

Article

Hierarchical Reinforcement Learning with Hungarian Assignment for Reliable Urban Smart Metering Under Cognitive Spectrum Access

Muhammed Al-Ali ^{1,*} , Esteban Inga ² , Juan Inga ³  and Elias Yaacoub ¹ 

¹ Department of Computer Science and Engineering, Qatar University, P.O. Box 2713, Doha, Qatar; elias@qu.edu.qa

² Smart Grids Research Group (GIREI), Universidad Politécnica Salesiana, Cuenca EC010102, Ecuador; einga@ups.edu.ec

³ Telecommunications and Telematic Research Group (GITEL), Universidad Politécnica Salesiana, Cuenca EC010102, Ecuador; jinga@ups.edu.ec

* Correspondence: ma0903261@qu.edu.qa

Highlights

What are the main findings?

- A two-timescale architecture couples a learned PPO cognitive layer with an exact Hungarian channel assignment for cellular smart metering.
- It improves strict per-meter satisfaction by 7 points over a Bayesian-belief baseline and by 19 points over Hungarian-only matching, with gains stable across seeds and sensitivity sweeps.

What are the implications of the main findings?

- Separating exact assignment from learned channel selection lets utilities use leased cellular spectrum more reliably and scalably than either alone.
- More reliable, lower-latency metering strengthens the grid-edge data underpinning smart-city energy services.

Abstract

Advanced metering infrastructure (AMI) is the sensing backbone of the smart grid, and its reliability underpins urban energy services such as state estimation, demand response, and distributed-energy integration. When AMI uses cellular spectrum leased through a cognitive mobile virtual network operator (C-MVNO), allocating channels to data aggregation points (DAPs) each frame is difficult because three uncertainties interact: imperfect spectrum sensing, time-varying and cross-channel-correlated primary-user activity, and stochastic urban propagation. Classical Hungarian assignment is optimal per frame but blind to primary-user dynamics, while cognitive-radio heuristics ignore queue state and cross-channel structure. We propose a two-timescale hierarchy that couples these established tools in a new way: a Proximal Policy Optimization (PPO) agent decides, once per epoch, which opportunistic channels to expose, and an exact Hungarian solver performs the per-frame DAP-to-channel assignment. To our knowledge this is the first coupling of a learned cognitive layer with exact Hungarian assignment for cognitive-radio resource allocation. On a 3GPP TR 38.901-compliant simulator, PPO significantly outperforms a Bayesian-belief baseline and the Hungarian-only configuration in delivery ratio, latency, and a strict per-meter satisfaction metric, and is robust across independent seeds and sensitivity sweeps. An architectural ablation shows the DAP tier is a precondition for viability, not merely an optimization.



Academic Editors: Javier Prieto and Pierluigi Siano

Received: 16 May 2026

Revised: 28 July 2026

Accepted: 29 July 2026

Published: 31 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: advanced metering infrastructure; smart metering sensor networks; cognitive spectrum access; dynamic spectrum access; reinforcement learning; Hungarian assignment; smart grid communications; sensed-data reliability

1. Introduction

Smart cities increasingly rely on data-driven, IoT-based infrastructure to manage energy and water resources efficiently, improve public services, and support sustainability. In this context, smart metering is a key enabling technology, providing fine-grained measurements that support monitoring, demand management, and data-informed decision-making across urban infrastructure. Prior studies have shown that real-time data from smart city applications can improve environmental, social, and economic outcomes, while supporting smarter governance and citizen participation [1]. In urban water management, smart metering, monitoring, and demand-management strategies have also been identified as effective tools for sustainable resource use, particularly when combined with public awareness, regulation, and context-specific planning [2]. More broadly, IoT-based smart city systems integrate communication technologies, standards, and performance metrics across domains such as energy, transport, healthcare, and logistics, but they also face important challenges related to interoperability, scalability, security, and implementation complexity [3,4].

Among these domains, the energy sector imposes particularly strict requirements on smart-metering communication, as metering data must be collected reliably, frequently, and at scale. Advanced metering infrastructure (AMI) forms the communication backbone that enables this functionality in modern electrical distribution grids. It supports the bidirectional exchange of energy-consumption data, demand-response commands, and grid-edge measurements between customer-side smart meters and utility back-end systems. As a result, AMI has become central to load forecasting, infrastructure planning, distributed-energy-resource integration, and electric-vehicle deployment [5–7]. As smart-meter penetration grows and reporting intervals shrink from hours to minutes, AMI is increasingly a large-scale communication and computing platform, not merely a measurement system. Therefore, AMI networks must satisfy quality-of-service (QoS) requirements across massive numbers of devices while remaining reliable, scalable, and cost-effective [8–11].

Cellular networks are an attractive backbone for AMI because they offer wide coverage and reuse existing infrastructure, but dedicated spectrum is expensive, and licensing is slow. A practical alternative that has emerged in the recent literature is the cognitive mobile virtual network operator (C-MVNO) arrangement, in which a utility leases a small slice of dedicated bandwidth from a host operator and supplements it through opportunistic access to channels left idle by primary users (PUs). The C-MVNO must decide, every frame, which opportunistic channels to attempt, and how to assign the available channel pool to a fleet of data aggregation points (DAPs) that serve clusters of smart meters [12]. Three sources of uncertainty interact in this decision: spectrum sensing is imperfect because of noise, fading, and limited integration time; primary-user activity is time-varying and often correlated across adjacent channels of the same legacy service; and the wireless propagation environment is stochastic, with shadowing and small-scale fading determining whether a chosen DAP–channel pair will actually deliver a packet at the required SINR.

Existing approaches address these uncertainties only partially. Classical bipartite-matching schedulers based on the Hungarian algorithm produce the optimal per-frame DAP-to-channel assignment for any given cost matrix [12]. However, they lack mechanisms to anticipate or avoid PU activity in the next epoch and do not exploit cross-channel

temporal correlation. Heuristic cognitive-radio policies such as energy-detection sense-then-transmit, threshold rules based on Bayesian PU posteriors, and queue-greedy access [13,14] operate on individual channels and ignore the queue state and the joint structure of the PU process. Recent deep-reinforcement-learning (DRL) approaches to dynamic spectrum access [15–19] learn the cognitive policy from interaction but operate at a single level: they couple channel selection and direct medium access in one policy, which forces the agent to relearn the assignment problem and limits scalability to fleets with many DAPs.

In this paper, we argue that different mechanisms best handle the per-frame assignment problem and the cognitive channel-selection problem. The assignment problem is a bipartite matching problem with known optimal classical solutions, and it is solved millions of times per second; embedding it within a learned policy is wasteful. The cognitive problem is sequential decision-making under partial observability, has no closed-form solution, and benefits from the statistical generalization that DRL provides. We therefore propose a hierarchical architecture that explicitly separates the two: a Proximal Policy Optimization (PPO) agent at the upper layer makes a discrete cognitive decision over the opportunistic channel pool every epoch, and a Hungarian solver at the lower layer performs the per-frame DAP-to-channel assignment using an SINR-driven cost matrix. The agent's state combines per-DAP backlog, queue age, recent SINR estimates, and noisy sensing readings; its reward credits deliver packets, penalize PU collisions, expired packets, and queue aging, and reward fairness across DAPs.

From a sensor-network perspective, the proposed architecture focuses on the reliability of large-scale metering data collection rather than only on the allocation of wireless resources. Smart meters act as distributed sensing nodes, DAPs operate as aggregation points, and the cognitive cellular layer determines whether sensed data can be delivered within latency and reliability constraints. It positions the contribution at the intersection of sensor-network operation, intelligent spectrum access, and smart-grid communication infrastructure.

We make the following contributions:

- We introduce a hierarchical resource-allocation architecture for smart-metering sensor networks under C-MVNO operation, combining a PPO-based cognitive layer with an exact Hungarian per-frame matcher. To the best of our knowledge, this is the first study to explicitly couple a learned cognitive layer with an exact assignment stage to improve the reliability of sensed data in cognitive spectrum access.
- We develop a reproducible 3GPP TR 38.901-compliant simulation framework that incorporates Urban-Macro NLOS path loss, log-normal shadow fading, Rayleigh small-scale fading, SINR estimation error, a correlated Gilbert–Elliott primary-user model, configurable false-alarm and missed-detection probabilities, packet deadlines, and per-smart meter accounting.
- We define a strict satisfaction metric that counts a smart meter as satisfied only when both its packet delivery ratio is at least 0.9, and its mean delivery latency is at or below the deadline threshold. It closes a loophole in aggregate metrics that allows policies to appear acceptable while, in fact, starving most smart meters.
- We evaluate the proposed agent against three baselines—random, a Bayesian-belief policy that maintains a posterior over PU occupancy, and the no-cognition Hungarian-only configuration that recovers the formulation of [12]—together with an ideal-sensing reference that has access to the true PU state at the start of each cognitive epoch. We additionally validate the result against a Deep Q-Network (DQN) trained on a discrete re-encoding of the action space, confirming that the cognitive benefit is not specific to the policy-gradient family.
- We perform an architectural ablation, comparing the proposed DAP-tier architecture with a no-DAP configuration in which each smart meter serves as its own aggregator

under the same scheduler and agent. Removing the DAP layer drops packet delivery ratio by 51% and satisfaction by 94%, indicating that the DAP tier is a precondition for AMI viability under C-MVNO operation rather than an optimization.

The rest of the paper is organized as follows. Section 2 reviews related work in AMI scheduling, cognitive-radio networking for IoT and smart grid, deep reinforcement learning for spectrum access, and hybrid classical–learning architectures, and positions our contribution within this literature. Section 3 formalizes the system model, including the topology, the PHY model, the primary-user dynamics, the queuing model, and the metrics. Section 4 presents the hierarchical architecture and its two algorithms. Section 5 details the simulation setup, baselines, and training configuration. Section 6 reports the experimental results. Section 7 concludes and indicates future work.

2. Related Work

2.1. Resource Allocation in AMI and Smart-Metering Networks

The architectural template adopted in this paper—smart meters aggregated at DAPs and DAPs forwarding to a base station over cellular spectrum—has been refined steadily over the last decade. Early work focused on DAP placement as a clustering problem [10], with k-means and graph-partitioning heuristics used to balance load across aggregators. As deployments grew, the focus shifted to QoS-aware scheduling on the constrained AMI uplink. Khan et al. [8,9,20] formulate cost-minimization problems that account for CPU, bandwidth, and buffer use at IoT gateways and propose machine-learning-assisted scheduling that meets latency and reliability targets while reducing infrastructure cost. Gallardo et al. [11] demonstrate a LoRa-based AMI architecture for residential smart grids that emphasizes low energy consumption and coverage. Orlando et al. [7] integrate AMI into a broader smart-grid IoT platform that supports both telemetry and control traffic.

Prior work on virtualized wireless-resource allocation has shown that opportunistic channel assignment can improve spectrum efficiency in smart-metering communications under C-MVNO-like operation [21]. This resource-virtualization perspective provides the conceptual background for the present architecture, where learning is introduced not to replace the exact assignment stage, but to determine which opportunistic channels should be exposed to the matcher under primary-user uncertainty.

Closest to this paper is the recent work by Jara et al. [12], which formulates the cellular-AMI uplink under C-MVNO operation as a bipartite matching problem. In each frame, DAPs are matched to channels drawn from a mixed pool of dedicated and opportunistic resources, with the cost matrix combining queue depth, instantaneous SINR, and a token-based fairness term. The Hungarian algorithm is used to solve the matching exactly. The authors model channel availability through a Markov process and report substantial gains over greedy and random baselines. Two design choices in [12] motivate the present work. First, the matcher operates on the realized channel state for the current frame and lacks a mechanism to anticipate primary-user activity or to avoid channels that are likely to become busy in the near future. Second, the cost matrix is fixed; it does not adapt to recent sensing outcomes or to the cross-channel correlation structure that arises when the same legacy service uses adjacent opportunistic channels. Both gaps are addressed by adding a learned cognitive layer above the matcher, as described in Section 4.

2.2. Cognitive Radio for IoT and Smart-Grid Networks

Cognitive radio enables secondary users to access channels left idle by primary users opportunistically and has been studied extensively as a means to address spectrum scarcity in IoT and smart-grid deployments [22–24]. Wang et al. [25] formulate the multi-band spectrum-sensing and power-allocation problem for cognitive smart-grid communication

as a constrained optimization that maximizes data rate subject to a primary-protection threshold. The authors of [26] propose LoRaCog, a cognitive extension of LoRaWAN in which gateways perform spectrum sensing and the network server schedules opportunistic transmissions on additional licensed channels; the scheme reduces packet rejection without increasing end-device energy. Salameh et al. [27] examine the integration of cognitive radio with NOMA in 5G-and-beyond networks and identify spectrum scarcity in unlicensed bands as a primary motivation for cognitive operation in massive-IoT settings.

