Universitat der Bundeswehr Munchen, Wiley Online Library on [24/06/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

using the `Stable-Baselines3` library [33]. Additionally, we utilize the `NetworkX` library [35] to represent and manipulate the underlying graph structure, as well as to perform pathfinding routines.

We consider two grid network configurations with $8 \times 6 = 48$ and $12 \times 24 = 288$ satellite nodes, respectively. The 48-node setup is supposed to represent the domain shown in Figure 1. Flows are generated by sampling source and destination nodes according to the heat map shown in Figure 3. The demand d_t is drawn uniformly from the interval $[0.1, 0.5]$.

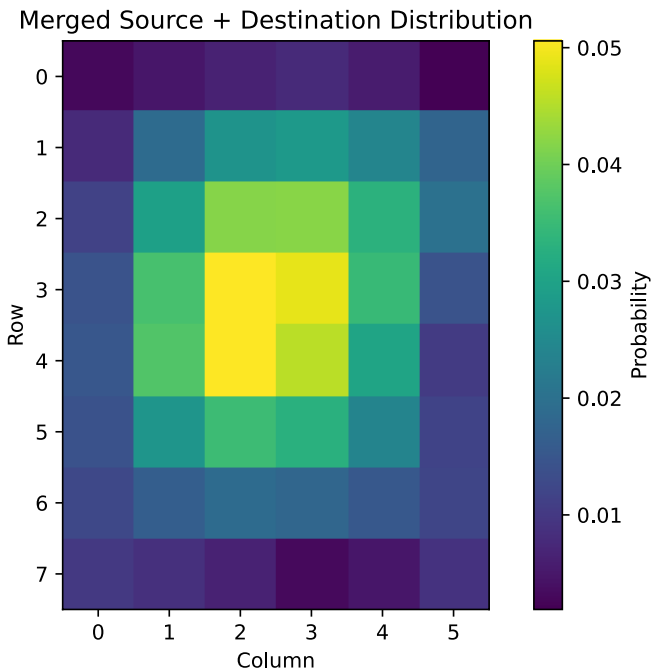

FIGURE 3 | Centered heat map (around center) for the 8×6 grid. Brighter areas indicate higher probability of node selection for source and destination sampling.

TABLE 3 | Hyperparameters of the SAC model using the simplified statistics-based observation space with $k = 4$ precomputed paths.

Hyperparameter	Value
Algorithm	SAC
Policy	MlpPolicy
State space size	$S = 5k + 5 = 25$ (see (18))
Policy architecture	[64, 64]
Critic architecture	[64, 64]
Action space size	$k = 4$
Activation function	ReLU
Learning rate	$1 \cdot 10^{-3}$
Buffer size	1, 000, 000
Batch size	512
Target update rate	0.005
Discount factor	0.999
Entropy coefficient	Automatic

At each time step, a new flow is sampled and a set of $k = 4$ candidate paths is computed based on the current network state, as described in Section 5.3. The agent observes path statistics and global utilization features as defined in Section 5.4 and selects a flow allocation accordingly.

Flow requests are continuously generated until the environment terminates due to saturation, that is, when no feasible path allocation exists without exceeding link capacities (see Section 5.1). Thus, there is no fixed number of flows per episode. Instead, the agent's goal is to route as much total demand as possible before failure.

The used hyperparameters for training are summarized in Table 3.

To investigate how different traffic patterns influence learning and performance, we compare two heat map configurations for the 8×6 grid. The first heat map emphasizes the central region of the network, encouraging flows between nodes in the middle area. The second heat map concentrates traffic at the edges, particularly in the top and bottom rows. The heat maps are visualized in Figures 3 and 4. These configurations are used to train and evaluate separate agents, allowing us to analyze generalization capabilities across spatially distinct traffic distributions.

