

Document Version

Final published version

Licence

CC BY

Citation (APA)

Orfanoudakis, S., Oliehoek, F. A., Palensky, P., & Vergara, P. P. (2026). Physics-informed reinforcement learning for large-scale EV smart charging considering distribution network voltage constraints. *Applied Energy*, 421, Article 128224. <https://doi.org/10.1016/j.apenergy.2026.128224>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

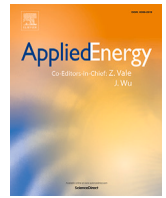
In case the licence states "Dutch Copyright Act (Article 25fa)", this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.



Physics-informed reinforcement learning for large-scale EV smart charging considering distribution network voltage constraints

Stavros Orfanoudakis^{a,*}, Frans A. Oliehoek^b, Peter Palensky^a, Pedro P. Vergara^a

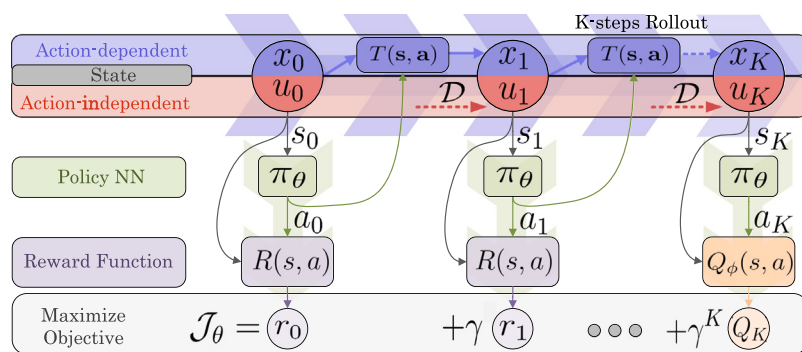
^a Delft University of Technology, Intelligent Electrical Power Grids Group, Delft, the Netherlands

^b Delft University of Technology, Sequential Decision Making Group, Delft, the Netherlands

HIGHLIGHTS

- Introduces a differentiable EV charging formulation with voltage-aware power flow.
- Embeds grid physics and battery dynamics directly into TD3 policy learning.
- Achieves faster learning and stronger voltage regulation under uncertainty.
- Scales physics-informed RL to city-scale EV charging networks.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:

Electric vehicles
Distribution networks
Physics-informed
Reinforcement learning
Vehicle-to-grid

ABSTRACT

Electric Vehicles (EVs) offer substantial flexibility for grid services, yet large-scale, uncoordinated charging can threaten voltage stability in distribution networks. Existing Reinforcement Learning (RL) approaches for smart charging often disregard physical grid constraints or have degraded performance for complex, large-scale tasks, limiting their scalability and real-world applicability. This paper introduces a physics-informed (PI) RL algorithm that integrates a differentiable power flow model and voltage-based reward design into the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm, enabling RL to better scale to a large number of EVs and deliver real-time voltage support while meeting EV user demands. The resulting PI-TD3 algorithm achieves faster convergence, improved sample efficiency, and reliable voltage magnitude regulation under uncertain and overloaded conditions. Benchmarks on the IEEE 34-bus and 123-bus networks show that the proposed PI-TD3 outperforms both model-free RL and optimization-based baselines in grid constraint management, user satisfaction, and economic metrics, even as the system scales to hundreds of EVs. These advances enable robust, scalable, and practical EV charging strategies that enhance grid resilience and support the operation of distribution networks.

1. Introduction

The growing penetration of electric vehicles (EVs) into distribution networks introduces significant challenges and opportunities for grid voltage regulation [1]. Uncoordinated EV charging can intensify

voltage violations, posing risks to grid stability, particularly during peak demand periods [2]. Meanwhile, coordinated EV charging strategies transform EVs into controllable distributed energy resources capable of alleviating these issues, especially by leveraging vehicle-to-grid (V2G)

* Corresponding author.

Email address: s.orfanoudakis@tudelft.nl (S. Orfanoudakis).

<https://doi.org/10.1016/j.apenergy.2026.128224>

Received 15 December 2025; Received in revised form 20 March 2026; Accepted 8 June 2026

Available online 15 June 2026

0306-2619/© 2026 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Table 1
Comparison of EV charging control methods under voltage and grid constraints.

Reference	Method	V2G	Grid Constraints	Comments	Grid	EV Chargers
[4]	Droop control	No	Phase voltage, unbalance	Cuts unbalance, reactive-only	10-bus	43
[5]	Metaheuristic Opt.	No	Voltage, THD	Improves voltage, needs data	IEEE 33-bus	10
[6]	Metaheuristic Opt.	Yes	Frequency, voltage	Better stability, offline retuning	5-bus	2
[7]	MPC	Yes	Frequency, voltage	Predictive control, high complexity	IEEE 39-bus	5
[8]	MPC	Yes	Frequency (islanded)	Inertia-like support, charging delays	–	1
[9]	MPC	Yes	Voltage, power limits	Lowers cost, needs aggregator/comms	8-bus	3
[11]	RL	No	Voltage	Fast coordination, training needed	IEEE 123-bus	–
[12]	RL	Yes	Voltage, generator limits	Cost + voltage co-optim., tuning burden	IEEE 33-bus	5
[13]	RL	Yes	Voltage, transformer limit	Protects grid/users, simplified env.	IEEE 33-bus	1
[14]	Safe RL	Yes	Voltage constraints	Explicit safety, higher complexity	IEEE 33-bus	4
[17]	Multi-agent RL	Yes	Transformer thermal	Reduces aging, complex training	1-bus	64
[19]	RL	No	Voltage magnitude	PI repr. learning	IEEE 119-bus	0
[20]	Physics-Informed RL	No	Voltage control	Transient voltage	IEEE 9-bus	0
[21]	Safe RL	No	Voltage control	PI exploration	–	1
[22]	RL	Yes	Power limits	Graph representation	–	1000
Ours	Physics-Informed RL	Yes	Voltage magnitude	Scalable and efficient	IEEE 123-bus	500

operation [3]. This dual potential highlights the crucial role of well-managed EV charging schemes, which can simultaneously maintain grid stability and facilitate the integration of renewable energy resources.

Early studies on voltage regulation via EV coordination typically employed heuristic methods [4] or stochastic mathematical optimization techniques. Examples include artificial bee colony optimization [5] and model predictive control (MPC), which schedule EV charging by accounting for uncertainties in renewable generation and load conditions [6]. Although MPC methods leverage forecasts to maintain voltage and frequency stability proactively, their effectiveness is often limited by uncertainty quantification inaccuracies [7] and computational complexity [8]. Additionally, pricing-based mechanisms have been proposed to incentivize EV charging behaviors beneficial for voltage support [9]. However, such approaches generally involve extensive offline computations and face scalability challenges, particularly when managing large EV fleets in real-time operational conditions.

To address scalability and modeling complexity challenges, reinforcement learning (RL) methods have emerged as effective solutions capable of real-time decision-making even for complex optimization tasks [10]. Model-free RL algorithms, such as Deep Q Networks, have been successfully applied in a two-layer framework to jointly optimize EV charging and Volt-VAR control [11]. Additionally, continuous-action algorithms like Deep Deterministic Policy Gradient (DDPG) have efficiently managed EV fleet charging while explicitly considering distribution network voltage stability [12] and incorporating battery degradation effects due to vehicle-to-grid (V2G) operations [13]. Safe RL approaches integrate constraints in the training process to enhance grid reliability and constraint satisfaction [14], whereas model-based RL techniques leverage learned transition dynamics to improve overall decision quality [15]. Furthermore, multi-agent RL strategies enable decentralized voltage control across network buses [16] and facilitate EV charging optimization to extend transformer lifetime [17]. Despite the advances summarized in Table 1, RL methods remain limited by high-dimensional, constrained, and stochastic decision spaces that degrade sample efficiency and reliability at large-scale [18], while classic optimization suffers from combinatorial explosion and nonconvex network physics, limiting tractable deployment in large EV charging systems.

Physics-informed learning methods have become prominent for enhancing machine learning robustness and accuracy by directly embedding domain-specific physical equations and constraints into training processes [23]. For instance, Physics-Informed Neural Networks (PINNs) explicitly embed domain-specific equations related to EV dynamics, such as battery state-of-charge (SoC) and power consumption [24,25], as well as networks' power flow equations [26], resulting in accurate supervised learning predictions, even with limited training data. Recent advancements have extended these techniques to complex spatiotemporal prediction tasks, such as citywide EV charging demand forecasting and dynamic pricing through physics-informed graph learning [27]. Similarly, graph neural networks combined with deep RL have leveraged physics-informed graph attention networks to address robust voltage control challenges under partial observability [19]. Moreover, other studies have integrated physics-based constraint layers into RL for transient voltage control [20], distributed voltage regulation using photovoltaic inverters [28], and enforcing safety constraints in action selection [21]. However, to the best of our knowledge, no existing approach explicitly embeds distribution network power flow and EV battery state-of-charge dynamics into the learning objective and parameter updates of an RL algorithm. Instead, prior works typically enforce physical constraints through action projections, constraint layers, or penalty-based reward shaping, resulting in a weak coupling between grid physics and the learning process. In contrast, for a method to be genuinely physics-informed, the governing physical equations, such as power flow, must be differentiable and directly integrated into the optimization pipeline, enabling backpropagation through the system dynamics. Approaches that incorporate physics solely through the reward function, therefore, do not constitute physics-informed RL (PI-RL) in a strict sense.