A recurring theme in the literature is that sensing is imperfect under realistic noise, fading, and integration constraints. Surveys by Nasser et al. [13] and Manco et al. [14] document the limitations of energy-detection sensing at low SNR and the role of cooperative and learning-based detectors in improving the missed-detection/false-alarm trade-off. Recent contributions extend cooperative sensing to intelligent-reflecting-surface-assisted [28] and full-duplex [29] settings. We adopt this view in our simulator: sensing is modeled as Bernoulli noise on top of the true PU state, with configurable false-alarm and missed-detection probabilities, and policies that consume the sensor output must do so under uncertainty.

2.3. Deep Reinforcement Learning for Spectrum Access

Reinforcement learning has been widely applied to dynamic spectrum access in cognitive radio networks. Wang et al. [15] introduce a usage-aware DRL scheme for interweaving cognitive-radio networks that explicitly learns temporal and cross-channel correlations in PU activity from historical traces, and reports improved access probability with limited PU collisions. Yuan et al. [16] integrate DQN with a graph-attention network in a multi-agent architecture for hybrid overlay-underlay CRNs with slicing, where DQN selects channel and power, and the graph-attention module aggregates neighbor information for local cooperation. Bai et al. [17] use centralized training and decentralized execution to optimize joint channel selection and power control for multi-user opportunistic spectrum access; the critic is augmented with a multi-head self-attention module that learns user-specific weights, and the reward is decomposed online into terms for throughput, collisions, and energy constraints. Tan et al. [18] model multi-user multi-channel DSA as a Dec-POMDP and adopt a centralized-offline-training, fully-distributed-online-execution framework using cooperative multi-agent DRL with deep recurrent Q-networks for each secondary user, addressing partial observability and learning a collision-avoiding joint policy without runtime signaling.

Other directions in this literature include independent-learner DRQN architectures that exploit temporal correlation in PU and SU activity [19], multi-agent UCB-Hoeffding exploration coupled with double DQN for cognitive UAV networks [30], deep Q-learning with guaranteed convergence in distributed and uncoordinated cognitive-radio operation [31], and energy-harvesting cognitive-IoT with jamming attacks treated through a DDQN with interference-aware bandit exploration [32]. Several recent contributions also embed DRL within broader 6G resource-management problems, including dueling-DQN sub-channel assignment combined with DDPG power allocation in mmWave massive MIMO-NOMA [33], and federated DRL for vehicular spectrum access with QoS-shaped rewards [34]. Recent surveys [35] synthesize these trends and highlight robustness to adversarial behavior, evaluation under realistic mobility and interference, and fairness analysis as remaining open issues.

What is consistent across this body of work is that DRL operates as a single-tier policy: it both selects channels and effectively performs the medium-access decision. As fleets grow—in our setting, hundreds to thousands of smart meters aggregated at a small number of DAPs—this monolithic design forces the agent to relearn a structure that can be solved

exactly by classical optimization, and it scales poorly because the joint action space grows combinatorially with the number of aggregators.

2.4. Hybrid Classical and Learning-Based Architectures

A small but growing literature investigates hybrids between exact or classical optimization and learning-based policies for resource allocation. Xu et al. [36] formulate a long-term system-cost minimization for an edge-computing cognitive-radio network and use Lyapunov optimization to convert the constrained long-term problem into a per-slot drift-plus-penalty minimization that PPO then solves; Lyapunov optimization plays the analytical role, and PPO learns the policy. Wang et al. [37] consider a multi-IRS-assisted sensing-enhanced spectrum-sharing network with physical-layer-security constraints and propose a hierarchical DRL scheme that combines D3QN and Soft Actor–Critic at different decision granularities, with an alternating-optimization mathematical benchmark used for verification. Farhi et al. [38] couple reinforcement learning with the Snake Optimizer metaheuristic to optimize spectrum utilization in cognitive radio networks. Cheggour and Loscri [39] hybridize RL with the Aquila Optimizer for frequency-resource management in 6G user-centric cell-free massive MIMO. Chigaba et al. [40] integrate NSGA-II with PPO for spectrum sharing in 6G CRNs, with NSGA-II generating multi-objective Pareto-optimal candidates and PPO refining policy in this space.

These hybrids share a common pattern: the classical or metaheuristic component handles a well-structured sub-problem (Lyapunov drift, alternating optimization, evolutionary search), and the learning component handles the residual stochastic decision-making. However, the available hybrid designs do not explicitly use exact Hungarian matching as the assignment layer beneath a learned cognitive-access policy for AMI sensor-network operation. The present work addresses this gap by separating sequential cognitive channel exposure from exact per-frame DAP-to-channel assignment.

2.5. Position of This Work

The literature surveyed above motivates four observations. First, AMI scheduling research has converged on a DAP-tiered architecture matched to channels by exact algorithms, but the resulting matchers do not exploit cognitive information about primary-user dynamics [8,12]. Second, cognitive-radio research has produced a rich set of sensing-and-access policies that handle imperfect detection but typically operate per channel and ignore queue and load dynamics [13,25]. Third, DRL for spectrum access has shown the value of learning cross-channel temporal correlation, but it does so with a single monolithic policy that absorbs the assignment problem and scales poorly to many aggregators [15,17,18]. Fourth, hybrids of learning and classical optimization in cognitive radio have so far paired DRL with Lyapunov, alternating optimization, or evolutionary search, but not with exact Hungarian matching [36,37,40]. The contribution of this paper is to close this gap through a hierarchical sensor-network architecture that preserves exact Hungarian matching at the per-frame assignment layer while using PPO to learn cognitive channel exposure under partial observability, cross-channel correlation, queue pressure, and latency constraints in the AMI-under-C-MVNO regime.

We emphasize what is and is not new. The individual components—Proximal Policy Optimization, the Hungarian assignment algorithm, and Bayesian occupancy estimation—are established, and we deliberately reuse them without modification. The novelty is their composition into a two-timescale hierarchy tailored to the AMI-under-C-MVNO problem, together with the empirical finding that this composition is what makes the system work. Three specific points distinguish the present work from the closest prior art. Relative to AMI Hungarian-matching schedulers [12], our matcher is driven by a learned cogni-

tive layer that anticipates and avoids primary-user activity, whereas the prior fixed-cost matcher operates only on the realized current-frame state. Relative to DRL-based dynamic spectrum access [15,17,18], which employs a single monolithic policy that must relearn the assignment problem and whose joint action space grows combinatorially with the number of aggregators, we separate the exactly solvable assignment from the learned cognitive decision, so the assignment remains optimal and the learned policy scales in the number of channels rather than the number of aggregators. Relative to hybrid learning-optimization architectures [36,37,40], which pair learning with Lyapunov drift, alternating optimization, or metaheuristic search, we are, to the best of our knowledge, the first to couple a learned cognitive layer with an exact Hungarian assignment for cognitive-radio resource allocation, and to demonstrate through an architectural ablation that the aggregation tier is a precondition for viability rather than an optimization.

Table 1 summarizes the most representative prior studies across these categories, their methodologies, their key findings or limitations, and how the present work improves upon or differs from them.

Table 1. Comparison with representative related studies. The present work is the only approach that couples a learned cognitive layer with an exact per-frame Hungarian assignment in the AMI-under-C-MVNO regime.

Study/Category	Methodology	Key Findings/Limitation	How This Work Differs
AMI Hungarian scheduling [12]	Exact bipartite (Hungarian) DAP-to-channel matching with a fixed cost matrix	Optimal per-frame assignment; DAP tiering improves scalability, but the matcher is blind to primary-user dynamics and cross-channel correlation	We retain the exact matcher but drive it with a learned cognitive layer that anticipates and avoids primary-user activity
Cognitive-radio sensing/access [13,25]	Energy detection, Bayesian occupancy posteriors, threshold or queue-greedy access	Handle imperfect sensing, but operate per channel and ignore queue state, load, and joint primary-user structure	Our agent conditions on backlog, queue age, sensing history, and SINR jointly, and exploits cross-channel correlation
DRL dynamic spectrum access [15,17,18]	Single monolithic DRL policy over channel selection and medium access	Learns cross-channel temporal structure, but absorbs the assignment problem and scales poorly as aggregators grow	We separate the exactly solvable assignment from the learned decision, so assignment stays optimal and learning scales in channels, not aggregators
Hybrid learning + optimization [36,37,40]	DRL paired with Lyapunov drift, alternating optimization, or metaheuristics	Combine learning with classical optimization, but not with exact assignment, and not for the AMI C-MVNO setting	First coupling of a learned cognitive layer with exact Hungarian assignment; ablation shows the DAP tier is a precondition for viability
This work	Two-timescale hierarchy: PPO cognitive layer + exact Hungarian assignment	Higher delivery and lower latency; strict satisfaction 0.702 vs. 0.633 (Bayesian) and 0.514 (Hungarian-only); robust across seeds and sweeps	—

Bold denotes the present work.

3. Problem Formulation and Methodology

3.1. Topology

We consider an AMI deployment over a square service area of side length L meters, served by a single base station (BS) located at its center. A fleet of N smart meters is distributed within the area according to a configurable spatial distribution. The smart meters are aggregated into $K \ll N$ DAPs, whose positions are determined by a placement strategy: k-means clustering of the smart meter locations, a regular grid, or a uniform random distribution. Each smart meter is assigned to its nearest DAP. Communication between smart meters and their associated DAP is treated as a logical pipe. In practice, this short-range connection may correspond to a wired link between a DAP and nearby smart meters, such as those on the same street, or to a low-power wireless technology such as LoRaWAN or a low-rate wireless personal-area network/6LoWPAN when wired deployment is impractical. Since the focus of this work is cellular spectrum access at the aggregation layer, the abstraction isolates the cellular aggregation bottleneck and avoids conflating short-range access effects with the cognitive spectrum-access decision studied here. The smart-meter-to-DAP links are assumed reliable and are not modeled at the PHY layer. By contrast, the DAP-to-BS link uses the cellular spectrum and is fully modeled. Figure 1 illustrates the topology under the default deployment with $N = 120$ smart meters, $K = 4$ DAPs, k-means placement, and an area side of $L = 1000$ m.

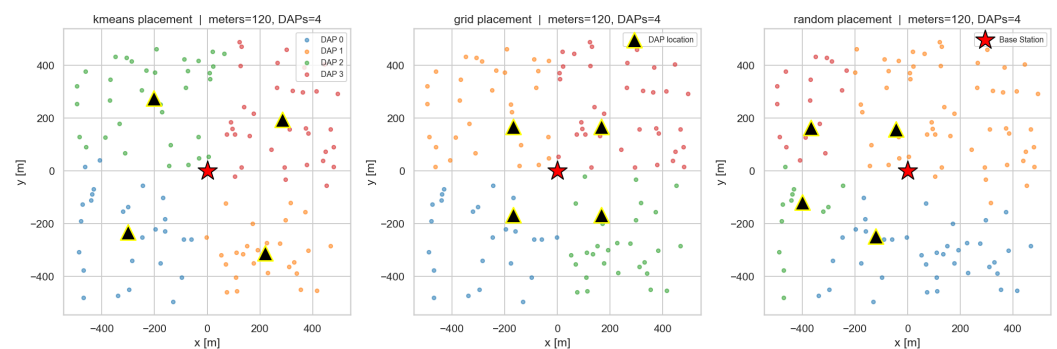


Figure 1. Service area topology with $N = 120$ smart meters (small markers), $K = 4$ DAPs (triangles, k-means placement), and one base station at the center (star). Meter color indicates DAP assignment.

3.2. Cellular Spectrum and Primary-User Model

The cellular pool consists of $C = C_{\text{ded}} + C_{\text{opp}}$ channels, of which C_{ded} are dedicated and always available to the C-MVNO, and C_{opp} are opportunistic and shared with primary users. Each opportunistic channel c has a binary occupancy state $S_c(t) \in \{0, 1\}$, with $S_c(t) = 1$ meaning the primary user is busy on channel c at frame t . The single-channel evolution follows a discrete-time two-state Markov (Gilbert–Elliott) process,

$$\Pr(S_c(t+1) = 1 \mid S_c(t) = 0) = p_{01}, \quad \Pr(S_c(t+1) = 0 \mid S_c(t) = 1) = p_{10}, \quad (1)$$

with stationary busy probability $\pi_b = p_{01} / (p_{01} + p_{10})$.

