

Review

A Decision-Centered Survey of Machine Learning for Routing in Ad Hoc Networks: MANETs, VANETs, and FANETs

Yue Liu  and Xue Jun Li * 

Department of Electrical and Electronic Engineering, Auckland University of Technology,
Auckland 1142, New Zealand; yue.liu@autuni.ac.nz

* Correspondence: xuejun.li@aut.ac.nz

Abstract

Machine learning is rapidly transforming routing research across mobile, vehicular, and flying ad hoc networks (MANETs, VANETs, and FANETs). However, cross-study comparison remains difficult because individual works focus on disparate objectives within inconsistent simulation environments. The existing reviews are often organized by algorithmic families, which include Q-learning, deep Q-networks, convolutional and graph networks, and multi-agent reinforcement learning. Consequently, such reviews clearly discuss the model used in a paper yet blur the object of study: the same algorithm may act at many points of the routing decision or replace a protocol such as AODV or GPSR. However, two natural questions arise: how does a design change the routing decision and which learning method implements the change? This survey takes the routing decision itself as the main focus: it unifies vehicular, UAV-swarm, and MANET studies in one decision-centered framework. Under this framework, we survey 92 papers on learning-based routing along five dimensions: (1) the learner's role and decision authority, (2) the network-state representation it consumes, (3) its information horizon, (4) its temporal horizon, present versus predicted future, and (5) its policy organization and coordination. We provide an evolutionary map, a classification of all 92 papers in the corpus, a mechanism-oriented comparison of trade-offs, and an evaluation audit of a focused 37-paper analytical core. Within this core, the reported results are based on fragmented simulation environments and self-selected baselines: mechanisms can improve performance, but the magnitude of these improvements remains uncertain. We conclude this survey with open challenges that reframe learning-based routing around generalization, prediction reliability, security, reproducible evaluation, and deployability rather than incremental packet-delivery gains.

Keywords: machine learning; reinforcement learning; deep reinforcement learning; routing protocols; mobile ad hoc networks (MANETs); vehicular ad hoc networks (VANETs); flying ad hoc networks (FANETs); multi-agent reinforcement learning; survey; taxonomy



Academic Editor: Kefeng Ji

Received: 23 August 2026

Revised: 22 September 2026

Accepted: 22 September 2026

Published: 25 September 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.

1. Introduction

Ad hoc networks enable mobile nodes to communicate by relaying information for one another, each node acting simultaneously as both host and router, with little or no reliance on fixed infrastructure. In their mobile form (mobile ad hoc networks, MANETs), their vehicular form (vehicular ad hoc networks, VANETs), and their aerial form (flying ad hoc networks, FANETs), they underpin applications from vehicular safety and intelligent transportation [1] through unmanned aerial vehicle (UAV) swarms [2,3] and disaster monitoring [4] to tactical [5] and public-safety [6] communication. Some deployments

include infrastructure—roadside units (RSUs) in vehicular networks [1]—but such coverage is typically partial rather than continuous, so the nodes must still route among themselves. A central challenge shared by these settings is routing (Figure 1). Nodes move and links can form and break within seconds. Unlike centrally coordinated designs such as software-defined architectures, whose logically centralized controller can maintain a global view of the topology [7], a purely ad hoc network has no persistent entity with such a view, so each node must make its routing decision based on local information. A path that is optimal at present may break before the next packet traverses it [5]. Classical protocols address this challenge through manually designed heuristics: reactive route discovery and repair in ad hoc on-demand distance vector (AODV) [8] and dynamic source routing (DSR) [9]; greedy geographic forwarding in greedy perimeter stateless routing (GPSR) [10]; and proactive link-state dissemination in optimized link state routing (OLSR) [11]. These protocols remain the baselines on which many learned routers are built [12–14]. Each encodes assumptions about the mobility, density, and traffic it was designed for, and its performance can degrade when those assumptions no longer hold.

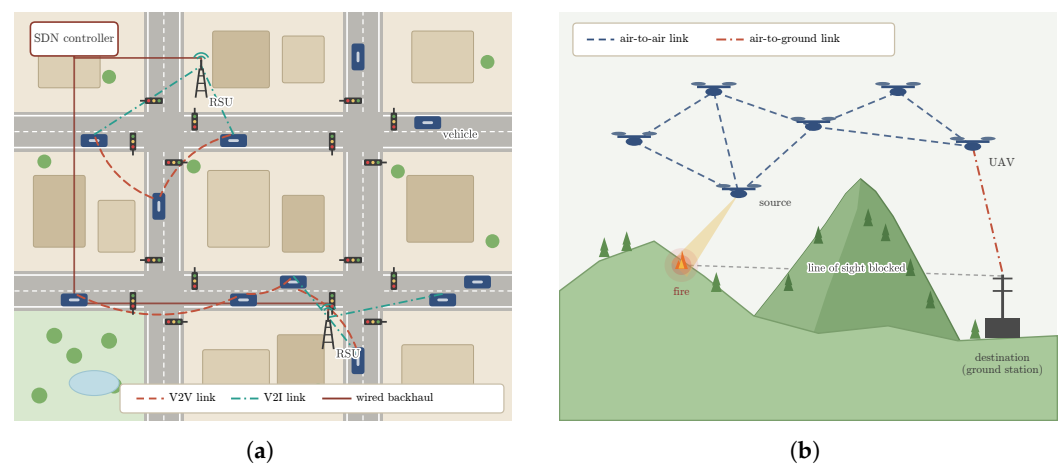


Figure 1. Representative ad hoc network scenarios from the surveyed corpus. (a) Urban VANET with SDN-assisted vehicle-to-infrastructure and vehicle-to-vehicle communication. (b) FANET disaster monitoring with air-to-air and air-to-ground UAV links.

Over the past two decades, learning-based routing research has grown considerably since early reinforcement-learning studies [15]. Machine learning has been applied to routing on the premise that a router can learn from experience what would otherwise have to be encoded in a manually designed rule. Reinforcement learning enables a node to estimate the long-run quality of a forwarding choice [13]; deep networks enable it to process a high-dimensional cross-layer state [16]; recurrent networks enable it to anticipate change [17]; and multi-agent learning enables multiple nodes to adapt in coordination under a shared policy [18]. This diversity has produced a substantial but fragmented body of work: individual studies address different routing problems in heterogeneous simulation environments, which complicates both synthesis and direct comparison.

A central methodological challenge in surveying learning-based routing is the choice of organizing principle. A natural way to do so is by technique: Q-learning, deep Q-networks (DQNs), long short-term memory (LSTM) networks, graph neural networks (GNNs), and multi-agent reinforcement learning (MARL). Such an approach primarily identifies which model a paper used but does not necessarily reveal how the model changes the routing decision. The same technique may serve contrasting roles in different papers. For example, reinforcement learning is embedded inside AODV to rank and switch among its routes in one paper [12] yet constitutes the forwarding policy itself, selecting each next hop from among admissible neighbors, in another [13]; an LSTM forecasts vehicle trajectories to select

relays in one design [19], predicts link stability in a second [17], and classifies malicious nodes, performing no forecasting, in a third [20]. An algorithm-centered taxonomy may therefore group studies that address different routing problems and separate studies that address the same problem with different techniques, obscuring the substantive progress in what the learner is asked to decide, to perceive, and to anticipate.

This difficulty is visible in the surveys that already exist (Table 1). One of them surveys artificial intelligence for vehicular communication at large, with routing as one application among many [21]; the other four do center on MANET routing but each does so within one method family or one routing task (bio-inspired energy-aware routing [22], Q-learning multicast routing [23], rule-based against AI-based protocols [24], and bio-inspired and learning protocols [25]), and each tabulates between seven and twenty-three designs against the 92 classified here. As Table 1 shows, none of the five combines a decision-level taxonomy that asks of each design which part of the routing decision the learner owns, a documented reconstruction of the relationships among designs, and a cross-scenario audit of evaluation practice. These three gaps motivate the contributions listed below: the taxonomy addresses the first, the evolutionary map the second, and the evaluation audit the third.

Table 1. The existing surveys nearest in topic and the present corpus: scope, organizing principle, and coverage of learning-based routing designs.

Survey	Scope	Organizing Principle	Learning-Based Routing Designs Covered	Decision-Level Framework
Safari et al., 2023 [25]	MANETs; bio-inspired and learning methods	by AI family (bio-inspired; learning automata and reinforcement learning)	23 protocols in two comparison tables (11 bio-inspired, 12 learning-based) [†]	AI-family taxonomy; no decision-level axes
Djihene et al., 2024 [22]	MANETs; energy-aware routing	by bio-inspired algorithm	8 bio-inspired energy-efficient routing protocols in one table [†]	algorithm overview; no decision-level axes
Al-Mashhadani and Karoui, 2025 [24]	MANETs; rule-based and AI-based protocols	rule-based protocols set against ML and DL protocols	10 ML and DL routing designs in one table [†]	rule-based versus learned contrast; no decision-level axes
Harshitha et al., 2026 [23]	MANETs; Q-learning multicast routing	protocol by protocol, on challenges and future directions	7 Q-learning multicast designs in one table [†]	single-task comparison; no decision-level axes
Wu et al., 2026 [21]	IoV and C-V2X; LLMs, agentic and embodied AI	by AI paradigm and vehicular communication task	not a routing survey; routing appears only as a network-optimization application	AI-paradigm and task taxonomy; routing is peripheral
This survey	MANETs, VANETs, and FANETs; every learning or predictive method that informs the routing decision	five decision dimensions (Section 3.1); documented lineages (Section 3.1)	92 studies classified (Section 3.8); 37 audited for evaluation practice (Section 5)	role, state, information horizon, prediction, coordination

[†] Counts are the entries that each survey itself tabulates; they are not independent counts of verified routing studies.

As a result, this survey is organized by the routing decision rather than the algorithms: Instead of ranking schemes by their reported performance or partitioning the literature by deployment setting, this survey considers vehicular networks, UAV swarms, and the broader MANET literature in a single decision-centered frame, with the scenario retained as a secondary dimension. We investigate the 92-paper corpus (studies in which a learned or predictive component directly informs a routing decision or monitors or governs the operation of a learned router) as a set of answers to five questions that a learned router must resolve: (1) what role learning plays in routing and how much decision authority it receives; (2) how the learner represents the network state on which routing decisions are made; (3) what part of the network the learner may observe and through what filter; (4) whether routing acts on the measured present or the predicted future; and (5) once many nodes learn concurrently, how the resulting policies are organized, shared, coordinated,

and trusted. These five questions become five analytical dimensions that serve as a stable coordinate system for the entire corpus. Within this framework we develop an evolutionary account, a comparative analysis, an examination of how implementations and comparisons are reported within the analytical core, and a set of open challenges and future directions. The specific contributions of this paper are the following:

- An evolutionary map (Section 3.1) that traces five mechanism-level trajectories through the corpus, distinguishing declared extensions, citation relationships, and inferred continuities and showing how mechanisms migrate between network families.
- A mechanism-based taxonomy (Section 3) organized around five analytical dimensions answering the five questions above: the role and decision authority of the learner; the network-state representation used by the learner; the information horizon that is available to the decision; the temporal horizon from measured present to predicted future; and policy organization and coordination. The taxonomy is accompanied by a complete primary classification of all 92 papers.
- A mechanism-oriented comparative analysis (Section 4) of the design trade-offs along each dimension, structured around a focused analytical core of 37 studies selected for extractability, influence, and coverage (Section 3.1).
- An analysis of evaluation practice (Section 5) across the 37 core papers that examines how the core papers report their implementations and against which baselines and performance metrics each scheme is compared and identifies what makes those comparisons difficult to conduct on a common basis.
- A set of open challenges and future directions (Section 6) that translates those trade-offs and the evaluation gaps documented in Section 5 into concrete problems for future research.

The remainder of this survey is organized as follows. Section 2 provides background on the network scenarios, the classical routing substrate, and the learning methods the corpus draws on; Sections 3–6 develop the contributions listed above in the order given; and Section 7 concludes the survey.

2. Background: Scenarios, Protocols, and Learning Methods

2.1. Network Scenarios and Their Routing Constraints

Although MANETs, VANETs, and FANETs all rely on multi-hop wireless forwarding under mobility, they impose different constraints on the routing decision. In general MANET settings, node mobility is not restricted by a fixed topology, links can appear and vanish unpredictably, energy and bandwidth are constrained, and no central controller is assumed to compute paths on the network's behalf. For the purposes of this survey, the MANET/general category also includes related settings whose nodes are mobile or intermittently connected, forward over peer nodes without a road or airspace structure that constrains motion, and lack fixed infrastructure: tactical and wireless sensor network (WSN) deployments that add jamming and adversaries [5], and delay-tolerant or opportunistic-network (DTN/OppNet) settings in which a contemporaneous end-to-end path may never exist, so packets must be stored and carried between encounters [26,27].

The VANET specializes the problem to vehicles and roadside infrastructure, and four properties are particularly relevant to the analysis in this survey. Firstly, motion follows the road network and traffic rules, which makes it partially predictable: a vehicle can move only along the road segments that are available to it [19]. Secondly, traffic density varies substantially from dense rush-hour traffic to sparsely populated roads, so a protocol tuned for one regime can degrade materially in the other [1,28]. Thirdly, RSUs at intersections offer fixed relays and a hierarchical tier above the vehicles [1,29]. Finally,

rich contextual information is often available: positions from GPS, digital road maps in map-aware designs, and historical mobility records such as taxi GPS traces [28,30]. Grid- and intersection-based state representations, RSU hierarchies, and traffic forecasting rely directly on these four properties (Sections 3.3 and 3.5).

FANETs, self-organizing networks of UAVs, introduce additional constraints. Nodes can move fast in three-dimensional space; neighbor changes and link opportunities can occur within seconds [4,7]; energy and onboard-resource budgets are tight [4,31]; some missions must operate where GPS is denied [31]; and fixed ground infrastructure may be unavailable, particularly in remote or emergency deployments. Consequently, a UAV's knowledge of its neighborhood can rapidly become outdated, making information-acquisition cost an important design constraint (Section 3.4) [4,32]. Under high mobility, links can change sufficiently rapidly that a decision based on the currently measured state may become stale before it is applied, which pushes aerial designs toward prediction (Section 3.5) [7,17] and, in some cases, multi-agent policy coordination (Section 3.6) [2,18].

These three families are the scenario columns of the master classification in Section 3.8. Vehicle-to-everything (V2X) studies are placed with VANETs, and mixed aerial-ground designs are assigned to the family of the nodes that execute the forwarding decision; a UAV-assisted vehicular scheme in which ground vehicles forward is therefore a VANET study.

2.2. The Classical Routing Substrate

Few schemes in the corpus operate entirely independently of the classical protocols. Learned routing is integrated into, embedded in, or benchmarked against a small set of classical protocol families, and the contribution of a learned scheme can usually be located at one or more identifiable decision points within a classical routing process. Four classical protocol families recur across the corpus together with one historically important bounded-probing mechanism.

Reactive discovery protocols find routes on demand. AODV [8] floods a route request through the network when a source needs a path, unicasts a route reply back along the discovered reverse path, and maintains the result with sequence numbers and link-failure notifications; DSR [9] also discovers routes on demand but records the complete route in the packet header and caches overheard routes. Their decision points are how widely to search, which discovered route to accept, when to abandon a degrading route, and which neighbor to relay through. These are the points that learned schemes most often occupy.

Proactive link-state protocols maintain routes continuously. OLSR [11] maintains topology through periodically flooded link-state messages, with flooding overhead reduced by delegating both link-state origination and rebroadcast to a heuristically chosen subset of neighbors, the multi-point relays (MPRs). MPR selection is itself a hand-designed rule, an idea subsequently adopted by later work that learns an MPR-like policy for multicast forwarding [33].

Geographic forwarding avoids maintaining end-to-end routes. GPSR [10] forwards each packet to the neighbor geographically closest to the destination, falling back to a perimeter mode along a planarized subgraph when greedy progress is blocked, and learns of its neighbors through periodic position beacons. Its decision points are how to score neighbors beyond raw distance, how to reduce reliance on perimeter forwarding (which can increase path length and latency), and how often to beacon. Together, these decision points make it a common scaffold in the aerial literature [17,34,35], where fast-changing neighborhoods can make beacons stale between updates.

Store-carry-forward protocols serve delay-tolerant settings. The probabilistic routing protocol using history of encounters and transitivity (PRoPHET) [36] and related protocols estimate from encounter history the probability that a node will eventually meet the

destination and replicate a packet only to an encountered carrier with a higher delivery predictability, a per-encounter forwarding decision that several studies in the corpus reformulate with classifiers or reinforcement learning [26,27].

Bounded QoS probing is the mechanism: ticket-based probing [37] searches for a path satisfying delay or bandwidth constraints by issuing a limited number of probing tickets, trading search cost against the chance of finding a feasible path. The maximum number of tickets is manually set, and the initial allocation issued for each discovery follows a fixed rule. This fixed allocation is the decision point that the earliest reinforcement-learning scheme in the corpus replaces with a learned policy [15].

Beneath most of these families lies a shared sensory substrate—the periodic hello or beacon exchange through which each node discovers its neighbors and obtains the state information they advertise (in the store-carry-forward family, encounter events play this role). Its rate and content set the trade-off between control overhead and the freshness and reach of information in an ad hoc network.

2.3. Learning Methods in the Corpus

The core learning methodologies in this body of work fall into two main categories: (1) reinforcement learning (RL) and (2) prediction and classification. The first category, RL, estimates value functions or optimizes policies from reward feedback generated through interaction with the routing environment [38]. Its dominant form in the corpus is tabular Q-learning, which represents state–action values in a lookup table typically updated online from per-hop feedback. This form is model-free and, for a sufficiently small state–action space, computationally inexpensive enough for on-node execution, but its convergence depends on sufficient exploration, repeated visitation of the relevant state–action pairs, appropriate learning-rate conditions, and sufficiently stationary dynamics. State-space size and topology-induced non-stationarity therefore recur as design constraints throughout this survey. Variants in the corpus include Monte Carlo RL [15] and SARSA-style updates [39,40]. Deep value networks, such as DQNs and their double and dueling variants, replace the lookup table with neural function approximation, which allows them to process higher-dimensional state representations, while experience replay and target networks are commonly used to stabilize training; both come at the expense of higher computational and training demands [16,41,42]. Policy-optimization methods, chiefly proximal policy optimization (PPO), learn the policy directly; multi-agent formulations range from learners that act independently but share a single Q-network [41] through game-theoretic updates (Nash-Q) [43] to multi-agent deep deterministic policy gradient (MADDPG) and the centralized-training decentralized-execution (CTDE) paradigm in which training may draw on joint or centrally available information while execution runs on local observations [2,18]. A smaller deep-RL line learns over relational or condition features rather than node identities, yielding forwarding policies that generalize across networks [44,45]. Related approaches at the boundary of this category include imitation learning, which learns a forwarding policy from demonstrations of a classical protocol [46], and inverse RL, which learns the reward function itself [47].

The second category is prediction and classification [48]: supervised artificial neural network (ANN) regressors, LSTM and other recurrent sequence models such as the gated recurrent unit (GRU), gradient-boosting ensembles, and random forests. These models are trained on historical data to forecast traffic, mobility, or link lifetime; the same architectures may also classify current conditions rather than forecast future states, so their analytical role is determined by how their outputs enter the routing decision rather than the model architecture alone. Unsupervised clustering is also used to support mobility and traffic prediction [49]. Beyond these two categories, the corpus also includes statistical and

analytical models, fuzzy fusion of link metrics, GNN topology encoders, and anomaly detection over learned policies.

Figure 2 arranges these methods by category, family, and representative algorithm. For this survey, however, the method family is less analytically informative than the decision role the method occupies, and Table 2 illustrates this distinction: the same family recurs across distinct roles, and the same role is filled by distinct families. This many-to-many structure forms the basis of the taxonomy developed in Section 3.

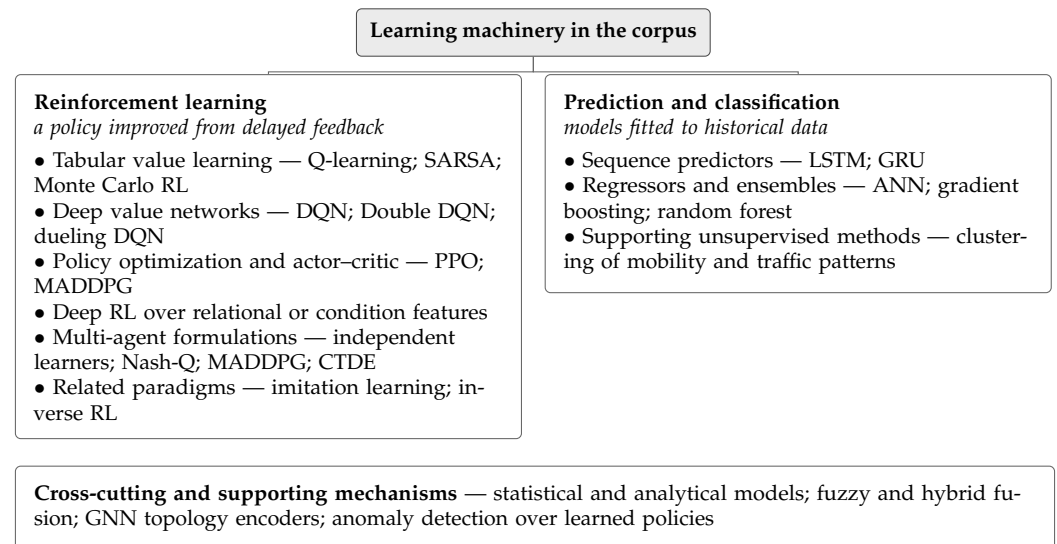


Figure 2. The learning methods of the corpus arranged by category, method family, and representative algorithm. Table 2 maps these families onto the routing decision roles they occupy.

Table 2. Learning-method families in the corpus and the decision roles they occupy (representative works; the same family recurs across roles and the same role across families).

Method Family	Decision Roles Occupied in the Corpus
Tabular value learning (Q-learning, SARSA, and Monte Carlo RL)	learn the initial probing-ticket allocation within a bounded discovery budget [15]; score and pre-emptively switch routes [12,50]; own the hop-by-hop forwarding choice [34,51]; learn over abstracted states—grids, intersections, road segments, RSU groups [28,29,52,53]; coordinate many agents [43]
Deep value networks (DQN and double/dueling DQN)	hop-by-hop forwarding on classical scaffolds [16,54]; manage a two-hop observation window [42]; learn over intersection states [55]; consume a learned forecast [31]; cooperate under jamming or queue/load pressure [5,56]; independent learners sharing one Q-network [41]
Policy optimization and actor-critic (PPO and MADDPG)	learned unicast-versus-broadcast switching inside a preserved CQ+ routing scaffold [18]; hop-by-hop forwarding on a classical geographic scaffold [57]; cooperative swarm routing as a stochastic game [2]; hierarchical cluster routing [3]
Deep RL over relational or condition features	identity-free forwarding policies that generalize across networks [44,45]; selection among a pool of learned and classical policies [14]
Sequence predictors (LSTM and GRU)	traffic forecasting [7] and joint mobility-traffic forecasting [58]; neighbor-trajectory prediction [19]; link-stability and link-quality forecasting [17,31]; present-tense malicious-node classification, with no forecasting [20]
Other supervised and unsupervised learners (ANN, gradient boosting, random forest, and clustering)	segment-traffic forecasting [1]; per-candidate delivery probability [59]; route-quality regression [60]; clustering-based prediction of vehicle movement and path capacity [49]; a random forest predictor beside a deep relay-selector [61]
Statistical and analytical models	link-lifetime estimation from vehicle kinematics [60] or queuing analysis and Kalman filtering [4]
Fuzzy and hybrid fusion	fusing or pre-filtering link metrics for learned selection [27,50]; fuzzy road-suitability scoring for a UAV-computed global route [62]
Imitation and inverse RL	imitate a classical protocol instead of hand-crafting a reward [46]; learn the reward function itself [47]
GNN encoders	topology encoding for multi-agent policies [3]
Anomaly detection over learned policies	runtime TD-error monitoring of a deployed policy for loss of trust [63]

2.4. From Background to the Five Decision Dimensions

Taken together, the scenario constraints of Section 2.1, the protocol decision points of Section 2.2, and the learning methods of Section 2.3 motivate the five analytical dimensions used in this survey. Classical protocols rely on manually specified rules and parameters, such as hello rates, route metrics, search budgets, and repair thresholds, that are set for expected operating conditions; a beaconing interval that is suited to highway mobility may be poorly suited to a hovering UAV swarm, and a hop-count metric ignores network load. Replacing such a rule with a learner raises the question of the learner's role and decision authority. Encoding vehicles, links, intersections, or regions as state raises the question of network-state representation. Acquiring neighbor or topology information raises the question of how far the decision may observe and at what control cost. Acting on forecasts rather than current measurements introduces the temporal horizon of the decision. Finally, deploying learners across many nodes raises the questions of how their policies are organized, coordinated, transferred, and trusted. Section 3 defines these five dimensions and applies them to the corpus.

3. A Mechanism-Based Taxonomy and Its Evolution

3.1. Rationale and Classification Methodology

An algorithm-family taxonomy is straightforward to construct, but it records which model was used rather than what the model changed about the routing decision. Table 2 shows why: Q-learning selects the maximum-value neighbor within a classical scaffold in one study [64] and, over city-grid states, selects the next grid toward the destination in another [28]; the same recurrent architecture forecasts traffic in one design [7] and classifies malicious nodes without forecasting in a second [20].