To overcome the scalability and efficiency shortfalls of existing methods (see Table 1), we propose an actual PI-RL algorithm¹ tailored to city-scale EV charging while supporting the distribution network's voltage

¹ Open-sourced code at: https://github.com/StavrosOrf/EV2Gym_PI-TD3, and https://github.com/distributionnetworksTUDelft/EV2Gym_PI-TD3.

magnitude limits. Rather than imposing physics through penalties or action filters, the proposed PI-RL embeds the power flow formulation and battery SoC dynamics into the training rollouts and reward. By embedding the power flow formulation via differentiable reward signals directly into the learning process, the algorithm obtains richer gradient information, accelerating convergence and improving constraint satisfaction, thereby making it more scalable to real-world EV dispatch problems. In detail, the proposed physics-informed Twin Delayed DDPG (PI-TD3) RL algorithm achieves higher sample efficiency, faster convergence, and fewer voltage magnitude violations in stochastic settings. Extensive experiments in EV2Gym [29] on benchmark IEEE distribution networks show that PI-TD3 scales to larger EV fleets while outperforming RL and optimization baselines, thereby enabling credible, real-time coordination of city-scale EV charging. The primary contributions of this work can be summarized as follows:

- A differentiable formulation of the EV smart charging problem with distribution network voltage magnitude constraints is developed, in which power-flow equations and EV battery state-of-charge dynamics are explicitly modeled in a form suitable for gradient-based optimization. In contrast to conventional model-free RL formulations, this enables direct gradient propagation with respect to charging actions while preserving the physical structure of the charging problem.
- The PI-TD3 algorithm is introduced by integrating the differentiable EV charging formulation into the RL training process through physics-aware rollouts. Extensive experiments on IEEE 34-bus and 123-bus distribution networks with up to hundreds of EVs demonstrate improved voltage magnitude regulation, faster convergence, and higher sample efficiency compared to model-free RL and optimization-based baselines.

2. The optimal EV charging problem

In this section, the optimal EV charging problem is formalized as both a mixed-integer nonlinear programming (MINLP) problem and a Markov decision process (MDP). These formulations capture the objectives of smart EV charging while limiting voltage magnitude violations.

2.1. Mathematical programming formulation

The optimal EV charging problem investigated in this work is formulated on a distribution network consisting of N buses, with the network topology described by the bus admittance matrix $\mathbf{Y} \in \mathbb{C}^{N \times N}$, as illustrated in Fig. 1. The problem is simulated over a discrete time horizon of T steps, $t \in \mathcal{T} = \{1, \dots, T\}$. At each time step, the Charge Point Operator (CPO) determines the charging and discharging power, $p_{i,t}^{\text{ch}}$ and $p_{i,t}^{\text{dis}}$, for each charging station $i \in \mathcal{I}$. Charging stations are geographically distributed and grouped according to the buses to which they are connected, indexed by $n \in \mathcal{N} = \{1, \dots, N\}$. For each bus $n \in \mathcal{N}$, let $\mathcal{I}_n \subset \mathcal{I}$ denote the set of charging stations associated with that bus. The distribution network is described by the admittance vector $\mathbf{Z} \in \mathbb{C}^N$ and the reduced admittance matrix $\mathbf{L} \in \mathbb{C}^{N \times N}$.

During operation, the Distribution System Operator (DSO) provides the CPO with real-time information including, active and reactive demands $p_{n,t}^L$ and $q_{n,t}^L$, as well as photovoltaic (PV) generation $p_{n,t}^{\text{PV}}$ at every bus $n \in \mathcal{N}$. Although the CPO does not have access to forecasts of future EV arrivals or network states, each EV, upon arrival at charging station i , communicates its expected departure time t_i^d and desired battery capacity at departure e_i^* , an assumption that is tested and approved by EV users given monetary incentives [30]. The real-time battery energy $e_{i,t}$ for every connected EV is assumed known, as is standard in V2G-enabled charging communication protocols [31].

Within this setting, the CPO seeks to maximize profit, satisfy user charging needs, while minimizing voltage magnitude violations and enforcing battery operational constraints. The expression in (1) defines the optimization objective: the first term penalizes voltage deviations

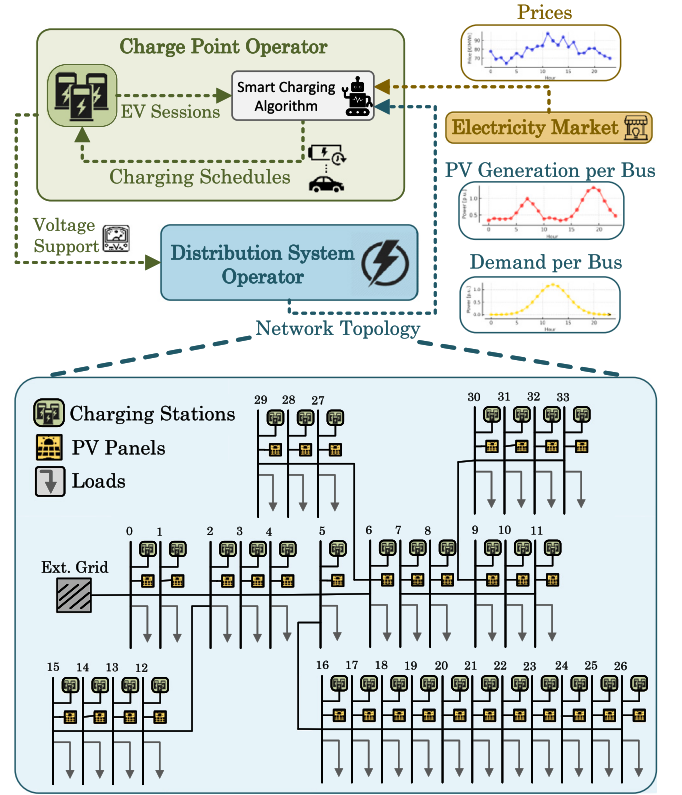


Fig. 1. Overview of the proposed problem setting illustrating an example distribution network with V2G charging stations, PV generation, and dynamic demand. The DSO provides real-time grid topology, demand, and PV generation data, while the CPO coordinates the charging of hundreds of EVs based on energy prices, respecting user constraints and enabling distribution network voltage support.

at each bus, the second term accounts for net profit from charging and discharging, and the third term penalizes deviations from user-specified energy requirements. Each term is weighted (λ_1 , λ_2 , and λ_3) to balance the priority of each objective. Therefore, the overall EV charging optimization problem is defined as:

$$\begin{aligned} \min_{p^{\text{ch}}, p^{\text{dis}}} \sum_{t \in \mathcal{T}} \left\{ \lambda_1 \sum_{n \in \mathcal{N}} \min \left[0, 0.05 - \left| 1 - V(p_{i,t}^{\text{ch}}, p_{i,t}^{\text{dis}})_{n,t} \right| \right] \right. \\ \left. + \sum_{i \in \mathcal{I}} \left[\lambda_2 \Delta t \left(\Pi_t^{\text{ch}} p_{i,t}^{\text{ch}} \omega_{i,t}^{\text{ch}} - \Pi_t^{\text{dis}} p_{i,t}^{\text{dis}} \omega_{i,t}^{\text{dis}} \right) \right. \right. \\ \left. \left. + \lambda_3 \sum_{j \in \mathcal{J}_i} \left(\sum_{s=t_i^a}^{t_i^d} (p_{i,s}^{\text{ch}} \omega_{i,s}^{\text{ch}} - p_{i,s}^{\text{dis}} \omega_{i,s}^{\text{dis}}) - e_i^* \right)^2 \right] \right\} \end{aligned} \quad (1)$$

Subject to:

$$p_{n,t}^{\text{EV}} = \sum_{i \in \mathcal{I}_n} (p_{i,t}^{\text{ch}} \omega_{i,t}^{\text{ch}} - p_{i,t}^{\text{dis}} \omega_{i,t}^{\text{dis}}) \quad \forall n \in \mathcal{N}, \forall t \in \mathcal{T} \quad (2)$$

$$s_{n,t} = (p_{n,t}^L + p_{n,t}^{\text{PV}} + p_{n,t}^{\text{EV}}) + q_{n,t}^L j \quad \forall n \in \mathcal{N}, \forall t \in \mathcal{T} \quad (3)$$

$$v_{n,t}^{(0)} = 1 + 0j \quad \forall n \in \mathcal{N}, \forall t \in \mathcal{T}, \quad (4)$$

$$v_{n,t}^{(k+1)} = Z_n + \sum_{n' \in \mathcal{N}} L_{nn'} \left(\frac{s_{n',t}}{v_{n',t}^{(k)}} \right) \quad \forall k, n \in \mathcal{N}, t \in \mathcal{T} \quad (5)$$

$$V(p^{\text{ch}}, p^{\text{dis}})_{n,t} = v_{n,t}^{(K)} \quad \forall n \in \mathcal{N}, t \in \mathcal{T}, \quad (6)$$

$$e_i \leq e_{i,t} \leq \bar{e}_i \quad \forall i \in I, \forall t \in \mathcal{T} \quad (7)$$

$$e_{i,t} = e_{i,t-1} + (p_{i,t}^{\text{ch}} \omega_{i,t}^{\text{ch}} + p_{i,t}^{\text{dis}} \omega_{i,t}^{\text{dis}}) \cdot \Delta t \quad \forall i \in I, \forall t \in \mathcal{T} \quad (8)$$

$$e_{i,t} = e_i^a \quad \text{if } t = t_i^a \quad \forall i \in I, \forall t \in \mathcal{T} \quad (9)$$

$$p_{i,t}^{\text{ch}} \leq p_{i,t}^{\text{ch}} \leq \bar{p}_i^{\text{ch}} \quad \forall i \in I, \forall t \in \mathcal{T} \quad (10)$$

$$p_{i,t}^{\text{dis}} \leq p_{i,t}^{\text{dis}} \leq \bar{p}_i^{\text{dis}} \quad \forall i \in I, \forall t \in \mathcal{T} \quad (11)$$

$$\omega_{i,t}^{\text{ch}} + \omega_{i,t}^{\text{dis}} \leq 1 \quad \forall i \in I, \forall t \in \mathcal{T} \quad (12)$$

