Optimizing EV Charging Routes in Inner London with Reinforcement Learning (Proximal Policy Optimization)

*Integrating tariff heterogeneity, nonlinear charging curves, and traffic simulation*

Hatim Shaherawala1[21054059], Vishesh Jakhar1[24058353], and Divykumar Patel1[24058610]

1 University of The West of England, Bristol BS16 1QY, United Kingdom

**Abstract:**

**Keywords:**

# 1 Introduction

The rapid adoption of electric vehicles (EVs) is reshaping transport and energy systems, driven by ambitious decarbonisation targets and urban air quality policies [8]. In London, this transition has been supported by an expansion of charging infrastructure [10], yet challenges persist. The charging landscape remains fragmented across multiple operators, with uneven station distribution [9] and tariffs that combine per-kWh energy rates, session fees, idle penalties, and membership discounts [12, 13]. For drivers, key concerns include minimising costs, reducing delays, and ensuring trip feasibility, while operators and regulators focus on infrastructure adequacy and grid stability. Existing optimisation efforts have often prioritised the grid perspective [1–3] but struggle to capture the complex, user-centred realities of urban charging.

Several gaps in the literature limit the applicability of current approaches to dense metropolitan contexts like Inner London. Many studies assume uniform or simplified tariffs, overlooking operator heterogeneity, idle fees, and subscription models [32, 13–15]. Charging dynamics are frequently simplified as well: constant or generic rates are used in place of vehicle-specific nonlinear charging curves [17, 33, 34]. Hybrid objectives balancing cost and time remain underexplored; weighted-sum and Pareto methods are sometimes applied, but their calibration is often ad hoc [35–37]. Reinforcement learning (RL) has shown promise in this domain [5–7, 21, 22], yet many implementations rely on oversimplified environments that neglect congestion, unrealistic station cycling, or other real-world constraints [16, 25, 41]. Reward design also poses difficulties: prior work often fixes objective priorities or under-utilises shaping and penalties [38–40]. Collectively, these gaps indicate that a new, driver-centric approach is needed. One that captures real-world variability in tariffs, charging behaviour, and urban constraints.

This thesis addresses these challenges by applying Proximal Policy Optimisation (PPO) to model EV charging decisions in Inner London. Using structured datasets, the study integrates operator-specific tariffs, detailed station metadata, and vehicle charging curves to simulate realistic charging costs and times. A custom RL environment was built using SUMO for traffic modelling and OpenAI Gym for training, with rules to prevent implausible behaviours such as repeated station cycling. Reward functions were designed for cost, time, and hybrid objectives, incorporating potential-based shaping and feasibility penalties. The methodology emphasises reproducibility through structured data inputs, fixed random seeds, and consistent evaluation protocols.

The research is guided by four questions:

Q1: How can reinforcement learning minimise total trip cost while ensuring state of charge (SoC) remains above a safe reserve threshold?

Q2: How can reinforcement learning minimise total trip duration while maintaining SoC for journey completion?

Q3: How can a hybrid reinforcement learning framework balance cost and time objectives under varying operational constraints?

Q4: How do cost-based, time-based, and hybrid optimisation objectives differ in shaping charging behaviour and route selection, and what trade-offs emerge between them?

By addressing these questions, the study contributes a reinforcement learning framework that captures the technical and behavioural complexities of EV charging in Inner London.

# 2 Literature Review

Before reinforcement learning gained prominence, EV charging optimisation was approached with deterministic and heuristic methods. Mixed-integer linear programming (MILP) can yield optimal solutions under fixed assumptions [19] but scales poorly for real-time, city-wide applications as the number of vehicles and stations grows. Metaheuristic techniques (e.g. genetic algorithms, particle swarm optimisation) improve scalability [20], yet early electric vehicle routing formulations still overlooked critical dynamics like traffic congestion, waiting queues, and nonlinear charging rates [16]. These limitations prompted a shift towards adaptive, data-driven methods, with RL offering a compelling alternative.