6.2 | Benchmark Comparison

To evaluate the performance of the proposed DRL-based method, we compare it to several routing baselines in a Monte Carlo simulation setting. For each run, flow requests are generated until the network reaches saturation, and the total

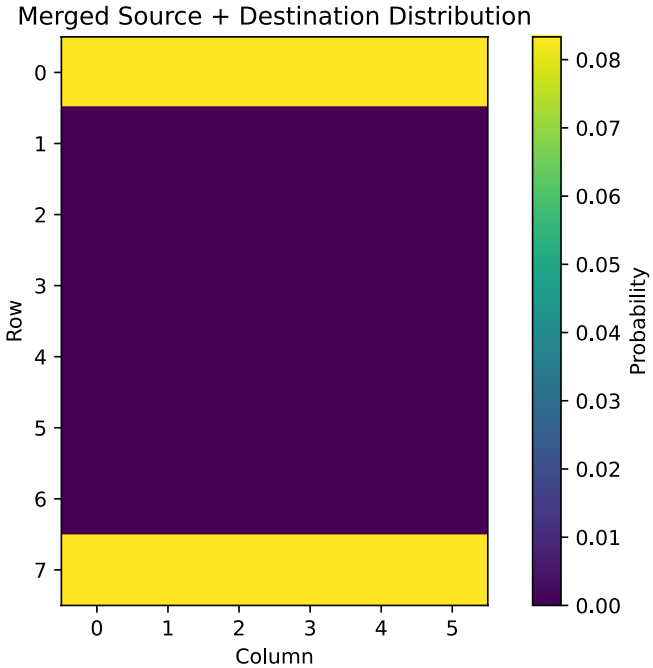

FIGURE 4 | Edge-focussed heat map for the 8×6 grid.

15420981, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sat.70043 by Universitat der Bundeswehr Munchen, Wiley Online Library on [24/06/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

successfully routed demand is recorded. This setup reflects the primary optimization goal of maximizing throughput before link overflow occurs (see Section 5.5).

The considered methods are as follows:

- **SPF**: Shortest-path-first, based on hop count.
- **LCF**: Least-congested-first, minimizing the sum of current link utilizations.
- **SM** $\alpha = 0.5$ and **SM** $\alpha = 10$: Softmin routing strategies based on exponential cost transformations (see Section 5.3).
- **EMP P1**: Equal multipath routing using candidate path set P1.
- **SAC P1**: DRL agent trained using the candidate path set P1.
- **SAC P2**: DRL agent trained using the candidate path set P2.

Figure 5 shows the distribution of total routed demand over 10,000 randomly sampled flow sequences on the 8×6 grid. The SAC agent using the interpretable candidate path set (P1) consistently outperforms all baseline approaches, including the rule-based methods and the legacy DRL setup (P2). This highlights the effectiveness of structured path selection for enabling better routing decisions under high load.

To further compare both SAC agents during training, Figure 6 shows the evolution of the average reward over time. Both agents were trained on the 8×6 grid using a centered traffic heat map. SAC P1 consistently achieves higher reward throughout training. This may be due to the higher consistency in performance across its candidate paths, which can simplify the learning task. SAC P1 enforces semantic separation between paths (e.g., shortest, least-congested, softmin variants), which may introduce more variability in action outcomes. While P2 is explicitly designed to reduce path overlap, it can be more challenging to learn an effective allocation policy in such a structured setting. Nonetheless, P1 offers better interpretability and shows strong inference performance under varying conditions.