In the objective function (1), the decision variables $p_{i,t}^{\text{ch}}$ and $p_{i,t}^{\text{dis}}$ define the charging and discharging power of each EV, while Π_i^{ch} and Π_i^{dis} designate the electricity price per kWh. In particular, the first term is for maintaining every bus's voltage between 0.95 and 1.05 pu, hence the 0.005 and 1 numbers in (1). Constraint (2) defines the total EV charging and discharging power $p_{n,t}^{\text{EV}}$ injected at bus n and time t , as the sum over all charging stations $i \in I_n$, where $p_{i,t}^{\text{ch}}$ and $p_{i,t}^{\text{dis}}$ denote the charging and discharging power, and $\omega_{i,t}^{\text{ch}}$, $\omega_{i,t}^{\text{dis}}$ are their respective binary activation variables. The total complex power injection $s_{n,t}$ at each bus (3) aggregates active load $p_{n,t}^L$, PV generation $p_{n,t}^{\text{PV}}$, EV charging power $p_{n,t}^{\text{EV}}$, and reactive load $q_{n,t}^L$. The voltage magnitude $v_{n,t}$ at each bus is computed iteratively [32]: initialization is set by (4); (5) updates the voltage using the reduced grid admittance parameters $\mathbf{Z}$ and $\mathbf{L}$, the total complex power $s_{n,t}$, and the previous voltage iterate; the process repeats for κ iterations, yielding the final voltage magnitude $V(p^{\text{ch}}, p^{\text{dis}})_{n,t}$ in (6). EVs' battery dynamics are enforced by (7)–(9): $e_{i,t}$ denotes the battery energy of the EV parked at charger i at time t , bounded by minimum and maximum values e_i and $\bar{e}_i$. The state is updated each timestep according to charging/discharging actions (8) and initialized to e_i^a at arrival time t_i^a (9). Charging and discharging power limits, $p_{i,t}^{\text{ch}}$, $\bar{p}_i^{\text{ch}}$, $p_{i,t}^{\text{dis}}$, and $\bar{p}_i^{\text{dis}}$, are imposed by (10) and (11). Finally, (12) ensures that a charger cannot charge and discharge simultaneously.

2.2. Markov decision processes for EV charging

The optimal EV charging problem can also be formulated as an MDP ($\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}$). At each time step $t \in \mathcal{T}$, the state vector $\mathbf{s}_t$ is given by

$$\mathbf{s}_t = [\sin(h_t), \cos(h_t), \Pi_t^{\text{ch}}, \mathbf{p}_t, \mathbf{q}_t, \text{SoC}_t, \mathbf{t}_t^{\text{left}}, \mathbf{b}_t], \quad (13)$$

where $\sin(h_t)$ and $\cos(h_t)$ represent the hour (h) of the day as cyclical features. The net active power ($p_{n,t}^{\text{PV}} - p_{n,t}^L$), $\mathbf{p}_t = [p_{1,t}, \dots, p_{N,t}]$ and reactive power $\mathbf{q}_t = [q_{1,t}, \dots, q_{N,t}]$ injections at each bus $n \in \mathcal{N}$, $\text{SoC}_t = [\text{SoC}_{1,t}, \dots, \text{SoC}_{|I|,t}]$ contain the state-of-charge of each connected EV with $\text{SoC}_{i,t} = e_{i,t}/\bar{e}_i$, $\mathbf{t}_t^{\text{left}} = [t_{1,t}^{\text{left}}, \dots, t_{|I|,t}^{\text{left}}]$ is the vector of remaining time to departure for each EV with $t_{i,t}^{\text{left}} = t_i^d - t$, $\mathbf{b}_t = [b_1, \dots, b_{|I|}]$ specifies the bus index to which each charger is connected, and Π_t^{ch} is the current electricity price.

The action vector at each time step is $\mathbf{a}_t = [a_{1,t}, \dots, a_{|I|,t}]^T \in [-1, 1]^{|I|}$, where $a_{i,t}$ is the normalized charging action ($a_{i,t} > 0$) or discharging ($a_{i,t} < 0$) for EV i , with $a_{i,t} = 0$ denoting no action. The transition function $\mathcal{P}$ determines the evolution of the system based on the chosen actions, grid power flow, and other unknown system dynamics, such as EV arrivals, PV generation, and load profiles.

The reward function $R(\cdot)$ is designed to closely mirror the objective of the mathematical programming formulation in (1), balancing voltage magnitude regulation, energy costs, and user satisfaction. Specifically, the reward r_t is defined as the outcome of the reward function:

$$R(\mathbf{s}_t, \mathbf{a}_t) = \lambda_1 \sum_{n \in \mathcal{N}} \min\{0, 0.05 - |1 - V_{n,t}(\cdot)|\} + \sum_{i \in I} [\lambda_2 \Delta t (\Pi_i^{\text{ch}} p_{i,t}^{\text{ch}} - \Pi_i^{\text{dis}} p_{i,t}^{\text{dis}}) + \lambda_3 \cdot \psi_{i,t}], \quad (14)$$

where the first term penalizes voltage magnitude violations at each bus. Note that the voltage magnitude at bus n and step t is described by (2)–(6), and ultimately $V_{n,t}(\mathbf{a}_t)$ is a function of charging actions. The second

term represents the net revenue from charging and discharging activities based on electricity prices Π_i^{ch} and Π_i^{dis} , and $\psi_{i,t}$ is a user satisfaction term that incentivizes each EV to maintain a minimum SoC as it approaches its departure time. Unlike the third term in the original mathematical programming objective (1), which is sparse and penalizes the energy shortfall only at the departure time, the RL reward uses a denser shaping signal. Specifically, it encourages the EV to begin charging as the departure time approaches, rather than waiting until the final step. In this way, the reward provides more informative feedback during training (while keeping the same optimal objective) by reflecting whether the current SoC is adequate relative to the remaining time before departure. The reward term is defined as:

$$\psi_{i,t} = \max\{0, \text{SoC}^* - \text{SoC}_{i,t}\} \cdot \mathbb{I}[t_{i,t}^{\text{left}} < \epsilon], \quad (15)$$

where $\text{SoC}_{i,t}$ is the current state-of-charge of EV i at time t , SoC^* is a target minimum SoC (e.g., 90%), ϵ is a threshold defining the proximity to departure, and $\mathbb{I}[\cdot]$ is the indicator function. The $\psi_{i,t}$ term penalizes the agent if any EV approaches departure with insufficient SoC, thereby encouraging timely charging to meet user expectations by the time of departure, while also providing a dense training signal. Here, as in the expression (1), the coefficients λ_1 , λ_2 , and λ_3 are chosen to match the weighting of the respective terms in the original MINLP formulation. This consistency ensures that the physics-informed RL agent optimizes towards the same operational goals as the mathematical programming approach.

2.3. RL for EV charging

RL can solve sequential decision-making problems expressed as MDPs, such as the EV charging problem described above, by learning a policy π that maps the observed state of the system to charging or discharging actions [10]. The agent's objective is to maximize the expected cumulative reward, mathematically expressed as:

$$J(\pi) = \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t r_t \right], \quad (16)$$

where r_t denotes the reward at time t , γ is a discount factor, and the expectation is taken over the stochastic evolution of the environment under the policy π . Furthermore, the state-action value function, or Q-function, is central to RL representing the expected cumulative reward obtained by taking action $\mathbf{a}_t$ in state $\mathbf{s}_t$, and subsequently following a policy π . The Q-function is recursively defined by the Bellman equation:

$$Q(\mathbf{s}_t, \mathbf{a}_t) = r_t + \gamma \mathbb{E}_{\mathbf{s}_{t+1} \sim P(\cdot|\mathbf{s}_t, \mathbf{a}_t)} [Q(\mathbf{s}_{t+1}, \pi(\mathbf{s}_{t+1}))], \quad (17)$$

where r_t denotes the immediate reward, and $P(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$ is the transition probability between states. This recursive relationship links current and future value estimates, allowing RL algorithms to iteratively improve their policies using only sampled transitions and rewards. As a result, near-optimal charging strategies can be learned even when the underlying system dynamics are partially known.

3. Physics-informed RL for EV charging

Unlike the standard MDP formulation presented above, which disregards distribution network constraints, the proposed framework leverages a differentiable power-flow formulation to enable grid-aware, more scalable policy learning through model-based rollouts and gradient-based optimization.