Therefore, we investigate the corpus as a set of recurring answers to the five questions: the role and decision authority of the learner, the network-state representation, the information horizon, the temporal horizon, and policy organization and coordination. Nearly every corpus design that acts on the future does so through an explicit forecast, so we refer to the temporal-horizon dimension simply as prediction. Figure 3 maps the declared extensions, citation links, and inferred continuations among the principal works.

Three fundamental properties of this analytical framework warrant explicit mention at the beginning. Firstly, the five dimensions are cross-cutting analytical descriptors rather than mutually exclusive classes: the same model can serve different dimensions in different papers, different models can answer the same dimension, and each design can in principle be characterized along all five. Secondly, for the countable summaries of Section 3.8 and Figure 4 only, we additionally assign each paper one primary-contribution label (its primary dimension hereafter), the dimension on which the paper's stated novelty most directly changes the routing mechanism. The label was determined sequentially from (1) the contribution explicitly claimed by the original authors, (2) the routing component directly changed by the proposed mechanism, judged by the state-action interface the design actually changes, and (3) the principal variable that its own evaluation manipulates; citation by later studies and historical importance were not used. Cases that remain ambiguous after these tests are ruled on individually in Section 3.7. Within a dimension a paper receives one position, not several labels; across dimensions it receives exactly one primary label, and any further relevance is recorded as secondary, retained in the qualitative analysis of Sections 3.2–3.7 but excluded from the frequency counts. Hybrid designs that combine, for instance, route selection, link-quality prediction, and adaptive forwarding are therefore placed by the component their authors present as the contribution, with the other components noted as secondary. Table 3 gives the operational test for each dimension and its boundary with its neighbors, and a worked example follows the classification rules

below. Thirdly, we classify by a paper's defining contribution, not by its most conspicuous algorithm: a paper in which the learner makes the next-hop choice is not automatically a learner-role paper because next-hop selection occurs across multiple dimensions; it belongs to this dimension only when a defining contribution is the relocation of decision authority from the classical protocol to the learner. Applying this rule to all 92 papers yields the distribution of Figure 4: the role and decision authority of the learner accounts for 45 papers, the temporal horizon (prediction) 18, policy organization and coordination 11, network-state representation 11, and the information horizon seven. Sections 3.2–3.6 trace the principal works of each dimension and the documented development relations among them. For the detailed mechanism comparison and evaluation audit, Sections 4 and 5 use a focused analytical core of 37 papers; how the corpus was assembled and how the core was drawn from it are set out at the end of this section. Figure 3 places the principal works by publication year, with three arrow types recording the relationships between them. Isolation is informative in its own right: PQR [60], for instance, has no edge to any work shown, arriving at similar mechanisms with no direct citation relationship. Such cases suggest that similar solutions may emerge independently and substantial scope for consolidation remains.

Table 3. Operational definitions of the five decision dimensions and their boundaries.

Dimension	Decision Question	A Paper is Primary Here When...	Boundary with Neighboring Dimensions
Role and decision authority	Which part of the routing decision does the learner own?	its defining contribution relocates a decision (a parameter, a ranking, a selection, or forwarding itself) from the classical protocol to the learner	learned next-hop selection alone is not sufficient since it occurs across all dimensions; what counts is the shift of authority
Network-state representation	What entity, described by which attributes, stands for the network at the decision point?	its defining contribution changes the unit of state (vehicle, grid cell, intersection, RSU group) or the attributes that describe it	concerns what the state signifies, not how far it reaches; a wider neighborhood with the same state unit is an information-horizon change
Information horizon	How much of the network, at what control cost, does the decision observe?	its defining contribution widens, narrows, reconstructs, or bounds what the decision observes or evaluates (one hop, two hops, sensed dynamics, or a geometrically or externally filtered set of candidate neighbors)	concerns the spatial reach of present information and its cost; a forecast of future state belongs to the temporal horizon even where it extends what is known
Temporal horizon (prediction)	Does the decision act on the measured present or on a predicted future?	an explicit forecast of a future network quantity, or a prospective representation of future network states, enters the routing decision, whatever model produces it	a wider present-time view is not prediction; a predictor of an adjacent quantity that never reaches the routing decision is not counted
Policy organization and coordination	When many nodes learn at once, how are their policies organized, shared, coordinated, and trusted?	its defining contribution is the organization of policies: sharing, centralized training with decentralized execution, size-agnostic generalization, or policy pools with rule-based fallback	many nodes independently running the same learner is not sufficient; the organization must be the defining contribution

A worked example. DeepCQ+ [18] shows why the rule is needed. Its learner decides only broadcast versus unicast inside the CQ+ scaffold (role) from a fixed-size encoding of best-neighbor statistics (state) using the one-hop tables that CQ+ already keeps (information horizon) on the measured present (temporal horizon) with one policy shared by every node, trained centrally and transferred unchanged across network sizes (coordination). Organized by action authority alone it would fall under the learner's role; its primary dimension is coordination because the shared size-agnostic policy is the contribution the paper presents as its own.

Classification procedure and its limits. Because the studies in which learning touches the routing decision are numerous and heterogeneous, a classification was needed to organize and count them, and the five dimensions were adopted so that the framework

concerns the routing decision rather than the algorithm. All 92 papers were classified by the first author alone through multiple full-text review passes with boundary cases re-read against the rules before the assignment was fixed. No second rater took part and no inter-rater statistic was computed, and secondary relevance was recorded only for the boundary cases of Section 3.7; both are limitations of the taxonomy, and no claim of formal inter-rater reliability is made. Two measures improve transparency and auditability: every boundary ruling is reported in Section 3.7, and Section 3.8 provides the complete primary assignment so that any alternative reading can be applied to the same corpus.

Corpus construction. Candidate studies were retrieved with an AI-assisted semantic literature search engine (Undermind) through eight iterative search paths, supplemented by forward and backward citation chasing (Appendix A); no publication-date restriction was imposed, and the retained studies span 2004 to 2026. A study was retained if a learned or predictive component directly informs a routing decision in a MANET, VANET, or FANET or monitors or governs the operation of a learned router; learning that serves only an adjacent function, such as medium-access control or resource allocation, was excluded. In a small number of studies the predictive or scoring component is a statistical or fuzzy model rather than a trained learner (for example, the fuzzy road-suitability scoring of [62]); these were retained because they occupy the same decision point as their learned counterparts, and the term learning-based is used in this survey to cover both. Preprint, conference, and journal versions of one study were merged and the published version is cited. The result is the 92-paper corpus classified in Section 3.8; Table 4 summarizes the steps.

Table 4. Corpus construction at a glance.

Step	Rule
Retrieval	AI-assisted semantic literature search (Undermind) in eight iterative search paths, plus forward and backward citation chasing; paths and studies per path in Table A1
Search vocabulary	learning-based routing; geographic, adaptive, and opportunistic forwarding; topology-aware and predictive routing; multi-agent and hybrid policies; routing-related sequence prediction across MANET, VANET, FANET, and UAV networks
Time window	none imposed; retained studies span 2004–2026
Inclusion criterion	a learned or predictive component directly informs a routing decision in a MANET, VANET, or FANET, or monitors or governs the operation of a learned router; learning that serves only an adjacent function excluded
Version rule	preprint, conference, and journal versions of one study merged; the published version cited
Analytical core	purposive selection by the first author: extractability (scenario, learned component, its interface with the routing decision, and at least one quantitative evaluation identifiable; other audit fields coded as not reported) as the decisive criterion; influence and venue as a preference; coverage of the principal works of the lineages reconstructed first from the full corpus, of all dimensions, and of all network families (37 papers)

The analytical core. Sections 4 and 5 concentrate on 37 papers selected by the first author for mechanism-level comparison on three considerations: extractability, the decisive one, meaning that the paper reports enough to identify the deployment scenario, the learned or predictive component, its interface with the routing decision, and at least one quantitative evaluation, the remaining audit fields of Section 5 being coded as not reported where necessary; influence and venue, with well-cited studies in well-regarded journals and conferences preferred as a preference rather than a threshold since the designs of 2024 and 2025 could not yet have accumulated citations; and coverage, the five lineages of Sections 3.2–3.6 having been reconstructed first from the full corpus and the core and then chosen to retain their principal works together with at least one design for every dimension and each network family. The core coincides with the designs examined in Section 4; Table A2 lists every corpus paper by its standing. The remaining 55 papers are classified in Section 3.8 and discussed in Section 3, where they extend a lineage, but their evaluations were not coded.

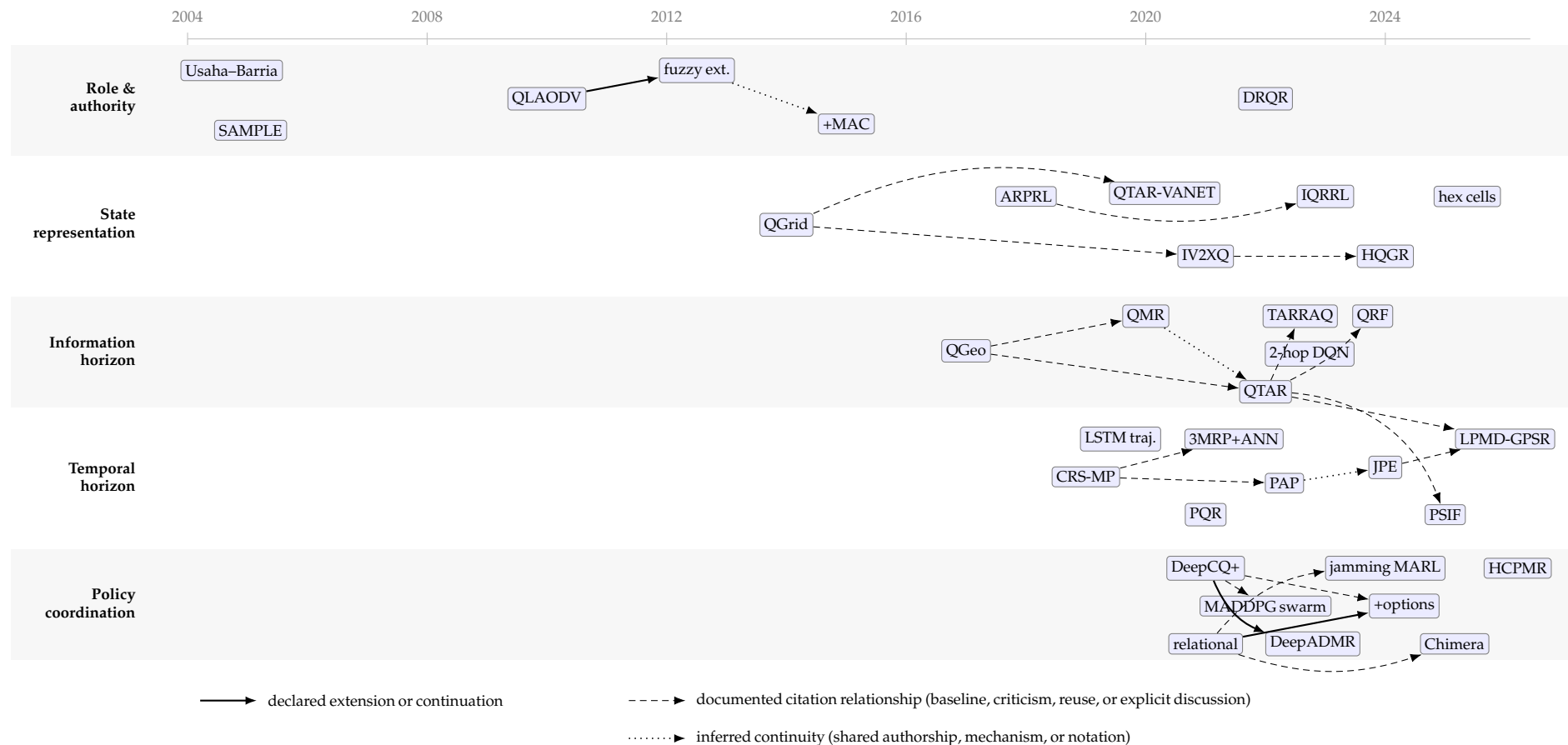


Figure 3. Developmental and citation relationships among the principal works of the five dimensions, placed by publication year, with the relations traced in Sections 3.2–3.6. Solid arrows mark declared extensions or continuations, dashed arrows documented citation relationships (baseline, criticism, reuse, or explicit discussion), and dotted arrows continuities inferred from shared authorship, mechanism, or notation. Principal works by lane, with their references: role and authority: Usaha-Barria [15], SAMPLE [13], QLAODV [12], fuzzy ext. [50], +MAC [65], DRQR [16]; state representation: QGrid [28,30], ARPRL [64], QTAR-VANET [29], IV2XQ [52], IQRRL [66], HQGR [53], hex cells [67]; information horizon: QGeo [34], QMR [68], QTAR [32], TARRAQ [4], 2-hop DQN [42], QRF [35]; temporal horizon: CRS-MP [1], LSTM traj. [19], PQR [60], 3MRP+ANN [59], PAP [7], JPE [58], PSIF [31], LPMD-GPSR [17]; policy coordination: DeepCQ+ [18], relational [44], MADDPG swarm [2], DeepADMR [63], jamming MARL [5], +options [45], Chimera [14], HCPMR [3].

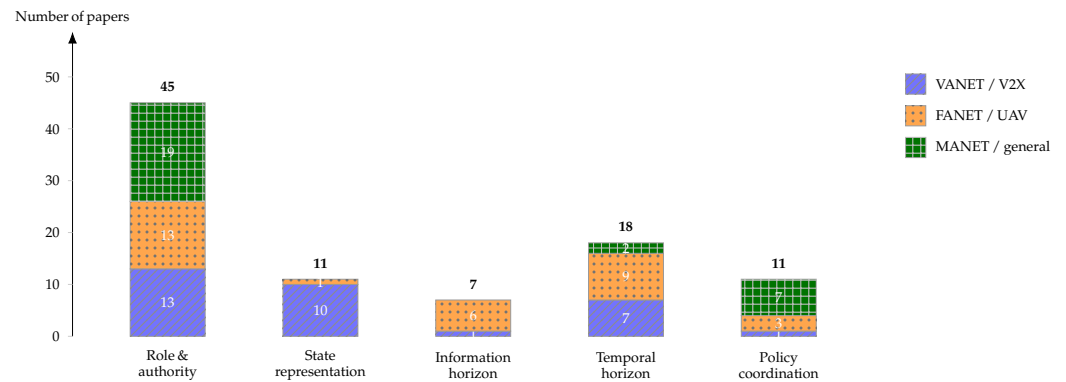


Figure 4. Distribution of the 92 papers across the five primary-contribution dimensions and three network scenarios.

3.2. Role and Decision Authority of the Learner

The first dimension examines the learner’s role and decision-making authority within the routing stack. Rather than focusing on model depth, it addresses where the line of responsibility is drawn between the classical protocol and the learning agent. At one extreme is the learning-assisted protocol in which the learner tunes a single parameter, scores links, ranks routes, or makes the hop-by-hop choice, while route discovery, route repair, and control signaling remain the responsibility of a classical protocol. At the opposite extreme is the learned forwarding policy, which may extend to a clean-slate design in which routing is posed directly as a learning problem and the policy itself is the forwarding logic, with the classical core reduced to a baseline or a feature source or removed entirely. With 45 of the 92 papers, this dimension is the largest class in the corpus. The most common defining contribution is still the extent of the routing decision assigned to the learner.

Both extremes were visible by the mid-2000s. At the assisted extreme, for example, Usaha and Barria [15] used a Monte Carlo reinforcement learning method to tune a single parameter (the initial probing-ticket allocation, chosen within a pre-specified maximum) of the bounded QoS search scheme of Section 2.2, leaving the discovery framework otherwise intact. At the learned-forwarding extreme, SAMPLE [13], a collaborative-RL routing scheme designed without a classical routing scaffold, chooses every packet’s next hop probabilistically from distributed Q-values, leaving no classical route-computation core for learning to “assist.” These two independent studies illustrate the two endpoints of this dimension.

The corpus’s most clearly documented lineage grew at the protocol-scaffold extreme in VANETs. Wu et al. embedded Q-learning into the AODV scaffold (QLAODV) to score path quality from vehicle mobility and available bandwidth and to switch to better routes pre-emptively [12]. Their subsequent extension replaced the hand-crafted combination of link metrics with fuzzy-logic fusion, moved Q-value selection into route discovery itself, and removed the hard dependence on position and physical-layer information, assessing links from hello message reception when positioning is unavailable [50]. The same group’s third step broadened the scope of learning beyond route selection to medium-access control, representing a shift toward broader system-level integration [65], a continuation that is visible in the shared authorship and the field-tested system rather than an explicit citation of its predecessor. Deep reinforcement learning subsequently appeared at both ends of the spectrum rather than in sequence. DRQR [16] retains AODV’s control machinery yet assigns the hop-by-hop forwarding decision to a deep Q-network in cognitive-radio MANETs (Section 4.1). By contrast, [44] and its continuation [45] extended the learned-forwarding approach with relational features and deep RL over temporally extended actions, reported to generalize across traffic patterns, connectivity regimes, and, in [45], node mobility. These

two works occupy the same end of the spectrum as [13], now using deep networks. Their defining generalization result is taken up under the fifth dimension (Section 3.6).

Around this core lineage, a substantial number of studies fall within the assisted category. Many vehicular schemes occupy it with different scaffolds: ARPRL treats each packet as an agent choosing the maximum-Q neighbor [64], and HAEQR augments QLAODV's selection with an explicit link-maintenance-time model [69]. RL-LAR integrates a per-vehicle Q-learner into the request zone of location-aided routing (LAR) [70], and a routing-table-free variant maintains no end-to-end route, relying only on an adaptively updated Q-table [71]. Others adapt the contention window of the routing protocol for low-power and lossy networks (RPLs) with Q-learning in vehicular deployments [72], drive named-data VANET forwarding with deep prioritized SARSA [40], and add a learned UAV-relay load-balancing layer above classical multipath routing [73]. Collectively, these studies broaden the variables and operational objectives considered by the learner, including mobility, link persistence, contention, and load, without changing the learner's broad functional role within the surrounding routing architecture. Deep networks enabled more complex decision functions without changing the learner's functional role within the routing stack: a deep Q-network (DQN) determines the grid-to-grid routing trajectory, while low-level packet forwarding remains greedy [74]; a convolutional network classifies the optimal next hop from a stacked neighbor-feature matrix on the GPSR scaffold [75]; and a security-oriented stack uses an ensemble-LSTM node classifier to blacklist malicious vehicles before a bio-inspired optimizer routes [20].

The non-vehicular literature covers similar levels of learner decision authority. In general MANET settings the learner variously replaces the hop-count metric with a cross-layer time metric [76] or groups neighbors to estimate delay and delivery [77]. In other schemes it adds mobility-, energy-, and link-quality-aware multipath choice [78], decouples exploration from data traffic in a hybrid on-demand/proactive Q-router [79], or learns a probabilistic multipath policy via SARSA [39]. In a further scheme it solves routing as a partially observable Markov decision process (POMDP) under incomplete information about neighbors' private parameters [80]. Across these studies the learner is given a different quantity to optimize, while the surrounding protocol continues to supply the candidate set. Among the deep variants, several give the learner direct control over next-hop selection while retaining classical signaling, neighbor discovery, or other routing support. Examples include prioritized-replay double-DQN forwarding [81], PPO-based single-hop selection replacing GPSR's greedy/perimeter rule [57], and deep-Q geographic routing that accounts for channel fading and stale beacons [54]. A GRU-DDPG controller emits continuous per-link weights under an SDN [82], and a DQN geographic policy serves high-speed robotic networks [83]. One robotic-network scheme learns QGeo's reward function by inverse RL instead of hand-crafting it [47]. DeepMPR replaces hand-designed multi-point-relay selection, familiar from OLSR, with a learned multicast forwarding policy [33]. Two opportunistic/delay-tolerant schemes learn the next forwarder through fuzzy-plus-Q-learning [27] or a supervised classifier over routing features [26]. In FANETs the dimension is just as densely populated: Q-FANET weights its Q-value updates by channel quality and episode recency [51]; a traffic-balancing Q-network penalizes queue backlog on GPSR [84]; QNGPSR reduces the use of perimeter mode, which can lengthen paths, by scoring neighbors with a Q-network [85]; AR-GAIL learns a forwarding policy using AODV as the expert [46]; a full-echo Q-router updates every neighbor's value and anneals its exploration [86]; and a further cluster adds adaptive Q-learning/AODV switching [87], rate control [88], distance-priority multi-Q updates that accelerate geographic route learning [89], dueling-DQN link-quality routing [90], resilient prioritized double-dueling forwarding under node failures [91], attention-based DQN over IP-encoded

positions [92], centralized weight-tuning for cluster-head merit [93], and a public-safety scheme in which UAVs relay for static victim clusters, a mixed aerial–ground setting that Section 3.8 accordingly files under MANET/general [6].

Only a few works occupy the learned-forwarding extreme: SAMPLE [13] and the relational forwarding-policy line [44,45]. Most studies remain within the assisted category. One boundary case, a scheme that learns to reposition a relay node while classical AODV continues to route [94], is retained in the assisted category: the learner does not select next hops, but its observations are router-observable signals, its reward is end-to-end delivery, and its actions determine whether the multi-hop path that AODV can use exists, so it shapes the routing decision through the connectivity it creates rather than the forwarding choice itself.

3.3. Network-State Representation Used by the Learner

The second dimension, the network-state representation used by the learner, focuses on what network entity, and which of its attributes, the learner represents as the state for its routing decisions: an individual vehicle or UAV, a local neighbor set, a grid cell, an intersection, a road segment, or a group of roadside units, described by attributes such as position, velocity, queue length, link lifetime, or residual energy. Here, “network state” refers to the information represented at the learner’s decision point, not to the complete global state space of the entire network. The choice has proved especially consequential in urban VANETs, where using individual vehicles as states scales poorly with network size: the state space grows with network size, slowing convergence [28], and vehicle movement can render vehicle-level Q-tables obsolete before they converge [53]. The papers on this dimension redefine the state representation to address this scalability problem, and what matters is not model depth (every scheme here is tabular Q-learning except for the single DQN variant [55]) but that the chosen unit governs the size of the state space, how faithfully road topology is expressed, how infrastructure enters the decision, and whether the table can be adapted online. There are 11 papers in this category, all of which target vehicular applications with the exception of the most recent one.

QGrid [28,30] made the foundational change by changing the unit of abstraction: the city is divided into uniform geographic grids, and the grids—not the vehicles—serve as the states of Q-learning. A table of learned Q-values, trained offline on historical taxi GPS traces by exploiting the day-to-day stability of traffic within each grid, selects the next grid toward the destination, while a non-learned heuristic picks the specific relay vehicle inside it. Two design choices were central to QGrid: the grid geometry and the offline static table. Subsequent studies modified both. IV2XQ [52] identifies four limitations of QGrid: grids ignore intersections and building occlusion; the offline table cannot observe network load; only a single fixed destination is supported; and always choosing the maximum-Q grid overloads popular regions. Benchmarking directly against QGrid, it re-grounds the abstraction in the road network itself: intersections become states, road segments become actions, and RSUs at intersections serve as relays, although the Q-table remains offline and congestion is handled by a fixed rule. HQGR [53] then applies the same treatment to IV2XQ: taking it as the main baseline, criticizing its slow convergence and a Q-table overhead that scales with the product of RSU count and neighbor count, and redesigning the hierarchy around RSU groups whose Q-tables are learned online. The three teams—QGrid’s [28,30], IV2XQ’s [52], and HQGR’s [53]—share no authors. This indicates that the progression was not confined to a single research group.

Two additional lines of research developed independently. QTAR, proposed in [29], should not be confused with the FANET protocol of the same name in Section 3.4, so we refer to it as QTAR-VANET where context does not disambiguate. QTAR-VANET also argues

against QGrid’s offline table—although the two are never compared empirically—running Q-learning fully online at two levels, between vehicles within a road segment and between RSUs across intersections. IQRRL [66] arrives at the same two-level division without a direct citation relationship to these studies. However, it allocates learning and explicit scoring in the opposite way to IV2XQ, suggesting that the division can arise independently. Its next intersection is chosen by explicit QoS scoring rather than learning, while RL is confined to selecting the relay vehicle within a segment. IQRRL instead cites ARPRL [64], its only learned baseline and the acknowledged source of its learning-rate setting. A cluster of related hierarchical designs elaborates the same idea: hello packets carry group Q-vectors and link rewards from which each RSU builds its group and local Q-tables [95]; a software-defined directional grid platform folds positions and historical trajectories into grid directionality and two-hop relay selection [96]; and a global-plus-local intersection scheme runs a central Q-learning server [97]. A DQN variant centers the state representation on the current intersection, augmenting it with packet details and a timestamp. It then selects the next relay intersection as its action, extending this abstraction into a deep learning framework [55].

The single non-vehicular entry provides particularly clear evidence that the principle is not confined to roads: a software-defined FANET protocol partitions the mission space into regular hexagonal cells and learns at the cell level, so the Q-table scales with the number of cells rather than the number of UAVs [67]. This is the first example in the corpus to apply the same state-abstraction principle to an aerial network, suggesting that the dimension is not intrinsically vehicular but appears to have emerged first in vehicular settings. Across the corpus, the represented unit ranges from individual vehicles to grids, intersections, road segments, RSU groups, and aerial cells, while the learner remained largely fixed. Section 4.2 examines what this re-engineering gains in convergence and costs in fidelity and adaptivity.