Many prior studies oversimplified key technical factors. Battery charging is nonlinear, power input tapers as the battery fills. Yet models often assume a constant rate, underestimating charging duration [17, 18]. Vehicle-specific differences are usually ignored due to limited data availability [33]; only recently have large datasets revealed broad heterogeneity in EV charging behaviour across models [34]. Real-world pricing is likewise complex, operators impose varying energy rates, session fees, idle penalties, and membership discounts [13, 32]. However, much of the earlier optimisation literature assumed uniform or flat pricing schemes [13–15, 32], which can yield unrealistic recommendations. Incorporating these tariff nuances (for example, discounted member-only rates vs. higher guest prices, or per-minute idle fees) can significantly change optimal charging decisions [15]. Any practical framework must therefore reflect such tariff heterogeneity and realistic charging dynamics, especially in dense urban settings.

RL offers the ability to learn adaptive policies through interaction with the environment, something static optimisations cannot achieve. Applications of RL in energy and transportation show it can reduce costs and improve operational efficiency [21, 22]. Early work on EV charging used value-based RL (e.g. Deep Q-Networks) to coordinate charging and reduce user cost peaks [5, 6], but the discrete action space in those approaches was not suitable for continuous routing decisions. Policy-gradient methods like Proximal Policy Optimization (PPO) can handle continuous actions [7] and are known for stable training; indeed, PPO has demonstrated strong performance on complex or multi-objective tasks in simulations [23]. Despite this promise, few studies to date have applied PPO for individual driver routing in dense cities, where variable tariffs, nonlinear charging curves, and traffic congestion all intersect. Other advanced RL algorithms (e.g., DDPG and SAC) have also been tested for EV charging control [31], but they similarly have not been deployed at the driver-route level in a congested urban context.

A realistic simulation environment is essential for training RL agents effectively. Many earlier studies made simplifying assumptions such as deterministic travel times or unconstrained charger availability, ignoring urban traffic and station queues [16]. In contrast, recent works stress simulation realism by integrating traffic models and operational constraints. Using tools like OpenAI Gym with the SUMO traffic simulator allows an agent to experience time-varying congestion and competition for charging stalls during training [41]. This added fidelity prevents overly optimistic strategies and yields policies that, for example, learn to avoid heavily congested routes or anticipate waiting times at busy stations, behaviours crucial for real-world feasibility.

Designing an effective reward function is another challenge when multiple objectives must be balanced. Past works often focused on a single goal (either cost or time) or used a fixed weighted sum of both [21, 22]. Some researchers even convert time into an equivalent monetary cost via a value-of-time factor to merge objectives into one metric [35, 37]. Alternatively, evolutionary and Pareto-based methods have been explored to avoid preset weights and generate a spectrum of cost–time trade-off solutions [36]. In practice, careful reward shaping is needed. One proven technique is potential-based shaping, which adds an extra incentive (e.g. rewarding reductions in remaining travel time or distance) without changing the optimal policy [38]. Strong penalty terms are also common: for instance, a very large negative reward if the vehicle depletes its battery en route trains the agent to avoid infeasible paths [39]. Effective use of shaping and penalties significantly improves learning efficiency and policy realism in EV charging simulations [38, 39]. Notably, a few multi-objective RL approaches for EV charging have been proposed in recent years [29, 30], but these remain limited in scope and have yet to tackle the full real-world complexity (heterogeneous prices, nonlinear charging, traffic congestion) that this study addresses.

In summary, the existing literature highlights the need for a driver-centric EV charging strategy that can adapt to complex urban conditions. No prior work fully integrates fine-grained tariff differences, realistic charging curves, and dynamic traffic into a single optimisation framework. The present study therefore applies PPO in a realistic Inner London simulation, explicitly integrating these real-world factors into a unified RL approach for EV routing and charging.