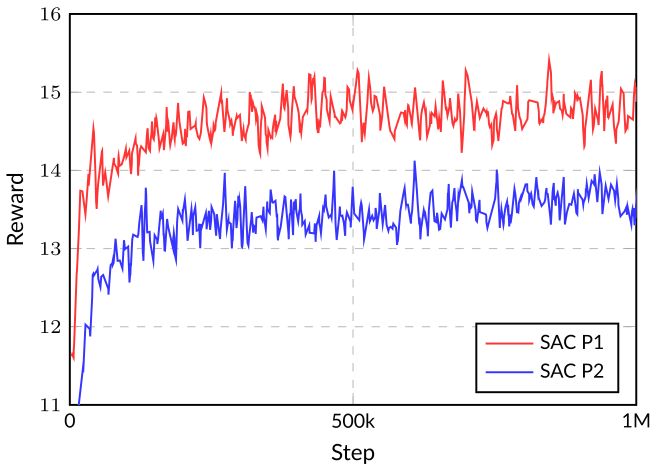

FIGURE 6 | Training curves showing the average reward over time for SAC agents using candidate path sets P1 and P2 on the 8×6 grid with a centered traffic heat map. SAC P1 consistently achieves higher reward during training, suggesting that the diverse set of candidate paths may simplify the learning task.

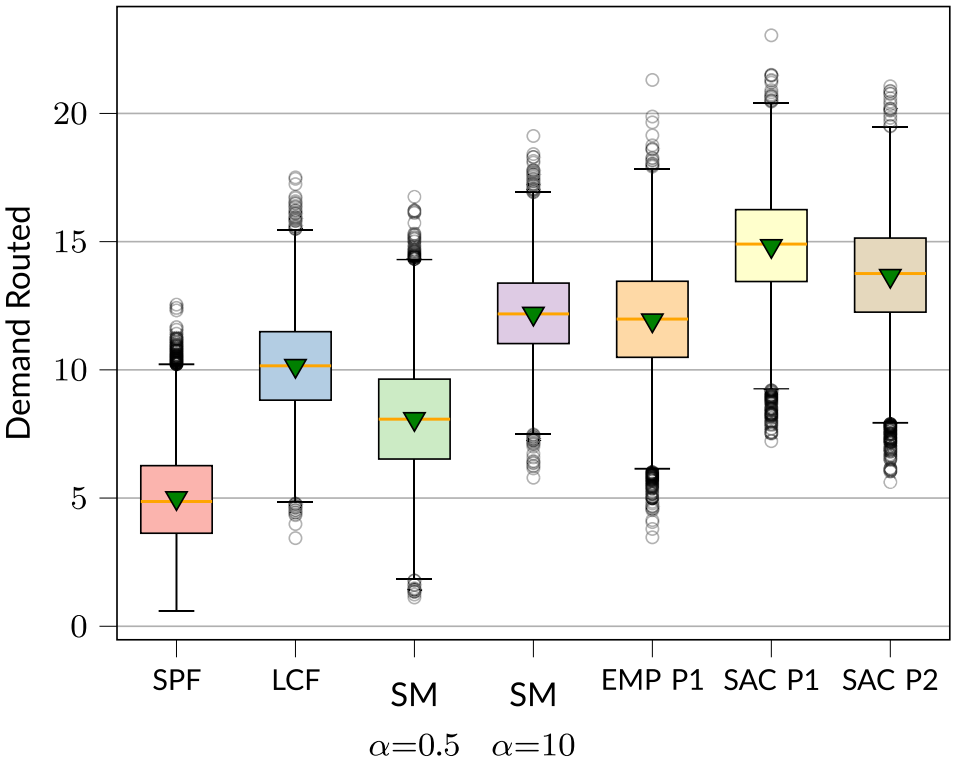

FIGURE 5 | Monte Carlo simulation: comparison of total routed demand for 8×6 grid.

15420981, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sat.70043 by Universitat der Bundeswehr Munchen, Wiley Online Library on [24/06/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

6.3 | Heat Map Comparison

To evaluate the robustness of the trained policy under different traffic patterns, we compare the performance of two SAC agents trained on different heat maps, while testing both on the same edge-focused traffic scenario. The experiment is conducted on an 8×6 grid network.

During evaluation, source and destination nodes are sampled exclusively from the top-most and bottom-most rows of the grid, simulating traffic concentrated at the network edges. **"SAC P1 edge"** was trained on this same edge-focused heat map (see Figure 4), whereas **"SAC P1 uniform"** was trained on a uniform heat map, where all nodes are equally likely to act as source or destination. Both models use the **P1** (see Section 5.3) candidate path selection method, which combines shortest path, least-congested path, and two softmin-based path variants.