3.4. The Information Horizon

The third dimension, the information horizon, focuses on how much of the network a learner can observe and the control effort required to access it rather than what the state signifies. The distinction is the observation boundary, that is, how much present information the decision acquires and, at its inner edge, how the set of neighbors it evaluates is delimited: a one-hop neighborhood, a two-hop neighborhood, sensed local topology dynamics, or a candidate set narrowed by geometric or external filtering. This dimension is most clearly delineated in FANETs, where information obtained through beacon exchanges can become outdated within seconds and expanding the observation range incurs additional control overhead [4,32]; five of its seven members form the most closely related group of studies in the corpus, a sixth [62] cites into it, and only the deep-learning member [42] stands apart.

The earliest work in this lineage is QGeo [34], framed for “unmanned robotic networks” spanning aerial and ground vehicles rather than FANETs proper. It recast one-hop geographic forwarding as distributed Q-learning: each node selects the next hop among its one-hop neighbors, guided by a reward based on packet travel speed that also accounts for link and location error, with periodic hello messages carrying position, current Q-value, link condition, and location error. Its closest successor in this corpus is QMR [68], which keeps the one-hop view but performs multi-objective optimization—jointly minimizing delay and energy—replacing QGeo’s fixed learning rate and two-level discount factor with a learning rate adapted to observed delay variation and a discount factor adapted to neighbor-set change; QGeo is the only protocol baseline in QMR’s reported simulations.

Within this lineage, the boundary was subsequently widened only once. QTAR [32] argues that a one-hop view can lead to blind paths and routing holes and extends the exchanged information to two-hop neighborhoods augmented with estimated link durations and delay, velocity, link-quality, and residual-energy information. It takes QGeo as a baseline rather than an acknowledged predecessor, and its adaptive learning-rate and discount-factor rules share their mathematical form with QMR's. Although this relationship is not explicitly discussed, the two studies use structurally similar update equations. TARRAQ [4] calls QTAR "a promising solution" and takes it as a baseline. Yet TARRAQ rejects QTAR's defining two-hop exchange as excessive overhead and quantifies the complexity gap. It reverts to one-hop information, compensating with adaptive sensing intervals and Kalman-filter prediction of each link's residual lifetime. The two protocols therefore address the same problem, topology awareness under rapid change, but not by the same method: TARRAQ substitutes temporal prediction for spatial breadth. Later work continued the substitution rather than resuming the expansion: LPMD-GPSR benchmarks against QGeo, QTAR, and QRF [35] yet abandons the two-hop widening, and PSIF cites QTAR yet does not position itself as an extension of this lineage; both predictive designs are treated under the prediction dimension (Section 3.5).

Meanwhile, a parallel line narrows the candidate space within a one-hop view: QRF [35] filters the candidate set geometrically into a spherical sector before Q-learning to counter the state-space growth it associates with QTAR's two-hop exchange. The one deep-learning member independently revisits the use of two-hop information: a DQN whose state stacks explicit one-hop and two-hop feature matrices while requiring only partial two-hop information, an accommodation of the incompleteness resulting from unstable links [42]. The sole vehicle-based member of this dimension extends the decision scope vertically instead of laterally. Specifically, a UAV-assisted VANET framework relies on an aerial tier to calculate a global route using fuzzy logic scoring and depth-first search, which then enables ground vehicles to filter out off-route or congested neighbor nodes when choosing their next hop [62].

The overall trajectory of this dimension is thus not monotone growth in what the router knows: within the main lineage the observation boundary expanded from one hop to two and then contracted again under overhead pressure, while the information supplied to the decision grew richer along other dimensions from measured neighbor tables to predicted link lifetimes and physically sensed surroundings.

3.5. Temporal Horizon: Measured Present Versus Predicted Future

The fourth dimension, the temporal horizon, asks whether the decision rests on the network as measured now or as it is predicted to become. This categorization depends on the temporal horizon of the routing input rather than the specific prediction model used. Papers belong in this category if they feed an explicit future-state forecast, or a prospective representation of future network states, into the routing decision regardless of whether that forecast is generated by an ANN, an LSTM, a statistical mobility model, a gradient-boosting regressor, or a clustering-based approach. (One boundary case instead projects into the future by planning across recorded encounter opportunities without generating an explicit forecast, as detailed in Section 3.7). Counted by defining contribution rather than algorithm, this dimension is the second-largest, with 18 papers.

The clearest early instantiations of this prediction-before-routing architecture are vehicular. CRS-MP [1] uses an SDN controller to train an ANN on real traffic-detector data to forecast the vehicle arrival rate on each road segment; RSUs and the base station convert that forecast through a stochastic traffic model into each request's success probability and expected delay "ahead of time," choosing between infrastructure-assisted

vehicle-to-infrastructure (V2I) and multi-hop vehicle-to-vehicle (V2V) service modes so as to minimize overall service delay. PQR [60] independently implements a related prediction-before-routing architecture with no direct citation link between the two papers: an acceleration-based statistical trajectory model estimates each link's remaining lifetime while a gradient-boosting model fed by online measurements predicts route quality, and the protocol switches routes before the current one breaks or degrades. Two further branches complete the vehicular core: an ANN-based forwarder [59] replaces the multi-metric forwarding criterion of the authors' own protocol 3MRP with per-candidate delivery-probability prediction, citing [1] only as background, and a delay-tolerant branch [19] has each neighboring vehicle predict its own trajectory with an LSTM, so a sender can transfer its data to the vehicle predicted to reach a base station earliest.

A similar prediction-before-routing architecture subsequently emerged in FANETs, although as a result of engineering reuse rather than explicit conceptual lineage. PAP [7], the earliest FANET entry in this prediction-before-routing line, uses an LSTM to forecast each UAV's packet arrival rate and folds the forecast into a routing decision factor together with queue backlog, distance, and hop count; it cites [1] for its SDN control architecture and channel model but not for the underlying prediction concept. JPE [58] is treated here as an inferred continuation of this line on the basis of shared authorship, notation, and forecasting framework. Five of JPE's seven authors overlap with those of PAP [7], whose notation and LSTM framework carry over. JPE criticizes PAP by name ("only traffic-related prediction is not rewarding to multi-metric sensitivity routing decisions") and retains it as a baseline. At the same time, prediction widens to joint mobility-and-traffic forecasting. Beyond that two-generation core the FANET line branches rather than continues: LPMD-GPSR [17] cites JPE only once to call its link-duration computation "relatively complex" and instead trains an LSTM to predict link stability directly from timestamped neighbor tables on the classical greedy scaffold, while PSIF [31], which in fact predates [17] and arrives at the same substitution outside the PAP–JPE authorial line, migrates prediction-driven routing to GPS-denied low-altitude airspace, fusing obstacle intensity from the onboard SLAM map with historical signal strength into an LSTM link-quality forecast consumed by a DQN forwarder.

This dimension's wider population organizes by what is predicted. Around the vehicular core sit an RSU-based system that predicts each vehicle's next turn and the next RSU on its path by clustering [49], a compatibility-aware scheme that predicts the connectivity duration between vehicles and quantizes it into programmable reliability classes [98], and an SDN scheme whose random forest unit forecasts each vehicle's next intersections and arrival times so that its deep relay-selector can favor relay vehicles whose predicted encounters are feasible [61]. In the aerial literature, PARRoT folds trajectory prediction into the reinforcement-learning update through a time-varying mobility-aware discount factor [99]; a prediction-driven 3D protocol switches its forwarding mode on a forecast of each UAV's future distance to the ground station [100]; and FANET schemes evaluate the future evolution of link states before relay selection [101], predict UAV motion to drive adaptive beaconing and DDQN forwarding [102], and anticipate future signal quality in a Q-learning router [103]. Two general works complete the dimension: a link-failure-resilient Q-router predicts each neighbor's remaining reachability time and folds it into a time-varying discount factor [104], and a delay-tolerant scheme builds a temporal-graph state whose survival action lets the agent defer a present forwarding opportunity in favor of a future encounter represented in the graph, planning over the future without explicitly forecasting it [105].

A broad tendency is observable across the vehicular and aerial prediction-based studies: prediction targets shift from aggregate traffic flows toward per-link lifetimes and

future link quality, the input space widens to physically sensed surroundings, and the forecast moves ever closer to the forwarding decision itself. Section 4.4 examines the benefit of computing a forecast before the routing decision and the impact of prediction errors.

3.6. Policy Organization and Coordination

The fifth dimension, policy organization and coordination, applies when a network contains many learners at once and asks how the resulting policies are organized, shared, coordinated, generalized, and deployed alongside classical rules. It concerns the lifecycle and governance of learned routing policies across nodes and operating conditions, including how policies are trained, shared, coordinated, transferred, selected, monitored, and replaced. It is broader than MARL: it covers a single size-agnostic policy shared under CTDE, explicit multi-agent cooperation, generalizable forwarding policies, and hybrid policy pools with rule-based fallback [2,14,18,45]. A paper belongs here when the organization of policies, not merely the fact that many nodes run the same algorithm, is its signature. This dimension is important because a network of learners raises problems that improved single-node reward design alone does not address: how many policies to maintain, how to keep them cooperating under a changing topology, how to generalize them to unseen network sizes, and which fallback mechanism should be used when a learned policy should no longer be trusted. It accounts for 11 papers.

Its starting point in this corpus, DeepCQ+ [18], already addressed several aspects of the question at once. A single policy trained with PPO is shared by every node (“one policy for all”) and generalizes across network sizes through a size-agnostic neighbor encoding (the mechanism is detailed in Section 4.5.1); yet the learner has a deliberately limited decision role, deciding only whether to broadcast or unicast each packet while next-hop choice, acknowledgement bookkeeping, and loop suppression remain the responsibility of the rule-based CQ+ protocol it is embedded in. Scalable generalization and learned-rule hybridity were thus present at this dimension’s origin in the form of one learned decision deliberately embedded inside a classical protocol.

Multi-agent studies then developed coordination further in different environments. For UAV swarms, [2] uses DeepCQ+ in three roles: as background, as the target of a named criticism (jointly with another learned scheme for ignoring channel quality and node load), and as an experimental baseline. It responds with a cooperative stochastic game solved by MADDPG with recurrent actor and critic networks. For tactical sensor networks under jamming, [5] proceeds in parallel—with no citation contact with DeepCQ+ or [2]—and addresses the corruption that jamming introduces into the learning signal itself; its place on this dimension rests on organizing distributed policies under such corrupted feedback. Alongside these, a geographic FANET routing scheme employs cooperative Q-learning with a Nash-Q update mechanism across single-hop agents [43]. A cooperative packet-routing algorithm makes each router a double-DQN agent over local and neighbor-queue state [56], and a vehicular scheme lets independent Q-learning agents share and update a single Q-network [41]. The relational forwarding-policy studies discussed under the first dimension are also relevant to this dimension: [44] and its same-team continuation [45], whose frozen-policy generalization spans device counts and, in [45], mobility models. The key contribution of these studies lies in policy organization and coordination rather than merely establishing learned-forwarding ownership. The relational and DeepCQ+ lines are linked at one point, where [45] cites the DeepCQ+ work among approaches confined to “relatively limited network scenarios with a few fixed flows and up to 50 devices.”

Chimera [14] integrates these approaches and reorganizes the hybrid. It engages [44] directly as one of its three learned baselines and the target of a named criticism (distributed training that can be trapped in local optima; an unrealistic assumption that nodes know

all others' real-time positions). Chimera addresses that criticism through its architecture rather than through increased model complexity: a policy pool in which nine learned policies and a classical protocol coexist. A device-agnostic vector of network conditions selects among the pooled policies. When conditions are too far from anything seen in training, the scheme falls back to an AODV-style traditional policy. A further member governs rather than routes: DeepADMR is not a router but a real-time anomaly detector that monitors the temporal-difference errors of a deployed DeepCQ+ policy and flags when the learned policy is operating outside its trusted operating regime [63]. This makes the underlying issue of trust along this dimension explicit. The most recent work on this dimension at the corpus cut-off, HCPMR [3], is a largely independent FANET branch: a GNN topology encoder with hierarchical multi-agent PPO within clusters and hand-crafted utilities between cluster leaders. Its experimental baselines (IQMR and CQMR) lie outside the corpus, and its only direct citation contact with the 10 works discussed above is a related-work citation of the Nash-Q geographic router [43]. Overall, the evolution across this dimension centers on the structural architecture of coexistence between learning and rules rather than whether they coexist as they have since the beginning.

3.7. Cross-Dimension Relationships and Boundary Rulings

Cross-dimension relevance. Because the five dimensions are cross-cutting descriptors rather than disjoint bins, many papers have a primary contribution on one dimension and secondary relevance on others; the primary assignment follows the defining-contribution rule of Section 3.1 and does not imply that a paper's contributions are confined to that dimension. DRQR is primary on the learner-role dimension since its defining contribution is relocating hop-by-hop selection into a DQN inside the AODV scaffold. It also widens the information the decision consumes to a cross-layer physical/MAC/network state [16]. JPE is primary on the prediction dimension for its joint forecast but also strengthens multi-metric path scoring [58]. PSIF is primary on the prediction dimension for its future-link-quality forecast yet simultaneously enriches the information that is available to the routing decision with physically sensed obstacle context, which is why it recurs in both the information and prediction discussions above [31]. LPMD-GPSR predicts link stability (temporal) but does so on a classical greedy scaffold, remaining partly within the assisted band of the learner-role dimension [17]. The two-level QTAR-VANET and IQRRRL are primary on state representation yet clearly also concern where learning sits since each also lets the learner control part of the forwarding [29,66]; the hexagonal-cell FANET scheme is primary on state representation but achieves scalability by decoupling Q-table size from the number of UAVs through its cell-level state [67].

Boundary rulings. The relational forwarding-policy line illustrates most sharply why the primary label must follow the defining contribution. Its learned forwarding policy places it near the far end of the learner-role dimension, but the contribution its authors present is a forwarding policy that generalizes across networks without retraining, which Table 3 lists under policy organization as size-agnostic generalization; we therefore assign that dimension as primary and record the learner-role placement as secondary [44,45]. SAMPLE receives the reverse assignment: it is assigned primarily to role and decision authority because its defining contribution is to make the learned policy itself the forwarding mechanism; its collaborative policy-sharing machinery is recorded as secondary [13].

The remaining boundary decisions mark where this classification departs from a by-algorithm reading. Four works are assigned primarily to the temporal-horizon dimension, although a by-model view would classify them as generic learned routing: the next-RSU predictor [49], the connectivity-duration predictor [98], the future-link-state evaluator [101], and the proactive link-failure predictor [104]. Each places an explicit future-state forecast

in front of the routing decision. The temporal-graph opportunistic learner is retained on the same dimension as a boundary case: it deploys no explicit predictor, but its temporal-graph state and survival action assign value to future encounter opportunities learned from temporal patterns [105]. One work is excluded from the prediction dimension for the opposite reason: the security stack whose LSTM is a present-tense malicious-node classifier, not a forecaster, is read under the learner-role dimension [20]. A Q-learner that selects the next hop is not primarily a learner-role paper when its distinctive move is a two-hop observation window [42]; and DeepMPR, whose signature is re-learning MPR-style relay selection for multicast forwarding, is read under the learner-role dimension with coordination secondary [33]. Where an isolated feature-based reading suggests placement on another dimension (e.g., QGeo and QMR as information horizon [34,68]; QTAR-VANET and IQRRL as state representation [29,66]), we retain the primary assignment determined by the defining-contribution rule of Section 3.1 and record the alternative relevance as secondary. The primary assignment is used only for countable classification; secondary relevance is retained for the cross-dimension trade-offs analyzed in Section 4.

3.8. Master Classification of the Corpus

Table 5 presents the primary-contribution classification of all 92 papers by dimension and scenario, and Figure 4 summarizes the distribution. The prediction dimension is the second-largest category. It is also largely a recent development, with 15 of its 18 papers published since 2020. Network-state representation is vehicular apart from one aerial entry, and the information horizon is aerial apart from one UAV-assisted vehicular design.

Scenario columns use three broad families; finer settings noted in the text (e.g., the delay-tolerant/opportunistic works [26,27,105]; the tactical sensor setting of [5]; and the general mobile-wireless partly stationary regime of [44,45]) are counted in the MANET/general column. QGeo [34], despite its unmanned-robotic framing, is kept in the FANET/UAV column, as are its successors, including the inverse-RL study built on it [47]. Scenario placement follows the deployment setting each study reports independently of which dimension carries the paper's defining contribution.

Table 5. Primary classification of all 92 papers by decision mechanism and network scenario. Each cell lists the papers whose primary dimension is that row.

Primary Dimension	VANET/V2X	FANET/UAV	MANET/General
Role and decision authority (45)	[12,20,40,50,64,65,69–75]	[46,47,51,84–93]	[6,13,15,16,26,27,33,39,54,57,76–83,94]
Network-state representation (11)	[28–30,52,53,55,66,95–97]	[67]	—
Information horizon (7)	[62]	[4,32,34,35,42,68]	—
Temporal horizon (18)	[1,19,49,59–61,98]	[7,17,31,58,99–103]	[104,105]
Policy organization and coordination (11)	[41]	[2,3,43]	[5,14,18,44,45,56,63]

4. Mechanism-Oriented Comparative Analysis

Building on the framework of Section 3, this section compares designs mechanism by mechanism rather than by algorithm in order to set out the trade-offs a designer faces along each dimension. The comparison is deliberately selective: it examines the 37 core papers drawn in Section 3.1 by the considerations stated there, whose evaluations Section 5 audits. Section 4.6 then compares the mechanisms against one another and names the trade-offs that design descriptions alone cannot settle.

4.1. From Protocol Assistance to Learning-Owned Routing Decisions

The question that has attracted the largest body of work is how much of the routing decision the learner is allowed to control. The reviewed designs form a spectrum rather than

discrete classes, from a learner that adjusts one parameter inside an otherwise unmodified protocol to a learner that constitutes the forwarding logic outright. Both extremes are represented, although most designs remain closer to the protocol-assisted end of the spectrum.

4.1.1. Learning as a Bounded Protocol Assistant

At the assisted extreme the learner only tunes a parameter of the protocol: in the ticket-budget scheme of Section 3.2, the protocol's routing framework and control procedures remain largely intact, and learning primarily replaces the heuristic selection of the initial search budget [15]. QLAODV extends the learner's role by a controlled increment: Q-values now score end-to-end path quality and trigger a pre-emptive route change, but route discovery still uses AODV's flooding-based procedure and each switch is confirmed by a dedicated unicast route-change request/reply exchange [12]. By contrast, SAMPLE occupies the opposite extreme as early as 2005, retaining no conventional explicit route-computation phase for learning to assist [13]. Notably, such clean-slate designs remain a minority in the reviewed corpus. Assigning the whole decision to the learner also makes the learner responsible for its training cost and failure modes and may remove an established rule-based fallback.

4.1.2. Learning Embedded in Route Scoring and Route Maintenance

In a large share of the corpus, the learner makes the selection while the protocol retains discovery, repair, and signaling. The work of Wu et al. [50,65], introduced in Section 3.2, shows the clearest progression within this group. In the first of the two, the learned link score moves from route maintenance into route discovery itself [50]. The final design [65] extends learning beyond routing by adding a second Q-learner for the medium-access transmission rate and accelerates convergence for newly joining nodes through transfer learning. It is also validated on a 10-vehicle physical testbed rather than in simulation alone, a form of validation that is rare among the core papers reviewed here. A parallel formulation defines each packet as an agent, each node as a state, and each neighbor as an action, with the Q-values maintained by the vehicles; each packet is always forwarded to the maximum-value neighbor. This is the standard form of learning-owned next-hop selection within an otherwise classical protocol [64].

Two further works show that this group is not tied to any one model: a general MANET scheme first groups neighbors, learns group-level delay and delivery estimates, and selects a group before a hop-count heuristic identifies the specific relay [77], and a deep variant embeds a local link-status vector into a DQN geographic forwarder for high-speed robotic networks, explicitly to compensate for the absence of link-state information in earlier geographic Q-routing [83]. These designs share an engineering compromise: learning is assigned to ranking decisions that are sensitive to mobility-induced topology changes, while the protocol retains the discovery, repair, and signaling functions that are easier to constrain and verify. The corresponding limitation is that the learner can only optimize among the routes, neighbors, or neighbor groups that the surrounding protocol makes available.

4.1.3. Learning as Direct Routing or Forwarding Control

At the learning-owned end of the spectrum, where the learner assumes direct control of the forwarding decision itself, deep models are particularly relevant. DRQR is an illustrative case because it is deep yet deliberately constrained: the deep Q-network owns the hop-by-hop choice over its cross-layer state, while AODV's recovery mechanisms are retained, so the learner gains authority over selection without discarding the protocol's repair path [16]. A PPO agent assumes more of the decision within a geographic-routing protocol, replacing GPSR's fixed greedy-and-perimeter rule with a learned single-hop policy that aims to avoid routing holes rather than to react to them [57].

Direct learned influence on forwarding decisions is not exclusive to reinforcement learning: a supervised model in opportunistic delay-tolerant networks supplies a delivery estimate that enters the forwarding rule, with the hand-crafted PROPHET predictability metric retained as both an input feature and a comparator [26]. The trade-off across these designs is consistent: relocating the decision into the learner can enable optimization beyond the constraints imposed by hand-designed forwarding rules and replace some hand-crafted rules, but it concentrates failure modes in a single model and may complicate behavioral verification and certification.

4.1.4. Comparison Across the Authority Spectrum

The spectrum can be compared along four deployment-relevant criteria. Protocol compatibility is highest at the assisted end, where [12,15] leave the standard stack recognizable and the learned part small and auditable. By design it is lowest at the clean-slate end, where [13] preserves no conventional route-discovery and route-computation stack, so interoperability with protocols built around that stack is correspondingly limited. Control authority moves in the opposite direction from a single tuned parameter through learned route ranking and maintenance [50,64] to primary control over the forwarding action within the retained scaffold [16,57]. Engineering practicality is monotone in neither direction. Within the audited core set, the clearest hardware evidence sits in the middle band. There, a physical multi-vehicle testbed and transfer learning for newly joining nodes [65] show learned selection on a classical scaffold to be the sub-band closest to real hardware. By contrast, the cited works at both extremes are validated in simulation alone. High representational capacity gives deeper designs in the middle and inward bands an advantage as their neural policies can handle high-dimensional cross-layer state inputs [16] for which tabular approaches scale poorly. However, deployment-level scalability still hinges on managing the training and certification overhead that lighter assisted designs are specifically built to avoid. Most reviewed studies balance these four factors: sufficient decision authority to seek improvements over hand-tuned metrics, enough compatibility and practicality for deployment, and no more autonomy than can be validated.

4.2. Spatially Structured Routing State in Urban VANETs

The second mechanism uses essentially the same learning method (tabular Q-learning in all but the single DQN variant [55]) and instead varies the spatial unit it reasons over. This is the most self-contained technical line of development in the corpus, and it is essentially a vehicular one because the urban road network creates the scalability problem while also providing a natural spatial structure for abstraction.

4.2.1. Vehicle-Level Routing and Its Instability

This motivation is clearer when examined in reverse, starting from studies that abandoned vehicle-centric state representation. A learner that models individual vehicles as states often struggles to stabilize in dense fast-changing urban traffic [28]. The designs in this section respond to that instability in different ways, and they are best compared by four consequences: how far each compresses the state space, how faithfully it preserves road topology, whether it can adapt online, and how heavily it relies on fixed infrastructure.

4.2.2. Grid-Based State Representation

QGrid introduces a uniform-grid abstraction, as mentioned in Section 3.3: a compact table over grid states, learned offline from taxi traces and pre-stored in each vehicle [28,30]. This abstraction provides substantial state-space compression and day-to-day stability, but it has two drawbacks. First, the grid-level state does not explicitly encode intersection-level road structure or local obstructions, limitations that its successors later made explicit [52]

(Section 3.3). Second, the offline table does not track live network load. QGrid achieves the strongest state compression on this dimension, but it does not adapt online. Figure 5 illustrates the resulting decision process. Q-values attach to grid transitions rather than vehicles, so the table's size is fixed by the map and the grid granularity. This compression enables offline training but leaves the macroscopic table unable to observe live load.

