Methodology

**Data Preparation**

The methodology relied on four structured data components: tariffs, charging stations, vehicle specifications with charging curves, and simulated user demands.

**Tariff data** were collected directly from operator websites (SureCharge, Believ etc.). The raw data exhibited heterogeneous formats, including per-kWh energy rates, session fees, idle charges, and membership discounts etc.

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

To enable consistent cost computation, we converted all rates to £/kWh and decomposed the tariffs into three structured datasets. This ensured realistic and provider-specific cost estimation in the simulation environment.

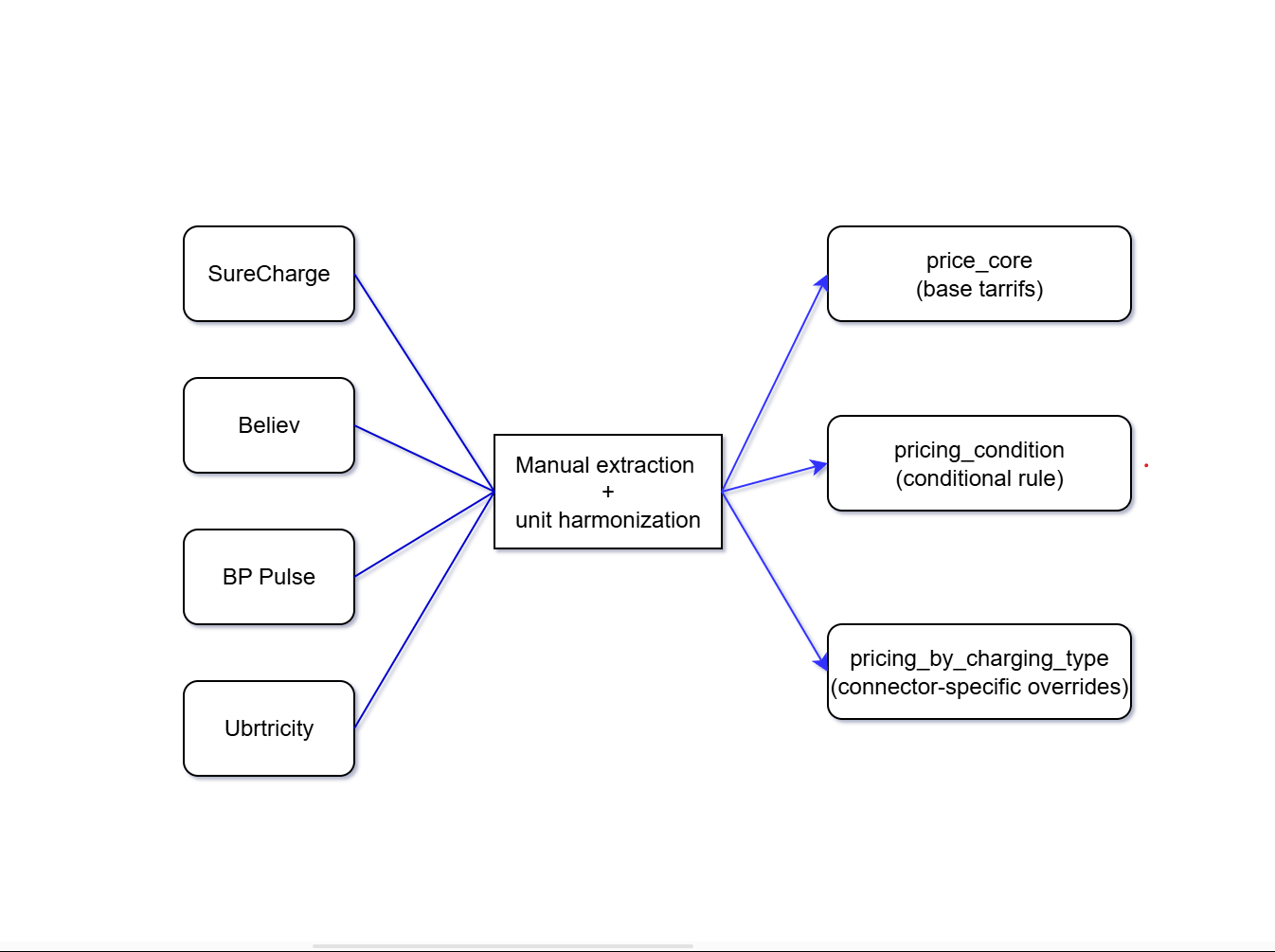


Figure 2 Conversion Process

**Charging station data** were obtained from the UK National ChargePoint Registry (NCR) [1]. After removing duplicates and fixing coordinate errors, we split the dataset into two layers: station-level metadata (for spatial queries) and connector-level features (for action feasibility in the RL environment).