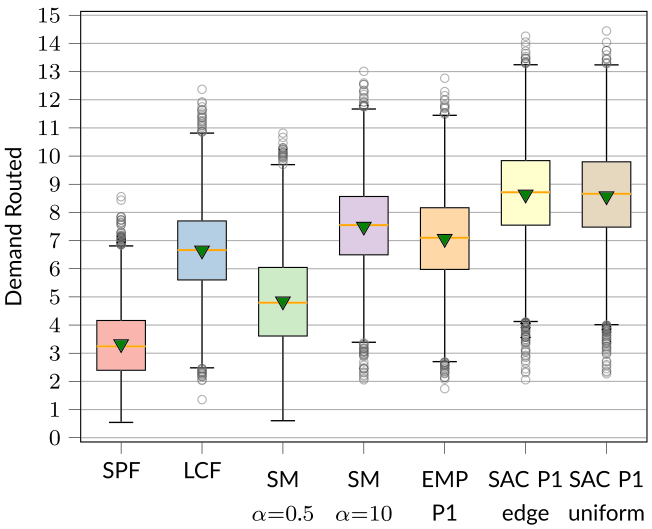

FIGURE 7 | Monte Carlo simulation on an 8×6 grid with edge-focused traffic (top and bottom rows only). "SAC P1 edge" is trained on the same edge-only heat map (see Figure 4), "SAC P1 uniform" is trained on a uniform heat map.

As shown in Figure 7, both SAC agents achieve comparable routing performance. This indicates that the learned policy generalizes well to new traffic distributions and is resilient to changes in the training heat map. The agent's ability to maintain performance under shifted traffic conditions supports its scalability and applicability to dynamic or unforeseen traffic scenarios.

6.4 | Larger Networks

While the main evaluation focused on routing within SCN subnetworks, the proposed DRL-based method is equally applicable to ground-based or centralized control of the full constellation. To assess its potential in larger scale domains, we extend our evaluation to a 24×12 grid topology, comprising 288 nodes. The corresponding heat map is shown in Figure 8.

It is important to note that this configuration does not replicate the precise layout of a Walker-type or polar-orbit constellation, as the orientation and directionality of ISLs are simplified. However, the grid structure provides a relevant approximation of path diversity, scale, and congestion effects in large space-borne or hybrid networks.

Figure 9 shows the total routed demand until saturation for different baseline and DRL methods, using the same Monte Carlo sampling procedure as in the smaller network. The SAC agent trained with the interpretable candidate path set (P2) again achieves the highest throughput, confirming that the approach scales well with network size and maintains a performance advantage even under increased path complexity and capacity pressure.

As shown, the proposed approach is scalable and can handle large constellation sizes. The state space, which is based on a compact, statistical representation of the network state, does not depend on the number of satellites in the constellation. Instead,

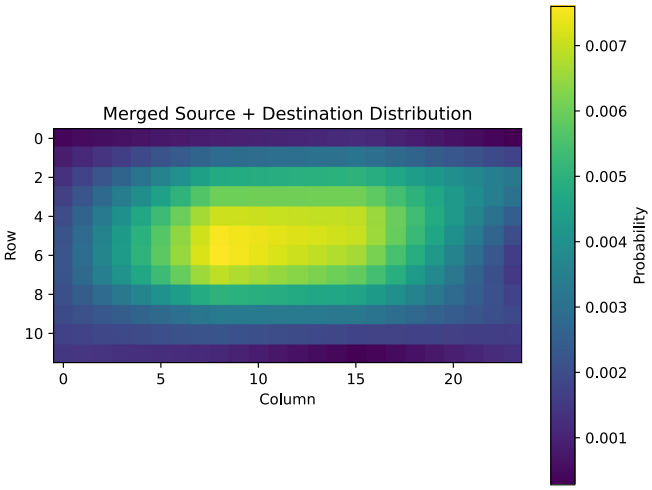

FIGURE 8 | Traffic heat map used for the 24×12 grid. Brighter areas indicate higher probability of node selection for source and destination sampling.

