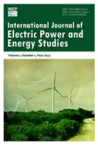
WEP Warwick Evans Publishing

International Journal of Electric Power and Energy Studies

ISSN: 3006-2004 (Print), ISSN: 3006-0826 (Online)
Volume 5, Number 3, Year 2026
DOI: https://doi.org/10.62051/ijepes.v5n3.02



Battery Health-Aware Multi-Objective Reinforcement Learning for Electric Vehicle Energy Consumption Optimization

Noah Lewis[1, *], Miles Baker[2]

[1] Department of Computer Science, College of Engineering, University of California, Davis, CA, USA

[2] Department of Computer Science and Engineering, University of California, Riverside, CA, USA

[*]Corresponding author: Noah Lewis

ABSTRACT

Thermal auxiliaries and regenerative braking form the controllable part of electric-vehicle (EV) energy use, but the settings that save energy are not always those that preserve the battery: regenerating into a cold pack promotes lithium plating, and every ampere-hour drawn from a hot pack ages it faster. We cast EV energy management as a three-objective problem-energy drawn from the battery, capacity fade and cabin discomfort-and learn a single preference-conditioned policy for the whole trade-off surface. The proposed preference-conditioned soft actor-critic (PC-SAC) jointly sets the regenerative-braking intensity, the cabin climate power and the battery thermal-management power. It combines decomposed twin critics and preference relabelling with two design elements that proved essential: exogenous re-centring of the learning costs and a physically consistent effort parameterisation of the thermal actions. Driving-scenario probabilities from a contrastive encoder complete the state. Experiments use 3,131 real-world trips (468 h, 27 drivers) from the Vehicle Energy Dataset, a powertrain model calibrated on measured battery power (trip-level R2 = 0.91) and a driver-disjoint test set of 853 trips. On the test drivers, one PC-SAC policy spans a wide trade-off: its health-oriented point changes capacity fade by −28.5 % at +6.2 % energy relative to a production-style controller, and its energy-oriented point changes energy by −21.8 % at the expense of comfort. Measured by its own training objective, it does not improve on the well-tuned production controller at the energy- and comfort-weighted preferences and reaches parity at the health-oriented one (ΔJ = +0.001 ± 0.031), which locates the value of learning in the battery-health region that the rules do not reach. Its front attains a hypervolume of 0.0235 ± 0.0015, against 0.0114 for three mode-specific agents without a battery objective and 0.0204 for three rule-based controllers. In ablations, removing the decomposed critics and preference relabelling lowers the hypervolume.

KEYWORDS

Electric Vehicles; Battery Degradation; Multi-Objective Reinforcement Learning; Soft Actor-Critic; Thermal Management; Regenerative Braking.

1. INTRODUCTION

Ambient temperature and driving conditions shape the energy use of electric vehicles (EVs) as much as vehicle design does. Fleet-level analyses report that cabin heating and cooling raise consumption markedly in cold and hot regions [1], and real-world measurements attribute large differences in consumption to driving behaviour and trip conditions [2], [3]. Battery degradation follows a different but overlapping set of drivers [4]: above room temperature, solid-electrolyte-interphase growth accelerates with every cycle, whereas charging below it promotes lithium plating [5]. In a vehicle,

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).
Published by Warwick Evans Publishing.

low-temperature charging happens mostly through regenerative braking, and the thermal state of the pack is set by the same thermal system that serves the cabin.

Regenerative-braking intensity, cabin climate power and battery thermal-management system (BTMS) power are therefore coupled decisions serving three conflicting goals: low energy use, slow capacity fade and passenger comfort. Optimisation-based energy management has long included battery wear through state-of-health-perceptive cost terms [6], and reinforcement learning (RL) has become a standard tool for vehicle energy management because it copes with stochastic driving without a trip forecast [7], [8]. Recent deep-RL strategies add battery-thermal and health terms to the reward [9], [10], but they collapse the objectives into a weighted sum fixed before training, so every new trade-off requires a new policy.

Wang *et al.* [11] recently paired contrastive driving-scenario recognition with RL for EV energy optimisation and exposed the energy–comfort trade-off through Eco, Balanced and Comfort operating modes. Extending this design to battery health raises two questions. Can a single agent offer a continuous trade-off that includes battery health and can be changed at run time? And does such an agent improve on production-style controllers when it is trained and tested on real-world driving rather than on standard cycles? This paper answers both questions with the following contributions.