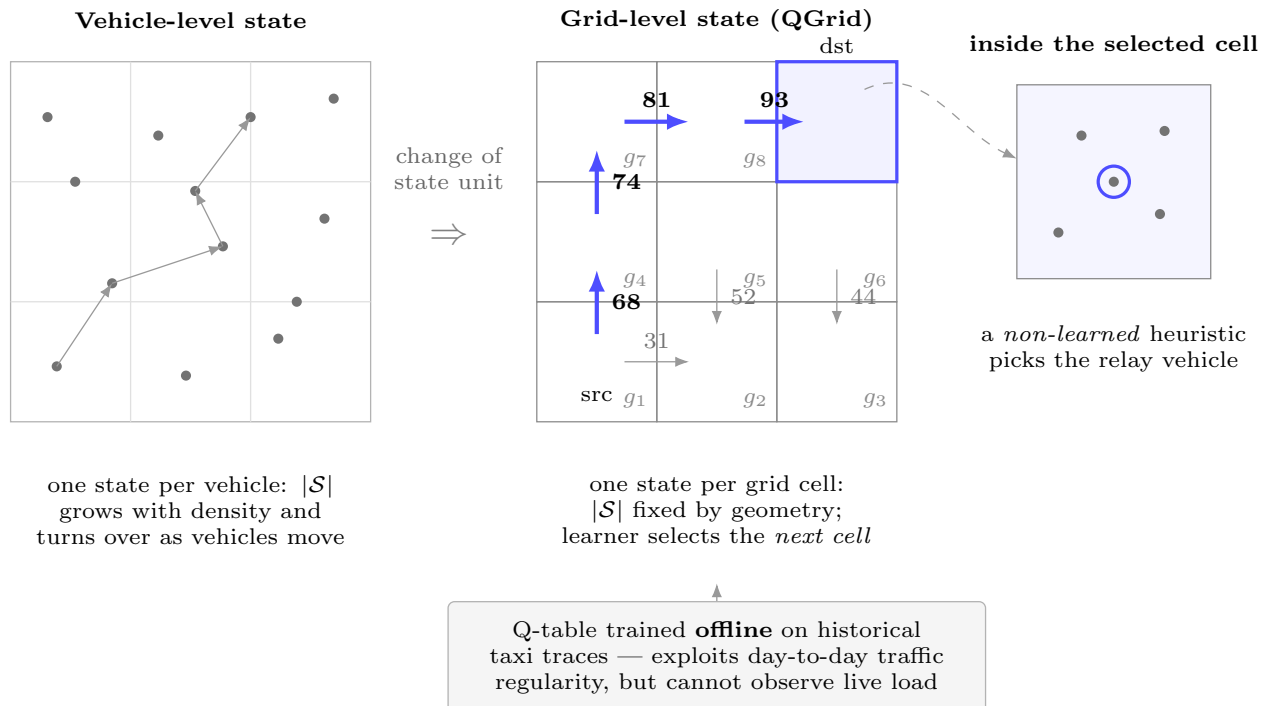


Figure 5. Grid-level Q-learning in QGrid: grids are the states, inter-grid movements the actions, and the highest-valued transitions form candidate paths.

4.2.3. Intersection- and Segment-Based State Representation

Subsequent designs restore road-topology fidelity by re-grounding state representation within the network structure itself (Section 3.3). In this setup, intersections serve as states while road segments act as actions, with Q-tables computed offline and traffic congestion managed via rule-based logic [52]. In a second design, a central server updates an intersection-level Q-table online from fresh traffic reports, while a local mechanism handles forwarding within each segment [97]. In the third, a QoS-scored intersection choice confines learning to multi-hop relay selection within the chosen segment, delegating intersection-level route computation to QoS scoring and Dijkstra-based classical computation [66]. Compared with the grid, these designs represent road topology more explicitly and, in the QoS-scored case, global path awareness, but their online adaptability differs substantially: [52] remains primarily offline, [97] adapts at the global intersection level, and [66] adapts locally within the selected segment. Fidelity improves across these designs, while online adaptability varies rather than improving uniformly.

4.2.4. Infrastructure-Aware and Grouped Hierarchy

The final architectures regain online adaptability through roadside infrastructure. One utilizes a two-level online learning framework with Q-tables that track real-time traffic both along road segments and across intersections [29]. The other is the grouped-RSU hierarchy of Section 3.3, which learns online at group and local levels, with a reported 20–34% delivery-ratio gain over the intersection scheme it takes as its baseline [53]. These two architectures provide the clearest online adaptation among the designs discussed

here. In their respective evaluations, they also report favorable packet-delivery performance. This adaptability, however, requires real-time infrastructure support and recurring infrastructure-assisted exchanges. The grouped hierarchy is explicitly designed to reduce the memory and communication burden relative to the ungrouped intersection scheme it builds on. The designs on this dimension therefore span a trade-off: from an offline grid with the strongest state compression and the least topology detail through intersection schemes of varying online adaptivity to grouped online hierarchies that track current traffic load but require infrastructure support. No design in the reviewed set yet combines maximal compression, full topology fidelity, and infrastructure-free online adaptation.

4.3. Topology Awareness and Its Control Overhead

The third mechanism concerns how much of the network the learner is allowed to observe and whether a wider view justifies the additional overhead it incurs. This dimension is most pronounced in flying ad hoc networks (FANETs), where three-dimensional mobility can sever communication links within seconds and the overhead of collecting topology information is significant compared to the payload data being carried [4]. The principal works on this dimension treat that overhead as an explicit object of study rather than a secondary concern.

4.3.1. One-Hop Local Awareness as the Baseline

The baseline designs adopt the narrowest usable information horizon. QGeo decides from information exchanged with immediate neighbors using its travel-speed reward [34]; QMR retains that one-hop exchange and one-hop actions but extends the objective optimized over them [68]. Together they establish that considerable performance can be achieved from one-hop information if the one-hop decision rule is well designed. This provides a reference point for the claims that follow: any benefit attributed to a wider information horizon must be shown to exceed what a well-designed one-hop scheme already achieves.

4.3.2. Two-Hop Topology-Aware Routing

The most explicit expansion of the information horizon is QTAR's two-hop neighborhood exchange with link-lifetime annotations, introduced in Section 3.4 to mitigate the blind paths and routing holes inherent in a strictly local view [32]. This is the widest explicit neighborhood on this dimension, and it is designed to reduce the likelihood of reaching dead ends. The overhead of this two-hop exchange is reassessed in the work discussed next. Figure 6 shows how the widened neighborhood is used in the design: the two-hop neighbor tables are the primary source of the learner's state, reward, and action inputs, so their maintenance cost cannot be removed without changing the mechanism itself.

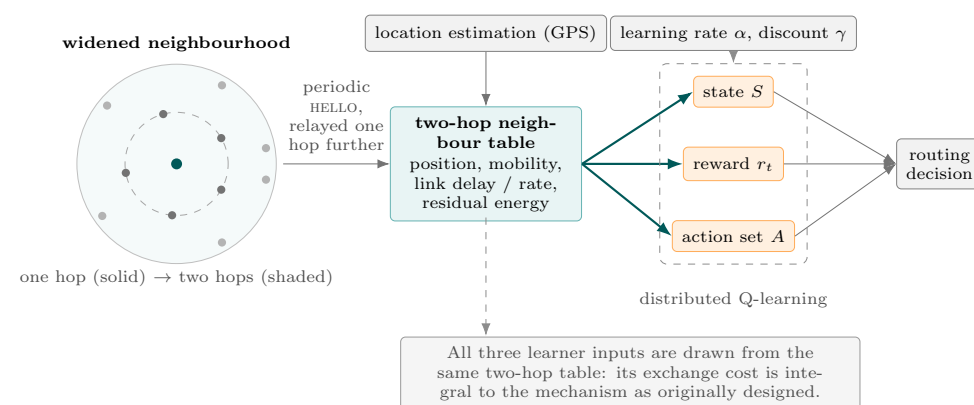


Figure 6. The QTAR framework: two-hop neighbor tables supply the state, reward, and action space of distributed Q-learning.

4.3.3. Predictive or Filtered Compensation for Limited Visibility

The most overhead-conscious strategies on this dimension manage visibility through distinct trade-offs. Treating the overhead of two-hop exchange as excessive under the evaluated conditions and noting that two-hop protocols generate the highest control overhead among baseline approaches, TARRAQ restricts its scope to one-hop information. To compensate, it incorporates a queuing-theoretic model of neighbor dynamics paired with a Kalman filter to predict residual link lifetimes, effectively substituting local computation for communication overhead to maintain link-stability awareness [4]. A different response filters rather than predicts: a spherical-coordinate scheme narrows the one-hop candidate set geometrically before Q-learning, controlling the state-space growth attributed to wider two-hop views [35]. In the comparisons examined here, when control overhead is included in the reported comparisons, widening the live observation horizon can be disadvantageous in a fast aerial topology. A second limitation concerns the effective size of the state space. A wider horizon enlarges the state the learner must generalize over, so two-hop visibility can slow convergence in addition to raising overhead. Figure 7 illustrates the substitution: the geometry of the communication range, swept at an estimated velocity, yields neighbor arrival and departure rates analytically, so topology awareness is computed rather than reported.

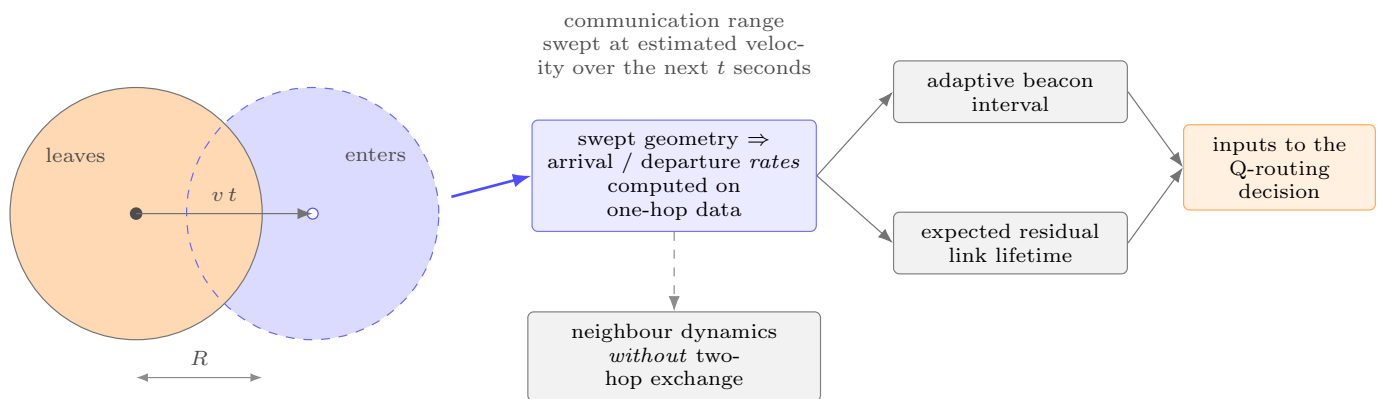


Figure 7. TARRAQ's analytical neighborhood model: the regions a UAV will enter or leave within the next t seconds from which neighbor dynamics are derived without two-hop exchange.

4.4. Routing on Predicted Network State

This technique routes on forecasts of routing-relevant network variables rather than solely on the network as measured. It is predominantly a development of the 2020s in this corpus and spans both vehicular and aerial settings. The designs differ in two respects: first, the quantity predicted and how directly that prediction informs the forwarding decision; second, the overhead the predictor itself introduces.

4.4.1. Traffic- and Service-Mode Prediction Before Routing

The earliest concrete form in this corpus predicts aggregate conditions and routes on the forecast. A software-defined scheme forecasts each road segment's vehicle-arrival rate, converts the forecast into each request's success probability and expected delay, and selects between infrastructure-assisted and multi-hop delivery ahead of time [1]. The prediction is coarse, a flow-level quantity, and is not applied at the level of the individual forwarding action: it informs a mode choice rather than a next hop. It nevertheless defines the coarse-grained end of the mechanism, which later designs couple more closely to the forwarding decision.

4.4.2. Predictive Forwarding and Relay-Level Decision Support

The prediction mechanism then shifts closer to individual hops. In this design, a neural forwarder evaluates candidate next hops based on their predicted end-to-end delivery probabilities, replacing the hand-crafted multi-metric aggregation rules of the authors' previous protocol with a learned delivery predictor [59]. The prediction is now per-candidate and directly gates relay selection. This coupling between prediction and forwarding is tighter than in the flow-level mode decision. It is also more demanding to implement since a forecast must now be produced and relied upon for every candidate at every hop.

4.4.3. Prediction-Driven Routing in FANETs

FANET designs apply prediction in the highly dynamic regime where short-lived topology makes proactive routing especially relevant. In one design, each UAV chooses between hop-by-hop forwarding and store-carry-forward on the basis of forecasts of its future distance to the ground station and packet-arrival conditions [100]. Another design feeds mobility prediction into adaptive beaconing and a double-DQN forwarder, where prediction error governs when beaconing is intensified or relaxed and the prediction-maintained neighbor table supplies the forwarder's state [102]. A third design forecasts each UAV's future packet-arrival rate with a recurrent network and incorporates it into the routing metric [7]. These works illustrate the intended benefit of the mechanism: proactivity in a topology that would otherwise be repaired only after it breaks. They also illustrate its characteristic overhead: an additional predictive component that must be trained or fitted, hosted within the node-controller architecture, and kept accurate on constrained airborne platforms.

4.4.4. Joint Prediction and Sensing-Enriched Forwarding

A joint-prediction design forecasts mobility and traffic together [58]. From the joint forecast it derives link-expiration times and buffer headroom. An entropy-based multi-metric scheme then weights them into a path choice. The design responds directly to the observation that predicting traffic alone is insufficient for a multi-metric decision. In GPS-denied low-altitude airspace another design fuses physically sensed obstacle intensity from onboard mapping with historical signal strength into a recurrent forecast of future link quality, which a deep-Q forwarder then takes as input [31]. Assessment of the prediction mechanism depends on whether its forecasts are validated against ground truth, and that validation is uneven: a minority report direct forecast checks (classification metrics [59], location-prediction error [102], and a sensing front-end evaluation [31]), while the rest infer forecast quality only from downstream routing performance. Figure 8 illustrates the extent of this coupling: prediction is the first stage of the pipeline, and the downstream routing quantities—metrics, elimination, and path choice—are all derived from its output.

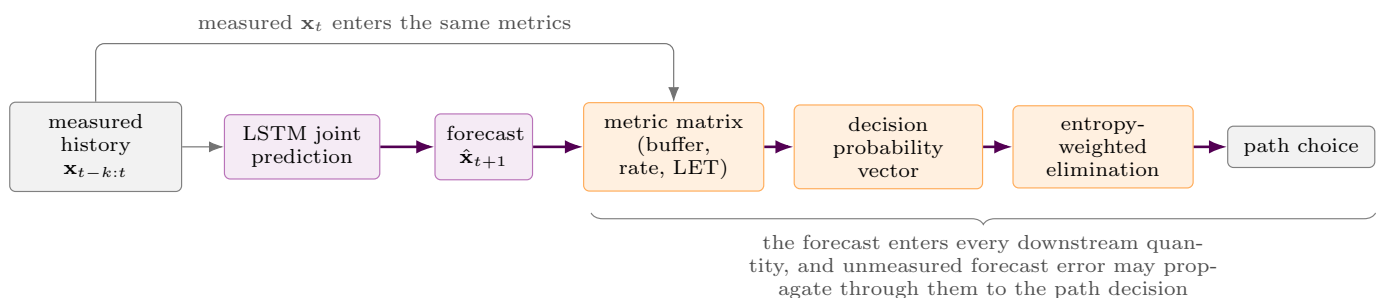


Figure 8. The prediction-to-decision chain of the JPE protocol: measured history and an LSTM joint forecast feed the metric matrix from which a decision probability vector and an entropy-weighted elimination gate the path choice.

4.4.5. Comparison of Prediction-Driven Designs

Two comparison criteria distinguish these designs. The first is what is predicted, which coincides with how granular the forecast is. A segment-level arrival rate [1] informs a coarse mode decision. A per-candidate delivery probability [59] and a per-UAV distance-to-station forecast [100] enter the per-packet decision directly. In [100] the choice is again between delivery modes (forward now or store-carry), but it is remade at each forwarding state rather than fixed in advance at the flow level. A per-UAV arrival rate enters the routing metric [7], and a per-link expiration time contributes to candidate elimination and path selection [58]. A per-link future-quality forecast [31] incorporates physically sensed obstacle intensity as a new input class. Granularity and decision coupling generally rise together: the finer the predicted quantity, the more tightly it can be coupled to the forwarding action and the more often it must be produced, although path-level designs [58] qualify the trend. The second dimension is the predictor and the overhead it introduces. The mechanism is defined by its use of future-state information rather than any particular prediction model. This is evident in the range of predictors used: a feed-forward network for traffic flow [1], recurrent and deep predictors for mobility, arrival, and link quality [7,31,58,102], and least-squares trajectory fitting for distance-to-station [100]. A common trade-off is nevertheless evident: each forecast adds a predictive component with a training or fitting cost and an inference cost at the node or controller hosting it. A further cost is the forecast error itself, which is often not directly observed and which the routing policy then acts on.

4.5. Coordinated and Multi-Agent Learned Routing

The fifth mechanism applies when many nodes act as learners, examining how the resulting policies are shared, coordinated, and scaled. As one of the most architecturally complex and operationally demanding mechanisms in the corpus, its comparison hinges on how designs address two challenges absent in single-agent environments: non-stationarity (arising because each agent's environment includes other learning agents) and generalization, which requires a policy trained on one network scale to operate effectively on another.

4.5.1. Shared-Policy Routing Across Many Nodes

The representative design addresses both challenges through a single shared policy. Under a centralized training with decentralized execution (CTDE) framework, DeepCQ+ trains a unified PPO policy shared across all nodes. By encoding state information into a fixed-size best-neighbor representation, a policy trained on 12-node networks transfers directly to test topologies ranging from five to 50 nodes without modification. Crucially, the learner is deliberately restricted to making broadcast-versus-unicast decisions within a rule-based protocol [18]. Figure 9 shows the architecture: the parameters are shared across all agents, so a single policy is used, and the fixed-size encoding keeps the per-agent observation interface constant in size as the network grows. Algorithm 1 sets out the division of functions between the learned policy and the rule-based protocol. Every step except the final broadcast-versus-unicast draw is inherited from that protocol.

4.5.2. Cooperative Routing in Vehicular and Aerial Networks

Alternative architectures vary in the degree to which they rely on explicit communication protocols rather than leveraging a shared representation. A vehicular scheme forgoes explicit joint-action coordination: it shares one Q-network across all agents but trains it by independent Q-learning for fully decentralized delay minimization with no global information [41]; an aerial scheme applies cooperative Q-learning with a Nash-style update across one-hop agents [43]; and a UAV-swarm scheme casts routing as a cooperative stochastic game solved with MADDPG whose recurrent actor and critic networks exploit

temporal continuity while rewarding signal quality, link-expiration time, and low queue backlog [2]. While these architectures can represent richer inter-agent dependencies, they incur significant coordination and training overhead. Furthermore, unless a scale-agnostic representation is explicitly incorporated, they provide less direct evidence of cross-size generalization than the framework of Section 4.5.1.

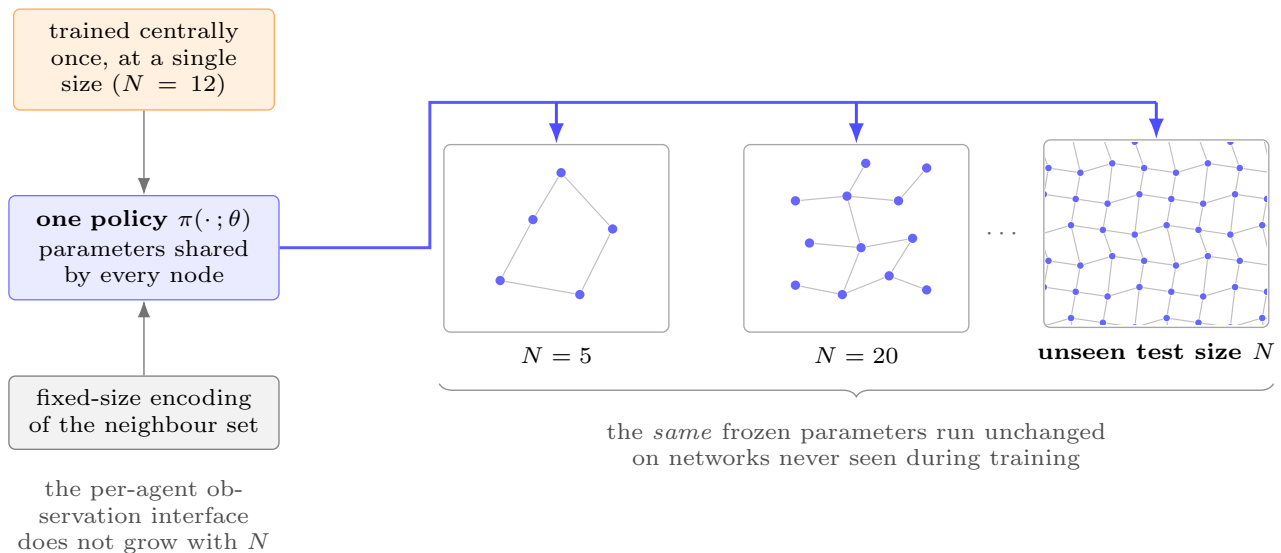


Figure 9. DeepCQ+ size-agnostic execution: one policy with shared parameters, trained centrally at a single network size and executed unchanged on network sizes not seen during training, with reported tests ranging from 5 to 50 nodes.

Algorithm 1 DeepCQ+ per-node forwarding loop: classical CQ+ bookkeeping handles acknowledgements, loops, and duplicates, while the learned policy decides only broadcast versus unicast.

```

1: receive incoming packet at node  $i$ 
2: if packet is an acknowledgement (ACK) then
3:   update the connectivity and quality statistics  $c$  and  $h$ 
4: else
5:   if packet has traversed a loop then
6:     drop packet without returning an ACK; continue
7:   end if
8:   if packet is already queued then
9:     recompute ACK statistics; drop packet but return an ACK; continue
10:  end if
11:  if packet is not a duplicate then
12:    add packet to the queue
13:  end if
14: end if
15: if no ACK is received for a transmission then
16:   leave  $c$  and  $h$  unchanged
17: end if
18: if queue is not empty then
19:   form the policy input  $o_t$  from local and neighbor statistics
20:   broadcast with probability  $\pi_\theta(a=1 | o_t)$ ; otherwise unicast to the CQ+-selected next hop with probability  $\pi_\theta(a=0 | o_t)$ 
21: end if

```

4.5.3. Distributed Learned Coordination Under Dynamic Conditions

Challenging deployment environments illustrate the overhead of agent coordination. For instance, a tactical multi-sink system that co-designs an intelligent jammer alongside a distributed multi-agent DQN routing policy illustrates the impact of interference on reinforcement learning. Packet corruption and lost feedback render the reward estimates required for effective coordination highly unreliable [5]. This system is a particularly demanding test of the mechanism because the jammer degrades not only packet forwarding but also the learning signal, and it shows that the reliability of coordinated learned routing is bounded by the reliability of the feedback channel used to train it. This dependence is most consequential in the adversarial and disconnected settings in which multi-agent routing is otherwise most attractive.

4.5.4. Toward Transferable Forwarding Policies

Finally, an important demonstration of generalization along this dimension stems from relational forwarding policies. By formulating packet forwarding as a fully learned policy over identity-agnostic relational features, this approach captures the underlying structural relationships between nodes rather than memorizing specific node identities [44]. Its continuation demonstrates a frozen policy generalizing from 25 to 100 devices and across mobility models without retraining [45]. Where DeepCQ+ achieves size-agnosticism by engineering the input format, this line achieves transferability by engineering the features, and it provides direct evidence within the corpus that a learned forwarding policy can transfer, in simulation, to network sizes and mobility conditions not used in training.

4.5.5. Comparison of Coordination Designs

These architectures align along a continuum of coordination intensity, with overhead scaling accordingly. At the weakest level of coordination, independent learners share a unified policy or network representation but omit explicit modeling of environment non-stationarity induced by co-adapting agents [41]. CTDE is a recurring compromise. Training may use joint or centrally available information, whereas each actor runs on local information at test time. Coordination is thus learned without requiring global information in deployment, as in the shared-policy [18] and actor–critic swarm [2] designs. The Nash-cooperative router instead performs coordination in a fully distributed manner through one-hop Nash-Q updates [43]. Explicit game-theoretic or actor–critic coordination [2] is among the most expressive designs here and captures interactions the independent learners miss, but it is also among the most training-intensive and is exposed to the reward corruption that adversarial conditions inject [5]. Scalability and generalization, finally, are orthogonal to coordination strength and have to be engineered separately: they come not from stronger coordination but a size-agnostic input encoding [18] or device-agnostic relational features [44,45]. Coordination and deployability are therefore independent design goals. Deployable systems must address both rather than assume that the first delivers the second.

4.6. Cross-Mechanism Synthesis and Comparative Insights

Examined collectively, these five mechanisms define a spectrum of trade-offs rather than a set of independent capabilities. Five central tensions structure the current landscape.

Assisted versus clean-slate control —The clearest tension runs the length of Section 4.1. Scaffold-based designs that retain AODV preserve familiar routing and control procedures together with a bounded inspectable learning role [12,50]. A deep learner can be given primary next-hop authority while those procedures remain in place [16]. A relational learned policy removes the conventional protocol scaffold from the forwarding-control role, pursuing end-to-end optimization and transferability [44]. The corpus remains weighted

toward the compatible end, and this distribution may reflect identifiable engineering considerations (training cost, certification difficulty, and an established fallback), although the corpus does not isolate a single cause. In the reviewed studies, claims of learner autonomy are better supported where generalization was engineered into the design.

Fine-grained versus coarse-grained state—Section 4.2 traced this trade-off stage by stage [28,52,53]. Abstraction and adaptivity have so far been difficult to achieve together: a gain in one has usually come at the cost of the other or greater reliance on infrastructure. The online grouped designs of Section 4.2.4 nonetheless attempt to combine the two.

Narrow versus wide information horizon—The tension examined in Section 4.3 has the most explicit quantitative treatment in the corpus. A two-hop horizon is designed to improve dead-end avoidance [32] but was measured as the option with the highest control overhead among the compared protocols. Later designs pursue part of the same robustness objective with lower communication overhead through analytical reconstruction from one-hop data [4] or geometric filtering of the candidate set [35]. These designs motivate the hypothesis that part of the robustness benefit of wider topology exchange can be recovered by shifting the burden from communication to local computation. The available evaluations, however, do not establish a common point beyond which additional topology exchange yields diminishing returns.