# 3 Methodology

## 3.1 Data Preparation

Four key data sources were compiled to drive the simulation:

**1. Tariffs**:Real world pricing data were collected from multiple charging network operators [44 -58] (see Fig. 1) and standardized into a unified format (converted into £/kWh units) to enable accurate cost calculations.

A screenshot of a computer

AI-generated content may be incorrect.A close-up of a message

AI-generated content may be incorrect.

Figure 1: Raw Tariff Information Examples

Tariffs were also split into three structured datasets. This ensured realistic and provider-specific cost estimation in the simulation environment (see Fig. 2).

A diagram of a manual selection

AI-generated content may be incorrect.

Figure 2 Conversion Process

**2. Stations:** the UK National Chargepoint Registry [42] provided the locations and technical details of public charging stations, which were cleaned and split into station-level metadata (e.g. location, operator, number of connectors) and connector-level attributes (power output, plug type, tariff info) for use in the simulation (see Fig. 3).

A diagram of a diagram

AI-generated content may be incorrect.

Figure 3 Splitting the Charging Data

**3. Vehicles:** Specifications (battery capacity and charging performance curves) for various EV models were taken from the Open EV Data project [43], so the simulator can model charging time based on each vehicle’s SoC and the charger power available.

A diagram of an electrical system

AI-generated content may be incorrect.

Figure 4 Splitting the EV Data

**4. User trips:** A set of simulated trip scenarios was generated, each specifying an origin, destination, and start time along with an initial SoC (with a set reserve margin), a particular EV model, the driver’s membership status (for tariff effects), and the chosen optimisation objective (cost, time, or hybrid).

These structured datasets ensured that the RL environment had access to rich, realistic information on costs, infrastructure, vehicle behaviour and driver preferences.

## 3.2 Environment Design

We implement a custom OpenAI Gym environment coupled with the SUMO microscopic traffic simulator. At each decision step the agent observes its current state and selects either to continue driving or to divert to a candidate charging station.

The environment is a Markov Decision Process (MDP) defined in Eq. (1).

**(1)**

Driving energy use and state-of-charge (SoC) updates follow Eq. (2) and travel time increments are obtained from SUMO per Eq. (3).

**(2)**

**(3)**

Charging dynamics respect connector limits and vehicle-specific charging curves per Eq. (4), with a fixed 3-minute overhead per session to discourage micro-charging.

**(4)**

**(4)**

**(4)**

Episodes terminate upon reaching the destination with SoC above reserve or upon depletion (stranding). Constraints mask invalid actions (e.g., revisiting the same station, charging immediately again before a cooldown) and limit the number of charges per trip. All random processes use fixed seeds for reproducibility.

Table 1: Environment Design Trade-Offs

|  |  |  |  |
| --- | --- | --- | --- |
| Aspect | Real-world baseline | Simulation choice | Rationale |
| Initial SoC | Often >50% | Sampled 10–30% | Ensures charging is frequently required, avoids trivial trips |
| Trip lengths | Many <10 km | Calibrated 12–25 km | 30–60% of trips require ≥1 charge, providing a learning signal |
| Charging | Overheads vary | Fixed 3-min per session | Penalises “nibbling” charges and station hopping |
| Traffic | Complex congestion patterns | SUMO microsim (vs. constant speed or multipliers) | Provides realistic congestion while keeping simulation deterministic |
| Station use | Drivers may revisit stations | No repeats, cooldowns, max charges | Prevents unrealistic cycling behaviour |
| Variability | High randomness in trips | Fixed seeds (environment + training) | Enables reproducibility and controlled comparisons |

### 3.3 Reward Design

We support three objectives: minimise cost, minimise time, and a hybrid of both using a value-of-time conversion. Per-step rewards are given in Eqs. (5–7).

**Cost objective (per step)**

**(5)**

**Time objective (per step)**

**(6)**

**Hybrid objective with value-of-time**

**(7)**

Potential-based shaping (Eq. (8)) grants a small bonus for reducing ETA between steps without changing the optimal policy.