- A battery-health-aware tri-objective formulation of EV energy management with joint regenerative-braking, cabin and BTMS control, driven by 3,131 real-world trips of the Vehicle Energy Dataset (VED) [12] and a powertrain model calibrated against measured battery power. The degradation cost combines the semi-empirical cycle-life model of [13] in severity-factor form [6] with a low-temperature plating branch.
- PC-SAC, a preference-conditioned soft actor-critic with decomposed twin critics and preference relabelling, made trainable by exogenous re-centring of the learning costs and by effort-type thermal actions that exclude physically meaningless conditioning; we document the failure modes that motivated both.
- A driver-disjoint evaluation against rule-based controllers and fixed-weight agents that follow the mode-based design discussed above, with hypervolume, seasonal and ablation analyses and a fleet-level projection based on workplace charging sessions.

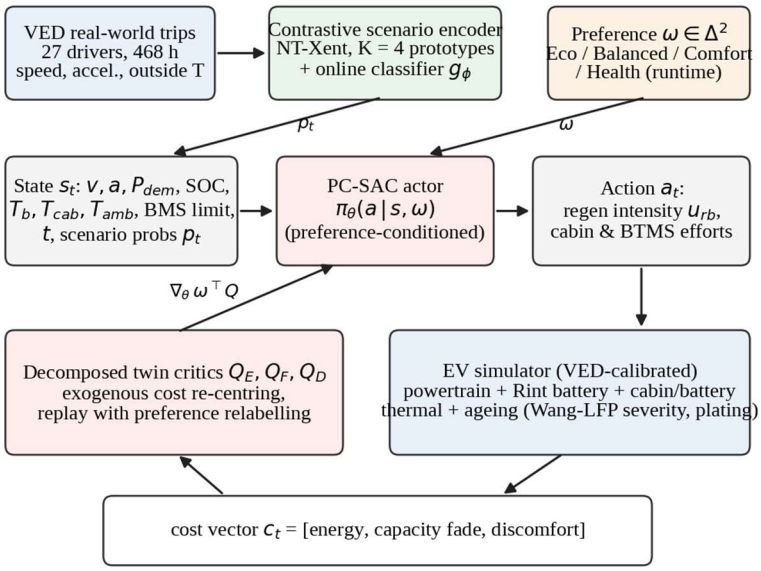


Fig. 1 Overview: a contrastive encoder supplies scenario probabilities, a preference-conditioned actor sets regeneration and thermal efforts, and decomposed critics learn from re-centred cost vectors with preference relabelling.

2. RELATED WORK

2.1. EV Energy Consumption and Thermal Loads

Physics-based consumption models combine longitudinal dynamics with powertrain and regeneration efficiencies and have been validated against on-road data [3]. Naturalistic datasets such as VED [12], which logs speed, auxiliary power and high-voltage battery signals of an Ann Arbor fleet over a full year, allow such models to be calibrated and stress-tested across seasons. Temperature enters both through cabin conditioning [1] and through the internal resistance of the pack, which we identify directly from VED voltage–current data with an equivalent-circuit representation [14].

2.2. Battery-Aware Energy Management and Reinforcement Learning

Ebbesen *et al.* [6] introduced a severity factor that weights ampere-hour throughput by its relative impact on battery life and used it inside an equivalent-consumption strategy. Model-free RL removes the need for a drive-cycle forecast: tabular Q-learning was applied early to hybrid powertrains [7], deep RL followed [8], and soft actor-critic (SAC) [15] has been used to enforce battery thermal and health constraints in hybrid buses [9]. Hierarchical RL separates strategic and tactical decisions [10]. Common to these works is a scalar reward whose weights are fixed before training.

2.3. Multi-Objective and Preference-Conditioned RL

Multi-objective RL (MORL) distinguishes single-policy methods, which solve one scalarisation, from multi-policy methods that approximate a coverage set [16], [17]. Conditioning one network on the preference vector and replaying experience under several preferences lets a single policy serve many trade-offs [18], [19]; PD-MORL extends the idea to continuous control [20]. Front coverage is commonly measured by the hypervolume indicator [21]. We are not aware of preference-conditioned MORL for the joint energy, battery-health and comfort management of EVs on real-world driving data. Carbon-aware offline reinforcement learning for coupled inventory and transportation decisions illustrates the methodological value of testing sequential policies against multiple operational objectives and hard resource constraints [28].

2.4. Learning under Exogenous Inputs

Vehicle energy management is input-driven: the speed trace evolves independently of the controller and dominates the return. Input-dependent baselines reduce the resulting variance without biasing the policy [22]; we adapt this idea to the decomposed critics of an off-policy, preference-conditioned agent.