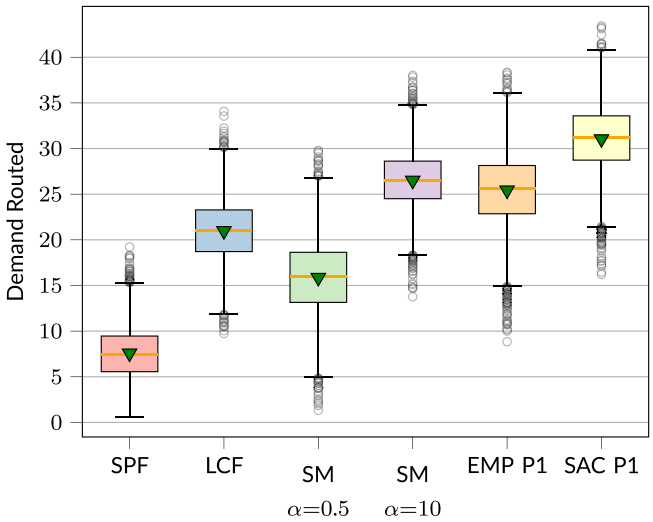

FIGURE 9 | Monte Carlo simulation: Comparison of total routed demand for 24×12 grid.

15420981, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sat.70043 by Universitat der Bundeswehr Munchen, Wiley Online Library on [24/06/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

it depends on the number of candidate paths (k) and the number of features used to represent each path. This means that the approach can be applied to constellations of varying sizes without significant modifications.

However, for much larger graphs, the number of possible paths increases, making it more challenging to compute the set of candidate paths. The computation required for the path preprocessing increases significantly. In addition, with the higher-dimensional state space, learning efficient policies becomes more difficult when training the agent. Finally, with increasingly large graphs, there is more variability in the flow distribution process. Approximating the MCFP is more of a challenge. To counter this, the number of paths, k, can be increased. This results in more expensive path computations as well as correspondingly larger state and action spaces, and thus larger models. Given the relatively small models used in the simulations, increasing k can be feasible depending on the scenario.

In summary, decreased performance gains have to be expected for significantly larger networks. While the approach may not be optimal for such extremely large graphs, its scalability and flexibility make it a valuable tool for optimizing flow distribution in a wide range of currently considered constellation sizes.

7 | Discussion and Implications

We proposed a novel approach for flow distribution in SCNs using a path-based SAC solution. Our results demonstrate that the proposed approach outperforms state-of-the-art methods in different scenarios and can effectively distribute flows to mitigate potential bottlenecks. The approach is unique in its architecture, using a SAC agent which decides how to distribute traffic efficiently on a set of candidate paths.

In this section, we summarize the strengths and limitations of the proposed approach and discuss its potential implications for real-world applications, particularly in the context of 6G networks.

7.1 | Strengths of the Proposed Method

The proposed method exhibits several key strengths that make it an attractive solution for flow distribution in SCNs. Notably, the approach demonstrates flexibility in learning complex traffic patterns and distributing flows with foresight, allowing it to anticipate and mitigate potential bottlenecks. The proposed method consistently outperforms state-of-the-art methods, including the least-congested path approach.

Furthermore, the improved performance achieved by the method could not be replicated by mimicking its behavior with rule-based decision-making. This demonstrates that the agent is actually learning effective policies and not generic strategies.

In addition, the observed complexity of the approach, even during training, poses no significant computational challenges. Given that we were able to achieve improved performance with smaller models indicates that further reduction of processing needs may be possible. Moreover, the method seamlessly integrates into SDN frameworks, making it a viable and practical solution for real-world deployment.