Reactive versus predictive routing—Across Section 4.4, prediction shifts routing from reactive decision-making toward proactive adaptation, and its reach has grown from flow-level traffic [1] through packet-arrival and joint forecasts [7,58] to physically sensed surroundings [31]. However, several designs add a separate predictive component, learned in some cases and analytical in others. The errors of that component are not always represented in the routing objective that consumes its forecasts. The benefit of prediction depends on forecast quality under the target operating conditions. The corpus more often supports that quality through downstream routing gains than direct validation.

Independent versus coordinated policies—In Section 4.5, multi-agent coordination captures interactions that independent learners do not explicitly model [2] and has begun to confront jammed or corrupted feedback [5], but non-stationarity and training cost impose substantial deployment burdens. Coordination and scalability reconcile most plausibly when the policy is engineered to reduce its dependence on the size or identity of the network it runs on.

The conclusion of the synthesis is that no mechanism dominates. Each addresses particular routing limitations while introducing trade-offs in compatibility, adaptivity, overhead, complexity, or scalability. Several systems reduce the combined burden: prediction reconstructed analytically from local data [4], coordination made size-agnostic by construction [18], and autonomy achieved through transferable features [44,45].

Taken together, this comparative analysis of the 37 core papers yields three principal conclusions, all resting on within-study evidence (Level 1 in the terms of Section 5.1). First, regarding maturity, the mechanisms exhibit uneven stages of development. Changes in the learner's role and decision authority (Section 4.1) account for the largest share of the corpus and have been the most repeatedly engineered. The VANET state-representation body of work (Section 4.2) is visibly cumulative, demonstrating multi-generational refinement. The information-horizon mechanism (Section 4.3) is the smallest dimension yet among the most explicit about the cost of information exchange. Prediction mechanisms (Section 4.4) remain weakly standardized, characterized by diverse architectures but few shared validation benchmarks. Multi-agent coordination (Section 4.5) is an expressive yet operationally demanding paradigm, making its strongest case where generalization is explicitly engineered rather than assumed.

Second, regarding deployment settings, the mechanisms exhibit an uneven distribution across network domains. Within this corpus, state representation advances are almost exclusively confined to VANETs, shaped by the structural constraints of urban road topologies. Conversely, the information-horizon trade-off is most explicitly demonstrated in the FANET literature, where rapid three-dimensional mobility drives frequent topology updates and hence heavy control overhead. Prediction mechanisms span both vehicular and aerial domains, particularly where traffic or mobility patterns display exploitable temporal structure. Multi-agent coordination concentrates within general MANET and tactical environments, where routing responsibilities are distributed among multiple autonomous agents. Learner role and decision authority is the only mechanism with double-digit representation in every network family. Consequently, the current evidence base for each mechanism remains closely tied to the network regimes in which it has been evaluated.

Third, and importantly for the subsequent discussion, several of the aforementioned trade-offs cannot be evaluated through design specifications alone. Several core empirical questions demand rigorous validation: whether expanding the information horizon genuinely underperforms relative to analytical reconstruction, whether predictive models maintain accuracy under live deployment, whether coordinated policies generalize beyond their training distribution, and whether clean-slate policies require safety fallbacks. Section 5 addresses the resulting gap between these mechanistic claims and their empirical demonstration.

4.7. Suggested Design Patterns: Classical Functions Replaced or Augmented by Learning

The comparison above may also be read as a preliminary set of design patterns that are intended for an engineer who must decide where in an existing protocol a learned module is most likely to be beneficial. Table 6 lists the classical functions that studies in the corpus have replaced or augmented with learning, the weakness of the classical rule that motivated each substitution, the form of the learned substitute, and the cost that the substitution introduces. Two regularities are observed. First, the substitutions concentrate on functions in which a classical protocol commits to a fixed rule whose assumptions may not hold across changing operating conditions: a route-discovery and repair cycle slower than the link breakages it must repair, a greedy geographic choice that encounters a void, a fixed beacon period under neighborhoods that change at varying rates, and a per-encounter forwarding heuristic whose parameters do not track the deployment. Second, the substitutions that have been evaluated repeatedly generally leave the surrounding protocol intact and replace or augment a single decision. This observation is consistent with the predominance of the assisted end of Section 4.1 in the corpus, and it suggests that the hybrid arrangements discussed in Section 6.4 may represent one of the more deployment-oriented patterns in the corpus rather than a compromise: the learned module is confined to the function it has been shown to improve, while the classical protocol continues to provide discovery, signaling, and a fallback. The last column of Table 6 records the kind of evidence that is available for each pattern, which is simulation in all cases but one. Whether these substitutions retain their benefit beyond the studies that introduced them remains to be established; this question is examined in Section 5.

Table 6. Design patterns in the corpus: classical routing functions replaced or augmented by learning, the weakness each substitution addresses, the cost it introduces, and the evidence the corpus offers for it.

Classical Function or Representation	Weakness Addressed	Learned Substitute in the Corpus	Cost Introduced	Evidence in the Corpus
Route ranking and maintenance within AODV	hop-count-only selection; re-pair only after a link breaks	learned ranking and switching among AODV routes, with discovery retained [12,50]; DQN next-hop selection on cross-layer state within the AODV scaffold [16]	state collection and training; re-tuning when the traffic pattern changes	simulation; vehicular hardware testbeds [50,65]
Greedy geographic forwarding (GPSR): the greedy choice	distance-only choice with no notion of link quality or the void ahead	Q-values attached to geographic transitions and learned from delivery feedback [34]	Q-values carried in beacons, which are also an attack surface (Section 6.4)	simulation
Greedy geographic forwarding (GPSR): stale neighbor information	neighbor positions age between beacons	link-stability prediction on the greedy scaffold [17]	predictor training; forecast error unmeasured	simulation
Geographic candidate set	candidate set and state space grow with the observation horizon	geometric filtering of the candidate set before learning [35]	the filter is hand-designed	simulation
Fixed-period beaconing	overhead at high density, staleness at high mobility	beacon interval adapted to predicted mobility [102]	prediction cost; assumptions about the mobility process	simulation
Two-hop topology exchange	control overhead that grows with density	partial two-hop information selected for the learner [42]; analytical reconstruction of wider topology information from one-hop data [4]	information withheld from the decision; assumptions about neighbor dynamics	simulation
Multi-point relay selection (OLSR)	relay heuristic fixed at design time	learned multicast forwarding policy in place of MPR selection [33]	training and policy distribution	simulation
Broadcast-versus-unicast choice (CQ+)	hand-tuned thresholds	one shared policy deciding broadcast or unicast within the CQ+ bookkeeping [18]	centralized training; a size-agnostic encoding is required	simulation
Search budget (ticket-based probing)	a fixed budget under varying load	learned allocation of the probing budget within a pre-specified maximum [15]	small learning cost; gain bounded by one parameter	simulation
Encounter-based forwarding (PProPHET)	hand-crafted delivery predictability	supervised delivery estimate entering the forwarding rule, with the classical metric retained as a feature [26]	training data from encounter histories	simulation
Vehicle-level routing state	state space grows and turns over with traffic	grid-cell abstraction with an offline Q-table [28]	offline table blind to live load	simulation on traffic traces
Vehicle-level routing state	as above	intersection- and RSU-group-level abstraction kept with roadside infrastructure [52,53]	dependence on infrastructure; tables to be kept current	simulation

5. Evaluation, Benchmarking, and Reproducibility

We audit the evaluation methodology of the same 37 core papers analyzed in Section 4 across ten dimensions: scenario, simulator, mobility source, channel and terrain fidelity, learning setup, baselines, metrics, statistical practice, reproducibility artifacts, and threats to comparability. Table 7 records the per-paper core of that audit (scenario, simulator, baselines and ablation, cost reporting, repeated-run or uncertainty reporting, and hardware validation), while the remaining dimensions, including reproducibility artifacts, are audited in the text of this section. The audit therefore characterizes reporting practices within this analytical core rather than estimating their prevalence across all 92 papers. The finding is that reported gains in the audited core set are not, at present, directly comparable across papers. This limitation is not unique to learned routing; the ad hoc networking community has documented the fragility of simulation-based protocol comparison for

roughly two decades [106,107]. Learning, however, adds further sources of variation that can complicate comparison: training conditions, random seeds, and hyperparameters.

Table 7. Evaluation profile of the 37 Section 4 core papers (membership rule in Section 3.1). In the baselines column, the ablation entry is y or n for whether the paper ablates its own mechanism; a parenthesized learned baseline is the authors' own prior scheme rather than an independent one, and in-paper reduced variants are counted as ablations rather than baselines. The cost and validation columns use a letter when an item is reported and a hyphen when it is not reported: o = control or routing overhead, e = energy, l = learning-specific cost; s = repeated-run or uncertainty-reporting practice, and t = hardware testbed.

Paper	Scenario	Simulator	Baselines: Classical/Learned/Ablation	Cost: Overhead/Energy/Learning	Validation: Statistics/Testbed
[15]	MANET	custom	Original ticket-based scheme, Flooding-based search/-/n	o-l	--
[13]	MANET	ns-2	AODV, DSR/-/n	o--	s-
[12]	VANET	ns-2	AODV, AODV-HPDF, NRD/-/n	o--	s-
[50]	VANET	ns-2 + testbed	AODV, AODV-L/(QLAODV)/y	--l	st
[65]	VANET	ns-2 + testbed	AODV-ETX, HLAR, minstrel autorate/-/n	o-l	st
[64]	VANET	QualNet	AODV, GPSR/QLAODV, QROUTING/n	o--	s-
[57]	MANET	custom	GPSR/-/n	--l	s-
[77]	MANET	OPNET	DSR, ARA, E-Ant-DSR/ANN, GA, ARA, E_ANT, FMRM/n	---	--
[16]	MANET	custom	CR-AODV, CRD/-/n	oel	--
[26]	DTN/ OppNet	The ONE	PRoPHET+/-/y	o--	s-
[83]	MANET	WSNet	GPSR/QGeo/n	---	--
[30]	VANET	custom	GPSR, HarpiaGrid/-/y	o--	--
[28]	VANET	custom	GPSR, HarpiaGrid, CBS_like/-/y	o--	--
[29]	VANET	QualNet	GPSR, LAR, GyTAR, iCar-II, RTAR/-/n	---	s-
[52]	VANET	OMNeT++	GPSR, RAVP/QGrid/n	o-l	--
[66]	VANET	ns-2	GyTAR, SRPMT/ARPRL/y	o-l	s-
[53]	VANET	OMNeT++	AODV/IV2XQ/y	o-l	s-
[97]	VANET	ns-2	GPSR/IV2XQ, QGrid/n	o--	--
[34]	FANET	ns-3	GPSR/QGrid/n	o--	s-
[68]	FANET	WSNet	-/QGeo/y	-e-	s-
[32]	FANET	MATLAB	GPSR/QGeo/n	oel	--
[4]	FANET	MATLAB	GPSR-EE-Hello, MPVR/QTAR/n	oel	s-
[35]	FANET	ns-2	-(QFAN), QTAR, QGeo/n	oe-	--
[1]	VANET	analytical (numerical)	-/-/y	---	s-
[59]	VANET	OMNeT++	GPSR, 3MRP/-/y	o-l	s-
[100]	FANET	custom	GPMOR, LAROD, AFP/-/n	---	--
[102]	FANET	custom	GPSR, PQR/QGeo/n	o--	--
[7]	FANET	custom	SPA, TARU/-/y	---	--
[58]	FANET	custom	Optimal, SPA/(PAP)/y	---	s-
[31]	FANET	custom	FX-AODV, TORP/QTAR/y	-el	-t
[43]	FANET	ns3-gym	GPSR/-/y	oe-	--
[2]	FANET	ns3-gym	-/DGATR, DeepCQ+, standard MADDPG/y	---	--
[41]	VANET	custom	AODV, DAODV, GPSR/-/n	---	s-
[18]	MANET	custom	-(CQ+, SRR)/n	o-l	--
[44]	static ad hoc	custom	Shortest path, Backpressure/-/n	--l	s-
[5]	tactical WSN	custom	SP/DQN-routing, QELAR/n	-el	--
[45]	MANET	custom	Oracle, Direct Transmission, Utility-Based, Seek-and-Focus/-/y	o-l	s-

5.1. The Audit: Scope, Dimensions, and Evidence Base

Table 7 provides the coded evidence base, aggregated by column, for the findings that follow. Under this coding, no paper reports overhead, energy, and learning cost together with both repeated-run or uncertainty reporting and hardware validation, and a common profile comprises delivery and delay, often with control or routing overhead, reported against a classical baseline in a paper-specific scenario, typically without learning-specific cost or a hardware testbed.

Throughout the audit we distinguish three levels of evidence because the single word “improvement” can refer to any of them. Level 1, within-study improvement: a scheme outperforms the baselines its own authors chose in the scenario and simulator its authors configured; this is the evidence almost every core paper provides. Level 2, cross-study comparability: two studies hold enough of the scenario, the baseline set, and the metric definitions in common for their reported numbers to be placed on one scale; the audit finds this condition met only within a few lineages that reuse a predecessor’s setup. Level 3, evidence of general superiority: a mechanism is shown to help across independently designed studies and scenarios; no mechanism in the audited core has yet reached this level. The mechanism comparison of Section 4.6 is based primarily on Level 1 evidence and is not to be read as evidence of general superiority; the benchmark gap of Section 5.6 is the limited availability of Level 2 evidence, and the open question of Section 6 is Level 3.

5.2. Scenario Realism and Simulation Credibility

Claims such as “a 20–34% improvement in packet delivery ratio” are interpretable only when contextualized by a specific baseline, scenario, and channel model. Across the 37 core papers, however, no standardized benchmark holds all three variables constant. The choice of simulation platform alone fragments comparison within the audited core set: 15 of the 37 studies rely on custom simulators, eight of which omit tooling details entirely, while the remainder disclose only the programming language or framework. Furthermore, one vehicular scheme is evaluated using an analytical numerical model rather than packet-level simulation [1]. Among the named tools, ns-2 is the most common (seven). By contrast, native ns-3 simulations appear in only three studies (one directly and two through the ns3-gym bridge [2,43]). The remaining studies are distributed across OMNeT++/Veins (three), WSNNet, QualNet, and MATLAB (two each), and The ONE and OPNET (one each). The mobility and traffic generators fragment the literature further, ranging from random waypoint in synthetic areas through Manhattan grids and real taxi traces [28,30] to bespoke three-dimensional UAV waypoint models. Node counts, speeds, transmission ranges, and channel models are chosen per paper. Consequently, most reported gains are conditioned on each paper’s chosen baseline and scenario; they do not, by themselves, establish how a scheme compares with the wider literature. These issues resemble the limitations identified two decades ago in the MANET-simulation-credibility literature—inconsistent scenarios, unstated assumptions, and irreproducible setups [106]—although the present corpus does not reproduce every problem in identical form. The resulting comparability limitation takes a different form in each of the three settings.

5.2.1. MANET Scenarios

Evaluations in the general MANET literature tend to be among the most abstract in the audited core, with few studies grounded in physical geographic maps or empirical mobility traces. Many rely on compact topologies whose scale supports proof-of-concept feasibility demonstrations rather than large-scale stress testing, such as a 36-node conference-room scenario contained within a 15 m × 15 m area [15]. Other studies evaluate 20-to-40-node networks in areas of 30-to-50 m [77]. Random waypoint in synthetic or weakly structured

simulation areas remains common in the MANET subset despite longstanding evidence that its transient behavior and initialization can materially change measured routing metrics [108]. Several MANET studies omit scenario details that are necessary for reproducible comparison, including density, mobility, traffic, and seed configuration.

5.2.2. VANET Scenarios

Vehicular evaluations achieve greater physical realism when they integrate microscopic traffic models or empirical mobility traces into the network simulation. This grounding matters because routing outcomes in VANETs are shaped not only by node motion but also road topology, intersection dynamics, vehicle density, and RSU placement. Some studies achieve this level of realism by training on real taxi GPS traces [28,30], but realism across the corpus is uneven: one intersection scheme is evaluated with a single common vehicle-speed setting on a single synthetic grid [97], another uses Barcelona for training and its first most detailed evaluation, treating transfer to Berlin and Rome as an explicit generalization test [59], and one vehicular protocol uses IEEE 802.11a rather than the vehicular-specific 802.11p because its simulator lacked an 802.11p module [64]. The audited VANET studies differ substantially in traffic–network coupling, map realism, and uncertainty reporting.

5.2.3. FANET Scenarios

Aerial evaluations show particularly limited scenario standardization among the audited core papers, and their evaluation practices reflect this. Several core aerial studies run in MATLAB with homogeneous synthetic three-dimensional mobility models (3D Gauss–Markov in [32] and 3D random waypoint in [4]) and free-space propagation; no real-UAV mobility trace or dedicated air-to-air channel model was identified in the audited core set, and one prediction study leaves its simulation platform unnamed [7]. Because FANET routing behavior is sensitive to three-dimensional mobility, channel variation, energy constraints, and—where relevant—mission traffic, results that omit flight patterns, node densities, and mission loads are difficult to interpret or transfer across studies, and the lack of a shared FANET routing benchmark means that many papers evaluate within study-specific airspaces and parameter regimes.

5.3. Metrics and Cost Reporting

A comparable evaluation would report along four metric groups: delivery and efficiency (delivery ratio, packet loss, throughput, and hop count); latency and stability (end-to-end delay, jitter, route lifetime, and recovery time); communication and resource cost (normalized routing load, beacon overhead, control-packet ratio, and energy); and learning-specific cost, hereafter learning cost (convergence episodes, training time, inference latency, model size, and retraining rate). Across the 37 papers, delivery ratio and end-to-end delay are the most consistently reported metrics, each appearing in 35 of the 37. As recorded in the cost column of Table 7, control or routing overhead appears in 22, energy in only eight, and any learning-specific cost (convergence, training time, inference latency, model size, or retraining) in only 16. The imbalance limits what the reported gains can establish because the unreported costs are those incurred by the Section 4 mechanisms. A two-hop information horizon should be assessed against its overhead [4,32]; a predictor should be judged by whether its gain offsets its training and inference burden, a burden only partly quantified in the cited studies [31,58]; and a multi-agent policy should be judged against its coordination cost [5,18]. A paper that omits overhead, energy, and learning cost therefore leaves the system-level cost of its delivery-ratio gain only partially characterized.

QGrid [28] provides an illustrative example. It reports delivery ratio, hop count, delay, forwarding count, and throughput for several learned hierarchical schemes against a bus-aided baseline, a relatively broad network-metric evaluation within the corpus, but it

does not quantify the associated control overhead, energy consumption, or learning cost, and its evaluation is confined to one trace-driven Shanghai dataset and a study-specific baseline set.

5.4. Baselines, Ablations, and Comparison Fairness

5.4.1. Baseline Strength

A mechanistic claim is calibrated by the benchmark against which it demonstrates superiority; however, most baselines in the audited core set evaluate progress within a specific lineage rather than across the broader field. Only 16 of the 37 papers benchmark against an independent learned scheme; the remaining papers rely on classical baselines, compare with prior schemes in the same design lineage, or lack an independent learned comparator. Such comparisons can help to isolate an incremental change within a lineage as, when the DeepCQ+ studies quantify their gain over the CQ+/SRR design, they build on [18], the joint-prediction FANET scheme over its earlier PAP predictor [58], and a fuzzy-constraint VANET scheme over its earlier QLAODV [50]. The comparative position of most designs relative to independently developed learned schemes remains weakly established in the audited core set. Five evaluate without a classical protocol in their reported numerical comparison sets [1,2,18,35,68]. A more demanding comparison would use three complementary baselines where available: a well-established and competitive classical protocol, a recent learned scheme, and a mechanism-matched one. A prediction-driven paper should ideally include both a prediction-free counterpart and another predictor and a multi-agent paper both a single-agent learner and a non-coordinated comparator [2,32,58]. Table 7 suggests that only a minority of the core set meets this stricter bar.

5.4.2. Ablation Requirements

Isolating performance gains that are attributable to a specific mechanism requires systematic ablation studies. While 15 of the 37 papers incorporate some form of ablation, several evaluate performance solely against simplified or reduced internal variants of their own proposed architecture. The informative ablations are mechanism-specific: comparing with a prediction-free variant or counterpart [7,58], examining reward construction and testing a shared policy across network sizes after training at one size [18], and replacing a learned component with the rule-only system it augments; the horizon question, by contrast, has so far been examined mainly through cross-protocol comparisons—one-hop TARRAQ against two-hop QTAR [4,32]—rather than by ablation within a single design. In the absence of such ablation studies—as is the case in the majority of papers—reported performance improvements cannot be definitively attributed to the proposed mechanism. Confounding factors, such as differences in baseline tuning or scenario selection, make it difficult to determine whether observed gains stem from the learning algorithm itself or merely from surrounding engineering choices.

What an informative ablation should contain follows from the five dimensions themselves because each dimension isolates a distinct design choice assigned to the learner. An ablation set that isolates a mechanism therefore includes (i) a mechanism-off control in which the learned component is replaced by the classical rule it displaces; (ii) a state-granularity control that varies the unit of state with a capacity-matched learner; (iii) an observation-horizon control that narrows or widens the observation boundary within the same design; (iv) a prediction control with measured-present, proposed-forecast, perturbed-forecast, confidence-gated, and, where traces permit, diagnostic oracle-future variants; and (v) coordination controls that vary one factor at a time. Two further controls apply to any learned router: a reward-component ablation that preserves the reward scale and computation- and tuning-budget-matched baselines. Reporting each variant with paired

seeds, confidence intervals, and effect sizes enables quantitative attribution of the observed effect; Table 8 specifies these controls as reporting items.

Table 8. A minimum reporting protocol for future learning-based routing studies.

Aspect	Variables to Report
Scenario	node count, density, speed, mobility model, road or airspace map, traffic model and load, duration
Wireless model	PHY/MAC standard, channel and propagation model, bandwidth, range, interference, transmission power, obstacle assumption
Routing baseline	a well-established classical baseline, a recent independent learned baseline, and a mechanism-matched baseline, with any missing category explicitly justified
Ablation: general	a mechanism-off control (the learned component replaced by the rule it displaces); scale-preserving reward-component ablation
Ablation: state and information controls	state-granularity and observation-horizon controls within the same design, with capacity-matched learners
Ablation: prediction controls	measured-present, perturbed-forecast, confidence-gated, and diagnostic oracle-future variants for predictive schemes
Ablation: coordination controls	single-factor controls (independent learners, shared policy, centralized training, communication) for coordinated schemes
Ablation: budget and statistical matching	computation- and tuning-budget-matched baselines; paired seeds, confidence intervals, and effect sizes across variants
Learning setup	state, action, reward, model architecture or table representation, hyperparameters, exploration or update policy
Training	episodes or update budget, training scenario, training-progress or convergence evidence, computation resources, retraining or online-update policy
Testing	unseen seeds and the full test scenario; for generalization claims, unseen density, mobility, traffic, topology, or channel conditions
Statistics	warm-up where applicable, repeat count and seed policy, confidence interval or variance, significance test where appropriate
Cost	control or routing overhead including beaconing and coordination traffic, inference latency, model or table size, energy, and training time
Reproducibility	source code, seeds, traces, simulator and dependency configuration, trained models or tables, baseline implementation details

Mechanism-specific items apply only where relevant.

5.4.3. Training–Test Separation and Generalization

Machine learning introduces a distinct evaluation pitfall: a policy may overfit to its training environment and subsequently undergo evaluation on that same distribution. This risk is present in the corpus, illustrated by offline Q-tables trained on historical traffic from a single fixed topology and evaluated strictly within that same scenario [52], leaving open the question of how well these reported gains would transfer to unseen environments. Table 8 lists the training-and-testing details that would allow a reader to assess that question. Generalization evidence in the corpus comes from papers that report explicit distribution shifts: a size-agnostic policy trained on 12 nodes and tested from five to 50 [18], a relational policy trained on 25 devices, updated no further during testing and evaluated on up to 100 [45], and a vehicular predictor trained on one city map and tested on two others [59]. These studies provide concrete examples of cross-distribution testing.

5.5. Reproducibility and Benchmark Infrastructure

Ensuring reproducibility in learned routing frameworks demands documentation beyond what classical protocols require. Because algorithmic behavior is governed by the complete learning paradigm (including state and reward formulation, network architecture, optimization and exploration schemes, random seeds, and hyperparameters), reproducibility requires these training parameters to be specified alongside the standard simulation configuration. A decision-loop pseudocode such as Algorithm 1 exposes the operational flow but not this learning configuration: it specifies the observation and action interfaces while leaving reward, architecture, hyperparameters, and training setup to surrounding text. Against that requirement the audited core does not consistently meet this standard:

only 19 of the 37 report at least one basic statistical practice (repeated runs, seeds, error bars, or confidence intervals), and only three contain any hardware-based validation. These are two 10-vehicle hardware testbeds from a single research group [50,65] and one UAV experiment that validates the sensing and link-prediction front-end while forwarding and routing remain simulated [31]; the remaining core studies report no hardware validation. Some building blocks for a shared infrastructure already exist within the corpus: two of the core papers run on the ns3-gym bridge between ns-3 and a reinforcement-learning environment [2,43]. Neither paper, however, reports releasing its own protocol implementation. What the audited core set lacks is consistent adoption of the reporting and reproducibility practices specified in Table 8.