To capture the practical case of legacy services occupying contiguous spectrum, we extend the per-channel process with cross-channel correlation. The opportunistic channels are partitioned into groups $\mathcal{G} = \{G_1, \dots, G_m\}$, and within each group the per-channel transitions are correlated by a coupling factor $\rho \in [0, 1]$. With probability ρ , all channels in a group share the same independent Markov draw; with probability $1 - \rho$, each channel evolves independently. It recovers the independent-channel case at $\rho = 0$ and the fully coupled case at $\rho = 1$. In our experiments, we use $\rho = 0.7$ and two groups of two channels each, modeling pairs of adjacent channels of the same primary service. Per-channel sensing

heuristics cannot exploit this correlation; learned policies that condition on cross-channel history can.

The C-MVNO performs imperfect spectrum sensing at the start of each frame. The sensor returns a binary observation $\hat{S}_c(t) \in \{0, 1\}$ on each opportunistic channel, with

$$\Pr(\hat{S}_c(t) = 1 \mid S_c(t) = 0) = p_{fa}, \quad \Pr(\hat{S}_c(t) = 0 \mid S_c(t) = 1) = p_{md}, \quad (2)$$

where p_{fa} is the false-alarm probability and p_{md} is the missed-detection probability. Both are configurable inputs to the simulator and are varied in the sensing-sweep experiment of Section 6.3.

A PU collision occurs whenever the C-MVNO transmits on an opportunistic channel that is, in fact, busy. PU collisions are penalized in the reward and are also reported separately as a system metric.

3.3. Wireless Propagation and SINR Model

We model the DAP-to-BS link with the 3GPP TR 38.901 Urban-Macro NLOS path-loss formula [41]. For a DAP at planar distance d from the BS, with DAP height h_{DAP} and BS height h_{BS} , the path loss in dB at carrier frequency f_c GHz is

$$PL_{NLOS}(d) = 13.54 + 39.08 \log_{10}(d_{3D}) + 20 \log_{10}(f_c) - 0.6 (h_{DAP} - 1.5), \quad (3)$$

where $d_{3D} = \sqrt{d^2 + (h_{BS} - h_{DAP})^2}$. To this we add log-normal shadow fading with standard deviation σ_{sh} dB and a Rayleigh small-scale fading term $|h|^2$, drawn each frame from an exponential distribution with a unit mean. The instantaneous received signal power at the BS, in linear scale, is therefore

$$P_r = P_t \cdot 10^{-(PL_{NLOS}(d) + X_\sigma)/10} \cdot |h|^2, \quad (4)$$

where P_t is the DAP transmit power and $X_\sigma \sim \mathcal{N}(0, \sigma_{sh}^2)$ is the shadow term.

The thermal noise floor at receiver bandwidth B Hz with noise figure F dB is $N_0 = kTB \cdot 10^{F/10}$, where k is Boltzmann's constant and T the reference temperature. The SINR for a DAP-channel pair (k, c) at frame t is

$$\gamma_{k,c}(t) = \frac{P_r^{(k,c)}(t)}{N_0 + I_{k,c}(t)}, \quad (5)$$

where $I_{k,c}(t)$ aggregates interference from any concurrent transmissions; in the single-cell single-MVNO setting with at most one transmission per channel per frame, $I_{k,c}(t) = 0$ in all our scenarios, and the operating point is interference-free at the level of the C-MVNO. A transmission on (k, c) at frame t succeeds at the link layer if $\gamma_{k,c}(t) \geq \gamma_{th}$, the SINR threshold. The C-MVNO does not observe the true SINR; instead, it observes a noisy estimate $\tilde{\gamma}_{k,c}(t)$ formed by adding zero-mean Gaussian noise with standard deviation σ_{est} dB to the true SINR in the dB domain, modeling channel-state-information (CSI) acquisition imperfections.

3.4. Traffic, Queuing, and Deadlines

The system has two scheduling scopes that should be distinguished at the outset. The first is the smart meter-to-DAP assignment, which is fixed at deployment time by the placement strategy in Section 3.1 and is not re-optimized at runtime. The second is the DAP-to-channel allocation, which is solved each frame by the lower layer of the proposed architecture (Section 4.1). Within a DAP, packets are drained from a first-in-first-out queue without per-smart meter prioritization; this reflects the homogeneity of AMI telemetry,

where no smart meter has a differentiated quality of service relative to its peers within the same aggregator.

Each smart meter independently generates a packet at frame t with probability p_{arr} . Packets are forwarded immediately over the logical short-range link to the smart meter's assigned DAP, where they enter a per-DAP first-in-first-out queue. Queues are unbounded; we never drop a packet because the buffer is full. Each packet j has an arrival timestamp τ_j and an age $a_j(t) = t - \tau_j$ in frames.

To model the AMI service-level agreement, it is assumed that late telemetry has limited operational value; packets older than a deadline D frames are expired. At the start of each frame, every queue is swept, and any packet j with $a_j(t) \geq D$ is removed and counted as expired. Expired packets are not subsequently delivered. The expiry sweep occurs before the cognitive decision and matching, so the queues the policies operate on contain only live packets.

When a DAP is matched to a channel and the link succeeds, it transmits a batch of up to B_{tx} packets from the head of its queue. Successful packets are recorded with their delivery latency $\ell_j = t - \tau_j$.

3.5. Per-Smart Meter Metrics

We track aggregate and per-smart meter outcomes. For each smart meter i , let A_i be its number of packet arrivals, D_i its number of delivered packets, E_i its number of expired packets, and $\bar{\ell}_i = (1/D_i) \sum_{j \in \text{delivered}_i} \ell_j$ its mean delivery latency (or undefined if $D_i = 0$). Its per-smart meter packet delivery ratio (PDR) is $\rho_i = D_i/A_i$.

Two aggregate metrics are central. Fleet PDR is the simple mean of per-smart meter PDRs, $\bar{\rho} = (1/N) \sum_i \rho_i$. Mean delivery latency over the fleet is computed only over delivered packets, $\bar{\ell} = \frac{1}{\sum_i D_i} \sum_i \sum_j \ell_j$.

The third metric is the satisfaction rate. We define a smart meter as *satisfied* if and only if both its mean latency is at or below a threshold L_{sat} frames and its PDR is at or above a minimum ρ_{min} :

$$\text{Sat}_i = \mathbf{1}[D_i > 0 \wedge \bar{\ell}_i \leq L_{\text{sat}} \wedge \rho_i \geq \rho_{\text{min}}]. \quad (6)$$

The fleet satisfaction rate is $\bar{S} = (1/N) \sum_i \text{Sat}_i$. The per-smart-meter PDR floor ρ_{min} closes a loophole present in latency-only definitions, where a smart meter that delivers a single packet quickly and loses the rest can appear satisfied. With $\rho_{\text{min}} = 0.9$, satisfaction indicates that a smart meter is genuinely delivering its data.

3.6. Reward Function

The reinforcement-learning agent operates on epochs of T_{ep} frames. The per-epoch reward credits successful deliveries, penalizes PU collisions, expired packets, and aged backlog, and rewards balanced load across DAPs:

$$r(t_{\text{ep}}) = w_d D_{\text{ep}} - w_c C_{\text{ep}} - w_e E_{\text{ep}} - w_q \bar{a}_{\text{ep}} + w_f J_{\text{ep}}, \quad (7)$$

where D_{ep} is the number of packets delivered in the epoch, C_{ep} the number of PU collisions, E_{ep} the number of packets expired, $\bar{a}_{\text{ep}}$ the mean queue age in frames at the end of the epoch, and J_{ep} Jain's fairness index across DAP delivery counts in the epoch. The weight vector $(w_d, w_c, w_e, w_q, w_f)$ is tuned once and held fixed across all experiments; values are reported in Section 5. The reward is multiplied by a global scale λ_r to keep its dynamic range comparable to the PPO normalizers.

4. Proposed Hierarchical Architecture

The decision problem decomposes naturally into two layers operating at different time scales and with different mathematical structures. Figure 2 depicts the architecture.

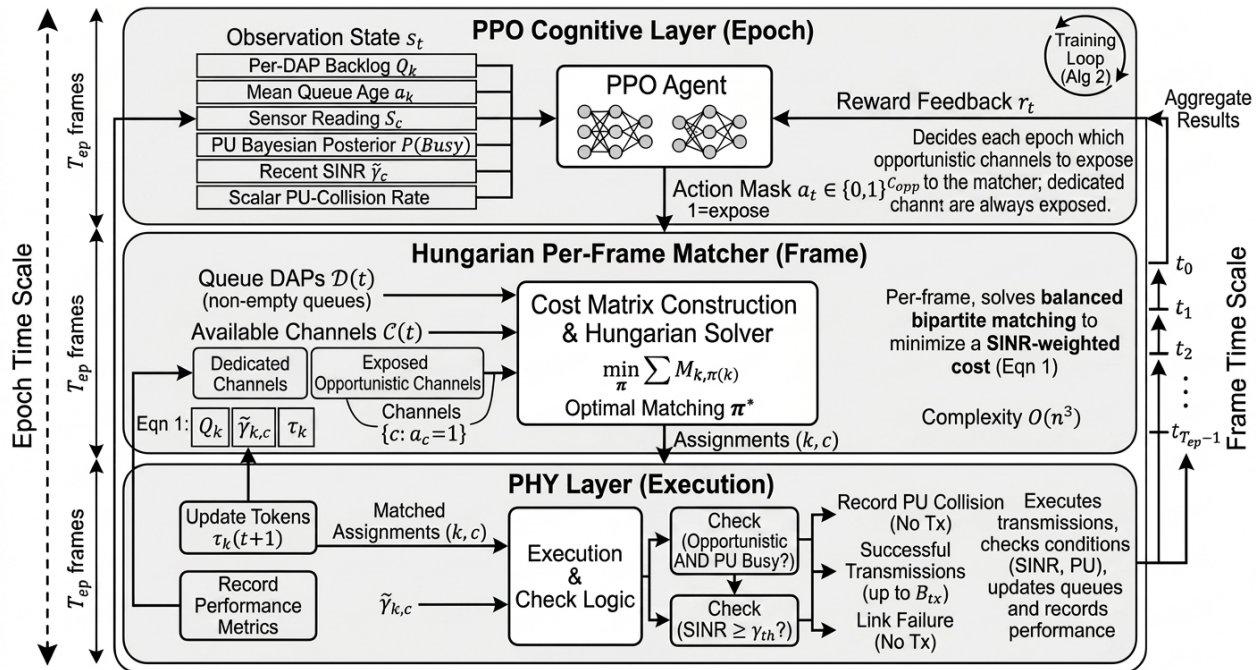


Figure 2. Hierarchical architecture. The PPO cognitive layer (**top**) decides, at each epoch, which opportunistic channels to expose to the matcher. The Hungarian per-frame matcher (**middle**) assigns DAPs to the available channel pool to minimize an SINR-weighted cost; π^* denotes the resulting optimal (cost-minimizing) assignment. The PHY layer (**bottom**) executes the matched transmissions over the cellular link.

4.1. Lower Layer: Per-Frame Hungarian Matching

At every frame t , given a pool of available channels (always-on dedicated channels plus opportunistic channels exposed by the cognitive layer), the lower layer assigns DAPs to channels by solving a balanced bipartite assignment. Let $\mathcal{D}(t)$ be the set of DAPs with non-empty queues at frame t and $\mathcal{C}(t)$ the set of currently exposed channels. We construct a square cost matrix $\mathbf{M}(t)$ of size $n \times n$ with $n = \max(|\mathcal{D}(t)|, |\mathcal{C}(t)|)$, padding with virtual rows or columns of zero-cost dummy entries as needed. Each real entry $M_{k,c}(t)$ combines a queue term, an SINR term, and a token-based fairness term:

$$M_{k,c}(t) = \alpha Q_k(t) + \beta (\gamma_{\max} - \tilde{\gamma}_{k,c}(t)) + \delta \tau_k(t), \quad (8)$$

where $Q_k(t)$ is the queue depth at DAP k , $\tilde{\gamma}_{k,c}(t)$ is the noisy SINR estimate, $\gamma_{\max}$ is a normalizing reference, $\tau_k(t)$ is a per-DAP token (incremented when the DAP is not served and reset on service), and (α, β, δ) are non-negative weights. Lower entries are preferred. The Hungarian algorithm [42] solves the assignment in $\mathcal{O}(n^3)$ time.