**Potential-based shaping and penalties**

**(8)**

**(8)**

**(8)**

All charge steps are net-negative (time + energy + overhead). Terminal bonuses/penalties reinforce feasibility: a significant positive reward on success and a large penalty on stranding. Small penalties discourage invalid choices and micro-actions.

Table 2: Summary of Design Choices

|  |  |  |
| --- | --- | --- |
| **Design choice** | **Implementation** | **Rationale** |
| **Success / failure signals** | +50 on trip completion, –200 on depletion | Rewards feasibility and strongly penalises stranding |
| **Infeasible actions** | –2 penalty per invalid station choice | Discourages wasted steps while keeping exploration possible |
| **Charging cost** | All charge steps net-negative (time + energy + overhead) | Ensures charging is necessary but never intrinsically rewarding |
| **Potential shaping** | ETA-based potential function | Provides denser feedback without biasing against cost minimisation |
| **Anti-dither penalties** | Small penalties for idle drive and micro-charges | Prevents oscillatory behaviour and “nibbling” charges |
| **Charging overhead** | Fixed 3 min per session | Reflects real-world setup delays and discourages frequent short sessions |
| **Efficiency & limits** | η=0.92\eta=0.92, connector caps | Models technical constraints, making some cheap stations unattractive |
| **Hybrid scaling** | Value of time £0.05/min | Keeps cost and time comparable, avoiding dominance of one metric |

### 3.4 Proximal Policy Optimization

We train policies with PPO (Stable‑Baselines3; no algorithmic modifications). The clipped objective and GAE follow Eqs. (9–10). Hyperparameters (rollout length, batch size, learning‑rate schedule, clipping, advantage settings, and network) are reported in Appendix A (Table A.1); training uses a single‑env DummyVecEnv, fixed seeds, and a KPI logger.

**Clipped surrogate objective**

**(9)**

**(9)**

**Generalized Advantage Estimation (GAE)**

**(10)**

Evaluation uses a fixed test set and seeds to compare cost, time and hybrid-optimised agents under identical conditions.

# 4 Training & Evaluation

# References

[1] Mullan, M., Harries, D., Bräunl, T., & Whitely, S. (2018). The technical, economic and commercial viability of the electric vehicle to grid concept. Energy Policy, 109, 403–417.

[2] Kiani, H., Hesami, K., Azarhooshang, A., Pirouzi, S., & Safaee, S. (2021). Adaptive robust operation of the active distribution network including renewable and flexible sources. Sustainable Energy, Grids and Networks, 26, 100476.

[3] Weckx, S., & Driesen, J. (2015). Load balancing with EV chargers and PV inverters in unbalanced distribution grids. IEEE Transactions on Sustainable Energy.

[4] Xydas, E., Marmaras, C., & Cipcigan, L. (2016). A data-driven approach for characterising the charging demand of electric vehicles: A UK case study. Applied Energy, 162, 763–771.

[5] Qian, T., Shao, C., Wang, X., & Shahidehpour, M. (2019). Deep reinforcement learning for EV charging navigation by coordinating smart grid and intelligent transportation system. IEEE Transactions on Smart Grid, 11(2), 1714–1723.

[6] Han, Y., Li, T., & Wang, Q. (2024). A DQN-based approach for large-scale EVs charging scheduling. Complex & Intelligent Systems, 10, 8319–8339.

[7] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimisation algorithms. arXiv preprint arXiv:1707.06347.

[8] Sovacool, B. K., Noel, L., Axsen, J., & Kempton, W. (2018). The neglected social dimensions to a vehicle-to-grid (V2G) transition: a critical and systematic review. Environmental Research Letters, 13(1), 013001.

[9] Nicholas, M., & Lutsey, N. (2020). Quantifying the EV charging infrastructure gap in the UK. ICCT Report.

[10] Transport for London. (2021). London’s 2030 Electric Vehicle Infrastructure Strategy. TfL.