5.6. A Minimum Reporting Protocol for Learning-Based Routing Studies

The methodological gaps identified in Table 7 motivate a minimum reporting protocol for future learned-routing studies to enhance comparability and reproducibility (Table 8). This protocol is intended as a scope-aware reporting checklist rather than a universal benchmark: it sets a floor of reported variables that allows two independent studies to be positioned on a common basis, mechanism-specific items apply only where relevant, and an omitted item should be explicitly justified.

Three core conclusions summarize this audit. First, the evidence of the audited core papers cannot yet resolve the five mechanism trade-offs of Section 4.6: a “superior” scheme is typically shown to outperform a specific baseline within a narrow scenario (Level 1 evidence), and comparable setups across studies (Level 2) remain the exception. Second, the methodological omissions also show a consistent pattern: outcome metrics are reported much more frequently than cost metrics. Finally, until future research adopts a minimum reporting protocol such as that outlined in Table 8, the audited results are best interpreted as study-specific evidence under particular evaluation settings rather than a basis for cross-paper performance rankings, while the taxonomy and analysis in Sections 3 and 4 serve as a map of the design space and its underlying trade-offs rather than a definitive hierarchy of mechanism performance.

6. Open Challenges and Future Directions

The five trade-offs of Section 4.6 and the evidence gaps of Section 5 together define the agenda below: generalization (Section 6.1) is a cross-cutting prerequisite; information budgeting (Section 6.2), scalable coordination (Section 6.3), and trustworthy deployment (Section 6.4) are the unresolved theoretical cores; and reproducible evaluation (Section 6.5) is the condition on which the others depend.

6.1. Generalization Beyond Training Scenarios

An important limitation of learned routing protocols is that the audited policies optimized for a specific data distribution rarely demonstrate transferability to unseen distributions. Node density, mobility model, traffic load, and topology are often held fixed or varied only narrowly between training and test, and where training and test are not held apart the reported gain may be optimistic, as the training–test overlap identified in Section 5 illustrates. Whether performance remains robust under substantial distribution shifts is unknown for much of the audited core set.

Several contributing factors can be identified, and they are not peculiar to routing. A learned router may face non-stationarity at two distinct levels: the topology and traffic change under it, and, under multi-agent learning, each learner’s environment also contains the other learners’ changing policies, so the fixed-point assumptions of single-agent learning no longer hold [109]. In decentralized designs its observation is often partial because the

information horizon of Section 4.3 is deliberately limited, so states that differ in their consequences may appear identical to the learner. Its reward can be sparse and delayed when it is tied to end-to-end delivery, whereas per-hop proxies, such as the per-action reward of DeepCQ+ [18], provide denser but more myopic feedback, and either choice shapes what the policy can generalize about. Finally, the training scenario is usually a single density, a mobility model, and a map, so the policy absorbs the regularities of that scenario, an offline table trained on one city’s traffic being the clearest case [52]. These factors motivate corresponding research directions, and the designs that generalize in the corpus often make representation an explicit design object through size-agnostic or relational encodings [18,45]. Outside the corpus, recent forwarding studies have begun to test generalization directly across graph sizes and unseen mobility scenarios through routing policies trained on a single graph [110] and continual adaptation across diverse mobile wireless environments [111].

Exceptions within the surveyed corpus demonstrate that generalization can be explicitly engineered rather than merely assumed. For instance, the size-agnostic and relational architectures discussed in Sections 3.6 and 4.5 deploy frozen policies across varying network sizes [18,45], with relational extensions further transferring across diverse mobility models. Additionally, device-agnostic condition vectors dynamically select among pooled policies as network regimes shift [14]. In these examples, the representation—invariance to network size and node identity—is an important contributor to generalization rather than training duration alone. The directions that follow from this lesson are encodings designed for invariance to size and identity, combined with transfer, meta-, and domain-randomized training, with continual or federated adaptation completing the program (Table 9). Such encodings include permutation-invariant representations of variable-size neighbor sets [112] alongside the graph-structured topology encodings the corpus has already begun to adopt [3]. The training machinery is established in machine learning and robotics but has not yet been imported into the learned routing of this corpus [113,114]. The goal is systematic evidence that a policy trained once remains reliable when the network moves outside its training distribution. Whether a design meets that goal can only be assessed by the out-of-distribution testing that the benchmark of Section 6.5 must make the default.

Table 9. Open challenges in learning-based routing: evidence from the full 92-paper classification and the 37-paper analytical core and future directions. The horizons represent the authors’ qualitative prioritization rather than predicted completion times.

Open Challenge	Horizon	What the Reviewed Evidence Shows	Future Directions	Depends on
Generalization beyond training scenarios	medium-term goal	Most audited policies are trained and tested on one distribution (e.g., [52]); size-agnostic and relational designs already generalize across scale [18,45], and a pooled selector adapts across regimes [14]	Size- and identity-invariant encodings; transfer, meta-, and domain-randomized training; continual and federated adaptation	a comparable benchmark (reproducible evaluation)
The information budget: observation cost and forecast trust	near-term bottleneck	The horizon widened once, and later designs reconstructed or filtered it to limit overhead [4,32,35]; prediction moved toward the forwarding decision [1,31,58], with direct forecast validation rare in the audited core set; control or routing overhead is reported in only 22/37 papers and learning cost in 16/37	Adaptive information horizon [102]; analytical reconstruction [4]; geometric filtering [35]; an explicit observation–computation–energy budget; uncertainty-aware prediction with confidence-gated fallback [14]; forecast-error monitoring analogous to the policy anomaly monitor of [63]; predictor–policy joint training, a step beyond the separately trained forecasts of [58]	overhead and learning-cost reporting

Table 9. Cont.

Open Challenge	Horizon	What the Reviewed Evidence Shows	Future Directions	Depends on
Scalable and certifiable multi-agent coordination	medium-term goal (scalability); long-term goal (certifiable robust coordination)	Multi-agent routing is expressive but constrained by non-stationarity, coordination cost, and corrupted feedback [5]	Communication-efficient MARL; learning stable under corrupted reward and feedback; certifiable or bounded policy behavior	generalization evidence; formal safety specifications
Trustworthy deployment: governance, monitoring, and robustness	long-term goal	Learned and rule-based policies co-exist [14,18] but their arbitration is hand-crafted or absent; malicious-node defenses exist [20]; the learned machinery itself is not systematically evaluated as a deliberate attack target	Principled learned/rule selection [14]; runtime monitoring with revocation triggered by drift or manipulation [63]; fallbacks with provable safety; poisoning defenses and adversarial training; explicit threat-model reporting	reproducible evaluation and generalization evidence
Reproducible and generalizable evaluation	near-term bottleneck	15/37 use a custom or unnamed simulator; 16/37 compare to an independent learned scheme; 2/37 validate routing on a hardware testbed (a third validates only its sensing front-end [31])	An open benchmark implementing the minimum protocol (Table 8), building on the ns3-gym bridge [2,43]; hardware-in-the-loop and digital-twin evaluation as the rung between simulation and field trials	none: it enables the comparisons the others require

6.2. The Information Budget: Observation Cost and Forecast Trust

The information horizon is the dimension in which the corpus has most directly confronted a trade-off, and it remains open precisely because the trade-off has not been settled on comparable ground. The horizon widened once and then retreated under measured overhead, with computation substituted for communication (Sections 3.4 and 4.3). The lineage's one deep-learning member has since reopened the two-hop design on deliberately partial two-hop information [42], evidence that the appropriate horizon remains an open design question. Yet Section 5 showed that control or routing overhead is reported in only 22 of the 37 core papers, so, for much of the audited core set, the very cost that decides this trade-off is left unreported.

The unresolved question is quantitative: how much observation is worth its overhead, under what mobility, and at what network scale. Adaptive beaconing [102], analytical reconstruction [4], and geometric filtering [35] are starting points to generalize rather than settled designs, and the state abstraction of Section 4.2 belongs to the same account since state compression is another means of reducing the cost of maintaining network awareness [28,53]. The budget, moreover, is not spent in packets alone: inference latency, model size, and on-node training cost bound what an embedded vehicular or energy-constrained UAV platform can support, and learning-specific cost of any kind is reported in only 16 of the 37 core papers. Priority-aware computation offloading, in which a deep reinforcement learner decides which tasks a resource-constrained node executes locally and which it defers, is the adjacent problem in which this budget has already been treated as a first-class objective [115]. The objective is a router that allocates observation, computation, and energy under an explicit budget, not one that fixes its horizon a priori. On the discipline side, this requires that overhead and energy be reported as first-class metrics so that the marginal value of an extra hop of information can be measured rather than assumed. The information budget entails a second trade-off: in several architectures, withheld observations are partially replaced by predictive forecasts (Section 4.4). This approach effectively defers the cost of information, paying for it through computational complexity and predictive error risk. However, the existing literature evaluates this forecasting risk even less consistently than it quantifies control overhead.

Section 4 traced prediction moving ever closer to the forwarding decision: from aggregate traffic [1] through per-link lifetimes [58] to physically sensed link quality [31]. The paper-by-paper reading of Section 4.4 found that validation of these forecasts is un-

even, with only a minority reporting direct checks against ground truth. The result is a largely unguarded failure mode: a wrong forecast is rarely represented in the routing objective, although some designs apply consistency-based elimination to the predicted quantities [58]. Most current designs feed the predictor's output into the decision with no explicit confidence estimate and no systematic recourse when it is wrong.

Making prediction-driven routing dependable requires providing the router with an estimate of forecast reliability. Four directions can be identified, largely unexplored in this corpus. First is uncertainty-aware prediction: forecasting a distribution or an uncertainty estimate rather than a point estimate so that the routing decision can weight the forecast by its reliability. Approximate Bayesian methods supply a model-based uncertainty signal [116], while conformal prediction offers finite-sample coverage guarantees under exchangeability assumptions [117]. Second is confidence-gated routing: falling back to a decision on the measured present when the forecast's confidence is low, a prediction-aware analogue of the rule-based fallback already used for out-of-distribution conditions in [14]. Third is online monitoring of forecast error so that a predictor whose accuracy has drifted under deployment is detected and down-weighted in the spirit of the anomaly monitor built for learned policies elsewhere in the corpus [63]. A fourth direction is end-to-end joint training of the predictor and the routing policy, removing the interface that the first three assume so that the forecast is optimized for the decision it serves rather than for its own accuracy. This is the step beyond the joint-prediction line the corpus has already begun, whose joint forecasts are still trained apart from the decision that consumes them [58]. Until the accuracy of a forecast is reported, bounded, and acted upon, the proactivity that prediction enables remains contingent on an assumption the surveyed studies rarely test. Figure 10 shows one form of the interface this challenge targets: sensed obstacle context and historical signal strength feed an LSTM forecast that the forwarding decision consumes directly, the natural attachment point for the confidence estimates and fallback paths this challenge calls for.

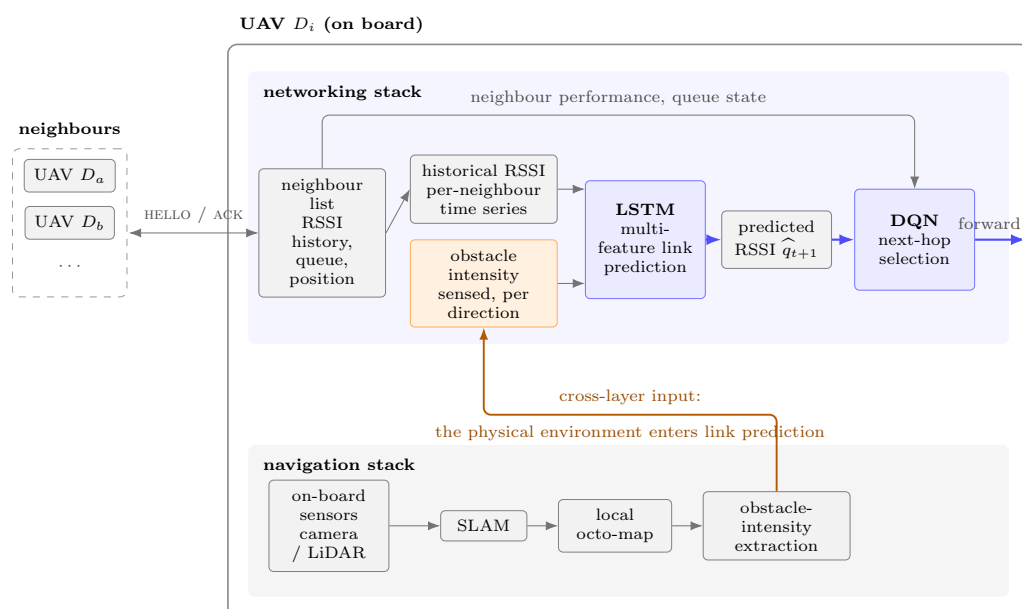


Figure 10. The PSIF pipeline: obstacle-intensity sensing and historical RSSI feed an LSTM link-quality forecast consumed by a DQN forwarding decision. The confidence estimate and fallback discussed in Section 6.2 would attach at the forecast-to-decision interface.

6.3. Scalable Multi-Agent Coordination

Section 4.6 identified multi-agent coordination as an expressive yet operationally demanding mechanism in the corpus. Distinct from general learned routing protocols, multi-agent systems incur unique coordination overheads: environment non-stationarity arises because each agent's effective transition and reward dynamics change as other agents update their policies, training complexity can increase sharply as the effective joint state–action space grows, and, where explicit coordination signaling is required, it consumes the limited bandwidth the protocol seeks to preserve. Furthermore, stress testing within the corpus highlights an important vulnerability of this mechanism: adversarial or degraded feedback corrupts the underlying learning signals that are essential for maintaining multi-agent coordination.

These challenges define specific future research directions: communication-efficient multi-agent learning with coordination costs that remain bounded or scale sublinearly with network size; learning frameworks that are robust to corrupted reward and feedback signals [5]; and policies whose performance can be formally certified or bounded rather than merely empirically measured. The latter goal aligns with shielded reinforcement learning in safety-critical control domains [118], although applying it to routing requires explicit safety specifications and tractable abstractions of network environments, prerequisites that are not yet well developed in the reviewed routing literature. Whether a deployed multi-agent policy can be trusted is a governance question shared with hybrid systems, and it is taken up in Section 6.4.

6.4. Trustworthy Deployment: Governance, Monitoring, and Adversarial Robustness

A recurring finding of this survey is that the core challenge is not in choosing between learned and rule-based systems but governing their coexistence. Sections 3 and 4 trace this design spectrum from learned decisions embedded within classical protocols [18] to fully learned forwarding schemes [13,45] to policy pools governed by rule-based fallbacks [14]. Figure 11 illustrates such an architecture, where learned policies and a classical fallback coexist by construction, orchestrated by a device-agnostic selector. An unresolved gap lies not in architectural coexistence itself but a rigorous theory of arbitration: defining precise conditions for transitioning between learned and rule-based execution, revoking untrustworthy policies, and formally guaranteeing fallback safety.

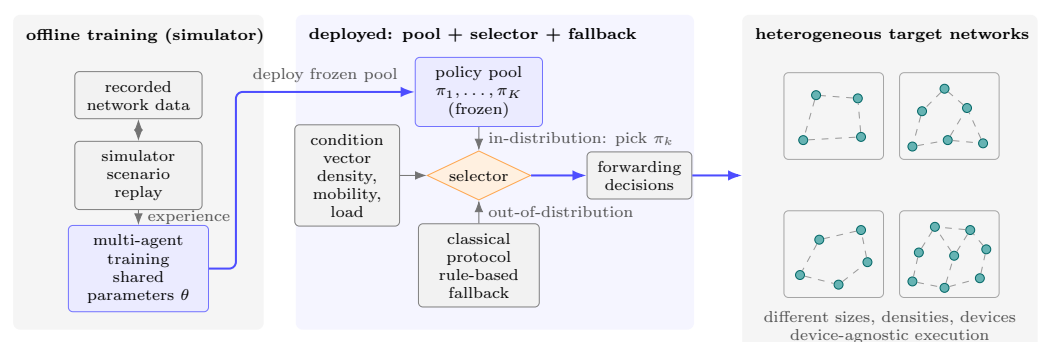


Figure 11. A hybrid routing system: offline-trained policies deployed as a pool, with a device-agnostic selector choosing among learned policies and a classical fallback.

This represents an important yet underdeveloped direction in the literature as it reconciles theoretical optimization with practical deployability. The existing work provides isolated components, such as condition-vector selectors with rule-based fallbacks for out-of-distribution scenarios [14] and anomaly monitors that detect when a deployed policy departs from its trustworthy operating regime [63]. A unified theoretical framework for their integration, however, remains lacking. The open problems are the governance layer

itself: a principled selector between learned and rule-based policies; runtime monitoring with safe reversible revocation; fallbacks with provable rather than assumed safety (the guarantee that runtime-assurance architectures provide elsewhere in safety-critical control [118]); and, underneath all of these, explicit trust criteria. The last of these is a principled account of when decision authority should be delegated to, or revoked from, a learned policy. In the audited core set, the governance layer is hand-crafted or absent. Figure 12 illustrates an initial component of such a governance layer: a nominal profile of the policy's own learning signal, built offline and compared online against deployment behavior, which provides monitoring but not yet the revocation and safety guarantees the governance question requires.

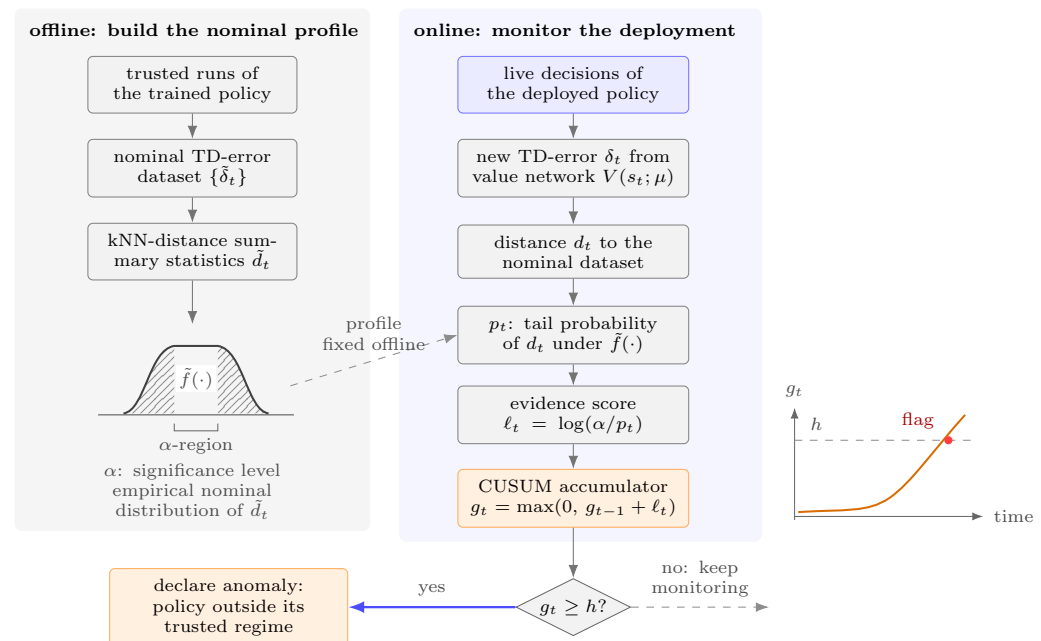


Figure 12. DeepADMR's two-phase anomaly detection, an initial monitoring component of a governance layer: offline nominal TD-error statistics support online CUSUM detection of a deployed policy leaving its nominal regime.

The governance question has an adversarial twin: a policy may leave its trusted operating regime not only through distributional drift but also deliberate adversarial manipulation. Learning adds or expands attack surfaces: observation inputs remain vulnerable as in classical routing, while the reward, training, and model layers introduce learning-specific targets, and the corpus has only begun to face them. Two of its studies provide initial evidence: a tactical study shows that jamming corrupts not only packets but the reward signal that trains the policy [5], and a security-first stack classifies malicious vehicles before routing at all [20]. Beyond these, the corpus's direct treatment of adversarial security remains limited. A significant gap in the audited core papers is that no study systematically evaluates attacks that deliberately target the learning mechanism itself. For example, in geographic routing schemes where Q-values are exchanged via hello messages [34], a forged beacon could corrupt not merely a single forwarding decision: it could poison the value functions learned by all neighboring nodes. Similarly, state representations derived from two-hop neighbor reports expose a vulnerability to adversarial perturbations [32], while globally shared policies across nodes could enable successful adversarial inputs to propagate across the multi-agent system. Outside the corpus the threat is already demonstrated: observation-perturbation attacks degrade neural policies in general [119], have been shown to disrupt an RL-based routing agent for tactical

mobile networks in particular [120], and are documented across learned wireless systems at large [121]. Figure 13 shows the exposed loop: state and reward flow from the environment into every agent's Q-network, so a jammer that corrupts this feedback attacks the learning itself rather than any single transmission.

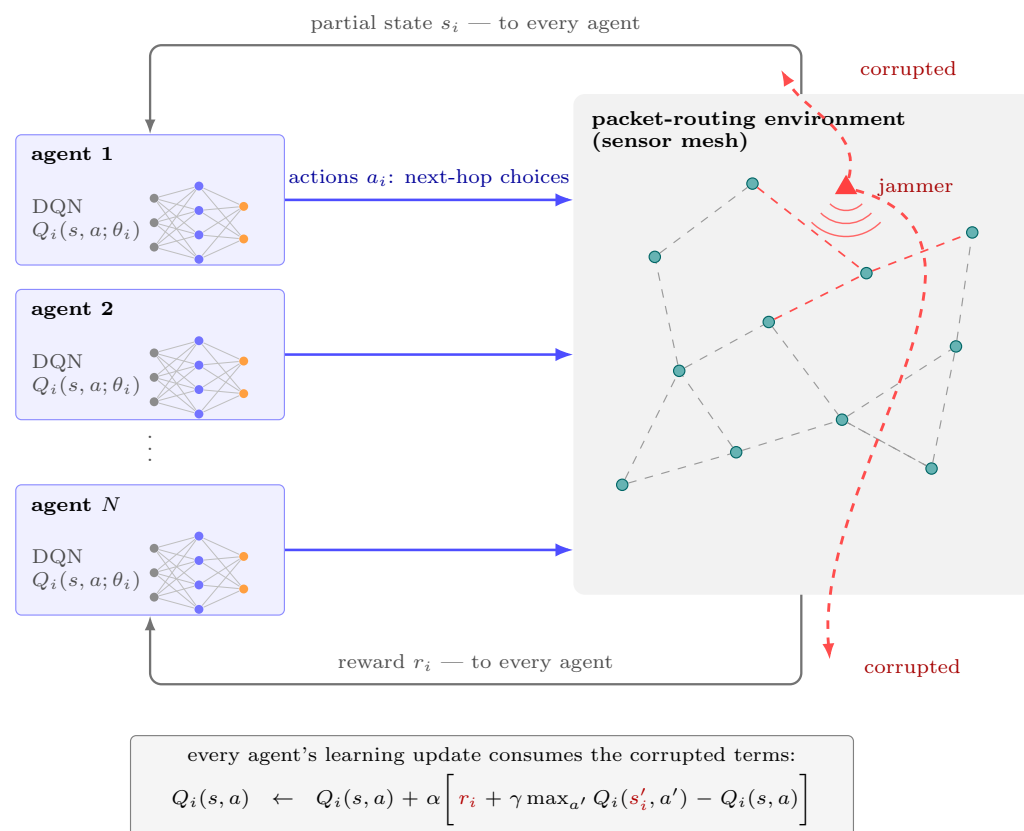


Figure 13. Multi-agent deep-RL packet routing under jamming: per-agent DQNs act on partial observations of an environment whose state and reward feedback an adversary can corrupt.

The directions follow the classic security decomposition, each reconsidered for a learned routing system: robust aggregation and poisoning defenses for value and observation exchange, adapted by analogy from Byzantine-tolerant gradient aggregation [122]; adversarial training and input sanitization for learned forwarding policies [123]; attack-aware extensions of the trust criteria above, so revocation can be triggered by evidence of manipulation and not only by drift; and explicit threat-model reporting, without which no security claim in this literature can be compared. The fact that the literature's existing defense mechanisms operate at distinct abstraction layers—signal processing and node behavior—suggests that a comprehensive solution may require a defense-in-depth architecture rather than a single security mechanism.

Security in these networks must also be accommodated within the budget of Section 6.2. An anomaly monitor of the kind shown in Figure 12 computes, for each new TD-error sample on a node, its distance to a nominal profile and updates a sequential test [63]; classifying malicious vehicles before routing may add a model inference for candidate or neighboring vehicles [20]. If transferred to an energy-constrained FANET, such mechanisms would compete with forwarding for computation, airtime, and battery, and neither paper quantifies this trade-off; conventional protections of the control traffic itself, such as authentication and encryption, are not learning mechanisms and lie outside the scope of this survey, although they draw on the same budget. The open question is therefore not whether learned routing can be secured but at what cost in delivery and

lifetime, and answering it requires the same reporting that the audit of Section 5 requires elsewhere: the overhead of the security mechanism measured in the same units as the overhead of routing so that a defense can be judged against the attack it prevents rather than assumed to be without cost.