A diagram of a diagram

AI-generated content may be incorrect.

Figure 3 Splitting the Charging Data

**Vehicle specifications** were sourced from Open EV Data v2 [2]. Charging curves were interpolated over 0–100% state of charge (SoC), expressed in kWh/km, and capped by station power limits to model charging times and SoC changes.

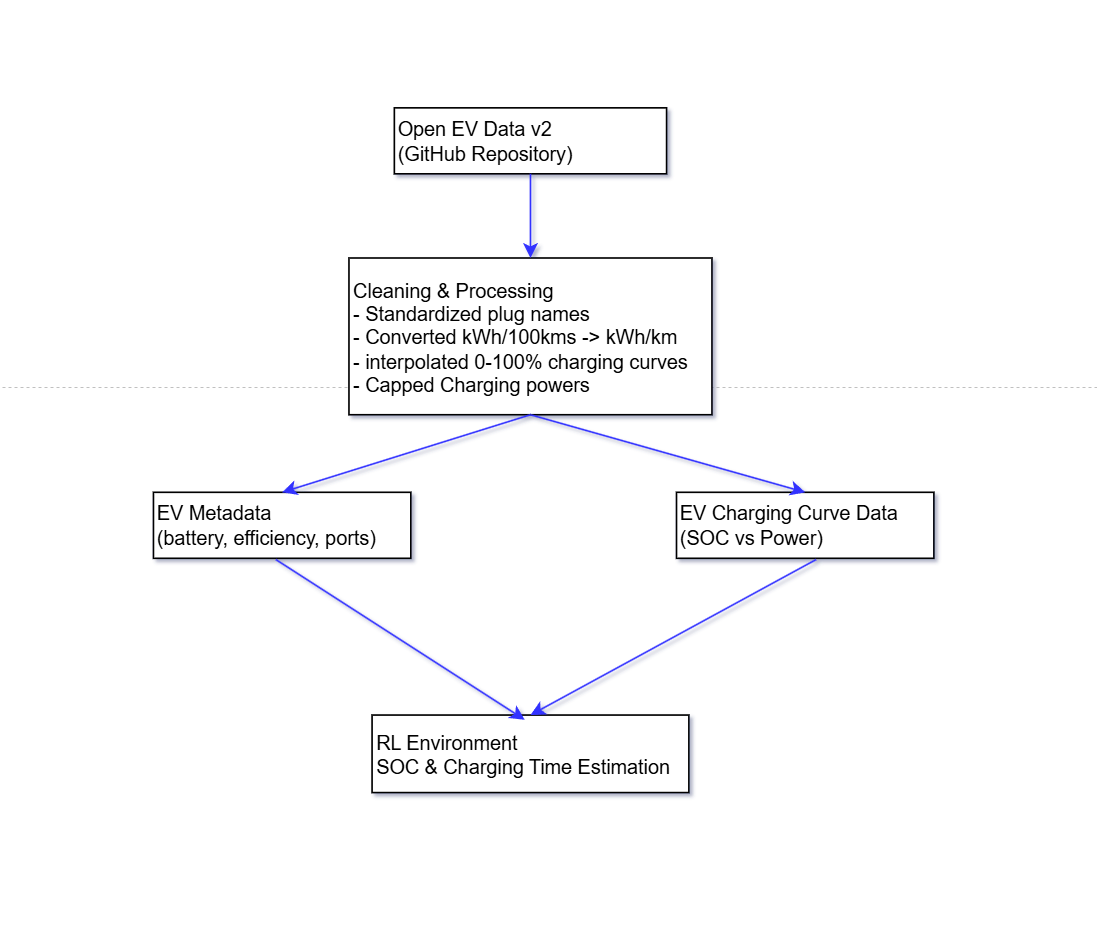


Figure 4 Splitting the EV Data

**Simulated users** were generated as structured trip instances combining user profiles, EV states, trip definitions, and behavioural constraints (see Table 1).

|  |  |
| --- | --- |
| **Feature Category** | **Examples** |
| User Profile | Membership type (Member/Payg), Subscription status, Sessions per month |
| Trip Definition | Origin-Destination coordinates, Departure time, Trip distance |
| EV State | Assigned EV model, Start SoC, Reserve SoC, Available energy |
| Preferences & Constraints | Optimization objective, max detour km |
| Energy Efficiency | Consumption in kWh/km, estimated required kWh per trip |

Table 1 Feature Categories & Examples

Together, these four components provided a unified and reproducible dataset for training and evaluating the RL agents. By integrating these, the environment captures both system-level constraints and realistic user behaviour.