7.2 | Limitations of Proposed Method

A key assumption underlying the approach is that distributing flows among a set of candidate paths will result in near-optimal performance. While the results show the capacity to route more demands, the given sets of paths are expected to represent a ceiling to the effectiveness of the approach. Naturally, as the analysis indicates, the potential improvement is also strongly scenario-dependent.

The learning of underlying traffic patterns is currently tied to static heat maps, which may not capture the full complexity of real-world traffic dynamics. Given the movement of the constellation and dynamically changing network topology, satellites may be faced with diverse traffic scenarios. More complex, time-dependent environments have to be considered. While the results demonstrate some degree of robustness and generalization, the effectiveness of the method may decrease for such environments and related models.

7.3 | Real World Application

The proposed approach has significant implications for real-world applications, particularly in the context of 6G NTN. With the increasing demand for high-speed and low-latency communications, 6G networks will require advanced flow distribution methods to ensure efficient and reliable data transfer. The proposed approach can be applied to 6G networks to optimize flow distribution and mitigate potential bottlenecks. As described in Section 3.2, it is compliant with proposed functional splits.

For example, the approach can be used to optimize flow distribution in a standard-compliant 6G satellite constellation. To this end, it can run on top of existing rule-based routing solutions. The proposed in-space distributed SDN architecture presents a promising baseline for autonomous satellite constellation designs with fast reactivity. Depending on the cluster size, routes can be adjusted in less than 50ms [2].

But for completely ground-based control, the proposed approach can represent an intriguing solution for efficient decision-making. As the results in the 288-satellite scenario show, the approach is also applicable for a complete constellation. For real-world application, this is particularly relevant, as operators may want to test ML-enabled routing extensions in ground-controlled constellations first, before committing to in-space network control.

Operator data will then enable further improvements and refinements of the approach. As it is data-driven, incorporating historical trends as inputs to improve the accuracy of traffic predictions is a logical extension. Moreover, additional advanced machine learning techniques to capture complex network dynamics can be applied to the general architecture if required.

15420981, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sat.70043 by Universitat der Bundeswehr Munchen, Wiley Online Library on [24/06/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

To this end, the approach can be easily integrated with other techniques, such as GNNs or attention-based schemes for preprocessing, to improve its performance and adaptability. These adaptions may be necessary given the potential challenges that may arise when applying the approach to different scenarios, such as handling nonstationary traffic patterns or time-varying network conditions.

While this work specifically focuses on the MCFP, the approach can be tailored to different objectives, such as resilient routing, by formulating the reward accordingly. This represents a flexible low-complexity solution in inference for routing and traffic engineering in satellite constellations.

Overall, the proposed approach represents a promising solution for optimizing flow distribution in 6G NTN. Its flexibility, modularity, and scalability make it an attractive solution for real-world applications, and its potential benefits in terms of increased throughput make it an interesting approach for future research and development.

8 | Conclusion

In this work, we extend the design of a SAC-based DRL approach for approximating solutions to the MCFP in SCNs, while maintaining relatively low computational complexity. Following a path-based decision-making strategy, a centralized agent effectively distributes incoming flow requests among a set of candidate paths. The approach is able to outperform rule-based state-of-the-art approaches in various scenarios. The provided analyses provide valuable insights into the architecture of the method, including the composition of path sets. Moreover, we demonstrate that the general performance improvement persists also in larger networks. Thus, while the method is primarily designed to integrate into the routing logic of in-orbit distributed SDN, it is also viable for global ground-based control entities. Overall, the combination of flexibility, adaptability, and performance makes the proposed method a promising solution for optimizing flow distribution in SCNs.

Acknowledgments

This work has been supported by the Space for 5G & 6G programme of the European Space Agency (ESA), activity code 5A. 079, https://connectivity.esa.int/projects/aicoms. Responsibility for the contents of this publication rests with the authors. Open Access funding enabled and organized by Projekt DEAL.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. T. Butash, P. Garland, and B. Evans, "Non-Geostationary Satellite Orbit Communications Satellite Constellations History," *International Journal of Satellite Communications and Networking* 39, no. 1 (2021): 1–5, https://doi.org/10.1002/sat.1375.