3. SYSTEM MODEL

3.1. Real-World Driving Data

VED [12] contains one year (Nov. 2017–Nov. 2018) of OBD-II logs from 383 vehicles in Ann Arbor, Michigan. We use the three battery-electric vehicles (BEVs), which record high-voltage current, voltage, state of charge (SOC), heater and air-conditioning power, and the speed and outside-air-temperature (OAT) traces of the 24 plug-in hybrids, which serve as additional human driving cycles for the simulated BEV. Trips are resampled to 1 Hz and smoothed with a 3-s moving average; trips with logging gaps above 15 s, shorter than 120 s or shorter than 0.5 km are removed. The corpus contains 3,131 trips (468 h, 16,903 km) from 27 vehicles. Vehicles-and hence drivers-are split disjointly into training (12 vehicles, 1,915 trips, 317 h), validation (3, 363 trips) and test sets (12 vehicles, 853 trips, 122 h, 4,312 km). The test set contains 308 cold trips (median OAT below 5 °C),

279 mild trips and 266 warm trips (20 °C and above). The initial SOC is the measured value for BEV trips and is drawn from the empirical BEV distribution otherwise.

3.2. Powertrain and Battery Model

Wheel power follows from the recorded speed v and acceleration a,

$$P_{\mathrm{w}} = (m_{\mathrm{eff}}a + mgC_{\mathrm{rr}} + \tfrac{1}{2}\rho C_{\mathrm{d}}Av^2)v \tag{1}$$

and the battery-side bus power combines traction, regeneration and auxiliary loads,

$$P_{\mathrm{bus}} = \min(P_{\mathrm{w}}^{+}/\eta_{\mathrm{d}}, P_{\max}) - u_{\mathrm{rb}} \min(\eta_{\mathrm{r}} f(v) P_{\mathrm{w}}^{-}, P_{\mathrm{BMS}}) + P_{\mathrm{aux}} + P_{\mathrm{cab}} + P_{\mathrm{btm}} \tag{2}$$

where $u_{\mathrm{rb}} \in [0, 1]$ is the regenerative-braking intensity, $f(v)$ fades regeneration out below 15 km/h, P_{BMS} is the charge-power limit of the battery-management system (BMS), and the remaining braking demand is met by the friction brakes. The pack is a resistive equivalent circuit [14] with current $I = (U_{\mathrm{oc}} - [U_{\mathrm{oc}}^2 - 4RP_{\mathrm{bus}}]^{1/2})/(2R)$, open-circuit voltage $U_{\mathrm{oc}}(\mathrm{SOC})$ and Arrhenius resistance $R(T_{\mathrm{b}}) = R_{25} \exp[k_{\mathrm{R}}(1/T_{\mathrm{b}} - 1/298.15\ \mathrm{K})]$.

Parameters were identified from the BEV logs (Table 1). Traction parameters were fitted by least squares on 10-s energy windows, with the measured heater and air-conditioning power as inputs and the physical coefficients bounded to published ranges for this vehicle class. $U_{\mathrm{oc}}(\mathrm{SOC})$ and $R(T)$ were fitted to measured voltage–current–SOC samples using OAT as a proxy for pack temperature, which gives $R_{25} = 29.6$ mΩ and an apparent activation energy of 42 kJ/mol. The ratio of energy to SOC change implies a usable capacity of 16.0 kWh (43 Ah). In cross-vehicle validation (parameters fitted on one BEV and tested on the others), trip energy is reproduced with $R^2 = 0.91$ and 0.88 and 10-s power with $R^2 = 0.78$ and 0.74 (Fig. 2a). A driver-specific bias of 24–20 % remains, mainly because VED does not record road grade.

Table 1. Identified and assumed model parameters

Quantity	Value
Mass / effective mass	1660 / 1743 kg
C_{rr}, $C_{\mathrm{d}}A$ (fitted, bounded)	0.012, 0.60 m^2
η_{d}, η_{r} (fitted)	0.92, 0.59
Auxiliary power P_{aux} (fitted)	481 W
R_{25}, activation energy (fitted)	29.6 mΩ, 42 kJ/mol
Usable capacity (identified)	16.0 kWh / 43 Ah
Trip-energy R^2, all / cross-vehicle	0.91 / 0.91, 0.88
C_{b}, hA_{b} (assumed)	180 kJ/K, 15 W/K
C_{cab}, UA (assumed; VED slope 102 W/K)	50 kJ/K, 100 W/K
Ageing: E_{a}, z, N_{eol}	31.7 kJ/mol, 0.55, 8102