Algorithm 1 formalizes one frame of the lower layer. The cognitive layer's decision enters only through the construction of $\mathcal{C}(t)$.

Algorithm 1 Per-frame Hungarian matching (lower layer)

Require: Set of DAPs $\mathcal{K}$, exposed channel set $\mathcal{C}(t)$, queue depths $\{Q_k(t)\}$, noisy SINR estimates $\{\tilde{\gamma}_{k,c}(t)\}$, tokens $\{\tau_k(t)\}$, weights (α, β, δ) , threshold γ_{th} , batch size B_{tx} .

- 1: Expire packets in all queues with age $\geq D$.
- 2: $\mathcal{D}(t) \leftarrow \{k \in \mathcal{K} : Q_k(t) > 0\}$.
- 3: Build cost matrix $\mathbf{M}(t)$ of size $n \times n$ with $n = \max(|\mathcal{D}(t)|, |\mathcal{C}(t)|)$, using (8) for real entries and zero-cost dummies for padded ones.
- 4: Solve $\pi^* = \arg \min_{\pi} \sum_k M_{k,\pi(k)}(t)$ by Hungarian algorithm.
- 5: **for** each real assignment $(k, c) = (k, \pi^*(k))$ **do**
- 6: Compute true SINR $\gamma_{k,c}(t)$ from (4)–(5).
- 7: **if** c is opportunistic and $S_c(t) = 1$ **then**
- 8: Record PU collision; do not transmit.
- 9: **else if** $\gamma_{k,c}(t) \geq \gamma_{th}$ **then**
- 10: Transmit up to B_{tx} packets from queue k ; record latencies.
- 11: $\tau_k(t+1) \leftarrow 0$.
- 12: **else**
- 13: Link failure; queue unchanged.
- 14: $\tau_k(t+1) \leftarrow \tau_k(t) + 1$.
- 15: **end if**
- 16: **end for**
- 17: **for** $k \in \mathcal{K} \setminus \mathcal{D}(t)$ or unassigned **do**
- 18: $\tau_k(t+1) \leftarrow \tau_k(t) + 1$.
- 19: **end for**

4.2. Upper Layer: PPO Cognitive Decision

The cognitive decision is made at the start of each epoch (a block of T_{ep} consecutive frames) and held fixed for that epoch. The action is a binary mask $\mathbf{a} \in \{0, 1\}^{C_{opp}}$, with $a_c = 1$, meaning that opportunistic channel c is exposed to the matcher for the next T_{ep} frames. Dedicated channels are always exposed and are not subject to the cognitive decision.

The state $\mathbf{s} \in \mathbb{R}^d$ provided to the agent at the start of each epoch concatenates:

- Per-DAP backlog Q_k , normalized by a reference value, K entries.
- Per-DAP mean queue age $\bar{a}_k$ in frames, normalized, K entries.
- Per-channel sensor reading $\hat{S}_c$ from the start of the epoch, C_{opp} entries.
- Per-channel running PU-busy estimate (Bayesian posterior of busy probability under the sensor model from (2) and the Markov dynamics of (1)), C_{opp} entries.
- Per-channel recent SINR estimate $\tilde{\gamma}_c$ at the most recent successful match, C_{opp} entries.
- One scalar capturing the most recent PU-collision rate.

The total state dimension is $d = 2K + 3C_{opp} + 1$, which evaluates to 21 at the default $K = 4$, $C_{opp} = 4$.

The agent is a PPO agent [43] with a multi-binary policy head producing C_{opp} independent Bernoulli factors. We use Stable-Baselines-3 [44] default hyperparameters except where noted in Section 5, with two-layer fully connected networks of 64 units per layer for both the policy and the value function. Training proceeds by interaction with the simulator: at each epoch step, the agent emits an action, the matcher executes T_{ep} frames under that action, and the cumulative reward (7) is returned. Algorithm 2 formalizes the training loop.

The hierarchy is one-directional: the upper layer commits an action that the lower layer executes for the next epoch, and the lower layer's outcomes (deliveries, collisions, expirations) feed back into the upper-layer state and reward. The Hungarian solver does not see the PU belief; that information is consumed solely by the cognitive layer. This separation allows the lower layer to remain optimal in its narrow role and the upper layer to specialize in cognitive decision-making.

Algorithm 2 Hierarchical training loop (upper layer)

Require: Initial policy π_θ , initial value function V_ϕ , total timesteps $T_{\max}$, epoch length T_{ep} , simulator ENV.

```

1: Initialize ENV to a fresh episode.
2: Initialize empty rollout buffer.
3: for  $t = 0, T_{\text{ep}}, 2T_{\text{ep}}, \dots, T_{\max}$  do
4:   Observe state  $\mathbf{s}_t$  from ENV.
5:   Sample action  $\mathbf{a}_t \sim \pi_\theta(\cdot \mid \mathbf{s}_t)$ .
6:   Run Algorithm 1 for  $T_{\text{ep}}$  frames with channel mask  $\mathbf{a}_t$ .
7:   Compute epoch reward  $r_t$  from (7).
8:   Store  $(\mathbf{s}_t, \mathbf{a}_t, r_t)$  in buffer.
9:   if episode ended (all queues drained or simulator horizon reached) then
10:     Reset ENV.
11:   end if
12:   if rollout buffer full then
13:     Compute advantages with generalized advantage estimation.
14:     Perform PPO clipped policy update on  $\pi_\theta$  and  $V_\phi$ .
15:     Empty rollout buffer.
16:   end if
17: end for
18: return  $\pi_\theta$ .

```

5. Simulation Setup

5.1. Default Configuration

Table 2 lists the default simulation parameters. The values are chosen to mirror a typical urban C-MVNO deployment: a 1 km² service area, 120 smart meters aggregated at four DAPs, two dedicated cellular channels, and four opportunistic channels at 2 GHz with 180 kHz bandwidth (one LTE physical resource block), 23 dBm DAP transmit power, 8 dB shadow fading, and 2 dB SINR estimation error. The PU process has a stationary busy probability of 0.4 with $\rho = 0.7$ correlation in two groups of two channels each. Sensing has false-alarm and missed-detection probabilities of 0.15–0.3 in the main configuration. Packets arrive at each smart meter with probability 0.10 per frame, expire at deadline $D = 12$ frames, and are deemed satisfied at $L_{\text{sat}} = 6$ frames mean latency with $\rho_{\min} = 0.9$. Each epoch is $T_{\text{ep}} = 4$ frames long, and an episode runs for 80 epochs.

Table 2. Default simulation parameters used in the main configuration of Section 6.

Parameter	Symbol	Value
Number of smart meters	N	120
Number of DAPs	K	4
Service-area side	L	1000 m
DAP placement	—	k-means
Dedicated channels	C_{ded}	2
Opportunistic channels	C_{opp}	4
Carrier frequency	f_c	2.0 GHz
Channel bandwidth	B	180 kHz
DAP transmit power	P_t	23 dBm
SINR threshold	γ_{th}	0 dB
Shadow std. dev.	σ_{sh}	8 dB
SINR estimation error	σ_{est}	2 dB
BS/DAP heights	$h_{\text{BS}}, h_{\text{DAP}}$	25 m, 10 m
PU station is busy prob.	π_b	0.4
PU transition prob. idle→busy	p_{01}	0.10

Table 2. Cont.

Parameter	Symbol	Value
PU transition prob. busy $\rightarrow$ idle	p_{10}	0.15
PU correlation factor	ρ	0.7
PU channel groups	$\mathcal{G}$	$\{(0, 1), (2, 3)\}$
False-alarm prob.	p_{fa}	0.15
Missed-detection prob.	p_{md}	0.30
Packet arrival prob.	p_{arr}	0.10
Batch size per match	B_{tx}	4
Frames per epoch	T_{ep}	4
Episode length	$T_{ep} \cdot 80$	80 epochs
Delivery deadline	D	12 frames
Satisfaction latency threshold	L_{sat}	6 frames
Satisfaction PDR floor	ρ_{min}	0.9
Reward weight: deliver	w_d	1.0
Reward weight: collision	w_c	0.8
Reward weight: expiry	w_e	2.0
Reward weight: queue age	w_q	0.15
Reward weight: fairness	w_f	0.4
Reward scale	λ_r	0.05
Cost weight: queue	α	1.0
Cost weight: SINR	β	0.4
Cost weight: token	δ	0.5

5.2. Baselines

We evaluate the proposed PPO agent against three baselines and one strong reference. All baselines share the same lower-layer Hungarian matcher and operate on the same state space.

Random. The cognitive layer samples a uniform random action $\mathbf{a} \in \{0, 1\}^{C_{opp}}$ each epoch. It is the lowest reasonable baseline.

Bayesian belief. The cognitive layer maintained, for each opportunistic channel c , a recursive Bayesian posterior $\hat{\pi}_{b,c}(t)$ of the busy probability, updated each frame using the sensor reading $\hat{S}_c(t)$ under the noise model of (2) and propagated forward through the Markov dynamics of (1). The agent exposes a channel if and only if $\hat{\pi}_{b,c}(t) < 1 - \theta$, where $\theta = 0.7$ is a tuned threshold. This baseline captures what a well-engineered classical cognitive-radio policy can do given perfect knowledge of the model parameters but only noisy observations of the PU state.

No cognition. All C_{opp} channels are exposed to the matcher at all times, without prior filtering based on sensing, channel availability, or learned access decisions. The assignment stage, therefore, attempts to allocate channels regardless of their current occupancy state. This baseline represents a non-cognitive spectrum-access strategy in which channel selection is determined only by the matching process, rather than by spectrum awareness.

Ideal sensing (strong reference). The cognitive layer simulates one Markov step of the PU process forward in time using a private random number generator, then uses the resulting predicted state to expose only channels predicted to be idle. This baseline has access to information no real policy can have: the true PU state at the start of the next cognitive epoch, with zero sensor noise. Crucially, however, this prediction is a single-frame snapshot, and the chosen channel mask is then held fixed for the entire epoch of $T_{ep} = 4$ frames as the PU state continues to evolve under the Markov dynamics. Ideal sensing, therefore, acts on perfect information at the start of the epoch but on increasingly stale information by the end of it. By contrast, the PPO agent observes a richer state—queue backlog, queue ages, noisy sensor readings, the running Bayesian PU posterior, and the

recent SINR history—and can learn to commit to channel choices that account for the multi-frame dynamics of the epoch. Ideal sensing is therefore best read as a strong reference that quantifies how much of PPO’s advantage can be attributed to access to PU information versus to the additional sequential context that PPO uses, and not as an upper bound on attainable satisfaction in this regime.

Per-frame clairvoyant (upper-bound oracle). To establish a genuine performance ceiling, we additionally evaluate a per-frame clairvoyant oracle that re-decides the channel mask *every frame* using the true current-frame PU state with zero sensor noise. Unlike ideal sensing, this oracle is not subject to the epoch-commitment constraint: it exposes exactly the channels that are idle in the current frame, re-deciding as the PU state evolves. It cannot be realized by any deployable policy—no real system observes the true PU state, and all deployable policies (including PPO) commit their cognitive decision for a full epoch—but it upper-bounds what any cognitive layer could achieve given the remaining physical-layer and capacity constraints of the system. The contrast between this oracle and the epoch-committed ideal-sensing reference isolates the cost of the epoch-commitment constraint itself.

This comparison separates three effects: the value of cognitive filtering, the value of sequential learning under imperfect sensing, and the extent to which queue and SINR context can compensate for stale or incomplete primary-user information.

5.3. Training and Evaluation Protocol

The PPO agent is trained for 200,000 timesteps in the main experiment, where one timestep corresponds to one cognitive-decision epoch. After training, the agent is evaluated on 30 fresh episodes with seeds disjointed from the training set. Baselines are evaluated under the same 30-seed protocol with the same fixed seed schedule, so all policies see identical PU realizations and arrival patterns. The sensing-sweep experiment trains a fresh PPO agent for each of 16 cells in the false-alarm-by-missed-detection grid, with 50,000 timesteps per cell and 10 episodes per cell. The scalability and placement experiments use the same training-from-scratch protocol per configuration. All runs use 42 seeds for environment initialization.