6.5. Reproducible and Generalizable Evaluation

Progress on these challenges is difficult to assess without a more consistent basis for comparison (Section 5), and further work should adopt the reporting framework of Table 8. The ns3-gym bridge, already used within the corpus [2,43], is a practical starting point for an open benchmark suite with standardized scenarios, classical and learned baselines, and released artifacts. The gap between simulation and field deployment [50,65] could be bridged by hardware-in-the-loop evaluation and a network digital twin: a continuously synchronized model of the deployed network that mirrors its topology, traffic, and channel state [124,125]. For a learned router, a twin could host the governance layer of Section 6.4. A candidate policy could be validated in shadow mode before being granted decision authority; faults, jamming, and adversarial observations could be injected in the twin to test the attack-aware trust criteria of Section 6.4; and continual learning could be confined to the twin, with a policy promoted only after it has been screened against a monitored operating regime before deployment. These roles are proposals rather than demonstrated capabilities, but their constituent elements are beginning to appear: digital-twin-enhanced reinforcement learning for network resource management [126], a connection-aware twin for MANETs in a 5G setting [127], and the synchronization cost of a vehicular twin formulated as an optimization problem [128,129]. High-fidelity prototypes exist for selected cellular and radio-planning functions [130] but do not yet imply packet-level routing fidelity; none of the 37 audited core studies reports a digital-twin-based routing evaluation; and synchronization under ad hoc mobility is expected to be more difficult, a research direction in its own right.

6.6. Emerging Learning Paradigms and Their Fit to the Routing Decision

The corpus is dominated by tabular and deep reinforcement learning and recurrent or feed-forward predictors, and the method families of Section 2.3 reflect what the routing literature of these three network families has used so far rather than what the wider learning literature now offers. Several learning paradigms that have advanced rapidly since 2023 are considered here in terms of the five questions of the framework.

Graph neural networks address the state-representation and information-horizon questions jointly. Their permutation-invariant aggregation over a neighborhood and tolerance of variable-size neighborhoods match the structure of a routing decision and are one reason why the relational design in the corpus generalizes across network sizes without retraining [45]. The same properties carry costs that the routing literature has only begun to quantify: each message-passing layer widens the aggregation radius and adds computation and, in a distributed implementation, may add a further round of information exchange; a highly dynamic topology may require frequent graph updates; and inference on an embedded node grows with the neighborhood size. A recent survey catalogues graph neural network designs for routing optimization and their open problems [131]. Attention-based encoders offer a related treatment of variable neighbor sets and have been applied to trajectory and resource decisions in multi-UAV systems [132]; this constitutes adjacent rather than routing evidence. Outside the corpus, two recent preprints indicate this direction: local routing policies trained on a single graph that generalize across random wireless topologies [110] and forwarding strategies for multi-hop mobile networks trained

with continual learning across scenarios [111]. Both are cited as recent adjacent evidence; the corpus was not reopened after the audit of Section 5 had been coded.

Large language models and foundation models for networking currently address a different question: not which next hop to select within milliseconds on an energy-constrained node but how to configure, monitor, and explain a network [133]. A recent vehicular survey maps large language models, agentic AI, and embodied AI onto beamforming, resource allocation, semantic communication, and network optimization and identifies model compression and cloud–edge–vehicle collaborative deployment as the adaptations that resource-constrained vehicles require [21]. Their inference cost and latency make them more plausible, for the present, for management and control functions than for per-packet forwarding, which is where Section 6.4 located the missing governance layer: interpreting anomaly reports, proposing selector rules, and generating test scenarios for the digital twins of Section 6.5. To our knowledge no study in the corpus applies a transformer or a language model to the routing decision itself, and the boundary between a learned forwarding policy and a foundation model that supervises it remains an open design question. Adjacent decision problems in the same networks indicate the same tendency: task-driven priority-aware computation offloading formulated as deep reinforcement learning over a hybrid action space [115] treats delay and energy as explicit decision costs, which is the form of cost accounting that Section 6.2 requires of learned routing.

6.7. Summary

Table 9 collects these challenges against the evidence in the corpus that motivates each and the directions that could address each challenge. This suggests a shift in evaluation emphasis: away from a single figure of merit measured in one scenario and toward the generalization, reliability, security, reproducibility, and trust required before a mechanism that helps in one study can be deployed. Beyond the scope of the surveyed corpus lie broader questions that this review deliberately sets aside: foundation-model paradigms for network control, which a recent survey of large language models, agentic AI, and embodied AI for vehicular communications maps in detail [21]; standardization pathways for integrating learned components into operational protocol stacks; and formal certification frameworks for aerial platforms. While unaddressed by the current literature, these issues are likely to become increasingly relevant as learned routing mechanisms transition from simulation to real-world deployment.

The five challenges are not equally urgent, and Table 9 marks each with a horizon and with what it depends on. Reproducible evaluation and the information budget are near-term bottlenecks: nothing else can be compared across papers until Level 2 evidence (Section 5.1) and cost reporting exist. Generalization beyond training scenarios and scalable coordination are medium-term goals since working designs exist and need extending to mobility, traffic, and topology shift. Trustworthy deployment and certifiable corruption-robust coordination are long-term goals because they need formal safety specifications the literature does not yet have. The order is one of dependency, not importance.

7. Conclusions

This survey has examined two decades of learning-based routing for mobile, vehicular, and aerial ad hoc networks not as a catalogue of algorithms but as the evolution of the routing decision. Across the 92-paper corpus, that decision moved along five dimensions. Learning occupied a spectrum of decision authority from tuning one search parameter inside an otherwise classical protocol [15] through scoring and selecting routes on a classical scaffold [12] to constituting the forwarding policy itself [13,45]. State representation shifted from individual vehicles to geographic grids, intersections, and groups of roadside

units [28,52,53]. The information horizon was widened in some designs, while later studies narrowed or computationally reconstructed it, partly to limit control overhead [4,32,35]. The temporal basis of the decision changed from the network as measured to the network as predicted [1,31,58]. Finally, policy organization expanded to include shared generalizable policies and regime-adaptive policy pools [14,18]. The reviewed studies indicate a shift from reactive repair toward more proactive, topology-aware, and generalizable learning, while governance remains an emerging requirement rather than an established characteristic, a development that is obscured when the literature is organized by model family.

Highlighting these architectural shifts is a core aim of the mechanism-based taxonomy: the five dimensions isolate papers that share a model architecture but address different decision problems and unify papers that address the same problem with different models.

The evaluation audit in Section 5 yields a conclusion that is both cautionary and clarifying. Within the audited core set, studies from all three domain settings report within-study performance benefits (Level 1 evidence in the terms of Section 5.1): within individual studies, learning-based mechanisms have been reported to govern forwarding decisions [13,45], to stabilize the learner in urban networks through restructured state spaces [28,53], to anticipate link disruptions through prediction [58,60], and to generalize across selected unseen network sizes and mobility conditions through shared policies [18,45]. However, what the audited core set cannot yet establish is a general performance advantage across studies (Level 3) because comparable setups (Level 2) are rare, primarily due to inconsistent evaluation scenarios, non-standardized baselines, and disparate metrics. Most papers in the audited core set lack an independent learned benchmark, while the cost metrics needed to assess the field's fundamental trade-offs are reported inconsistently or omitted entirely. This survey maps the field's trajectory along five dimensions of the routing decision, exposes its design tensions, and identifies the evaluation limitations that prevent these tensions from being resolved. Addressing them requires greater emphasis on systematic evaluation, out-of-distribution generalization, security, and deployment assurance, alongside continued architectural development. Future progress in learning-based routing may depend not on packet-delivery gains alone but five under-addressed capabilities: out-of-distribution generalization, explicit information budgeting and trustworthy prediction, scalable multi-agent coordination, governable and adversarially robust deployment, and reproducible generalizable evaluation.

Developing an open benchmark with standardized scenarios, cost-sensitive metrics, learned baselines, and publicly available implementations would facilitate the transition from study-specific demonstrations to a cumulative and comparable evidence base. Such a benchmark would enable systematic assessment of the conditions under which learning-based routing is effective and the performance–cost trade-offs that it entails.

Author Contributions: Conceptualization, Y.L.; methodology, Y.L.; validation, Y.L.; investigation, Y.L.; formal analysis, Y.L.; data curation, Y.L.; visualization, Y.L.; writing—original draft preparation, Y.L.; writing—review and editing, X.J.L.; supervision, X.J.L.; project administration, X.J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new primary experimental data were generated. The classification of the 92 studies and the coded evaluation audit of the 37 core studies that support this review are contained in Tables 5 and 7 and Appendix A; the underlying coding sheets are available from the corresponding author upon reasonable request.

Acknowledgments: During the preparation of this manuscript, the authors used Undermind (undermind.ai) for literature search assistance and Claude Code (Anthropic, version 2.1) for figure preparation, formatting, and grammar checking. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

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

Appendix A. Corpus Retrieval Paths and the Studies in the Analytical Core

The 92-paper corpus was not retrieved by a single query. It was assembled through the eight search paths of Table A1, run iteratively with the semantic literature search engine described in Section 3.1 and supplemented by forward and backward citation chasing. Most studies were reached through more than one path; Table A1 lists each study once under its main entry path and gives in parentheses the number of studies that the path reached in total. Searches were run through this engine alone; no separate search of Scopus, Web of Science, IEEE Xplore, or the ACM Digital Library was made. Because the corpus was assembled by iterative semantic retrieval rather than under a systematic-review protocol, no PRISMA-style count of candidates screened at each step was kept and none is claimed. The main-entry counts of S1 to S7 sum to 92; S8 served as a recall check on forwarding vocabulary and reached 42 studies, all of which had already entered through another path, so it has no main entry. Table A2 lists the 37-paper analytical core by the role that qualified each member (Section 3.1) and, for completeness, the 55 studies that are classified in Table 5 but whose evaluations were not coded.

Table A1. Search paths used to assemble the corpus, with the studies that entered through each path (main entry path; in parentheses, the number of studies reached by the path at all).

Path	Main Vocabulary or Entry Point	Studies	References (Main Entry)
S1	AI, machine learning, reinforcement learning, or deep reinforcement learning + routing + MANET, VANET, FANET, or UAV	9 (10)	[12,47,50,57,64,65,69,73,74]
S2	VANET + intersection, road segment, grid, RSU, or traffic-aware routing	11 (19)	[28–30,52,53,55,62,66,95–97]
S3	FANET or UAV + geographic, topology-aware, or adaptive routing	19 (33)	[4,32,34,35,42,46,51,67,68,84–93]
S4	MANET + Q-routing, QoS, adaptive, opportunistic, or packet forwarding	16 (26)	[6,13,15,16,26,27,39,76–81,83,104,105]
S5	predictive routing, trajectory, link stability, route quality, LSTM, or ANN	17 (19)	[1,7,17,19,31,49,58–61,75,98–103]
S6	multi-agent centralized training with decentralized execution, shared policy, generalized, or hybrid routing	12 (12)	[2,3,5,14,18,33,41,43–45,56,63]
S7	semantic searches for 2023–2026 work and citation chasing from retrieved studies	8 (32)	[20,40,54,70–72,82,94]
S8	geographic forwarding, adaptive forwarding, proactive rerouting, mobility-aware, or opportunistic routing	0 (42)	none as main entry; the 42 studies reached are [4,13,15,16,20,26–30,32,34,35,42,44,45,47,50,52,57,59,60,63–66,68,74–77,80,84–86,89,92,96,98,99,104,105]

Table A2. Standing of every corpus study with respect to the 37-paper analytical core.

Standing	Studies	References
Core: principal work of a lineage (Figure 3)	28	[1,2,4,5,7,12,13,15,16,18,28–32,34,35,44,45,50,52,53,58,59,64–66,68]
Core: reference design of a mechanism not represented by a lineage member	9	[26,41,43,57,77,83,97,100,102]
Corpus only: classified in Table 5, evaluation not coded	55	[3,6,14,17,19,20,27,33,39,40,42,46,47,49,51,54–56,60–63,67,69–76,78–82,84–96,98,99,101,103–105]

References

1. Tang, Y.; Cheng, N.; Wu, W.; Wang, M.; Dai, Y.; Shen, X. Delay-Minimization Routing for Heterogeneous VANETs with Machine Learning Based Mobility Prediction. *IEEE Trans. Veh. Technol.* **2019**, *68*, 3967–3979. [CrossRef]
2. Qiu, X.; Xu, L.; Wang, P.; Yang, Y.; Liao, Z. A Data-Driven Packet Routing Algorithm for an Unmanned Aerial Vehicle Swarm: A Multi-Agent Reinforcement Learning Approach. *IEEE Wirel. Commun. Lett.* **2022**, *11*, 2160–2164. [CrossRef]
3. Agrawal, J.; Kumar, A.; Alam, M.M.; Arafat, M.Y. HCPMR: A Hierarchically Coordinated Proximal Multi-Hop Routing Scheme for FANETs in Mission-Critical Environments. *IEEE Access* **2026**, *14*, 36505–36522. [CrossRef]
4. Cui, Y.; Zhang, Q.; Feng, Z.; Wei, Z.; Shi, C.; Yang, H. Topology-Aware Resilient Routing Protocol for FANETs: An Adaptive Q-Learning Approach. *IEEE Internet Things J.* **2022**, *9*, 18632–18649. [CrossRef]

5. Okine, A.A.; Adam, N.; Naeem, F.; Kaddoum, G. Multi-Agent Deep Reinforcement Learning for Packet Routing in Tactical Mobile Sensor Networks. *IEEE Trans. Netw. Serv. Manag.* **2024**, *21*, 2155–2169. [\[CrossRef\]](#)
6. Minhas, H.I.; Ahmad, R.; Ahmed, W.; Waheed, M.; Alam, M.; Gul, S.T. A Reinforcement Learning Routing Protocol for UAV Aided Public Safety Networks. *Sensors* **2021**, *21*, 4121. [\[CrossRef\]](#)
7. Zhang, M.; Dong, C.; Yang, P.; Tao, T.; Wu, Q.; Quek, T.Q.S. Adaptive Routing Design for Flying Ad Hoc Networks. *IEEE Commun. Lett.* **2022**, *26*, 1438–1442. [\[CrossRef\]](#)
8. Perkins, C.E.; Royer, E.M. Ad-hoc On-Demand Distance Vector Routing. In *Proceedings of the WMCSA'99, Second IEEE Workshop on Mobile Computing Systems and Applications*; IEEE: New York, NY, USA, 1999; pp. 90–100. [\[CrossRef\]](#)
9. Johnson, D.B.; Maltz, D.A. Dynamic Source Routing in Ad Hoc Wireless Networks. In *Mobile Computing*; Springer: Berlin/Heidelberg, Germany, 1996; pp. 153–181. [\[CrossRef\]](#)
10. Karp, B.; Kung, H.T. GPSR: Greedy Perimeter Stateless Routing for Wireless Networks. In *Proceedings of the 6th Annual International Conference on Mobile Computing and Networking (MobiCom 2000)*; ACM: New York, NY, USA, 2000; pp. 243–254. [\[CrossRef\]](#)
11. Clausen, T.; Jacquet, P. *Optimized Link State Routing Protocol (OLSR)*; Technical Report RFC 3626; IETF: Wilmington, DE, USA, 2003. [\[CrossRef\]](#)
12. Wu, C.; Kumekawa, K.; Kato, T. Distributed Reinforcement Learning Approach for Vehicular Ad Hoc Networks. *IEICE Trans. Commun.* **2010**, *93*, 1431–1442. [\[CrossRef\]](#)
13. Dowling, J.; Curran, E.; Cunningham, R.; Cahill, V. Using feedback in collaborative reinforcement learning to adaptively optimize MANET routing. *IEEE Trans. Syst. Man. Cybern.-Part A Syst. Hum.* **2005**, *35*, 360–372. [\[CrossRef\]](#)
14. Fan, Y.; Xie, P.; Zhang, Y.; Liu, L.; Ma, H. Chimera: Pioneering Generalized and Adaptable Intelligent Routing in MANETs. *IEEE Trans. Cogn. Commun. Netw.* **2025**, *11*, 2027–2042. [\[CrossRef\]](#)
15. Usaha, W.; Barria, J. A reinforcement learning ticket-based probing path discovery scheme for MANETs. *Ad. Hoc Netw.* **2004**, *2*, 319–334. [\[CrossRef\]](#)
16. Tran, T.; Nguyen, T.V.; Shim, K.; da Costa, D.B.; An, B. A Deep Reinforcement Learning-Based QoS Routing Protocol Exploiting Cross-Layer Design in Cognitive Radio Mobile Ad Hoc Networks. *IEEE Trans. Veh. Technol.* **2022**, *71*, 13165–13181. [\[CrossRef\]](#)
17. Dong, S.; Tang, Z. LPMD-GPSR: LSTM-based link stability prediction and multi-parameter decision-making for adaptive routing in FANETs. *Comput. Commun.* **2026**, *250*, 108458. [\[CrossRef\]](#)
18. Kaviani, S.; Ryu, B.; Ahmed, E.; Larson, K.A.; Le, A.; Yahja, A.; Kim, J.H. DeepCQ+: Robust and Scalable Routing with Multi-Agent Deep Reinforcement Learning for Highly Dynamic Networks. In *Proceedings of the MILCOM 2021—2021 IEEE Military Communications Conference (MILCOM)*; IEEE: New York, NY, USA, 2021; pp. 31–36. [\[CrossRef\]](#)
19. Jin, Z.; Xu, Y.; Zhang, X.R.; Wang, J.; Zhang, L. Trajectory-prediction based relay scheme for time-sensitive data communication in VANETs. *KSII Trans. Internet Inf. Syst.* **2020**, *14*, 3399–3419. [\[CrossRef\]](#)
20. Bagirathan, K.; Saravanan, N.; Vijayabhaskar, K.; C, S. An Intelligent Recurrent Neural Network Driven Secured Routing Protocol for Vehicular Ad Hoc Networks. *Knowl.-Based Syst.* **2025**, *317*, 113371. [\[CrossRef\]](#)
21. Wu, Q.; Zhang, W.; Fan, P.; Wang, K.; Fan, Q.; Chen, W.; Mao, G.; Letaief, K.B. From LLMs to Agentic and Embodied AI for Next-Generation Intelligent Vehicular Communications: A Comprehensive Survey. *IEEE Commun. Surv. Tutor.* **2026**, *28*, 6983–7020. [\[CrossRef\]](#)
22. Djihene, A.; Amal, B.; Ali, K. Enhance Energy Using Bio-Inspired Algorithms in MANET: An Overview. In *Proceedings of the 2024 2nd International Conference on Electrical Engineering and Automatic Control (ICEEAC)*; IEEE: New York, NY, USA, 2024; pp. 1–6. [\[CrossRef\]](#)
23. Harshitha, V.S.; Surekha, V.R.; Ananthi, V. A Comprehensive Review on Q-Learning Based Optimal Routing Protocol for MANET. In *Proceedings of the 2026 International Conference on Machine Learning and Autonomous Systems (ICMLAS)*; IEEE: New York, NY, USA, 2026; pp. 1754–1758. [\[CrossRef\]](#)
24. Al-Mashhadani, M.J.; Karoui, K. Rule-Based and AI-Based Routing Protocols for Mobile Ad Hoc Networks: A Comparative Review. In *Proceedings of the 2025 7th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (ICHORA)*; IEEE: New York, NY, USA, 2025; pp. 1–12. [\[CrossRef\]](#)
25. Safari, F.; Savic, I.; Kunze, H.; Ernst, J.B.; Gillis, D. A Review of AI-based MANET Routing Protocols. In *Proceedings of the 2023 19th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*; IEEE: New York, NY, USA, 2023; pp. 43–50. [\[CrossRef\]](#)
26. Sharma, D.; Dhurandher, S.K.; Woungang, I.; Srivastava, R.; Mohananey, A.; Rodrigues, J. A Machine Learning-Based Protocol for Efficient Routing in Opportunistic Networks. *IEEE Syst. J.* **2018**, *12*, 2207–2213. [\[CrossRef\]](#)
27. Dhurandher, S.K.; Singh, J.; Obaidat, M.; Woungang, I.; Srivastava, S.; Rodrigues, J. Reinforcement Learning-Based Routing Protocol for Opportunistic Networks. In *Proceedings of the ICC 2020—2020 IEEE International Conference on Communications (ICC)*; IEEE: New York, NY, USA, 2020; pp. 1–6. [\[CrossRef\]](#)
28. Li, F.; Song, X.; Chen, H.; Li, X.; Wang, Y. Hierarchical Routing for Vehicular Ad Hoc Networks via Reinforcement Learning. *IEEE Trans. Veh. Technol.* **2019**, *68*, 1852–1865. [\[CrossRef\]](#)

29. Wu, J.; Fang, M.; Li, H.; Li, X. RSU-Assisted Traffic-Aware Routing Based on Reinforcement Learning for Urban Vanets. *IEEE Access* **2020**, *8*, 5733–5748. [\[CrossRef\]](#)
30. Li, R.; Li, F.; Li, X.; Wang, Y. QGrid: Q-learning based routing protocol for vehicular ad hoc networks. In *Proceedings of the 2014 IEEE 33rd International Performance Computing and Communications Conference (IPCCC)*; IEEE: New York, NY, USA, 2014; pp. 1–8. [\[CrossRef\]](#)
31. Chong, J.; Jia, X.; Yang, Z. Toward Routing in Low-Altitude Drone Networks: A Physical Sensing-Aided Intelligent Forwarding Mechanism with Deep Learning. *IEEE Internet Things J.* **2025**, *12*, 25442–25456. [\[CrossRef\]](#)
32. Arafat, M.Y.; Moh, S. A Q-Learning-Based Topology-Aware Routing Protocol for Flying Ad Hoc Networks. *IEEE Internet Things J.* **2022**, *9*, 1985–2000. [\[CrossRef\]](#)
33. Kaviani, S.; Ryu, B.; Ahmed, E.; Kim, D.; Kim, J.; Spiker, C.; Harnden, B. DeepMPR: Enhancing Opportunistic Routing in Wireless Networks via Multi-Agent Deep Reinforcement Learning. In *Proceedings of the MILCOM 2023—2023 IEEE Military Communications Conference (MILCOM)*; IEEE: New York, NY, USA, 2023; pp. 51–56. [\[CrossRef\]](#)
34. Jung, W.; Yim, J.; Ko, Y.B. QGeo: Q-Learning-Based Geographic Ad Hoc Routing Protocol for Unmanned Robotic Networks. *IEEE Commun. Lett.* **2017**, *21*, 2258–2261. [\[CrossRef\]](#)
35. Hosseinzadeh, M.; Ali, S.; Ionescu-Feleaga, L.; Ionescu, B.S.; Yousefpoor, M.S.; Yousefpoor, E.; Ahmed, O.H.; Rahmani, A.M.; Mehmood, A. A novel Q-learning-based routing scheme using an intelligent filtering algorithm for flying ad hoc networks (FANETs). *J. King Saud Univ.-Comput. Inf. Sci.* **2023**, *35*, 101817. [\[CrossRef\]](#)
36. Lindgren, A.; Doria, A.; Schelén, O. Probabilistic Routing in Intermittently Connected Networks. *ACM SIGMOBILE Mob. Comput. Commun. Rev.* **2003**, *7*, 19–20. [\[CrossRef\]](#)
37. Chen, S.; Nahrstedt, K. Distributed Quality-of-Service Routing in Ad Hoc Networks. *IEEE J. Sel. Areas Commun.* **1999**, *17*, 1488–1505. [\[CrossRef\]](#)
38. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*, 2nd ed.; MIT Press: Cambridge, MA, USA, 2018.
39. Glam, A.; Farbman, B.; Shleifer, A. RRP: Reinforced Routing Policy Architecture for MANET Routing. In *Proceedings of the 2019 IEEE International Conference on Microwaves, Antennas, Communications and Electronic Systems (COMCAS)*; IEEE: New York, NY, USA, 2019; pp. 1–6. [\[CrossRef\]](#)
40. Gui, Y.; Li, P.; Wang, P.; Liu, L.; Cao, R.; Zhang, L. Improving Mobility in NDN-Based VANET: A Deep Reinforcement Learning Approach with Deep Prioritization. *IEEE Trans. Intell. Transp. Syst.* **2026**, *27*, 1458–1470. [\[CrossRef\]](#)
41. Lu, C.; Wang, Z.; Ding, W.; Li, G.; Liu, S.; Cheng, L. MARVEL: Multi-agent reinforcement learning for VANET delay minimization. *China Commun.* **2021**, *18*, 1–11. [\[CrossRef\]](#)
42. Lin, D.; Peng, T.; Zuo, P.; Wang, W. Deep-Reinforcement-Learning-Based Intelligent Routing Strategy for FANETs. *Symmetry* **2022**, *14*, 1787. [\[CrossRef\]](#)
43. Qiu, X.; Xie, Y.; Wang, Y.; Ye, L.; Yang, Y. QLGR: A Q-learning-based Geographic FANET Routing Algorithm Based on Multi-agent Reinforcement Learning. *KSII Trans. Internet Inf. Syst.* **2021**, *15*, 4244–4274. [\[CrossRef\]](#)
44. Manfredi, V.; Wolfe, A.P.; Wang, B.; Zhang, X. Relational Deep Reinforcement Learning for Routing in Wireless Networks. In *Proceedings of the 2021 IEEE 22nd International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*; IEEE: New York, NY, USA, 2021; pp. 159–168. [\[CrossRef\]](#)
45. Manfredi, V.; Wolfe, A.P.; Zhang, X.; Wang, B. Learning an adaptive forwarding strategy for mobile wireless networks: Resource usage vs. latency. *Mach. Learn.* **2024**, *113*, 7157–7193. [\[CrossRef\]](#)
46. Liu, J.; Wang, Q.; Xu, Y. AR-GAIL: Adaptive routing protocol for FANETs using generative adversarial imitation learning. *Comput. Netw.* **2022**, *218*, 109382. [\[CrossRef\]](#)
47. Jin, W.; Gu, R.; Ji, Y. Reward Function Learning for Q-learning-Based Geographic Routing Protocol. *IEEE Commun. Lett.* **2019**, *23*, 1236–1239. [\[CrossRef\]](#)
48. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
49. Lai, W.; Lin, M.T.; Yang, Y.H. A Machine Learning System for Routing Decision-Making in Urban Vehicular Ad Hoc Networks. *Int. J. Distrib. Sens. Netw.* **2015**, *11*, 374391. [\[CrossRef\]](#)
50. Wu, C.; Ohzahata, S.; Kato, T. Flexible, Portable, and Practicable Solution for Routing in VANETs: A Fuzzy Constraint Q-Learning Approach. *IEEE Trans. Veh. Technol.* **2013**, *62*, 4251–4263. [\[CrossRef\]](#)
51. da Costa, L.A.L.F.; Kunst, R.; Pignaton de Freitas, E. Q-FANET: Improved Q-learning based routing protocol for FANETs. *Comput. Netw.* **2021**, *198*, 108379. [\[CrossRef\]](#)
52. Luo, L.; Sheng, L.; Yu, H.; Sun, G. Intersection-Based V2X Routing via Reinforcement Learning in Vehicular Ad Hoc Networks. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 5446–5459. [\[CrossRef\]](#)
53. Yang, Q.; Yoo, S.J. Hierarchical Reinforcement Learning-Based Routing Algorithm with Grouped RSU in Urban VANETs. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 10131–10146. [\[CrossRef\]](#)
54. Seo, J.; Choi, Y.; Jeon, S.E.; Chae, S.; Hong, J.P. Learning-Based Geographic Routing for Delay-Limited Multihop Wireless Networks. *IEEE Sens. J.* **2024**, *24*, 42163–42171. [\[CrossRef\]](#)