3.1. From model-free to physics-informed RL

In classic model-free RL, the environment is treated as a black box, and state transitions are learned solely from sampled experience. In contrast, PI-RL leverages known, differentiable components of the system (physics), such as the SoC update for each EV. This allows for more

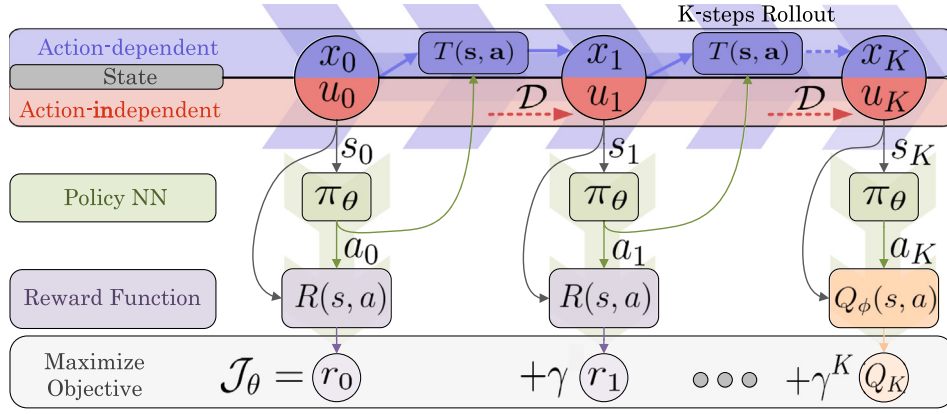


Fig. 2. The policy network π_θ generates actions that, together with the known environment dynamics (e.g., SoC and voltage updates) and sampled exogenous variables (e.g., loads, prices, and EV arrivals), are used to simulate K -step rollouts through the differentiable transition $T(s, a)$ and reward $R(s, a)$ functions. This enables direct gradient propagation from cumulative rewards back through the rollout.

accurate and efficient learning by directly modeling the underlying EV battery dynamics. To ensure that EV battery constraints (7) and (8) are satisfied at every step, the SoC transition update is implemented as a piecewise, differentiable function:

$$\text{SoC}_{i,t+1} = \begin{cases} 1, & \text{if } x_{i,t+1} > 1 \\ \text{SoC}_i, & \text{if } x_{i,t+1} < \text{SoC}_i \\ x_{i,t+1}, & \text{otherwise} \end{cases} \quad (18)$$

where $x_{i,t+1} = \text{SoC}_{i,t} + \frac{\Delta t a_{i,t} \bar{p}_{i,t}}{\bar{e}_i}$, $a_{i,t}$ is the charging/discharging action and SoC_i is the minimum SoC while doing V2G discharging. For clarity of presentation, the charging efficiency factor has been excluded from (18); nevertheless, it can be incorporated in a straightforward manner without necessitating any modification to the algorithm. Thus, (18) accurately and completely describes the EV battery transition given any charging action.

Some aspects of the transition are unknown or stochastic and are independent of the actions taken, e.g., future demand at each bus, electricity price signals, and the arrival and departure of new EVs. These elements are difficult to model or forecast, but can be sampled from historical data. In practice during the training phase, these unknown exogenous variables are sampled from a replay buffer $\mathcal{D}$, which stores past system trajectories, so that model-based rollouts are grounded in realistic scenarios. Therefore, as shown in Fig. 2, the state $s_t = \{\mathbf{x}_t, \mathbf{u}_t\}$ is defined as the combination of action-dependent variables $\mathbf{x} = \{\text{SoC}_t\}$, and the action-independent variables $\mathbf{u} = \{\sin(h_t), \cos(h_t), \Pi_t^{\text{ch}}, \mathbf{p}_t, \mathbf{q}_t, \mathbf{t}_t^{\text{left}}, \mathbf{b}_t\}$.

The reward function $R(s, \mathbf{a})$ for this problem is fully known and differentiable, as defined in (14). All variables required for its computation, such as voltage magnitudes, charging and discharging power, electricity prices, and user satisfaction terms, are either included in the state, sampled as exogenous variables, or, in the case of network parameters like the grid admittance matrix, remain constant and are hardcoded throughout the simulations. As a result, for any given state-action pair, the reward can be deterministically computed.

With the transition function $T(s, \mathbf{a})$ fully specified, using sampled exogenous trajectories from $\mathcal{D}$ for the unknown variables, any state in a trajectory can be recursively computed as $s_{t+1} = T(s_t, \mathbf{a}_t)$. Also, given the deterministic reward function $R(s, \mathbf{a})$, it becomes possible to efficiently simulate future trajectories and compute the corresponding rewards for any sequence of actions. Using this approach, the Bellman equation can be rolled out over K steps as a K -step expansion:

$$Q(s_t, \mathbf{a}_t) = R(s_t, \mathbf{a}_t) + \gamma R(s_{t+1}, \mathbf{a}_{t+1}) + \dots + \gamma^{K-1} R(s_{t+K-1}, \mathbf{a}_{t+K-1}) + \gamma^K Q(s_{t+K}, \mathbf{a}_{t+K}), \quad (19)$$

where $s_{t+1} = T(s_t, \mathbf{a}_t)$ and $\mathbf{a}_{t+1} = \pi_\theta(s_{t+1})$ is the output of the actor policy neural network π with parameters θ .

Since both $R(s, \mathbf{a})$ and $T(s, \mathbf{a})$ are known for the sampled trajectories, the actor network π_θ can be optimized by directly backpropagating gradients through these simulated rollouts. The policy gradient update can thus be computed as:

$$\nabla_\theta J(\theta) \approx \mathbb{E}_{\tau \sim \mathcal{D}} \left[\sum_{j=0}^{K-1} \gamma^j \nabla_\theta R(s_{t+j}, \pi_\theta(s_{t+j})) + \gamma^K \nabla_\theta Q_\phi(s_{t+K}, \pi_\theta(s_{t+K})) \right], \quad (20)$$

where τ denotes a trajectory segment sampled from the replay buffer. During training, length- K trajectories $\{\mathbf{u}_{t:t+K-1}\}$ are sampled from the replay buffer $\mathcal{D}$ and treated as fixed within each rollout, so that the trajectory evolves deterministically under the known, (piecewise) differentiable transition and reward models. This rollout process, together with the calculation of the total cumulative objective, is illustrated in Fig. 2. For example, a three-step trajectory of exogenous variables $\tau = \{\mathbf{u}_t, \mathbf{u}_{t+1}, \mathbf{u}_{t+2}\}$ is sampled from $\mathcal{D}$ and held fixed. The policy π_θ then outputs $\{\mathbf{a}_t, \mathbf{a}_{t+1}, \mathbf{a}_{t+2}\}$, the states evolve via $s_{t+j+1} = T(s_{t+j}, \mathbf{a}_{t+j}; \mathbf{u}_{t+j})$, rewards are computed deterministically, and gradients flow only through (π_θ, T, R) with no backpropagation through the sampled τ , which is treated as constant during the rollout. This approach provides direct and informative gradient feedback based on the “physics” of the system described by $R(s, \mathbf{a})$ and $T(s, \mathbf{a})$, thereby enabling the policy to directly learn how charging decisions influence grid voltages, SoC evolution, and long-term operational rewards.

3.2. Physics-informed TD3 for EV charging

The proposed physics-informed formulation for EV charging can be integrated with a range of RL algorithms for continuous control [33], such as Soft Actor Critic (SAC) and DDPG. However, TD3 [34] was selected as the backbone due to its superior performance in the EV charging setting [22], where long horizons, continuous actions, and voltage magnitude violation penalties pose particular challenges. The proposed PI-TD3 algorithm is described in Algorithm 1. At the start of each epoch, new transitions are collected by executing the current policy with exploration noise (lines 4–6). During each training iteration, a mini-batch of K -step trajectories $\{\tau_i\}_{i=1}^B$ is sampled from the replay buffer $\mathcal{D}$ (line 7). The twin critics Q_{ϕ_1} and Q_{ϕ_2} are updated using target values computed from the replayed transitions (lines 8–10), minimizing mean-squared error. The actor network π_θ is then updated (line 11) with the policy gradient derived from the multi-step rollout objective defined in (20), where gradients flow through the differentiable transition and reward models. Target networks are softly updated to ensure training stability (line 12).

Algorithm 1 Physics-Informed TD3.

```

1: Initialize critics  $Q_{\phi_1}, Q_{\phi_2}$  and actor  $\pi_{\theta}$  with parameters  $\phi_1, \phi_2, \theta$ 
2: Initialize target networks:  $\phi'_1 \leftarrow \phi_1, \phi'_2 \leftarrow \phi_2, \theta' \leftarrow \theta$ 
3: for  $t = 1$  to  $T$  do
4:   Select action with exploration noise:  $\mathbf{a}_t \sim \pi_{\theta}(\mathbf{s}_t) + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma)$ 
5:   Execute  $\mathbf{a}_t$  in environment, observe reward  $r_t$  and next state  $\mathbf{s}_{t+1}$ 
6:   Store  $(\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1})$  in replay buffer  $D$ 
7:   Sample mini-batch  $\tau_{i=1}^B$  of length- $K$  trajectories from  $D$ ,  $\tau_i =$ 
       $(\mathbf{s}_{0,i}, \mathbf{a}_{0,i}, r_{0,i}, \dots, \mathbf{s}_{K,i}, \mathbf{a}_{K,i}, r_{K,i})$ 
8:    $\tilde{\mathbf{a}}_{1,i} \leftarrow \pi_{\theta'}(\mathbf{s}_{1,i}) + \epsilon, \epsilon \leftarrow \text{clip}(\mathcal{N}(0, \sigma), -c, c)$ 
9:    $y_i \leftarrow r_{0,i} + \gamma \min_{j \in \{1,2\}} Q_{\phi'_j}(\mathbf{s}_{0,i}, \tilde{\mathbf{a}}_{1,i})$ 
10:  Update critics by minimizing, for  $j \in \{1,2\}$ :
11:     $\phi_j \leftarrow \arg \min \phi_j \sum_{i=1}^B (y_i - Q_{\phi_j}(\mathbf{s}_{0,i}, \mathbf{a}_{0,i}))^2$ 
12:  Update actor using  $\nabla_{\theta} J(\theta)$  from (20)
13:  Update target networks:
14:     $\phi'_j \leftarrow \tau \phi_j + (1 - \tau) \phi'_j, j \in \{1,2\}$ 
15:     $\theta' \leftarrow \tau \theta + (1 - \tau) \theta'$ 
16: end for