For the architectural ablation (Section 6.7), we train PPO twice: once on the standard with DAPs configuration at 200,000 timesteps, and once on a no-DAPs configuration at 200,000 timesteps. The no-DAPs configuration is identical, except that $K = N$ (every smart meter is its own aggregator). The DAP transmit power is reduced to 14 dBm (typical IoT endpoint), and the DAP height is reduced to 1.5 m. The same Hungarian matcher is used in both modes. Using the same scheduler in both modes is an intentional worst-case assumption for the no-DAPs configuration, because real direct-to-cellular AMI deployments would use a random-access medium access control protocol. We use the same scheduler in both configurations to keep all variables, except the architectural choice, constant.

The DQN baseline (Section 6.8) is trained on a discrete re-encoding of the action space. The native MultiBinary(C_{opp}) action space has $2^{C_{\text{opp}}} = 16$ patterns, so we wrap the environment to expose Discrete(16) and map each integer to its corresponding binary mask. The DQN is trained for 300,000 timesteps on this wrapped environment.

The simulator and training code are implemented in Python 3.11 with PyTorch 2.5.1 and Stable-Baselines-3 [44]. The PPO and DQN hyperparameters are listed in Table 3; any values not listed there are the Stable-Baselines-3 defaults.

Reproducibility. The study uses no external dataset; all data is generated by the simulator from the parameters in Table 2, which together with the hyperparameters in Table 3 fully determine every reported result. The code is organized into a simulator core (the environment, the 3GPP physical-layer model, the correlated primary-user model, and the

DAP-placement routines), a policy module (the learned agent and the four baselines behind a common interface), a training-and-evaluation module, and a set of experiment notebooks that generate each table and figure in this paper. Every experiment fixes explicit random seeds for the environment, the policy, and the evaluation schedule, so runs are reproducible. The complete code and notebooks are available from the authors on reasonable request.

Table 3. Reinforcement-learning hyperparameters. Values not listed are Stable-Baselines-3 defaults.

Hyperparameter	PPO	DQN
Policy/value network	2 layers $\times$ 64 units	2 layers $\times$ 64 units
Activation	Tanh	ReLU
Learning rate	3×10^{-4}	3×10^{-4}
Discount factor γ	0.97	0.99
Rollout length (n_{steps})	1024	—
Minibatch size	64	128
Epochs per update	8	—
GAE λ	0.95	—
Entropy coefficient	0.01	—
Value-function coefficient	0.5	—
Clip range	0.2	—
Replay buffer size	—	50,000
Learning starts	—	2000
Target update interval	—	1000
Exploration fraction	—	0.3
Final exploration ϵ	—	0.05
Training timesteps (main)	200,000	300,000

6. Results and Discussion

6.1. Main Comparison

Table 4 summarizes the main comparison. To provide statistical rigor, PPO is trained from scratch under five independent random seeds and every metric is reported as the mean over seeds with its 95% confidence interval; the deployable and reference policies are evaluated under the same fixed 30-episode schedule. PPO achieves the strongest performance among all deployable policies across the headline metrics: satisfaction rate 0.702 ± 0.016 , PDR 0.963 ± 0.003 , mean latency 4.34 ± 0.15 frames, and the lowest PU-collision count among the cognitive policies. PPO's satisfaction advantage over the epoch-committed ideal-sensing reference (0.665 ± 0.020) and over the Bayesian baseline (0.633) is statistically significant, as the confidence intervals do not overlap. The two deterministic baselines (Bayesian belief and no-cognition) yield identical results across policy seeds under the fixed evaluation schedule and therefore have zero-width intervals. The per-frame clairvoyant oracle attains 0.972 satisfaction, far above every deployable policy, confirming that it is a genuine upper bound; we discuss its interpretation below.

Table 4. Main comparison. PPO metrics are reported as mean $\pm$ 95% confidence interval over five independent training seeds; each seed is evaluated over 30 episodes. The per-frame clairvoyant oracle is an unrealizable upper bound (see text). Default configuration of Table 2.

Policy	Reward	PDR	Sat.	Latency
Per-frame clairvoyant (oracle)	196.35	0.994	0.972	1.42
PPO (proposed)	172.98 $\pm$ 2.14	0.963 $\pm$ 0.003	0.702 $\pm$ 0.016	4.34 $\pm$ 0.15
Ideal sensing (reference)	172.00 $\pm$ 1.11	0.960 $\pm$ 0.002	0.665 $\pm$ 0.020	4.73 $\pm$ 0.11
Bayes belief	156.50	0.940	0.633	4.70
No cognition	146.58	0.925	0.514	5.10
Random	147.56 $\pm$ 5.29	0.923 $\pm$ 0.010	0.432 $\pm$ 0.021	6.42 $\pm$ 0.17

Bold indicates the best value among the deployable policies (the per-frame clairvoyant oracle is an unrealizable upper bound and is excluded from the comparison).

The ordering among the deployable policies is as expected. The Bayesian-belief baseline lies between PPO and the non-cognitive policies: it outperforms the no-cognition Hungarian-only configuration by roughly 12 percentage points on satisfaction, showing that even a well-engineered classical cognitive policy adds substantial value over matching alone, and PPO improves on it by a further 7 percentage points. The three-level structure of the table—oracle far above, PPO and the ideal-sensing reference in the middle, non-cognitive policies below—resolves an apparent paradox that the single-frame reference alone would raise: ideal sensing has perfect PU information yet scores below PPO. The reason is the epoch-commitment constraint. Ideal sensing acts on a perfect single-frame prediction that it then holds fixed for the full four-frame epoch, during which the PU state continues to evolve; PPO, observing per-DAP backlog, queue ages, the running Bayesian posterior, and recent SINR history, learns to commit to channel choices that account for the within-epoch dynamics and therefore slightly exceeds the epoch-committed reference. When the epoch-commitment constraint is lifted entirely (the per-frame clairvoyant oracle, which re-decides every frame), performance jumps to 0.972 satisfaction, confirming that the constraint—not a lack of information—is what separates the reference from the ceiling. The epoch-length sensitivity analysis in Section 6.9 quantifies this effect directly. For smart-metering sensor networks, PPO’s lower PU-collision count is operationally relevant because fewer collisions translate into fewer lost transmission opportunities for time-sensitive metering packets. Figure 3 visualizes the three core metrics across the deployable policies.

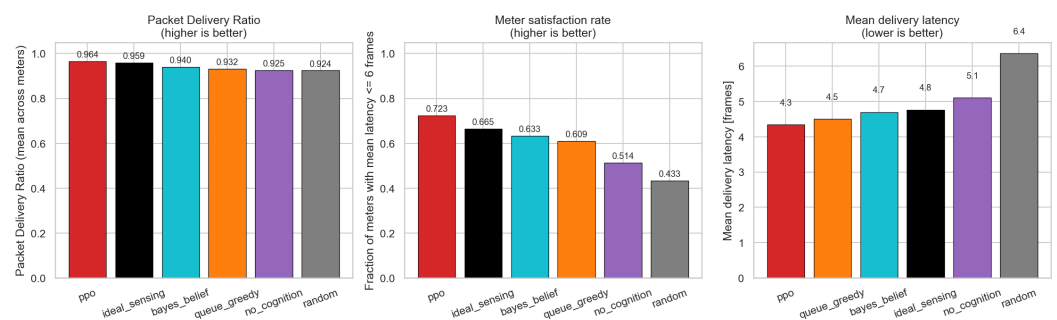


Figure 3. Headline three-metric comparison. From left to right: PDR, satisfaction rate, and mean delivery latency. PPO (red) leads on every metric.

Figure 4 shows the PPO training curve for the main run. Episode reward rises from 159.5 in the first five episodes to 180.0 in the last five; satisfaction climbs from 0.4–0.5 early to a stable 0.73 by mid-training, and latency falls from 6+ frames to 3.6 frames in the last five episodes. Convergence is monotonic to within seed noise.

6.2. Cognitive Avoidance Behavior

To verify that PPO has learned a meaningful cognitive policy and is not simply benefiting from incidental correlations, we tabulate its action distribution conditional on the true PU state. Across the full evaluation, PPO attempts only 14.7% of channel frames when the primary user is busy and 69.2% when the channel is idle. This $4.7\times$ discrimination ratio indicates that the policy learned to use sensing history and state context to avoid channels likely to be occupied. The mean number of channels attempted per epoch is 2.02 of 4, indicating that PPO actively chooses subsets of channels to expose rather than defaulting to expose all. Figure 5 visualizes the joint distribution.

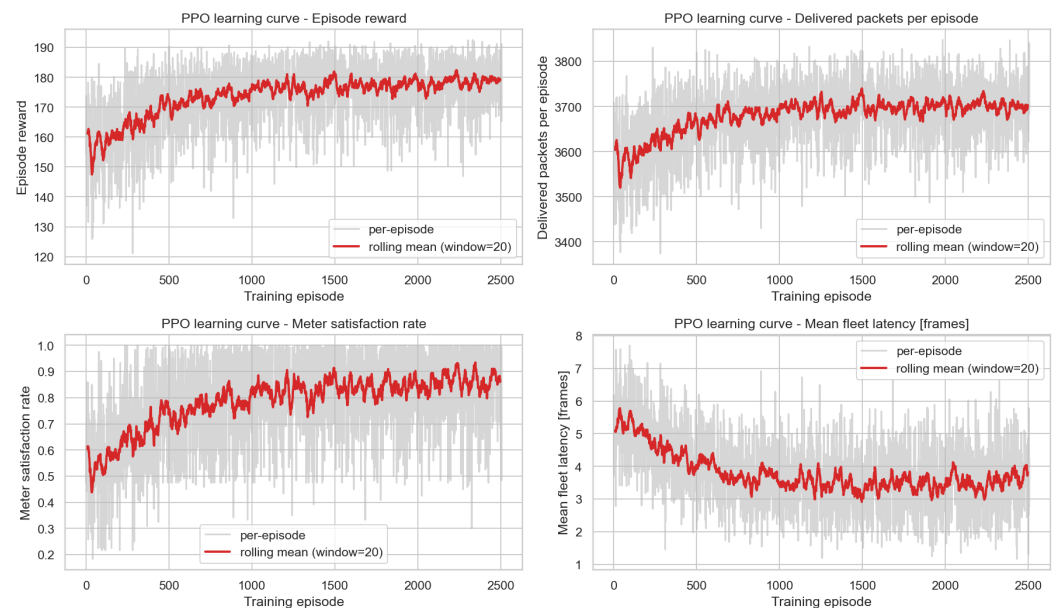


Figure 4. PPO training dynamics over 200,000 timesteps. Convergence is reached in approximately 1500 episodes.

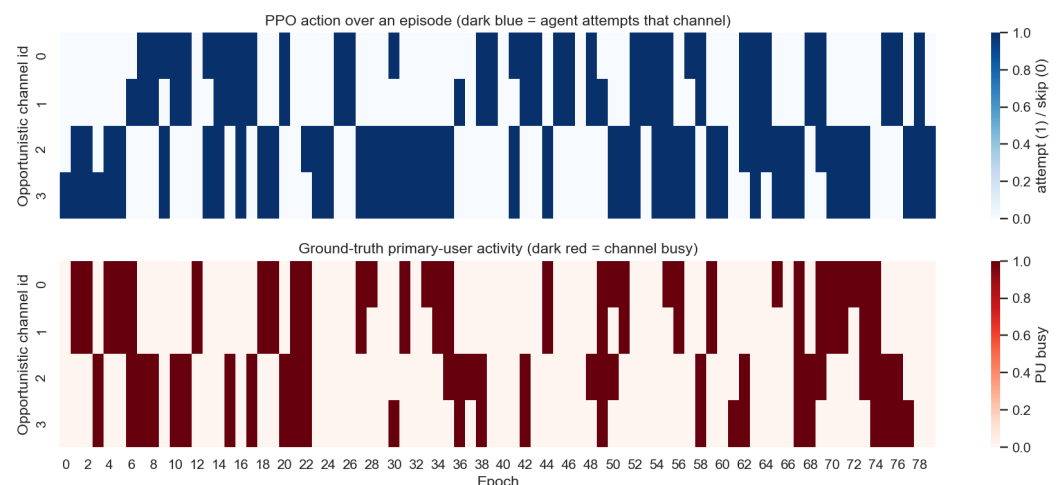


Figure 5. PPO action distribution conditional on the true PU state. The agent attempts only 14.7% of channels when the PU is busy versus 69.2% when the channel is idle, a $4.7\times$ discrimination ratio.