55. Song, Y.; Yen, C.; Hsieh, Y.H.; Kuo, C.H.; Chang, I.C. An Intersection-Based Traffic Awareness Routing Protocol in VANETs Using Deep Reinforcement Learning. *Wirel. Pers. Commun.* **2024**, *138*, 659–683. [\[CrossRef\]](#)
56. Modi, A.; Shah, R.; Jain, K.; Verma, R.; Shorey, R.; Saran, H. Multi-Agent Packet Routing (MAPR): Co-Operative Packet Routing Algorithm with Multi-Agent Reinforcement Learning. In *Proceedings of the 2023 15th International Conference on COMMunication Systems & NETWORKS (COMSNETS)*; IEEE: New York, NY, USA, 2023; pp. 722–730. [\[CrossRef\]](#)
57. Bai, Y.; Zhang, X.; Yu, D.; Li, S.; Wang, Y.; Lei, S.; Tian, Z. A Deep Reinforcement Learning-Based Geographic Packet Routing Optimization. *IEEE Access* **2022**, *10*, 108785–108796. [\[CrossRef\]](#)
58. Zhang, M.; Cheng, H.; Yang, P.; Dong, C.; Zhao, H.; Wu, Q.; Quek, T.Q. Adaptive Routing Design for Flying Ad Hoc Networks: A Joint Prediction Approach. *IEEE Trans. Veh. Technol.* **2024**, *73*, 2593–2604. [\[CrossRef\]](#)
59. Cárdenas, L.L.; Mezher, A.M.; Bautista, P.A.B.; León, J.P.A.; Igartua, M.A. A Multimetric Predictive ANN-Based Routing Protocol for Vehicular Ad Hoc Networks. *IEEE Access* **2021**, *9*, 86037–86053. [\[CrossRef\]](#)
60. Xu, W.; Ji, X.; Zhang, C.; Zhang, B.; Wang, Y.; Wang, X.; Wang, Y.; Wang, J.; Liu, B. PQR: Prediction-supported Quality-aware Routing for Uninterrupted Vehicle Communication. In *Proceedings of the 2021 IEEE/ACM 29th International Symposium on Quality of Service (IWQOS)*; IEEE: New York, NY, USA, 2021; pp. 1–10. [\[CrossRef\]](#)
61. Yen, C.; Jhang, Y.S.; Hsieh, Y.H.; Chen, Y.C.; Kuo, C.H.; Chang, I.C. An Integrated DQN and RF Packet Routing Framework for the V2X Network. *Electronics* **2024**, *13*, 2099. [\[CrossRef\]](#)
62. Jiang, S.; Huang, Z.; Ji, Y. Adaptive UAV-Assisted Geographic Routing with Q-Learning in VANET. *IEEE Commun. Lett.* **2021**, *25*, 1358–1362. [\[CrossRef\]](#)
63. Yahja, A.; Kaviani, S.; Ryu, B.; Kim, J.H.; Larson, K. DeepADMR: A Deep Learning based Anomaly Detection for MANET Routing. In *Proceedings of the MILCOM 2022—2022 IEEE Military Communications Conference (MILCOM)*; IEEE: New York, NY, USA, 2022; pp. 412–417. [\[CrossRef\]](#)
64. Wu, J.; Fang, M.; Li, X. Reinforcement Learning Based Mobility Adaptive Routing for Vehicular Ad-Hoc Networks. *Wireless Personal Communication* **2018**, *101*, 2143–2171. [\[CrossRef\]](#)
65. Wu, C.; Ji, Y.; Liu, F.; Ohzahata, S.; Kato, T. Toward Practical and Intelligent Routing in Vehicular Ad Hoc Networks. *IEEE Trans. Veh. Technol.* **2015**, *64*, 5503–5519. [\[CrossRef\]](#)
66. Rui, L.; Yan, Z.; Tan, Z.; Gao, Z.; Yang, Y.; Chen, X.; Liu, H. An Intersection-Based QoS Routing for Vehicular Ad Hoc Networks with Reinforcement Learning. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 9068–9083. [\[CrossRef\]](#)
67. Bouziane, N.; Doukha, Z.; Kimri, F.; Djouama, M. SQBRP-SDFANET: A scalable Q-learning-based routing protocol for SD-FANETs. *Ad. Hoc Netw.* **2025**, *178*, 103913. [\[CrossRef\]](#)
68. Liu, J.; Wang, Q.; He, C.; Jaffrès-Runser, K.; Xu, Y.; Li, Z.; Xu, Y. QMR: Q-learning based Multi-objective optimization Routing protocol for Flying Ad Hoc Networks. *Comput. Commun.* **2020**, *150*, 304–316. [\[CrossRef\]](#)
69. Yang, X.; Zhang, W.; Lu, H.; Zhao, L. V2V Routing in VANET Based on Heuristic Q-Learning. *Int. J. Comput. Commun. Control* **2020**, *15*, 3928. [\[CrossRef\]](#)
70. Kumar, A.; Tyagi, S.; Dixit, P.; Tyagi, S. Simulation-Based Evaluation of Reinforcement Learning-Enhanced Location-Aware Routing in Urban Vehicular Ad-hoc Networks (VANETs). *J. Curr. Sci. Technol.* **2026**, *16*, 180. [\[CrossRef\]](#)
71. Jiang, Y.; Zhu, J.; Yang, K. Environment-Aware Adaptive Reinforcement Learning-Based Routing for Vehicular Ad Hoc Networks. *Sensors* **2023**, *24*, 40. [\[CrossRef\]](#)
72. al Mutoki, S.M.M.; Al-sharhanee, K.; Faisal, N.; Alkhayyat, A.; Abbas, F. Mobility Based Improved Q-Learning Approach for RPL Routing Based Vehicular Adhoc Networks. In *Proceedings of the 2023 6th International Conference on Engineering Technology and Its Applications (IICETA)*; IEEE: New York, NY, USA, 2023; pp. 623–628. [\[CrossRef\]](#)
73. Roh, B.; Han, M.; Ham, J.; Kim, K.I. Q-LBR: Q-Learning Based Load Balancing Routing for UAV-Assisted VANET. *Sensors* **2020**, *20*, 5685. [\[CrossRef\]](#)
74. Chen, W.; Ding, X.; Zheng, H.; Yang, W.; Fan, Y. Stable V2V routing protocol based on deep reinforcement learning in VANETs. In *Proceedings of the The 2nd International Conference on Distributed Sensing and Intelligent Systems (ICDSIS 2021)*; IET: London, UK, 2021; pp. 196–206. [\[CrossRef\]](#)
75. Chen, J.; Paul, R.; Choi, Y.J. An Efficient Neural Network-Based Next -Hop Selection Strategy for Multi-hop VANETs. In *Proceedings of the 2021 International Conference on Information Networking (ICOIN)*; IEEE: New York, NY, USA, 2021; pp. 699–702. [\[CrossRef\]](#)
76. Russell, B.; Littman, M.; Trappe, W. Integrating machine learning in ad hoc routing: A wireless adaptive routing protocol. *Int. J. Commun. Syst.* **2011**, *24*, 950–966. [\[CrossRef\]](#)
77. Ghaffari, A. Real-time routing algorithm for mobile ad hoc networks using reinforcement learning and heuristic algorithms. *Wirel. Netw.* **2016**, *23*, 703–714. [\[CrossRef\]](#)
78. Tilwari, V.; Dimyati, K.; Hindia, M.N.; Fattouh, A.; Amiri, I. Mobility, Residual Energy, and Link Quality Aware Multipath Routing in MANETs with Q-learning Algorithm. *Appl. Sci.* **2019**, *9*, 1582. [\[CrossRef\]](#)

79. Hendriks, T.P.M.; Camelo, M.; Latré, S. Q2-Routing: A Qos-aware Q-Routing algorithm for Wireless Ad Hoc Networks. In *Proceedings of the 2018 14th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*; IEEE: New York, NY, USA, 2018; pp. 108–115. [\[CrossRef\]](#)
80. Nurmi, P. Reinforcement Learning for Routing in Ad Hoc Networks. In *Proceedings of the 2007 5th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks and Workshops*; IEEE: New York, NY, USA, 2007; pp. 1–8. [\[CrossRef\]](#)
81. Cai, J.; Wang, C.; Lei, M.; Zhao, M. An Intelligent Routing Algorithm Based on Prioritized Replay Double DQN for MANET. In *Proceedings of the 2020 IEEE 92nd Vehicular Technology Conference (VTC2020-Fall)*; IEEE: New York, NY, USA, 2020; pp. 1–5. [\[CrossRef\]](#)
82. Al-Mashhadani, M.J.; Karoui, K. A Distributed QoS-Aware MANET Routing Protocol Based on Deep Reinforcement Learning and SDN Integration. *Concurr. Comput. Pract. Exp.* **2025**, *37*, e70381. [\[CrossRef\]](#)
83. Liu, J.; Wang, Q.; He, C.; Xu, Y. ARdeep: Adaptive and Reliable Routing Protocol for Mobile Robotic Networks with Deep Reinforcement Learning. In *Proceedings of the 2020 IEEE 45th Conference on Local Computer Networks (LCN)*; IEEE: New York, NY, USA, 2020; pp. 465–468. [\[CrossRef\]](#)
84. Chen, Y.; Lyu, N.; Song, G.; Yang, B.; Jiang, X. A traffic-aware Q-network enhanced routing protocol based on GPSR for unmanned aerial vehicle ad-hoc networks. *Front. Inf. Technol. Electron. Eng.* **2020**, *21*, 1308–1320. [\[CrossRef\]](#)
85. Lyu, N.; Song, G.; Yang, B.; Cheng, Y. QNGPSR: A Q-Network Enhanced Geographic Ad-Hoc Routing Protocol Based on GPSR. In *Proceedings of the 2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)*; IEEE: New York, NY, USA, 2018; pp. 1–6. [\[CrossRef\]](#)
86. Rovira-Sugranes, A.; Afghah, F.; Qu, J.; Razi, A. Fully-Echoed Q-Routing with Simulated Annealing Inference for Flying Adhoc Networks. *IEEE Trans. Netw. Sci. Eng.* **2021**, *8*, 2223–2234. [\[CrossRef\]](#)
87. Sun, C.; Hou, L.; Yu, S.; Shu, J. HQA: Hybrid Q-learning and AODV multi-path routing algorithm for Flying Ad-hoc Networks. *Veh. Commun.* **2025**, *55*, 100947. [\[CrossRef\]](#)
88. Tho, M.C.; Ly, N.T.H.; Binh, L.H.; Vo, T.T. QLR-FANET: A Q-learning and rate control-based routing protocol for flying ad hoc network. *ETRI J.* **2025**, *47*, 1015–1027. [\[CrossRef\]](#)
89. Wei, C.; Wang, Y.; Wang, X.; Tang, Y. QFAGR: A Q-learning-based Fast Adaptive Geographic Routing Protocol for Flying Ad hoc Networks. In *Proceedings of the GLOBECOM 2023—2023 IEEE Global Communications Conference*; IEEE: New York, NY, USA, 2023; pp. 4613–4618. [\[CrossRef\]](#)
90. Zhang, Y.; Qiu, H. Delay-Aware and Link-Quality-Aware Geographical Routing Protocol for UANET via Dueling Deep Q-Network. *Sensors* **2023**, *23*, 3024. [\[CrossRef\]](#)
91. Pang, Y.; Dong, F.; Huang, R.; He, Q.; Shi, Z.; Chen, Z. A Resilient Packet Routing Approach Based on Deep Reinforcement Learning. In *Proceedings of the 2024 IEEE 24th International Conference on Communication Technology (ICCT)*; IEEE: New York, NY, USA, 2024; pp. 741–747. [\[CrossRef\]](#)
92. Park, C.; Lee, S.; Joo, H.; Kim, H. Empowering Adaptive Geolocation-Based Routing for UAV Networks with Reinforcement Learning. *Drones* **2023**, *7*, 387. [\[CrossRef\]](#)
93. Hosseinzadeh, M.; Tanveer, J.; Rahmani, A.M.; Aurangzeb, K.; Yousefpoor, E.; Yousefpoor, M.S.; Darwesh, A.; Lee, S.-W.; Fazlali, M. A Q-learning-based smart clustering routing method in flying Ad Hoc networks. *J. King Saud Univ.-Comput. Inf. Sci.* **2024**, *36*, 101894. [\[CrossRef\]](#)
94. Karacelebi, C.; Onur, E.; Sahin, Y. Routing-Aware RL for Mobile Relay Navigation in Disconnected Ad Hoc Networks. In *Proceedings of the 2025 21st International Conference on Network and Service Management (CNSM)*; IEEE: New York, NY, USA, 2025; pp. 1–6. [\[CrossRef\]](#)
95. Yang, Q.; Yoo, S. Grouped Intersection-based Routing using Reinforcement Learning for Urban VANETs. In *Proceedings of the 2022 13th International Conference on Information and Communication Technology Convergence (ICTC)*; IEEE: New York, NY, USA, 2022; pp. 1855–1858. [\[CrossRef\]](#)
96. Yang, C.; Yen, C.; Chang, I.C. A Software-Defined Directional Q-Learning Grid-Based Routing Platform and Its Two-Hop Trajectory-Based Routing Algorithm for Vehicular Ad Hoc Networks. *Sensors* **2022**, *22*, 8222. [\[CrossRef\]](#)
97. Khan, M.U.; Hosseinzadeh, M.; Mosavi, A. An Intersection-Based Routing Scheme Using Q-Learning in Vehicular Ad Hoc Networks for Traffic Management in the Intelligent Transportation System. *Mathematics* **2022**, *10*, 3731. [\[CrossRef\]](#)
98. Kumbhar, F.; Shin, S.Y. Novel Vehicular Compatibility-Based Ad Hoc Message Routing Scheme in the Internet of Vehicles Using Machine Learning. *IEEE Internet Things J.* **2022**, *9*, 2817–2828. [\[CrossRef\]](#)
99. Sliwa, B.; Schöler, C.; Patchou, M.; Wietfeld, C. PARROT: Predictive Ad-hoc Routing Fueled by Reinforcement Learning and Trajectory Knowledge. In *Proceedings of the 2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring)*; IEEE: New York, NY, USA, 2021; pp. 1–7. [\[CrossRef\]](#)
100. Zhang, M.; Dong, C.; Feng, S.; Guan, X.; Chen, H.; Wu, Q. Adaptive 3D routing protocol for flying ad hoc networks based on prediction-driven Q-learning. *China Commun.* **2022**, *19*, 302–317. [\[CrossRef\]](#)

101. Xu, M.; Xia, Y.; Liu, W.; Huang, D. Reinforcement-Learning-Based Geographic Routing Considering Future Evolution of Link States for UAV Networks. *Drones* **2026**, *10*, 150. [\[CrossRef\]](#)
102. Liu, C.; Wang, Y.; Wang, Q. PARouting: Prediction-supported adaptive routing protocol for FANETs with deep reinforcement learning. *Int. J. Intell. Netw.* **2023**, *4*, 113–121. [\[CrossRef\]](#)
103. Lau, W.J.; Lim, J.M.Y.; Chong, C.Y.; Shen, H.N.; Ooi, T.W.M. AQR-FANET: An Anticipatory Q-Learning-based Routing Protocol for FANETs. In *Proceedings of the 2023 IEEE 16th Malaysia International Conference on Communication (MICC)*; IEEE: New York, NY, USA, 2023; pp. 6–11. [\[CrossRef\]](#)
104. Oddi, G.; Macone, D.; Pietrabissa, A.; Liberati, F. A proactive link-failure resilient routing protocol for MANETs based on reinforcement learning. In *Proceedings of the 2012 20th Mediterranean Conference on Control & Automation (MED)*; IEEE: New York, NY, USA, 2012; pp. 1259–1264. [\[CrossRef\]](#)
105. Visca, J.; Baliosian, J. RL4dtn: Q-Learning for Opportunistic Networks. *Future Internet* **2022**, *14*, 348. [\[CrossRef\]](#)
106. Kurkowski, S.; Camp, T.; Colagrosso, M. MANET Simulation Studies: The Incredibles. *ACM SIGMOBILE Mob. Comput. Commun. Rev.* **2005**, *9*, 50–61. [\[CrossRef\]](#)
107. Andel, T.R.; Yasinsac, A. On the Credibility of MANET Simulations. *Computer* **2006**, *39*, 48–54. [\[CrossRef\]](#)
108. Yoon, J.; Liu, M.; Noble, B. Random Waypoint Considered Harmful. In *Proceedings of the IEEE INFOCOM 2003—Twenty-Second Annual Joint Conference of the IEEE Computer and Communications Societies*; IEEE: New York, NY, USA, 2003; pp. 1312–1321. [\[CrossRef\]](#)
109. Hernandez-Leal, P.; Kartal, B.; Taylor, M.E. A Survey and Critique of Multiagent Deep Reinforcement Learning. *Auton. Agents Multi-Agent Syst.* **2019**, *33*, 750–797. [\[CrossRef\]](#)
110. Chen, Y.F.; Lin, S.; Arora, A. Learning from A Single Graph is All You Need for Near-Shortest Path Routing in Wireless Networks. *arXiv* **2023**, arXiv:2308.09829. [\[CrossRef\]](#)
111. Park, C.; Manfredi, V.; Zhang, X.; Liu, C.; Wolfe, A.P.; Song, D.; Tasneem, S.; Wang, B. Continual Learning to Generalize Forwarding Strategies for Diverse Mobile Wireless Networks. *arXiv* **2025**, arXiv:2509.23913. [\[CrossRef\]](#)
112. Zaheer, M.; Kottur, S.; Ravanbakhsh, S.; Póczos, B.; Salakhutdinov, R.; Smola, A.J. Deep Sets. *Proc. Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3391–3401.
113. Finn, C.; Abbeel, P.; Levine, S. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. *Proc. Mach. Learn. Res.* **2017**, *70*, 1126–1135.
114. Tobin, J.; Fong, R.; Ray, A.; Schneider, J.; Zaremba, W.; Abbeel, P. Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World. In *Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*; IEEE: New York, NY, USA, 2017; pp. 23–30. [\[CrossRef\]](#)
115. Hao, H.; Xu, C.; Zhang, W.; Yang, S.; Muntean, G.M. Task-Driven Priority-Aware Computation Offloading Using Deep Reinforcement Learning. *IEEE Trans. Wirel. Commun.* **2025**, *24*, 8114–8128. [\[CrossRef\]](#)
116. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *Proc. Mach. Learn. Res.* **2016**, *48*, 1050–1059.
117. Angelopoulos, A.N.; Bates, S. Conformal Prediction: A Gentle Introduction. *Found. Trends Mach. Learn.* **2023**, *16*, 494–591. [\[CrossRef\]](#)
118. Alshiekh, M.; Bloem, R.; Ehlers, R.; Könighofer, B.; Niekum, S.; Topcu, U. Safe Reinforcement Learning via Shielding. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*; AAAI Press: Washington, DC, USA, 2018; pp. 2669–2678. [\[CrossRef\]](#)
119. Huang, S.; Papernot, N.; Goodfellow, I.; Duan, Y.; Abbeel, P. Adversarial Attacks on Neural Network Policies. In *Proceedings of the 5th International Conference on Learning Representations (ICLR 2017), Workshop Track*, Toulon, France, 24–26 April 2017. Available online: <https://openreview.net/forum?id=ryvIRyBKI> (accessed on 1 September 2026).
120. Loevenich, J.F.; Bode, J.; Hürten, T.; Liberto, L.; Spelter, F.; Rettore, P.H.; Lopes, R.R.F. Adversarial Attacks Against Reinforcement Learning Based Tactical Networks: A Case Study. In *Proceedings of the MILCOM 2022—2022 IEEE Military Communications Conference (MILCOM)*; IEEE: New York, NY, USA, 2022; pp. 986–992. [\[CrossRef\]](#)
121. Adesina, D.; Hsieh, C.C.; Sagduyu, Y.E.; Qian, L. Adversarial Machine Learning in Wireless Communications Using RF Data: A Review. *IEEE Commun. Surv. Tutor.* **2023**, *25*, 77–100. [\[CrossRef\]](#)
122. Blanchard, P.; El Mhamdi, E.M.; Guerraoui, R.; Stainer, J. Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent. In *Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017)*; NeurIPS: San Diego, CA, USA, 2017; pp. 119–129.
123. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*, Vancouver, BC, Canada, 30 April–3 May 2018. Available online: <https://openreview.net/forum?id=rjzIBfZAb> (accessed on 1 September 2026).
124. Wu, Y.; Zhang, K.; Zhang, Y. Digital Twin Networks: A Survey. *IEEE Internet Things J.* **2021**, *8*, 13789–13804. [\[CrossRef\]](#)

125. Khan, L.U.; Han, Z.; Saad, W.; Hossain, E.; Guizani, M.; Hong, C.S. Digital Twin of Wireless Systems: Overview, Taxonomy, Challenges, and Opportunities. *IEEE Commun. Surv. Tutor.* **2022**, *24*, 2230–2254. [[CrossRef](#)]
126. Cheng, N.; Wang, X.; Li, Z.; Yin, Z.; Luan, T.H.; Shen, X. Toward Enhanced Reinforcement Learning-Based Resource Management via Digital Twin: Opportunities, Applications, and Challenges. *IEEE Netw.* **2025**, *39*, 189–196. [[CrossRef](#)]
127. Jesús-Azabal, M.; Zhang, Z.; Gao, B.; Yang, J.; Soares, V.N.G.J. Connection-Aware Digital Twin for Mobile Adhoc Networks in the 5G Era. *Future Internet* **2024**, *16*, 399. [[CrossRef](#)]
128. Zheng, J.; Luan, T.H.; Zhang, Y.; Li, R.; Hui, Y.; Gao, L.; Dong, M. Data Synchronization in Vehicular Digital Twin Network: A Game Theoretic Approach. *IEEE Trans. Wirel. Commun.* **2023**, *22*, 7635–7647. [[CrossRef](#)]
129. Hui, Y.; Li, Y.; Cheng, N.; Li, C.; Zhou, C.; Su, Z.; Chen, R. Prevent Deception: On-Demand Data Synchronization for Vehicle Digital Twins. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 182–195. [[CrossRef](#)]
130. Lin, X.; Kundu, L.; Dick, C.; Obiodu, E.; Mostak, T.; Flaxman, M. 6G Digital Twin Networks: From Theory to Practice. *IEEE Commun. Mag.* **2023**, *61*, 72–78. [[CrossRef](#)]
131. Jiang, W.; Han, H.; Zhang, Y.; Wang, J.; He, M.; Gu, W.; Mu, J.; Cheng, X. Graph Neural Networks for Routing Optimization: Challenges and Opportunities. *Sustainability* **2024**, *16*, 9239. [[CrossRef](#)]
132. Feng, Z.; Wu, D.; Huang, M.; Yuen, C. Graph-Attention-Based Reinforcement Learning for Trajectory Design and Resource Assignment in Multi-UAV-Assisted Communication. *IEEE Internet Things J.* **2024**, *11*, 27421–27434. [[CrossRef](#)]
133. Huang, Y.; Du, H.; Zhang, X.; Niyato, D.; Kang, J.; Xiong, Z.; Wang, S.; Huang, T. Large Language Models for Networking: Applications, Enabling Techniques, and Challenges. *IEEE Netw.* **2025**, *39*, 235–242. [[CrossRef](#)]

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.