```

This physics-informed formulation enables the proposed PI-TD3 algorithm to scale to hundreds of EVs while remaining closely tied to the underlying grid physics. In contrast to existing related studies that typically incorporate physics through penalty terms, action filtering, or constraint-enforcement layers, the proposed method directly embeds differentiable power-flow Eqs. (2)–(6) and EV battery dynamics into the training rollouts and reward function (14). As a result, the policy can exploit richer gradient information on how charging actions affect both network voltages and EV SoC evolution, thereby accelerating convergence and improving constraint satisfaction. In addition, model-based rollouts increase sample efficiency by reducing reliance on costly interactions with the environment. Unlike classical optimization methods, which become increasingly difficult to solve as the number of EV decision variables grows, the proposed approach leverages neural policies together with differentiable physical models, leading to a more tractable learning process for large fleets. Therefore, compared with existing physics-guided and conventional RL approaches, the proposed PI-TD3 algorithm achieves more stable and scalable learning, making it suitable for real-world deployment in large urban charging networks.

4. Experimental results

In this section, the performance of the proposed PI-TD3 algorithm is systematically evaluated. Average results across multiple scenarios are reported, and additional ablation experiments are conducted to isolate the impact of key components of the proposed method. Detailed analyses of representative case studies are provided, generalization under different grid loading conditions is examined, and scalability is further demonstrated on a substantially larger test system.

4.1. Experimental setup

To assess the effectiveness of the proposed PI-TD3 algorithm, experiments were conducted on a modified IEEE 34-bus distribution network, with each bus hosting 4–5 V2G-enabled charging stations for a total of 150 charging points. EV arrivals and departures followed realistic daily and weekly patterns derived from the ElaadNL dataset. All scenario generation was performed using the EV2Gym simulator [29], which models EV fleet behavior, grid topology, and user sessions with high fidelity, and was enhanced with the RL-ADN [35] power flow module for fast, accurate voltage magnitude calculations. The simulator included detailed EV parameters based on field data, such as battery capacity, charging rates, and efficiency. The following weights ($\lambda_1 = -5 \times 10^4$, $\lambda_2 = 1$, $\lambda_3 = -10$) were used in the problem formulation (1) to reflect the practical priority of the three objectives and to place them on a comparable optimization

scale. In particular, λ_1 was assigned the largest magnitude because the first term is associated with voltage magnitude violations, which are operationally critical and must be penalized much more strongly than the economic objective or the EV charging-completion objective. By contrast, λ_2 and λ_3 were selected to balance charging costs and user service quality, while ensuring safe network operation. Each scenario consisted of 300 steps, each representing 15min of simulated time. In every step, the operator had to determine in real-time the charging action for all EVs. Following standard operation and safety procedures, a voltage limit of lower limit 0.95 and higher limit 1.05 was selected. This setup ensures a realistic benchmarking environment for RL algorithms under operational scenarios representative of real-world distribution networks.

All RL experiments were conducted on the DelftBlue high-performance computing cluster [36]. Each RL algorithm was trained independently until convergence, with training durations ranging from 5 to 10 hours for simpler scenarios, and up to 48 hours for larger grids. To ensure a fair comparison, the default hyperparameter settings recommended in the literature for each baseline [37] were used, including learning rates, discount factors, batch sizes, and exploration noise levels. Training data are generated through interaction with EV2Gym, while communication delays and interruptions are important practical issues left for future work. All models were implemented in PyTorch and trained using the Adam or AdamW optimizer. Performance was averaged across multiple random seeds to assess statistical robustness, and convergence was monitored by tracking moving averages of episode returns.

4.2. Baseline methods & evaluation metrics

To benchmark the performance of the proposed PI-TD3 algorithm, several representative baselines and state-of-the-art algorithms for EV smart charging were selected. These include: (i) Charge as Fast as Possible (CAFAP), a simple heuristic in which each EV is charged at maximum rate immediately upon connection [29]; (ii) a no-charging reference (No Charging); (iii) three widely used model-free RL algorithms, Soft Actor Critic (SAC), Proximal Policy Optimization (PPO), and standard TD3 [37]; and (iv) an MPC method that assumes perfect knowledge of future system states and EV demands [38]. While the MPC is not feasible in practical deployments (since it assumes perfect future information), it provides a useful upper bound on achievable performance under ideal information. All the RL algorithms used the same state and reward formulations to have a fair comparison with the proposed PI-TD3. Note that, to the best of our knowledge, no existing PI-RL algorithm for EV smart charging explicitly integrates differentiable power-flow and battery dynamics into the policy optimization process. As a result, direct comparisons with alternative differentiable PI-RL baselines are not possible, and the evaluation necessarily focuses on representative model-free RL, optimization-based, and heuristic approaches.

All algorithms were evaluated in a deliberately overloaded network scenario, where the distribution grid operates under high load conditions and dense EV integration. This scenario was designed to rigorously test the robustness of each method, as voltage magnitude violations may occur even in the absence of active charging (No Charging baseline). The resulting environment poses a challenging benchmark for coordinating large-scale EV charging while maintaining voltage stability.

Evaluation metrics were selected to reflect the multi-objective nature of the optimization problem in (1). These include total charging cost, average user satisfaction (quantified as the ratio of SoC at departure to target SoC for each EV), and three distinct voltage magnitude violation metrics: total voltage magnitude violations per bus over the evaluation, the number of steps with at least one voltage magnitude violation (Total V.V. per step), and the aggregate absolute per-unit voltage magnitude violations across all buses. Additionally, to further characterize the performance of each approach, total energy charged and discharged by the EV fleet and the average execution time per step were recorded.

Table 2

Average performance of the best trained models over 50 evaluation scenarios on the IEEE 34-bus network with 150 EV chargers.

Algorithm	Costs [€]	User Satisfaction [%]	Total V.V. per bus [-]	Total V.V. per step [-]	Voltage Violation [p.u.]	Tot. Energy Ch. [MWh]	Tot. Energy Dis. [MWh]	Step time [sec/step]
CAFAP	-2545 ± 1581	100.0 ± 0.0	169.5 ± 157.1	36.5 ± 29.4	-0.790 ± 0.991	12.6 ± 0.6	0.00 ± 0.00	0.003
No Charging	0 ± 0	52.1 ± 1.4	106.6 ± 123.0	24.5 ± 25.1	-0.430 ± 0.655	0.0 ± 0.0	0.00 ± 0.00	0.002
SAC	-291 ± 189	57.7 ± 1.3	111.7 ± 125.7	25.6 ± 25.5	-0.458 ± 0.679	20.3 ± 0.9	18.87 ± 0.92	0.017
PPO	-905 ± 583	69.4 ± 1.1	121.4 ± 132.4	27.7 ± 26.6	-0.505 ± 0.737	6.4 ± 0.4	1.80 ± 0.17	0.007
TD3	-1900 ± 1267	90.2 ± 4.5	127.8 ± 134.4	30.0 ± 27.5	-0.482 ± 0.692	12.6 ± 0.6	2.58 ± 1.06	0.009
PI-TD3 (Ours)	-2025 ± 1314	95.6 ± 3.6	104.2 ± 126.4	25.3 ± 27.4	-0.364 ± 0.586	14.3 ± 0.7	2.86 ± 1.04	0.010
MPC	-1640 ± 1203	99.9 ± 0.6	98.7 ± 125.3	24.3 ± 27.8	-0.321 ± 0.545	26.0 ± 2.0	13.46 ± 1.97	—

4.3. Comparison with baseline algorithms

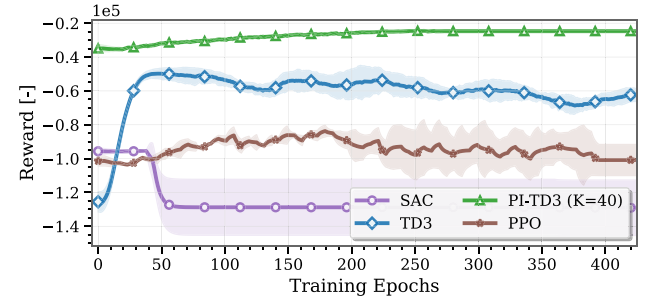
Table 2 summarizes the average performance and standard deviation of all algorithms in 50 scenarios. Notably, the proposed PI-TD3 algorithm achieves a favorable balance among all operational objectives (costs, user satisfaction, and voltage violations). PI-TD3 delivers 14.3 MWh of total energy to the EV fleet, with a user satisfaction rate of 95.6%, ensuring nearly all charging requirements are met. This satisfaction level is within 4% of the MPC (which achieves 99.9%) but outperforms all other RL baselines by at least 5% (TD3: 90.2%, PPO: 69.4%, SAC: 57.7%). In terms of voltage regulation, PI-TD3 reduces the total number of voltage violations per bus to 104.2, which is a 20% improvement over TD3 (127.8), and 15% lower than PPO (121.4). Compared to the MPC, PI-TD3's voltage violation is only 6% higher, indicating near-optimal grid support even without perfect future knowledge. For total voltage violations per step, PI-TD3 achieves 25.3 violations, 15% lower than TD3 (30.0), and only 4% higher than MPC (24.3). The average absolute per-unit voltage magnitude violation is also 15% lower for PI-TD3 compared to TD3.