6.3. Robustness to Sensing Imperfection

We sweep the false-alarm and missed-detection probabilities each over $\{0.02, 0.10, 0.20, 0.30\}$, training a fresh PPO agent at 50,000 timesteps for each of the 16 cells. Figure 6 shows the resulting satisfaction-rate heatmaps for PPO, the Bayesian baseline, and the ideal-sensing reference.

PPO shows satisfaction in the range 0.55–0.76 across the entire grid, with a mean of 0.65. Bayesian belief degrades steeply with sensor quality, ranging from 0.78 in the easiest cell (low false alarms and low missed detections) to 0.46 in the hardest cell, with a mean of 0.55. The PPO heatmap varies less across sensing conditions, indicating lower sensitivity to sensing degradation than the Bayesian-belief baseline. The ideal-sensing reference is, by design, invariant across the grid: it bypasses the sensor entirely, so the false-alarm and missed-detection probabilities do not affect its decisions.



Figure 6. Satisfaction rate across the false-alarm by missed-detection sensing grid. **Left:** PPO. **Middle:** Bayes-belief. **Right:** ideal-sensing reference (flat by construction, since the policy bypasses the sensor). PPO holds 0.55–0.76 satisfaction across the grid; Bayes-belief degrades from 0.78 at top-left to 0.46 at bottom-right.

6.4. Scalability

We sweep the number of smart meters from 80 to 5000 while holding the total system load constant by scaling the per-smart-meter arrival rate inversely with N . Figure 7 shows the satisfaction rate of PPO and the strongest causal baselines.

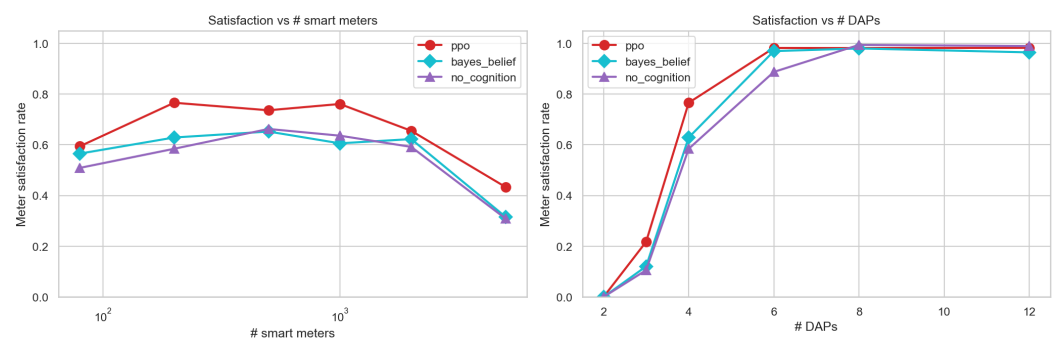


Figure 7. Scalability sweep. **Left:** satisfaction vs. number of smart meters from 80 to 5000. **Right:** satisfaction vs. number of DAPs from 2 to 12. PPO maintains a satisfaction-rate lead of 8–12 percentage points across both sweeps and degrades gracefully at the extremes.

PPO maintains satisfaction at 0.59 (80 smart meters), 0.77 (200 smart meters), and 0.73 (500 smart meters), with the lead over Bayesian belief widening from 3 to 14 percentage points as the system grows. At 1000 smart meters, PPO holds 0.76; at 5000 smart meters, PPO degrades to 0.43 but still leads Bayesian belief (0.32) and no-cognition (0.31) by 11–12 percentage points. The DAP sweep (right panel of Figure 7) covers $K \in \{2, 3, 4, 6, 8, 12\}$ at fixed total load. Below $K = 4$, the system lacks sufficient aggregation capacity, and all policies collapse to near-zero satisfaction. From $K = 4$ onwards, PPO leads, and from $K = 6$ onwards, all cognitive policies converge as the system becomes capacity-rich and the cognitive decision becomes less critical.

6.5. Per-Smart Meter Latency Distribution

Figure 8 shows the cumulative distribution of per-smart meter mean latency, pooled over 10 episodes for each of PPO, Bayesian belief, ideal sensing, and no-cognition.

The dashed vertical line at $L_{\text{sat}} = 6$ frames marks the satisfaction threshold; the satisfaction rate of each policy equals the fraction of smart meters whose curve lies to the left of this line.

The PPO curve dominates Bayesian belief and no-cognition almost everywhere; only at the extreme tail (above eight frames) does no-cognition catch up, because under that policy, a fraction of smart meters are perpetually starved by PU collisions and accumulate large latencies. Per-smart meter pooled statistics (mean PDR 0.951, P5 0.786, P95 1.000 for PPO) confirm that the distribution is tight at the high-PDR end, with PPO and ideal sensing closely matching.

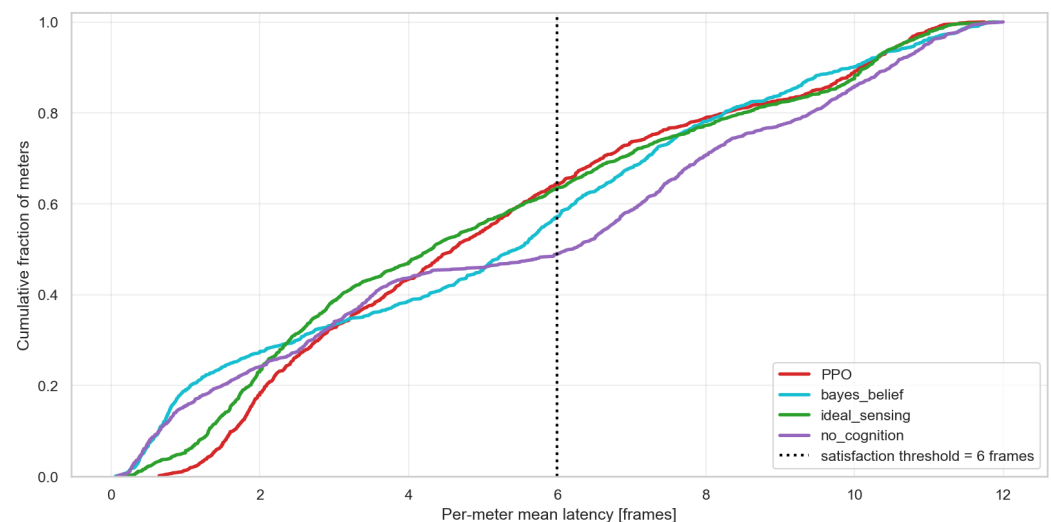


Figure 8. Cumulative distribution of per-smart meter mean delivery latency, pooled over 10 evaluation episodes per policy. The dashed vertical line marks the satisfaction threshold $L_{\text{sat}} = 6$ frames. PPO (red) and ideal sensing (green) dominate Bayesian belief (cyan) and no-cognition (purple) across the whole range.

6.6. Per-DAP Fairness and Energy Efficiency

Per-DAP delivery counts are balanced across all evaluated policies, with Jain's fairness index above 0.98 in every case (PPO: 0.992, Bayesian belief: 0.989, ideal sensing: 0.992, no-cognition: 0.991, random: 0.984). The Hungarian matcher's load-balancing structure, reinforced by the token-based fairness term in (8), prevents persistent imbalance between DAPs.

Energy efficiency, defined as delivered packets divided by attempted transmissions, clearly separates the policies: PPO achieves 2.99, ideal sensing 2.63, Bayesian belief 2.02, and no-cognition 1.85. PPO delivers more packets while making fewer attempts—it has both the highest throughput and the lowest attempt count among policies that actually use the cognitive layer.

6.7. Architectural Ablation: DAPs Versus No DAPs

To quantify the value of the DAP tier, we compare the proposed architecture with a no-DAP configuration in which each smart meter serves as its own aggregator. The setup is described in Section 5.3: the same Hungarian scheduler, the same PPO agent, and the same channel pool; only the DAP tier is removed, and the transmit-power and height parameters are reduced to smart meter-grade values.

The headline numbers in Table 5 are unambiguous. Removing the DAP tier reduces PDR from 0.960 to 0.458 (a 52% drop) and satisfaction from 0.668 to 0.042 (a 94% drop) under PPO, and similar collapses occur across all other policies. Episode reward goes from +171 with DAPs to −121 without, a swing of 292 reward units, dominated by the

heavy expiry penalty in (7): in the no-DAPs configuration, roughly half of all packets expire because the matcher cannot serve more than C aggregators per frame and we now have $N \gg C$ candidate transmitters. Figure 9 pairs the metrics side by side.

Table 5. Architectural ablation: with-DAP versus no-DAP configurations, both trained for 200,000 PPO timesteps, 30 evaluation episodes.

Mode	Policy	Reward	PDR	Sat.	Latency	PU Coll.
With DAPs	PPO	171.61	0.960	0.668	4.48	153.1
	Bayes belief	150.77	0.930	0.567	5.01	244.7
	Ideal sensing	171.17	0.959	0.665	4.76	139.9
	No cognition	146.58	0.925	0.514	5.10	250.2
	Random	148.34	0.924	0.433	6.35	208.4
No DAPs	PPO	−121.59	0.458	0.042	4.37	357.4
	Bayes belief	−123.72	0.459	0.054	4.26	422.0
	Ideal sensing	−108.30	0.469	0.018	4.42	186.8
	No cognition	−121.57	0.468	0.063	4.22	506.0
	Random	−154.53	0.395	0.022	4.56	251.5

Bold indicates the best value for each metric within each configuration block.

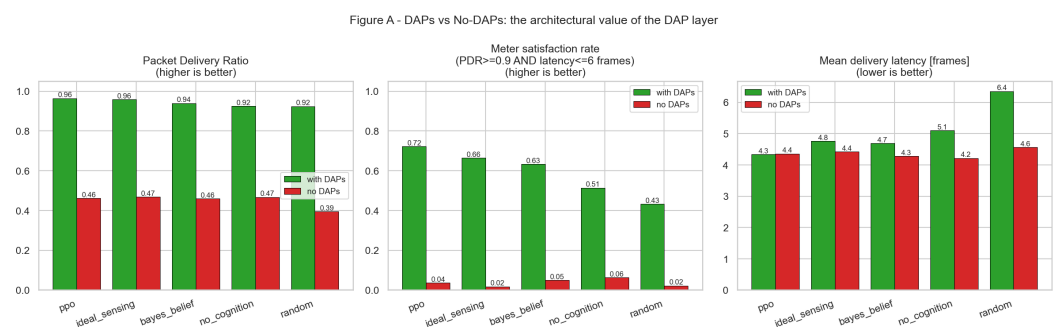


Figure 9. Architectural ablation: with-DAPs (green bars) versus no-DAPs (red bars), for each policy. **Left:** PDR. **Middle:** satisfaction. **Right:** mean delivery latency. The DAP tier is essential: every policy collapses without it.

The result is robust across policies: even the ideal-sensing reference, which has access to the true PU state at the start of each epoch, achieves only 0.02 satisfaction without DAPs. It confirms that the bottleneck without the DAP tier is not cognitive (which channels to use) but capacity (most aggregators cannot be scheduled at all in any given frame). PPO learned to do as much as possible in the no-DAP configuration: it reduced PU collisions from 506 (no-cognition) to 357 and reduced expiry penalties by attempting fewer doomed transmissions. The reward improved from −146 at the start of training to −99 at the end, demonstrating that learning happened. These results indicate that learning alone cannot compensate for an aggregation architecture that lacks sufficient service capacity within the deadline.

It suggests a clear engineering recommendation: in cellular AMI deployments under C-MVNO operation, the DAP tier should be treated as a structural requirement for viable AMI operation under the evaluated C-MVNO conditions, rather than as a secondary optimization. The cognitive layer is what differentiates good performance from excellent within the with-DAP regime, but it is not a substitute for the relay tier itself.

6.8. Cross-Algorithm Validation: DQN

To check that the cognitive benefit of PPO is not specific to the policy-gradient family, we additionally trained a Deep Q-Network (DQN) [45] on a discrete re-encoding of the

action space. With $C_{\text{opp}} = 4$ opportunistic channels, there are $2^{C_{\text{opp}}} = 16$ binary masks; we wrap the environment to expose a Discrete(16) action space and translate each integer back into the corresponding binary mask before stepping the underlying matcher. The DQN agent and the underlying environment are otherwise identical to the main configuration. The DQN is trained for 300,000 timesteps. Table 6 compares PPO and DQN head-to-head against the leading classical baselines.