**Environment Design**

The charging decision-making task was modelled as a Markov Decision Process (MDP),

where **S** is the state space, **A** the action space, **P** the transition dynamics, **R** the reward function, and the discount factor.

**State representation.** Each state contained the vehicle’s SoC, position, remaining distance, and features of up to k candidate stations (connector type, power, tariff, detour). Telemetry such as visited stations and last charging time was also tracked for analysis.

**Action space.** The agent could select one of the candidate stations or continue driving. Invalid options (repeat visits, cooldown violations, exceeding charge limits) were masked before routing, with a low-SoC override ensuring stations were always available in emergencies

**Transitions.** Driving reduced SoC in proportion to distance:

Where is energy consumption andbattery capacity. Travel time was obtained from SUMO’s microscopic simulation:

**.**

Charging followed nonlinear battery curves capped by connector power, with efficiency and a fixed 3-minute overhead:

**Termination.** Episodes ended when the destination was reached or if SoC fell below a reserve threshold

**Constraints and reproducibility.** Rules limited charges per trip, enforced minimum gaps between sessions, and prohibited repeat stations. Fixed seeds across environment and training components ensured reproducibility of results.

**Design trade-offs.** Simplifications were introduced to ensure meaningful charging decisions. Table 2 summarises the key adjustments and their effects.

|  |  |  |  |
| --- | --- | --- | --- |
| **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 |

Table 2 Environment Design Trade-offs

Together, these design elements produced an RL environment that was both computationally feasible and sufficiently realistic to evaluate charging strategies under urban driving conditions. While simplified in some respects, the use of SUMO-backed traffic and structured user demands ensured that the environment captured the essential challenges of EV route planning in Inner London.

**Reward Design**

The environment supported three optimization objectives. Terminal bonuses and penalties enforced feasibility, while charging was strictly net-negative. Potential-based shaping rewarded driving progress, and small anti-dither penalties reduced oscillatory behaviours. Table 3 summarises the key design choices, their implementation, and the rationale for each.

**1. Cost minimisation**

where ​ is the charging session cost in GBP.

**2. Time minimisation**

where is detour time and is charging duration.

**3. Hybrid objective**

Where is total journey time and cost in GBP, with denominators normalising scales to prevent dominance.

**4. Shaping.** To provide denser feedback, potential-based shaping rewarded reductions in estimated time-to-arrive (ETA):

Where  **. 60**. Shaping was neutral during charging, so only forward progress yielded positive signals.

**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 |

Table 3 Summary of Design Choices

**Proximal Policy Optimization**

Policy learning used **Proximal Policy Optimisation (PPO)** from Stable-Baselines3, chosen for its stability and efficiency. The algorithm was left unmodified; the study’s contribution lay in environment, reward, and evaluation design.

The clipped surrogate objective is defined as:

Where is the probability ratio and is the advantage estimate.

Advantages were computed using Generalised Advantage Estimation (GAE) {CITATION NEEDED}:

Where is the temporal-difference error.

**Training setup.** Fixed seeds were applied across PPO, numpy, torch, and the environment to ensure reproducibility. Short smoke runs confirmed that KPI trajectories matched baseline expectations, providing a stable reference point for later experiments.

**Design choice rationale.**

* **PPO** was chosen for robustness and wide adoption in RL research.
* **No algorithm modifications** kept the focus on environment and reward design rather than algorithmic novelty.
* **GAE** reduced variance in advantage estimates, accelerating learning.
* **Reproducibility controls** ensured fair comparisons across reward definitions and environment settings.

**Evaluation overview.** All trained policies were assessed on fixed evaluation subsets under consistent seeds, tracking key performance indicators such as success rate, journey time, and charging cost. Full evaluation design and results are presented in Section 4.

Overall, the methodology combined structured datasets, a custom RL environment, carefully designed rewards, and reproducible PPO training, enabling the evaluation of EV charging strategies under realistic Inner London traffic conditions.

References

1. UK Department for Transport: National Chargepoint Registry (NCR) dataset [archived version, supplied on request]. (2025). Provided to author via email communication, March 2025.
2. KilowattApp: Open EV Data v2. GitHub. Available at: https://github.com/KilowattApp/open-ev-data/tree/master/data/v2 (n.d.).
3. Aslan Yıldız, Ö.; Sarıçiçek, İ.; Yazıcı, A.: A Reinforcement Learning-Based Solution for the Capacitated Electric Vehicle Routing Problem from the Last-Mile Delivery Perspective. Appl. Sci. 15(3), 1068 (2025). <https://doi.org/10.3390/app15031068>