PI-TD3 also maintains competitive charging cost performance, with total costs 19% lower than TD3 and only 23% below MPC. Although CAFAP minimizes user dissatisfaction (charging everyone at full speed), it results in the worst voltage magnitude violations (over 50% higher than PI-TD3) and higher costs. Classic RL methods (SAC, PPO) either sacrifice user satisfaction or grid reliability, as reflected in their lower performance across at least one key metric. Meanwhile, the proposed PI-TD3 can be executed in real-time, requiring only an average 10 ms per step, while an equivalent MPC-based method would take a few minutes to generate an optimal solution given the scale and the complexity of this MINLP. Overall, PI-TD3 is the only method that closely matches the MPC in all three objectives (user satisfaction, voltage magnitude regulation, and operational cost), demonstrating the advantage of embedding physical knowledge into the RL training process for large-scale, grid-aware EV charging.

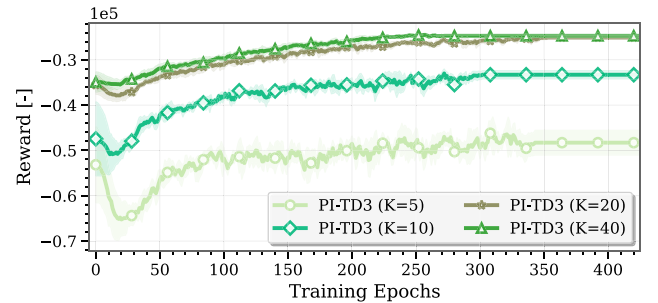
4.4. Sample efficiency and convergence analysis

To evaluate the training efficiency and learning dynamics of PI-TD3, convergence curves and rollout ablation results are compared against state-of-the-art model-free RL baselines. Fig. 3(a) compares the convergence behavior of PI-TD3 and model-free RL algorithms. PI-TD3 rapidly achieves a maximum reward above -0.3×10^5 within the first 75 epochs, whereas model-free TD3 plateaus near -0.5×10^5 , and PPO and SAC remain below -0.8×10^5 throughout training. The incorporation of physical knowledge enables PI-TD3 to reach stable, near-optimal policies approximately four times faster than TD3 and with considerably reduced variance across training runs. This proves a marked improvement in sample efficiency and robustness, making PI-TD3 substantially more suitable for large-scale EV charging control where rapid and reliable learning is critical.

The impact of the rollout horizon K on the performance of PI-TD3 is investigated in Fig. 3(b). As K increases, the learning curve improves: PI-TD3 with $K = 5$ exhibits noticeably lower final rewards and slower convergence, demonstrating that short rollouts do not provide sufficient



(a) Convergence curves for PI-TD3 and model-free RL algorithms.

(b) Effect of rollout horizon K on PI-TD3 training, highlighting the existence of an optimal range for K that balances stability and learning speed.**Fig. 3.** Evaluation of reward performance comparison during training averaged over five random seeds.

gradient information to fully leverage the physics-based environment. Increasing K to 10 and then 20 leads to substantial improvements, both in terms of the speed of convergence and the final achieved reward. Notably, the gap between $K = 20$ and $K = 40$ becomes minimal, with both configurations converging to a similar reward level close to -0.2×10^5 . This plateau indicates that, beyond a certain point, further increasing the rollout horizon offers diminishing returns, as the policy already benefits from sufficiently long, informative trajectories. Moreover, using extremely large K may introduce practical drawbacks, such as higher computational costs and an increased risk of accumulating modeling errors or numerical instability over long rollouts. Therefore, careful selection of the rollout horizon is crucial for achieving optimal trade-offs between gradient quality, learning efficiency, and computational tractability in PI-RL.

4.5. Node-level and system voltage analysis

To provide a detailed view of how each algorithm operates in a realistic scenario, Fig. 4 examines the voltage and charging profiles at bus 22 over a representative evaluation day. As shown in Fig. 4(a), the proposed PI-TD3 algorithm substantially reduces both the frequency

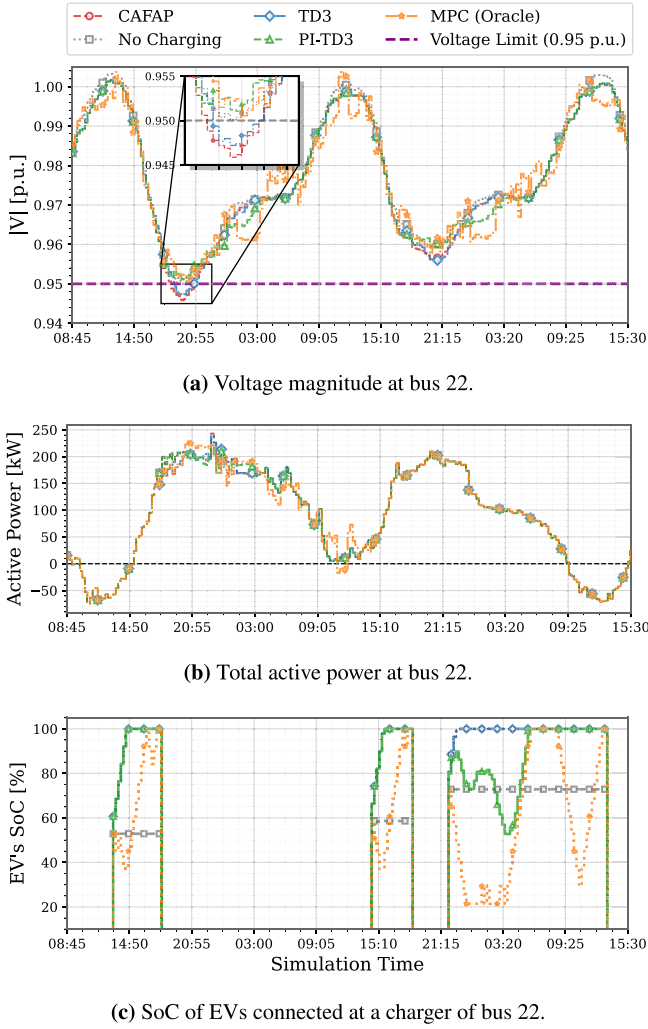


Fig. 4. Impact of EV charging algorithm on the bus voltage magnitude and in EV charging. Bus 22 is demonstrated here as a node where applying the proposed algorithm effectively mitigates voltage limit violations.

and severity of voltage limit violations at this critical bus. In particular, PI-TD3 maintains the voltage magnitude above the 0.95 p.u. threshold, compared to TD3 and CAFAP baselines which fail to do so. Fig. 4(b) presents the total active power drawn at bus 22. While all

algorithms yield similar aggregate profiles, PI-TD3 selectively modulates the charging load, especially during periods of heightened grid stress, to mitigate voltage violations. This dynamic response highlights the PI-TD3's ability to maintain high charging throughput without sacrificing grid stability. The charging schedules of individual EVs, as depicted in Fig. 4(c), demonstrate the diverse and adaptive strategies enabled by PI-TD3. The SoC trajectories reveal that, unlike heuristic or purely model-free baselines, PI-TD3 achieves 100% user satisfaction at bus 22, ensuring all EVs depart fully charged even under congested conditions.

System-wide results are summarized in Fig. 5. Here, the voltage magnitude distributions across all 33 buses (excluding the reference bus) show that PI-TD3 achieves similar median voltages and violation rates as the MPC, despite lacking access to future information. Specifically, PI-TD3 outperforms TD3 by reducing the average per-bus voltage magnitude violations by a noticeable margin, and narrows the gap with the MPC lower bound. However, some violation events persist across all algorithms due to the intentionally overloaded grid design, underscoring the challenging nature of the test scenario. Overall, these results confirm that PI-TD3 delivers robust, grid-compliant EV charging, achieving an advantageous trade-off between energy delivery, cost savings, and voltage magnitude regulation when compared to state-of-the-art RL and heuristic baselines.

4.6. Generalization to unseen load profiles

To assess the robustness of the proposed PI-TD3 algorithm, a generalization study was conducted using modified IEEE 34-bus networks with load scaling factors from $0.5\times$ up to $1.25\times$ nominal demand. The PI-TD3 and classic TD3 agents were exclusively trained on the nominal grid ($1.0\times$ load) and evaluated directly on all other load scenarios, providing an out-of-distribution generalization benchmark.

Fig. 6 summarizes the performance across four key metrics. As shown in Fig. 6(a), PI-TD3 maintains top performance in total reward, with values nearly indistinguishable from the MPC in all but the most extreme loads. For grid reliability (Fig. 6(b)), PI-TD3 reduces the number of time steps with voltage magnitude violations compared to standard TD3, particularly as the network becomes more stressed. In terms of user experience, Fig. 6(c) shows that PI-TD3 keeps average user satisfaction above 90% across the entire range, a level matched only by MPC and CAFAP, and exceeding the TD3 baseline by 5–15% as load increases. Regarding total profits (Fig. 6(d)), PI-TD3 is more cost-efficient than the simple baselines. However, PI-TD3 is also very close to the MPC when evaluated on cases with higher loads than the one it was trained on.