Table 6. Cross-algorithm validation against the DQN trained on a Discrete(16) re-encoding of the action space, 300,000 timesteps. To keep the two learned agents directly comparable, both PPO and DQN values here are from a single matched training run rather than the multi-seed means of Table 4. Bayesian belief and no-cognition are shown for context.

Policy	Reward	PDR	Sat.	Latency
PPO	175.13	0.964	0.723	4.34
DQN	171.37	0.962	0.704	4.16
Bayes belief	156.50	0.940	0.633	4.70
No cognition	146.58	0.925	0.514	5.10

Bold indicates the best value for each metric.

The DQN comes within 2.1% of PPO on reward, ties on PDR, lies 1.9 percentage points below PPO on satisfaction, and slightly beats PPO on mean latency. Both PPO and the DQN dominate Bayesian belief and no-cognition. Figure 10 visualizes the comparison.

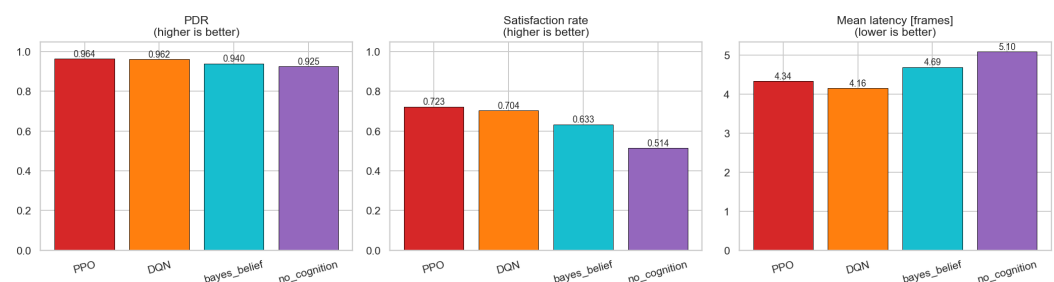


Figure 10. Cross-algorithm validation. PPO (red) and DQN (orange) both clearly outperform Bayesian belief (cyan) and no cognition (purple). PPO leads DQN slightly on reward, PDR, and satisfaction; DQN leads PPO slightly on mean latency.

The fact that an entirely different RL algorithm, trained on a different action representation, recovers nearly the same performance provides additional evidence that the observed cognitive gain is architectural rather than specific to PPO. This demonstrates that the cognitive benefit is a property of the architecture (the cognitive layer and Hungarian matcher), not an artifact of any particular learning algorithm.

6.9. Sensitivity Analysis

We examine the sensitivity of the result to three parameters central to the paper's argument: the primary-user correlation ρ , the CSI estimation error σ_{est} , and the epoch length T_{ep} . For each parameter we retrain PPO from scratch at each swept value and compare against the Bayesian baseline; to keep the additional computation tractable we use a single training seed per sweep point while retaining the full multi-seed confidence intervals on the main comparison of Table 4. Figure 11 reports the satisfaction rate for all three sweeps.

Three observations follow. First, PPO leads the Bayesian baseline at every level of primary-user correlation (panel a), by a margin of +0.08 to +0.14 in satisfaction; the advantage is present across the range rather than confined to the default $\rho = 0.7$, confirming that the result is not an artifact of a single correlation setting. Second, PPO is robust to

CSI degradation (panel b): its satisfaction is essentially unchanged for estimation errors up to 2 dB—below this level the error rarely alters the assignment relative to the fading margin, so the trained policies coincide—and declines only modestly at 4 and 6 dB, while the Bayesian baseline degrades more sharply, so that PPO's advantage *widens* from +0.09 to +0.15 at 6 dB. Third, the epoch-length sweep (panel c) directly quantifies the epoch-commitment effect discussed in Section 6.1: PPO satisfaction is highest at the shortest commitment window ($T_{ep} = 1$, where the cognitive decision is refreshed every frame) and declines as the window lengthens to $T_{ep} = 8$. This is consistent with, and provides an internal explanation for, the gap between the epoch-committed ideal-sensing reference and the per-frame clairvoyant oracle: shortening the commitment window moves any policy toward the per-frame regime, in which perfect current-frame information can be exploited immediately.

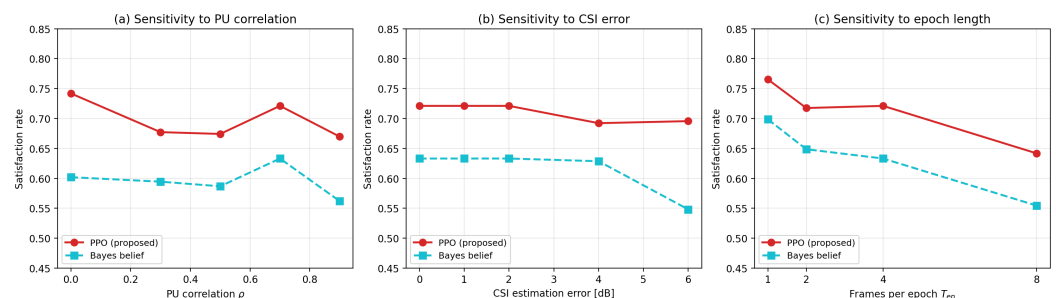


Figure 11. Sensitivity of satisfaction rate to (a) primary-user correlation ρ , (b) CSI estimation error σ_{est} , and (c) epoch length T_{ep} . PPO (red) retains its advantage over the Bayesian baseline (cyan) across the full range of all three parameters.

Three observations integrate the experimental evidence. First, the cognitive layer pays for itself across every operating regime tested: 200–5000 smart meters, $K = 4$ to $K = 12$ DAPs, false-alarm and missed-detection probabilities up to 0.30, and three different DAP placement strategies. Second, the value of the cognitive layer rests on the architectural prerequisite of having a DAP tier in the first place; cognition cannot rescue an architecture whose underlying scheduling problem is infeasible. Third, the result is robust across reinforcement-learning families so that the contribution can be read as architectural rather than algorithmic.

This study has three main limitations, each connected to a concrete aspect of practical AMI deployment in urban environments. First, the smart-meter-to-DAP segment is modeled as a reliable logical pipe and is not explicitly simulated at the physical or medium-access layers. This abstraction isolates the cellular aggregation bottleneck (the DAP-to-base-station uplink), which is the scarce and contested resource under C-MVNO operation, but it omits the short-range access technologies a real deployment would use. Practical options behave differently: LoRaWAN imposes duty-cycle limits and ALOHA-style contention that would add meter-to-DAP loss and latency; LTE-M and NB-IoT provide scheduled or repetition-based access with higher reliability but additional latency from coverage-enhancement repetitions. In all three cases the effect is an additional loss-and-delay stage upstream of the DAP queue, raising the effective arrival burstiness at the DAP without altering the cellular-uplink allocation problem studied here; a joint model of the short-range tier with a contention-based medium-access protocol, together with DAP buffering, is a natural extension. Second, the traffic, sensing, and primary-user parameters are simulation-driven and should be calibrated against field measurements from utility deployments before operational use; in particular, the lower-layer matcher uses SINR estimates with 2 dB Gaussian error, and substantially more impaired channel-state feedback would require the cognitive layer to act more conservatively (the CSI-error sensitivity anal-

ysis in Section 6.9 quantifies this dependence). Third, the evaluation assumes a single-cell C-MVNO setting; a multi-cell urban rollout would introduce inter-cell interference and require coordination of the cognitive decision across base stations, which we leave to future work. We also note that the no-DAP comparison in the architectural ablation uses the same Hungarian scheduler in both modes, which is a worst-case assumption for the no-DAP configuration relative to a random-access deployment; we adopt it because it isolates the DAP tier as the single variable, with all else fixed.

To support reproducibility, all simulation parameters are reported explicitly, the training and evaluation protocols use disjoint seeds, and comparisons are performed under matched primary-user and packet-arrival realizations across policies.

7. Conclusions

This paper presented a hierarchical resource-allocation architecture for cellular smart-metering sensor networks under C-MVNO operation. The two layers handle the parts of the problem for which they are best suited: the matcher optimally resolves the bipartite assignment at each frame, and the learned cognitive layer sequentially decides which channels to expose to the matcher. Across an extensive evaluation in a 3GPP TR 38.901-compliant simulator with imperfect sensing, correlated primary-user activity, packet deadlines, and per-smart meter accounting, the proposed architecture outperforms a Bayesian-belief baseline by 7 percentage points in satisfaction rate and the original Hungarian-only formulation by 19 points, and exceeds an ideal-sensing reference that has access to the true primary-user state at the start of each cognitive epoch, on account of the richer sequential context the learned policy exploits; a per-frame clairvoyant oracle, unconstrained by epoch commitment, bounds the achievable satisfaction from above and confirms that the gap to the reference is a cost of commitment rather than of information. The result is statistically significant across five independent seeds, robust to sensing and CSI degradation, scales gracefully from 200 to 5000 smart meters, generalizes across DAP placement strategies, and reproduces under DQN as a second learned baseline. An architectural ablation establishes that the DAP tier is a precondition for AMI viability under C-MVNO operation, not an optimization that smarter spectrum decisions can substitute. For urban infrastructure management, these findings indicate that a utility scaling metering across a dense deployment under leased spectrum should treat the aggregation tier as mandatory and the cognitive layer as the lever that turns an adequate system into a reliable one—reliability that the grid-edge visibility, demand response, and distributed-energy integration of smart-city energy services depend upon.

Several directions extend this work. The most immediate is to study multi-cell C-MVNO operation, where multiple base stations contend for overlapping opportunistic spectrum and the cognitive decision must be coordinated. A second is to replace the same-Hungarian no-DAPs comparison in the architectural ablation with a contention-based medium-access baseline (LTE-M PRACH, NB-IoT) for a more realistic direct-to-cellular reference. A third is to add a power-control degree of freedom on the DAP-to-BS link, which would expand the upper-layer action space and allow the agent to trade between attempting more channels at lower power and fewer at higher power. Then, a federated training scheme that lets multiple C-MVNOs train cognitive agents on their own deployments while sharing model weights, but not raw sensing data, would address the privacy and trust concerns identified in the AMI security literature.

Overall, the results suggest that reliable sensed-data delivery in future smart-metering infrastructures will require joint design of aggregation architecture, cognitive spectrum access, and exact per-frame resource assignment.

Author Contributions: Conceptualization, M.A.-A. and E.Y.; methodology, M.A.-A. and E.Y.; software, M.A.-A.; validation, M.A.-A., J.I., E.I. and E.Y.; formal analysis, M.A.-A.; investigation, M.A.-A.; writing—original draft preparation, M.A.-A.; writing—review and editing, M.A.-A., J.I., E.I. and E.Y.; supervision, E.Y. and E.I. All authors have read and agreed to the published version of the manuscript.

Funding: This work was made possible by Qatar University research grant no. QUICG-CENG-26/27-1148. The findings herein reflect the work and are solely the responsibility of the authors.

Data Availability Statement: The simulator source code and the experiment notebooks that produced all results in this paper are available from the authors on reasonable request.