[11] Robinson, C., Blythe, P., Bell, M., Hübner, Y., & Hill, G. (2013). Analysis of electric vehicle driver recharging demand profiles and subsequent impacts on the electricity distribution network. Energy Policy, 61, 337–348.

[12] Figenbaum, E., & Kolbenstvedt, M. (2016). Learning from Norwegian battery electric and plug-in hybrid vehicle users: Results from a survey of vehicle owners. Institute of Transport Economics Report.

[13] Hardman, S., Jenn, A., Tal, G., Axsen, J., Beard, G., Daina, N., ... & Turrentine, T. (2018). A review of consumer preferences of and interactions with electric vehicle charging infrastructure. Transportation Research Part D, 62, 508–523.

[14] Zavvos, E., Gerding, E. H., & Brede, M. (2021). A comprehensive game-theoretic model for EV charging station competition. IEEE Transactions on Intelligent Transportation Systems.

[15] Lin, R., Chu, H., Gao, J., & Chen, H. (2024). Charging management and pricing strategy of electric vehicle charging station based on mean field game theory. Asian Journal of Control, 26(2), 803-813.

[16] Montoya, A., Guéret, C., Mendoza, J. E., & Villegas, J. G. (2017). The electric vehicle routing problem with nonlinear charging function. Transportation Research Part B, 103, 87–110.

[17] Froger, A., Mendoza, J. E., Jabali, O., & Laporte, G. (2019). Improved formulations and algorithmic components for the EV routing problem with nonlinear charging functions. Computers & Operations Research, 104, 256–294.

[18] Luo, C., Huang, Y. F., & Gupta, V. (2019). Stochastic dynamic programming for EV charging station placement in urban areas. IEEE Transactions on Smart Grid, 10(2), 2272–2282.

[19] He, F., Wu, D., Yin, Y., & Guan, Y. (2013). Optimal deployment of public charging stations for plug-in hybrid electric vehicles. Transportation Research Part B: Methodological, 47, 87–101.

[20] García-Álvarez, J., González, M. A., & Vela, C. R. (2018). Metaheuristics for solving a real-world electric vehicle charging scheduling problem. Applied Soft Computing, 65, 292–306.

[21] Perera, A. T. D., & Kamalaruban, P. (2021). Applications of reinforcement learning in energy systems. Renewable and Sustainable Energy Reviews, 137, 110618.

[22] Xu, B., Rubenis, A., & Long, C. (2024). Reinforcement learning based smart charging for electric vehicle fleet. In International Symposium on Intelligent Technology for Future Transportation (pp. 375–383).

[23] Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., & Wu, Y. (2022). The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games. NeurIPS 2022.

[24] Brockman, G., et al. (2016). OpenAI Gym. arXiv preprint arXiv:1606.01540.

[25] Behrisch, M., Bieker, L., Erdmann, J., & Krajzewicz, D. (2011). SUMO – Simulation of Urban Mobility: An Overview. Proceedings of SIMUL 2011, 55–60.

[26] Gunasekaran, R., Mohanraj, M.R., Senthilkumar, R., & Kamalakannan, R.S. (2025). Enhancing electric vehicle charging station utilization by reducing users waiting times through WOA-GLNN approach. Electrical Engineering.

[27] Zhang, L., Ji, Y., Li, X., Huang, Z., Cui, D., Chen, H., Gong, J., Breer, F., Junker, M., & Sauer, D.U. (2025). Multi-objective charging scheduling for electric vehicles at charging stations with renewable energy generation. Green Energy and Intelligent Transportation, 4(4), 100283.

[28] Cui, F., Lin, X., Zhang, R., & Yang, Q. (2022). Multi-objective optimal scheduling of charging stations based on deep reinforcement learning. Frontiers in Energy Research, 10, 1042882.

[29] Lin, R., Chu, H., Gao, J., & Chen, H. (2025). A Multi‑Objective Evolutionary Reinforcement Learning framework for optimizing EV charging strategies. Applied Energy, in press.