Notably, PI-TD3 attains these results without retraining or fine-tuning on the new conditions, highlighting its ability to generalize robustly to previously unseen load profiles. By leveraging physical

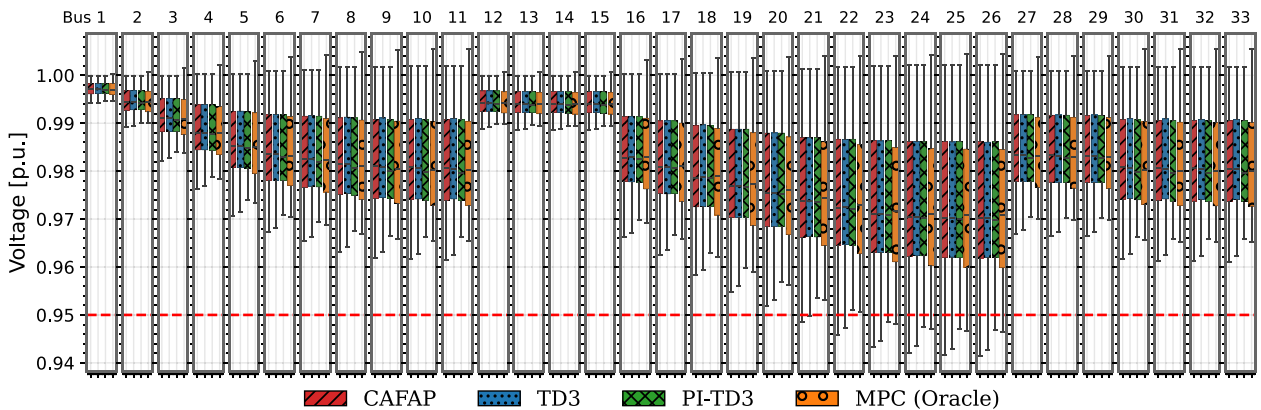


Fig. 5. Distribution of bus voltage magnitudes for each charging algorithm. Box plots show median, interquartile range, and full voltage range over an experimental evaluation scenario; the red dashed line indicates the operational limit (0.95 p.u.).

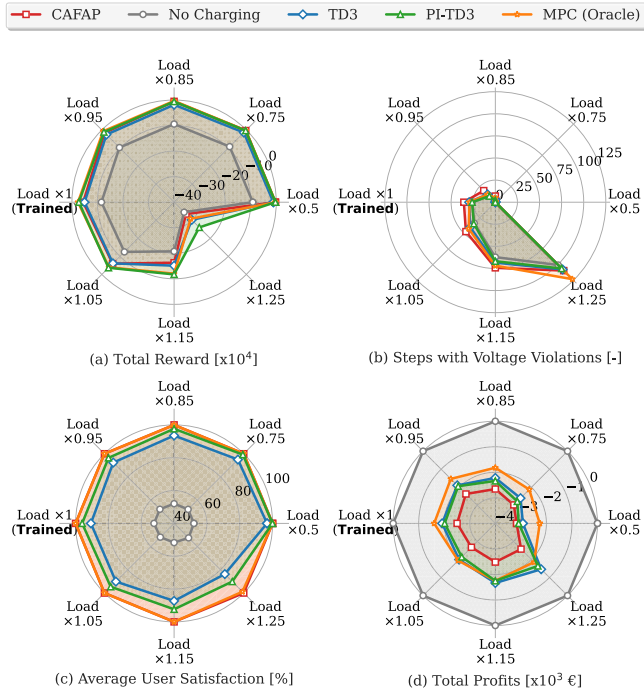


Fig. 6. Generalization and robustness of charging algorithms to varying grid loading. Each radar plot shows performance: (a) total reward, (b) voltage violation steps, (c) average user satisfaction, and (d) total profits, for load multipliers from $0.5\times$ to $1.5\times$ on the IEEE 34-bus network. The PI-TD3 agent was trained on the reference scenario and evaluated out-of-distribution, demonstrating adaptability.

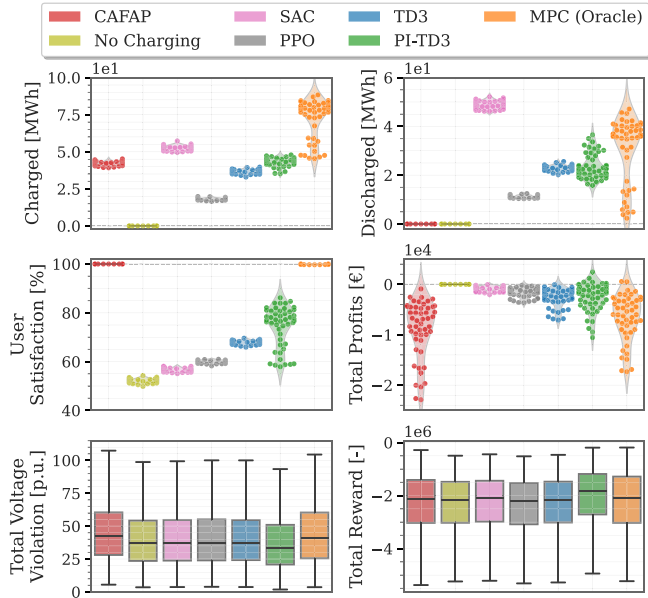


Fig. 7. Performance comparison of charging algorithms on the IEEE 123-bus network with 500 EVs. Metrics shown include energy charged/discharged, user satisfaction, profits, total voltage violations, and total reward. PI-TD3 matches or surpasses the MPC in most metrics and maintains robust grid operation at scale.

knowledge and differentiable rollouts, PI-TD3 learns policies that remain grid-compliant and economically efficient under a wide spectrum of practical operating conditions, outperforming all model-free and heuristic alternatives in these challenging out-of-distribution tests.

4.7. Case study of a larger grid with 500 EVs

To further evaluate the scalability and robustness of the proposed PI-TD3 algorithm, experiments were conducted on the IEEE 123-bus network with 500 distributed EV charging points, a significant increase in both network and EV fleet size. For each algorithm, 50 independent experimental scenarios were randomly selected. In Fig. 7, the first two rows display violin plots, where each dot represents the outcome of a single scenario, thereby visualizing both the overall result distribution and the deviation across trials. These plots reveal that PI-TD3 and the MPC consistently achieve the highest user satisfaction, and the most favorable profits, while classic RL algorithms and heuristics exhibit less diverse performances. The final row of Fig. 7, presents box plots summarizing the total voltage magnitude violations (p.u.) and total rewards. Here, PI-TD3 matches or slightly outperforms MPC in both metrics, achieving median total voltage magnitude violations below 30 and maintaining a consistently higher total reward than all other methods. Notice that the MPC algorithm fails to achieve satisfactory results due to the non-convexity and scale of the problem, even when knowing the future (oracle). The box plots show that the distribution of voltage magnitude violations and rewards for PI-TD3 is notably narrower, indicating reliability and stability in grid operation at this scale. Overall, these results demonstrate that PI-TD3 not only achieves strong average performance but also maintains low variability and robust grid support even as the problem size and complexity are greatly increased.

5. Conclusion

This work introduced PI-TD3, a PI-RL algorithm for large-scale EV smart charging that supports the voltage magnitude of the distribution network. By directly integrating differentiable power-flow equations and EV battery dynamics into the RL training process, the proposed approach provides richer learning signals than conventional model-free RL. On the IEEE 34-bus system with 150 EV chargers, PI-TD3 achieved 95.6% user satisfaction, compared with 90.2% for TD3, 69.4% for PPO, and 57.7% for SAC, while remaining close to the MPC benchmark at 99.9%. At the same time, PI-TD3 reduced total voltage violations per bus to 104.2, corresponding to a 20% improvement over TD3 (127.8) and remaining only 6% above MPC (98.7). In addition, total voltage violations per step were reduced to 25.3, compared with 30.0 for TD3, and the average execution time remained suitable for real-time control at about 10 ms per step.

The proposed method also generalized well under different loading conditions and remained effective when scaled to the IEEE 123-bus system with 500 EV chargers, showing its potential for large-scale deployment in real distribution networks. Overall, these results indicate that physics-informed RL is a promising approach for real-time coordination of EV charging in complex power systems. At the same time, practical aspects such as communication disturbances and data privacy remain important topics for future research and should be further investigated to support real-world implementation.

CRediT authorship contribution statement

Stavros Orfanoudakis: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Conceptualization. **Frans A. Oliehoek:** Writing – review & editing. **Peter Palensky:** Supervision, Funding acquisition. **Pedro P. Vergara:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The study was funded by the DrIVe2X research and innovation project, supported by the European Commission with grant number 101056934.

Data availability

Data and code are open source.