Acknowledgments: During the preparation of this manuscript, the authors used generative AI tools to assist with language editing, drafting and refining portions of the text, and supporting the implementation of the simulation and analysis code. All AI-assisted content was reviewed, verified, and edited by the authors, who take full responsibility for the accuracy and integrity of the final manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kaluarachchi, Y. Implementing Data-Driven Smart City Applications for Future Cities. *Smart Cities* **2022**, *5*, 455–474. <https://doi.org/10.3390/smartcities5020025>.
2. Bouramdane, A. Optimal Water Management Strategies: Paving the Way for Sustainability in Smart Cities. *Smart Cities* **2023**, *6*, 2849–2882. <https://doi.org/10.3390/smartcities6050128>.
3. Zaman, M.; Puryear, N.; Abdelwahed, S.; Zohrabi, N. A Review of IoT-Based Smart City Development and Management. *Smart Cities* **2024**, *7*, 1462–1501. <https://doi.org/10.3390/smartcities7030061>.
4. Silva, N.; Castro, R.; Ferrão, P. Smart Grids in the Context of Smart Cities: A Literature Review and Gap Analysis. *Energies* **2025**, *18*, 1186. <https://doi.org/10.3390/en18051186>.
5. Chen, Z.; Amani, A.M.; Yu, X.; Jalili, M. Control and Optimisation of Power Grids Using Smart Meter Data: A Review. *Sensors* **2023**, *23*, 2118. <https://doi.org/10.3390/s23042118>.
6. Phing, C.; Kiong, T.; Koh, S.; Abedin, T.; Tak, Y.; Yusaf, T.; Yaw, M. A Brief Review on Ancillary Services from Advanced Metering Infrastructure (ASAMI) for Distributed Renewable Energy Network. *J. Adv. Res. Appl. Sci. Eng. Technol.* **2024**, *41*. <https://doi.org/10.37934/araset.41.2.4361>.
7. Orlando, M.; Estebsari, A.; Pons, E.; Pau, M.; Quer, S.; Poncino, M.; Bottaccioli, L.; Patti, E. A Smart Meter Infrastructure for Smart Grid IoT Applications. *IEEE Internet Things J.* **2022**, *9*, 12529–12541. <https://doi.org/10.1109/JIOT.2021.3137596>.
8. Khan, A.; Shirazi, S.; Adeel, M.; Assam, M.; Ghadi, Y.; Mohamed, H.; Xie, Y. A QoS-Aware Data Aggregation Strategy for Resource Constrained IoT-Enabled AMI Network in Smart Grid. *IEEE Access* **2023**, *11*, 98988–99004. <https://doi.org/10.1109/ACCESS.2023.3312552>.
9. Khan, A.; Umar, A.; Shirazi, S.; Ishaq, W.; Shah, M.; Assam, M.; Mohamed, A. QoS-Aware Cost Minimization Strategy for AMI Applications in Smart Grid Using Cloud Computing. *Sensors* **2022**, *22*, 4969. <https://doi.org/10.3390/s22134969>.
10. Gallardo, J.; Ahmed, M.; Jara, N. Clustering Algorithm-Based Network Planning for Advanced Metering Infrastructure in Smart Grid. *IEEE Access* **2021**, *9*, 48992–49006. <https://doi.org/10.1109/ACCESS.2021.3068752>.
11. Gallardo, J.; Ahmed, M.; Jara, N. LoRa IoT-Based Architecture for Advanced Metering Infrastructure in Residential Smart Grid. *IEEE Access* **2021**, *9*, 124295–124312. <https://doi.org/10.1109/ACCESS.2021.3110873>.
12. Jara, C.; Inga, J.; Inga, E. Opportunistic Allocation of Resources for Smart Metering Considering Fixed and Random Wireless Channels. *Sensors* **2025**, *25*, 2570. <https://doi.org/10.3390/s25082570>.
13. Nasser, A.; Hassan, H.A.; Chaaya, J.; Mansour, A.; Yao, K. Spectrum Sensing for Cognitive Radio: Recent Advances and Future Challenge. *Sensors* **2021**, *21*, 2408. <https://doi.org/10.3390/s21072408>.
14. Manco, J.; Dayoub, I.; Nafkha, A.; Alibakhshikenari, M.; Thameur, H. Spectrum Sensing Using Software Defined Radio for Cognitive Radio Networks: A Survey. *IEEE Access* **2022**, *10*, 131887–131908. <https://doi.org/10.1109/ACCESS.2022.3229739>.
15. Wang, X.; Teraki, Y.; Umehira, M.; Zhou, H.; Ji, Y. A Usage Aware Dynamic Spectrum Access Scheme for Interweave Cognitive Radio Network by Exploiting Deep Reinforcement Learning. *Sensors* **2022**, *22*, 6949. <https://doi.org/10.3390/s22186949>.
16. Yuan, S.; Zhang, Y.; Tang, T.; Cheng, Z.; Guo, D. Graph convolutional reinforcement learning for resource allocation in hybrid overlay-underlay cognitive radio network with network slicing. *IET Commun.* **2022**, *17*, 215–227. <https://doi.org/10.1049/cmu2.12527>.

17. Bai, W.; Zheng, G.; Xia, W.; Mu, Y.; Xue, Y. Multi-User Opportunistic Spectrum Access for Cognitive Radio Networks Based on Multi-Head Self-Attention and Multi-Agent Deep Reinforcement Learning. *Sensors* **2025**, *25*, 2025. <https://doi.org/10.3390/s25072025>.
18. Tan, X.; Zhou, L.; Wang, H.; Sun, Y.; Zhao, H.; Seet, B.C.; Wei, J.; Leung, V. Cooperative Multi-Agent Reinforcement-Learning-Based Distributed Dynamic Spectrum Access in Cognitive Radio Networks. *IEEE Internet Things J.* **2021**, *9*, 19477–19488. <https://doi.org/10.1109/JIOT.2022.3168296>.
19. Liu, S.; Pan, C.; Zhang, C.; Yang, F.; Song, J. Dynamic Spectrum Sharing Based on Deep Reinforcement Learning in Mobile Communication Systems. *Sensors* **2023**, *23*, 2622. <https://doi.org/10.3390/s23052622>.
20. Khan, A.; Umar, A.; Munir, A.; Shirazi, S.; Khan, M.; Adnan, M. A QoS-Aware Machine Learning-Based Framework for AMI Applications in Smart Grids. *Energies* **2021**, *14*, 8171. <https://doi.org/10.3390/en14238171>.
21. Inga, E.; Inga, J.; Hincapie, R. Maximizing Resource Efficiency in Wireless Networks through Virtualization and Opportunistic Channel Allocation. *Sensors* **2023**, *23*, 3949. <https://doi.org/10.3390/s23083949>.
22. Fernando, X.; Lăzăroiu, G. Spectrum Sensing, Clustering Algorithms, and Energy-Harvesting Technology for Cognitive-Radio-Based Internet-of-Things Networks. *Sensors* **2023**, *23*, 7792. <https://doi.org/10.3390/s23187792>.
23. Khasawneh, M.; Azab, A.; Alrabaaee, S.; Sakkal, H.; Bakhit, H. Convergence of IoT and Cognitive Radio Networks: A Survey of Applications, Techniques, and Challenges. *IEEE Access* **2023**, *11*, 71097–71112. <https://doi.org/10.1109/ACCESS.2023.3294091>.
24. Nurelmadina, N.; Hasan, M.; Memon, I.; Saeed, R.; Ariffin, K.; Ali, E.; Mokhtar, R.; Islam, S.; Hossain, E.; Hassan, M. A Systematic Review on Cognitive Radio in Low Power Wide Area Network for Industrial IoT Applications. *Sustainability* **2021**, *13*, 338. <https://doi.org/10.3390/su13010338>.
25. Wang, J.; Jiang, W.; Wang, H.; Huang, Y.; Chen, R.; Lin, R. Multiband Spectrum Sensing and Power Allocation for a Cognitive Radio-Enabled Smart Grid. *Sensors* **2021**, *21*, 8384. <https://doi.org/10.3390/s21248384>.
26. Salika, F.; Nasser, A.; Mroue, M.; Parrein, B.; Mansour, A. LoRaCog: A Protocol for Cognitive Radio-Based LoRa Network. *Sensors* **2022**, *22*, 3885. <https://doi.org/10.3390/s22103885>.
27. Salameh, H.B.; Abdel-Razeq, S.; Al-Obiedollah, H. Integration of Cognitive Radio Technology in NOMA-Based B5G Networks: State of the Art, Challenges, and Enabling Technologies. *IEEE Access* **2023**, *11*, 12949–12962. <https://doi.org/10.1109/ACCESS.2023.3242645>.
28. Ge, J.; Liang, Y.C.; Wang, S.; Sun, C. RIS-Assisted Cooperative Spectrum Sensing for Cognitive Radio Networks. *IEEE Trans. Wirel. Commun.* **2024**, *23*, 12547–12562. <https://doi.org/10.1109/TWC.2024.3393516>.
29. Ivanov, A.; Tonchev, K.; Poulkov, V.; Manolova, A. Probabilistic Spectrum Sensing Based on Feature Detection for 6G Cognitive Radio: A Survey. *IEEE Access* **2021**, *9*, 116994–117026. <https://doi.org/10.1109/ACCESS.2021.3106235>.
30. Jiang, W.; Yu, W.; Wang, W.; Huang, T. Multi-Agent Reinforcement Learning for Joint Cooperative Spectrum Sensing and Channel Access in Cognitive UAV Networks. *Sensors* **2021**, *22*, 1651. <https://doi.org/10.3390/s22041651>.
31. Tondwalkar, A.; Kwasinski, A. A Deep Q-Learning Algorithm With Guaranteed Convergence for Distributed and Uncoordinated Operation of Cognitive Radios. *IEEE Access* **2025**, *13*, 19678–19693. <https://doi.org/10.1109/ACCESS.2025.3534657>.
32. Abdolkhani, N.; Khalek, N.; Hamouda, W. Deep Reinforcement Learning for EH-Enabled Cognitive-IoT Under Jamming Attacks. *IEEE Internet Things J.* **2024**, *11*, 40800–40813. <https://doi.org/10.1109/JIOT.2024.3457012>.
33. Wang, M.; Liu, X.; Wang, F.; Liu, Y.; Qiu, T.; Jin, M. Spectrum-efficient user grouping and resource allocation based on deep reinforcement learning for mmWave massive MIMO-NOMA systems. *Sci. Rep.* **2024**, *14*, 8884. <https://doi.org/10.1038/s41598-024-59241-x>.
34. Zhu, R.; Li, M.; Liu, H.; Liu, L.; Ma, M. Federated Deep Reinforcement Learning-Based Spectrum Access Algorithm With Warranty Contract in Intelligent Transportation Systems. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 1178–1190. <https://doi.org/10.1109/TITS.2022.3179442>.
35. Gbenga-Ilori, A.; Imoize, A.; Noor, K.; Adebolu-Ololade, P. Artificial Intelligence Empowering Dynamic Spectrum Access in Advanced Wireless Communications: A Comprehensive Overview. *AI* **2025**, *6*, 126. <https://doi.org/10.3390/ai6060126>.
36. Xu, C.; Zhang, P.; Yu, H. Lyapunov-Guided Resource Allocation and Task Scheduling for Edge Computing Cognitive Radio Networks via Deep Reinforcement Learning. *IEEE Sensors J.* **2025**, *25*, 12253–12264. <https://doi.org/10.1109/JSEN.2025.3542972>.
37. Wang, L.; Wu, W.; Zhou, F.; Wu, Q.; Dobre, O.; Quek, T. Hybrid Hierarchical DRL Enabled Resource Allocation for Secure Transmission in Multi-IRS-Assisted Sensing-Enhanced Spectrum Sharing Networks. *IEEE Trans. Wirel. Commun.* **2023**, *23*, 6330–6346. <https://doi.org/10.1109/TWC.2023.3330999>.
38. Farhi, H.; Messai, A.; Berghout, T. Enhancing Spectrum Utilization in Cognitive Radio Networks Using Reinforcement Learning with Snake Optimizer: A Meta-Heuristic Approach. *Electronics* **2025**, *14*, 2525. <https://doi.org/10.3390/electronics14132525>.
39. Cheggour, S.; Loscri, V. Frequency Resource Management in 6G User-Centric CFmMIMO: A Hybrid Reinforcement Learning and Metaheuristic Approach. *arXiv* **2025**, arXiv:2505.22443. <https://doi.org/10.48550/arXiv.2505.22443>.
40. Chigaba, A.; Nleya, S.; Velepini, M.; Dube, S. A Multi-Objective Genetic Algorithm–Deep Reinforcement Learning Framework for Spectrum Sharing in 6G Cognitive Radio Networks. *Appl. Sci.* **2025**, *15*, 9758. <https://doi.org/10.3390/app15179758>.

41. 3GPP. *Study on Channel Model for Frequencies from 0.5 to 100 GHz*; Technical Report TR 38.901, Release 17; 3rd Generation Partnership Project (3GPP): Valbonne, France, 2022.
42. Kuhn, H.W. The Hungarian Method for the Assignment Problem. *Nav. Res. Logist. Q.* **1955**, *2*, 83–97. <https://doi.org/10.1002/nav.3800020109>.
43. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal Policy Optimization Algorithms. *arXiv* **2017**, arXiv:1707.06347.
44. Raffin, A.; Hill, A.; Gleave, A.; Kanervisto, A.; Ernestus, M.; Dormann, N. Stable-Baselines3: Reliable Reinforcement Learning Implementations. *J. Mach. Learn. Res.* **2021**, *22*, 1–8.
45. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level Control through Deep Reinforcement Learning. *Nature* **2015**, *518*, 529–533. <https://doi.org/10.1038/nature14236>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.