[30] Jia, R., Gao, K., Cui, S., et al. (2025). A multi‑objective RL‑based velocity optimization for electric trucks, considering battery degradation mitigation. Transportation Research Part E, in press.

[31] Yan, Z., Liu, Q., Wang, J., & Li, B. (2025). Deep reinforcement learning-based plug-in EV charging scheduling under uncertainty: a constrained SAC vs DDPG comparison. Energy.

[32] Visaria, A.A., Jensen, A.F., Thorhauge, M., Mabit, S.L. (2022). User preferences for EV charging, pricing schemes, and charging infrastructure. *Transportation Research Part A*, 165, 120–143.

[33] Baek, K., Lee, E., Kim, J. (2024). A dataset for multi-faceted analysis of electric vehicle charging transactions. *Scientific Data*, 11, 262.

[34] Zhan, W., Liao, Y., Deng, J., Wang, Z., Yeh, S. (2025). Large-scale empirical study of electric vehicle usage patterns and charging infrastructure needs. *npj Sustainable Mobility and Transport*, 2, 9.

[35] Alam, M.R., Guo, Z. (2023). Charging infrastructure planning for ride-sourcing electric vehicles considering drivers’ value of time. *Transportation Letters*, 15(6), 573–583.

[36] Kemper, N., Heider, M., Pietruschka, D., Hähner, J. (2025). Multi-objective and neuroevolutionary reinforcement learning for electric vehicle charging and load management. *Applied Energy*, 391, 125890.

[37] Dorsey, E., et al. (2025). Optimizing EV charging infrastructure deployment considering value-of-time. *Working Paper*.

[38] Lv, K., Pei, X., Chen, C., Xu, J. (2022). A safe and efficient lane change decision-making strategy of autonomous driving based on deep reinforcement learning. *Mathematics*, 10(9), 1551.

[39] Dorokhova, M., Ballif, C., Wyrsch, N. (2021). Routing of electric vehicles with intermediary charging stations: a reinforcement learning approach. *Frontiers in Big Data*, 4, 586481.

[40] Hu, Z., Wan, K., Gao, X., Zhai, Y. (2019). A dynamic adjusting reward function method for deep reinforcement learning with adjustable parameters. *Mathematical Problems in Engineering*, 2019, 7619483.

[41] Liu, S., Wang, Y., Chen, X., Fu, Y., Di, X. (2022). SMART-eFlo: An integrated SUMO-Gym framework for multi-agent reinforcement learning in electric fleet management. In *Proc. 25th IEEE Int. Conf. Intelligent Transportation Systems (ITSC)*, 2884–2890. IEEE.

**Appendix A: PPO Hyperparameters**

Table A.1 lists the hyperparameters used in the training script. Values reflect the exact settings in the code; they can be adjusted for ablation without changing the main text.

**Table A.1 PPO training hyperparameters**

|  |  |  |
| --- | --- | --- |
| Parameter | Value | Notes |
| Vector env | DummyVecEnv (1 env) | Single-process training |
| n\_steps | 4096 | Rollout length per update |
| batch\_size | 2048 | Minibatch size |
| n\_epochs | 10 | Optimisation epochs per update |
| gamma | 0.995 | Discount factor |
| gae\_lambda | 0.95 | GAE parameter |
| clip\_range | 0.2 | Policy clip |
| clip\_range\_vf | 0.2 | Value function clip |
| ent\_coef | 0.01 | Entropy regularisation |
| vf\_coef | 0.7 | Value loss weight |
| target\_kl | 0.02 | Early‑stop heuristic |
| max\_grad\_norm | 0.5 | Gradient clipping |
| learning rate | Linear decay from 1e‑4, clamp to **7.5e−5** after 100k steps | Clamp to 7.5e‑5 after 100k steps |
| policy network | MLP [256, 256] (actor & critic) | ReLU activations; orthogonal init |
| seeding | Fixed seeds (env + PPO + numpy/torch) | Reproducibility |
| logging | Episode KPI callback | Writes cost, time, success/stranding, charges |