2. M. M. H. Roth, H. Brandt, H. Bischl, D. F. Piñas, and G. Acar, "IDLB: An SDN-Based Load-Balancing Routing Protocol for Autonomous Satellite Constellation Networks," *International Journal of Satellite Communications and Networking* 42, no. 2 (2026): 109–127, https://doi.org/10.1002/sat.70004.

3. 3GPP. Specification # 38.821. Tech. Rep. 3GPP TR 38.821, Technical Specification Group Radio Access Network; 2023.

4. L. Bertaux, S. Medjiah, P. Berthou, et al., "Software Defined Networking and Virtualization for Broadband Satellite Networks," *IEEE Communications Magazine* 53, no. 3 (2015): 54–60, https://doi.org/10.1109/MCOM.2015.7060482.

5. F. Bannour, S. Souihi, and A. Mellouk, "Distributed SDN Control: Survey, Taxonomy, and Challenges," *IEEE Communication Surveys and Tutorials* 20, no. 1 (2018): 333–354, https://doi.org/10.1109/COMST.2017.2782482.

6. A. Papa, T. De Cola, P. Vizarreta, M. He, C. Mas-Machuca, and W. Kellerer, "Design and Evaluation of Reconfigurable SDN LEO Constellations," *IEEE Transactions on Network and Service Management* 17, no. 3 (2020): 1432–1445, https://doi.org/10.1109/TNSM.2020.2993400.

7. C. H. Papadimitriou and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity* (Courier Corporation, 1998).

8. C. Larsson, *Design of Modern Communication Networks: Methods and Applications* (Academic Press, 2014).

9. M. M. Roth, T. Jerkovits, A. Hegde, T. Delamotte, and A. Knopp, "Deep Reinforcement Learning for Adaptive Traffic Engineering in Satellite Constellation Networks," in *In 2025 12th Advanced Satellite Multimedia Systems Conference and the 18th Signal Processing for Space Communications Workshop (ASMS/SPSC)* (IEEE, 2025), 1–8.

10. C. Zhang, P. Patras, and H. Haddadi, "Deep Learning in Mobile and Wireless Networking: A Survey," *IEEE Communication Surveys and Tutorials* 21, no. 3 (2019): 2224–2287, https://doi.org/10.1109/COMST.2019.2904897.

11. J. Xie, F. R. Yu, T. Huang, et al., "A Survey of Machine Learning Techniques Applied to Software Defined Networking (SDN): Research Issues and Challenges," *IEEE Communication Surveys and Tutorials* 21, no. 1 (2019): 393–430, https://doi.org/10.1109/COMST.2018.2866942.

12. M. Corici, H. Buhr, H. Zope, and M. Zaboub, "An SDN-Based Solution for Mega-Constellation Routing," in *In 2024 IEEE 35th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)* (IEEE, 2024), 1–6.

13. K. Salimifard and S. Bigharaz, "The Multicommodity Network Flow Problem: State of the Art Classification, Applications, and Solution Methods," *Operational Research* 22, no. 1 (2022): 1–47, https://doi.org/10.1007/s12351-020-00564-8.

14. P. Grislain, N. Pelissier, F. Lamothe, et al., "Rethinking LEO Constellations Routing With the Unsplittable Multi-Commodity Flows Problem," in *In 2022 11th Advanced Satellite Multimedia Systems Conference and the 17th Signal Processing for Space Communications Workshop (ASMS/SPSC)* (IEEE, 2022), 1–8.

15. P. Zuo, C. Wang, Z. Yao, S. Hou, and H. Jiang, "An Intelligent Routing Algorithm for LEO Satellites Based on Deep Reinforcement Learning," in *In 2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall)* (IEEE, 2021), 1–5.

16. A. Hegde, M. Roth, and H. Bischl, "Routing in Low Earth Orbit Satellite Networks: Constrained DQN-Based Approach," in *In 2025 IEEE Wireless Communications and Networking Conference (WCNC)* (IEEE, 2025), 1–6.