References

- [1] Stiasny J, Zufferey T, Pareschi G, Toffanin D, Hug G, Boulouchos K. Sensitivity analysis of electric vehicle impact on low-voltage distribution grids. *Electr Power Syst Res* 2021;191:106696. <https://doi.org/10.1016/j.epsr.2020.106696>
- [2] Hasan KN, Muttuqi KM, Borboa P, Scira J, Zhang Z, Leishman M. Distribution network voltage analysis with data-driven electric vehicle load profiles. *Sustain Energy Grids Netw* 2023;36:101216. <https://doi.org/10.1016/j.segan.2023.101216>
- [3] Mattos MM, Archetti JAG, Bitencourt LDA, Wallberg A, Castellucci V, Dias BH, de Oliveira JG. Analysis of voltage control using V2G technology to support low voltage distribution networks. *IET Gener Transm Distrib* 2024;18(6):1133–57. <https://doi.org/10.1049/gtd2.13066>
- [4] Knezović K, Marinelli M. Phase-wise enhanced voltage support from electric vehicles in a Danish low-voltage distribution grid. *Electr Power Syst Res* 2016;140:274–83. <https://doi.org/10.1016/j.epsr.2016.06.015>
- [5] K. M. K. M, Rajagopal K. Enhancing voltage control and regulation in smart microgrids through deep learning - optimized EV reactive power management. *Energy Rep* 2025;13:1095–107. <https://doi.org/10.1016/j.egy.2024.12.072>
- [6] Elalfy DA, Gouda E, Kotb MF, Bureš V, Sedhom BE. Frequency and voltage regulation enhancement for microgrids with electric vehicles based on red panda optimizer. *Energy Convers Manag* X 2025;25:100872. <https://doi.org/10.1016/j.ecmx.2025.100872>
- [7] Khan B, Ullah Z, Grusso G. Enhancing grid stability through physics-informed machine learning integrated-model predictive control for electric vehicle disturbance management. *World Electr Veh J* 2025;16(6). <https://doi.org/10.3390/wevj16060292>
- [8] Ke S, Yang J, Chen L, Fan P, Shi X, Li G, Wu F. A frequency control strategy for EV stations based on MPC-VSG in islanded microgrids. *IEEE Trans Ind Inform* 2024;20(2):1819–31. <https://doi.org/10.1109/TII.2023.3281658>
- [9] Singh S, Verma M. Smart charging schedule of plug-in electric vehicles for voltage support: a prosumer-centric approach. *Sustain Energy Grids Netw* 2023;33:100972. <https://doi.org/10.1016/j.segan.2022.100972>
- [10] Sutton RS, Barto AG. Reinforcement learning: an introduction. Bradford Books; 2018.
- [11] Yang Q, Wang G, Sadeghi A, Giannakis GB, Sun J. Two-timescale voltage control in distribution grids using deep reinforcement learning. *IEEE Trans Smart Grid* 2020;11(3):2313–23. <https://doi.org/10.1109/TSG.2019.2951769>
- [12] Liu D, Zeng P, Cui S, Song C. Deep reinforcement learning for charging scheduling of electric vehicles considering distribution network voltage stability. *Sensors* 2023;23(3):1618. <https://doi.org/10.3390/s23031618>
- [13] Shibli MM, Ismail LS, Massoud AM. Electric vehicles charging management using deep reinforcement learning considering vehicle-to-grid operation and battery degradation. *Energy Rep* 2023;10:494–509. <https://doi.org/10.1016/j.egy.2023.07.008>
- [14] Fan J, Liebman A, Wang H. Safety-aware reinforcement learning for electric vehicle charging station management in distribution network. In: 2024 IEEE Power & Energy Society General Meeting (PESGM); 2024. p. 1–5. <https://doi.org/10.1109/PESGM51994.2024.10688856>
- [15] Hossain RR, Yin T, Du Y, Bienstock D, Zussman G. Efficient learning of power grid voltage control strategies via model-based deep reinforcement learning. *Mach Learn* 2024;113:2675–700. <https://doi.org/10.1007/s10994-023-06422-w>
- [16] Hu D, Ye Z, Gao Y, Ye Z, Peng Y, Yu N. Multi-agent deep reinforcement learning for voltage control with coordinated active and reactive power optimization. *IEEE Trans Smart Grid* 2022;13(6):4873–86. <https://doi.org/10.1109/TSG.2022.3185975>
- [17] Li S, Hu W, Cao D, Zhang Z, Huang Q, Chen Z, Blaabjerg F. EV charging strategy considering transformer lifetime via evolutionary curriculum learning-based multiagent deep reinforcement learning. *IEEE Trans Smart Grid* 2022;13(4):2774–87. <https://doi.org/10.1109/TSG.2022.3167021>
- [18] Dulac-Arnold G, Levine N, Mankowitz DJ, Li J, Paduraru C, Goyal S, Hester T. Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. *Mach Learn* 2021;110:2419–68. <https://doi.org/10.1007/s10994-021-05961-4>
- [19] Cao D, Zhao J, Hu J, Pei Y, Huang Q, Chen Z, Hu W. Physics-informed graphical representation-enabled deep reinforcement learning for robust distribution system voltage control. *IEEE Trans Smart Grid* 2024;15(1):233–46. <https://doi.org/10.1109/TSG.2023.3267069>
- [20] Gao J, Chen S, Li X, Zhang J. Transient voltage control based on physics-informed reinforcement learning. *IEEE J Radio Freq Identif* 2022;6:905–10. <https://doi.org/10.1109/JRFID.2022.3213895>
- [21] Biswas A, Acquarone M, Wang H, Miretti F, Misul DA, Emadi A. Safe reinforcement learning for energy management of electrified vehicle with novel physics-informed exploration strategy. *IEEE Trans Transp Electrif* 2024;10(4):9814–28. <https://doi.org/10.1109/TTE.2024.3361462>
- [22] Orfanoudakis S, Robu V, Salazar EM, Palensky P, Vergara PP. Scalable reinforcement learning for large-scale coordination of electric vehicles using graph neural networks. *Commun Eng* 2025;4(1):118. <https://doi.org/10.1038/s44172-025-00457-8>
- [23] Karniadakis GE, Kevrekidis IG, Lu L, Perdikaris P, Wang S, Yang L. Physics-informed machine learning. *Nat Rev Phys* 2021;3:422–40. <https://doi.org/10.1038/s42254-021-00314-5>
- [24] Kamal NF, Sharida A, Bayhan S, Abu-Rub H, Alnuweiri H. Enhancing electric vehicle charging predictions: a physics-informed neural network approach. In: IECON 2024 - 50th annual conference of the IEEE industrial electronics society; 2024. p. 1–6. <https://doi.org/10.1109/IECON55916.2024.10905642>
- [25] Lim H, Lee JW, Boyack J, Choi JB. EV-PINN: a physics-informed neural network for predicting electric vehicle dynamics. *arXiv:2411.14691*. 2024.
- [26] Kaseb Z, Orfanoudakis S, Vergara PP, Palensky P. Adaptive informed deep neural networks for power flow analysis. *arXiv:2412.02659*. 2025.
- [27] Kuang H, Qu H, Deng K, Li J. A physics-informed graph learning approach for citywide electric vehicle charging demand prediction and pricing. *Appl Energy* 2024;363:123059. <https://doi.org/10.1016/j.apenergy.2024.123059>
- [28] Zhang B, Cao D, Hu W, Ghias AM, Chen Z. Physics-informed multi-agent deep reinforcement learning enabled distributed voltage control for active distribution network using PV inverters. *Int J Electr Power Energy Syst* 2024;155:109641. <https://doi.org/10.1016/j.jepes.2023.109641>
- [29] Orfanoudakis S, Diaz-Londono C, Emre Yilmaz Y, Palensky P, Vergara PP. EV2Gym: a flexible V2G simulator for EV smart charging research and benchmarking. *IEEE Trans Intell Transp Syst* 2025;26(2):2410–21. <https://doi.org/10.1109/tits.2024.3510945>
- [30] Mehdizadeh M, Nordfjaern T, Klöckner CA. Estimating financial compensation and minimum guaranteed charge for vehicle-to-grid technology. *Energy Policy* 2023;180:113649. <https://doi.org/10.1016/j.enpol.2023.113649>. <https://www.sciencedirect.com/science/article/pii/S0301421523002343>
- [31] Open Charge Point Protocol (OCPP) 2.1. Edition 1 Technical report & protocol specification. Open Charge Alliance (Jan 2025). <https://openchargealliance.org>.
- [32] Giraldo JS, Montoya OD, Vergara PP, Milano F. A fixed-point current injection power flow for electric distribution systems using laurent series. *Electr Power Syst Res* 2022;211:108326. <https://doi.org/10.1016/j.epsr.2022.108326>
- [33] Xing E, Luk V, Oh J. Stabilizing reinforcement learning in differentiable multi-physics simulation. *arXiv:2412.12089*. 2024.
- [34] Fujimoto S, van Hoof H, Meger D. Addressing function approximation error in actor-critic methods. In: Dy J, Krause A, editors. *Proceedings of the 35th international conference on machine learning*. 80. PMLR; 2018. p. 1587–96. *Proceedings of Machine Learning Research*.
- [35] Hou S, Gao S, Xia W, Salazar Duque EM, Palensky P, Vergara PP. RL-ADN: a high-performance deep reinforcement learning environment for optimal energy storage systems dispatch in active distribution networks. *Energy AI* 2025;19:100457. <https://doi.org/10.1016/j.egyai.2024.100457>
- [36] Delft High Performance Computing Centre (DHPC). DelftBlue supercomputer (phase 2); 2024. <https://www.tudelft.nl/dhpc/ark:/44463/DelftBluePhase2>
- [37] Raffin A, Hill A, Gleave A, Kanervisto A, Ernestus M, Dormann N. Stable-Baselines3: reliable reinforcement learning implementations. *J Mach Learn Res* 2021;22(268):1–8. <http://jmlr.org/papers/v22/20-1364.html>
- [38] Diaz-Londono C, Orfanoudakis S, Vergara PP, Palensky P, Ruiz F, Grusso G. Open source algorithms for maximizing V2G flexibility based on model predictive control. *Electr Power Syst Res* 2026;250:112082. <https://doi.org/10.1016/j.epsr.2025.112082>. <https://www.sciencedirect.com/science/article/pii/S0378779625006704>