17. R. E. Ali, B. Erman, E. Baştuğ, B. Cilli. Hierarchical Deep Double Q-Routing. Tech. Rep. arXiv:1910.04041, arXiv; 2020.

18. K. C. Tsai, L. Fan, L. C. Wang, R. Lent, and Z. Han, "Multi-Commodity Flow Routing for Large-Scale LEO Satellite Networks Using Deep Reinforcement Learning," in *In 2022 IEEE Wireless*

15420981, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sat.70043 by Universitat der Bundeswehr Munchen, Wiley Online Library on [24/06/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Communications and Networking Conference (WCNC) (IEEE, 2022), 626–631.

19. S. Kumar, G. Bansal, and V. S. Shekhawat, "A Machine Learning Approach for Traffic Flow Provisioning in Software Defined Networks," in *In 2020 International Conference on Information Networking (ICOIN)* (IEEE, 2020), 602–607.

20. A. Cigliano and F. Zampognaro, "A Machine Learning Approach for Routing in Satellite Mega-Constellations," in *In 2020 International Symposium on Advanced Electrical and Communication Technologies (ISAECT)* (IEEE, 2020), 1–6.

21. F. Geyer and G. Carle, "Learning and Generating Distributed Routing Protocols Using Graph-Based Deep Learning," in *Big-DAMA'18* (Association for Computing Machinery, 2018), 40–45.

22. H. Wang, Y. Ran, L. Zhao, J. Wang, J. Luo, and T. Zhang, "GRouting: Dynamic Routing for LEO Satellite Networks With Graph-Based Deep Reinforcement Learning," in *In 2021 4th International Conference on Hot Information-Centric Networking (HotICN)* (IEEE, 2021), 123–128.

23. J. G. Walker, "Satellite Constellations," *Journal of the British Interplanetary Society* 37 (1984): 559.

24. Met Office, *Cartopy: A Cartographic Python Library With a Matplotlib Interface* (Exeter, 2010).

25. R. Campana, C. Amatetti, and A. Vanelli-Coralli, "O-RAN Based Non-terrestrial Networks: Trends and Challenges," in *In 2023 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)* (IEEE, 2023), 264–269.

26. 3GPP. Specification # 23.501. Tech. Rep. 3GPP TS 23.501, Technical Specification Group Services and System Aspects; 2025.

27. 3GPP. Specification # 23.700–29. Tech. Rep. 3GPP TR 23.700–29, Technical Specification Group Services and System Aspects; 2024.

28. R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, 2018).

29. C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning. 4* (Springer, 2006).

30. J. Boyan, M. Littman. Packet Routing in Dynamically Changing Networks: A Reinforcement Learning Approach. In: 6. Morgan-Kaufmann 1993.

31. T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning With a Stochastic Actor," in *International Conference on Machine Learning* (Pmlr, 2018), 1861–1870.

32. T. Haarnoja, A. Zhou, K. Hartikainen, et al., "Soft Actor-Critic Algorithms and Applications," arXiv preprint arXiv:1812.05905, arXiv; 2019.

33. A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-Baselines3: Reliable Reinforcement Learning Implementations," *Journal of Machine Learning Research* 22, no. 268 (2021): 1–8.

34. M. Towers, A. Kwiatkowski, J. Terry, et al., "Gymnasium: A Standard Interface for Reinforcement Learning Environments," arXiv preprint arXiv:2407.17032, arXiv; 2024.

35. A. Hagberg, P. J. Swart, D. A. Schult. Exploring Network Structure, Dynamics, and Function Using NetworkX. Tech. Rep. LA-UR-08-05495; LA-UR-08-5495, Los Alamos National Laboratory (LANL), Los Alamos, NM (United States); 2008.

15420981, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sat.70043 by Universitat der Bundeswehr Munchen, Wiley Online Library on [24/06/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License