3.3. Thermal Models

Battery and cabin temperatures follow lumped energy balances,

$$C_b \mathrm{d}T_b/\mathrm{d}t = I^2R + Q_{btm} - hA_b(T_b - T_{amb}) \tag{3}$$

$$C_{cab} \mathrm{d}T_{cab}/\mathrm{d}t = UA(T_{amb} - T_{cab}) + Q_{cab} + Q_{int} \tag{4}$$

with the values in Table 1; UA = 100 W/K matches the slope of measured heater power against OAT in VED (102 W/K). Cabin heating uses a heat pump whose coefficient of performance (COP) falls linearly from 2.5 at 10 °C to 1.0 at −10 °C (5 kW thermal maximum); cooling uses COP 2.5 (4 kW). The BTMS heats resistively or cools with COP 2.5, up to 3 kW thermal in either direction. The BMS charge limit falls from 2.5C above 20 °C to 0.6C at 0 °C and 0.25C at −10 °C and is derated linearly above 90 % SOC.

3.4. Battery Degradation Model

Capacity fade is computed with the semi-empirical cycle-life model of Wang *et al.* [13], whose rate constant depends on the C-rate c and the temperature T,

$$k(c, T) = B(c) \exp[-(E_a - 370.3c)/(R_g T)] \tag{5}$$

with E_a = 31.7 kJ/mol, a C-rate-dependent pre-factor $B(c)$ and a throughput exponent z = 0.55. Following [6], instantaneous wear is expressed through the severity factor $\sigma = (k/k_{nom})^{1/z}$ relative to 1C and 25 °C, and the fade increment of one step is

$$\Delta q = 20\% \cdot \sigma |I| \Delta t \,/\, (3600\, Q\, N_{eol}) \tag{6}$$

where N_{eol} is the number of capacity throughputs that cause 20 % fade at nominal conditions. Eq. (5) describes only the high-temperature branch. Because charging below about 25 °C causes lithium plating with an inverse Arrhenius dependence [5], charge throughput below T_{ref} = 25 °C is evaluated at T_{ref} and multiplied by $\exp[E_a(1/T - 1/T_{ref})/R_g]$. Mirroring the activation energy is an assumption made in the absence of cell-specific plating data. The resulting V-shaped severity (Fig. 2b) makes regeneration into a cold pack an order of magnitude more damaging than discharge, while all throughput becomes more harmful as the pack warms.

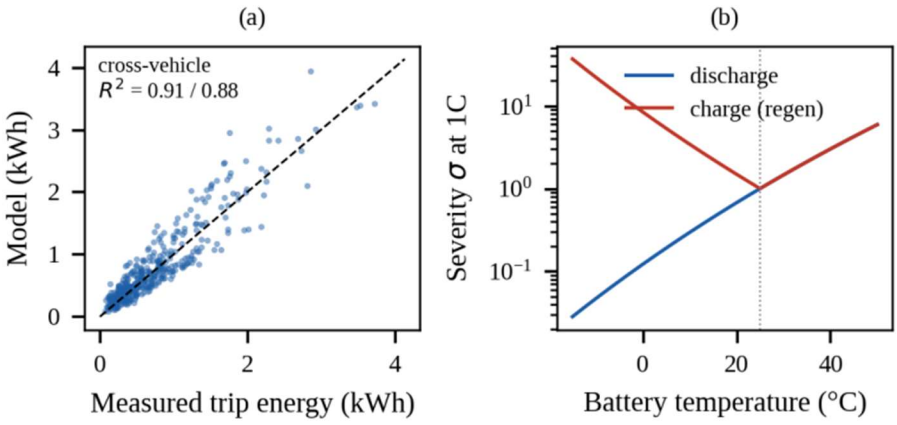


Fig. 2 (a) Measured versus modelled trip energy of the VED BEVs in cross-vehicle validation. (b) Ageing severity σ at 1C for discharge and for charge (regeneration) versus battery temperature.

4. METHOD

4.1. Driving-Scenario Recognition

The agent benefits from knowing whether it is in congested stop-and-go traffic, where regeneration events are frequent, or on an expressway, where they are rare. We adopt the contrastive scenario-recognition module of Wang *et al.* [11] and apply it to VED. Each driving state is the 60-s history of speed and acceleration (120 inputs); two views are generated with Gaussian noise ($\sigma = 0.1$), temporal shifting (±5 s) and feature masking ($p = 0.2$); an encoder with hidden widths 256 and 128 and a 64-dimensional projection head is trained with the NT-Xent loss [23] (temperature 0.07, batch 256); k-means on the 128-dimensional embeddings yields K prototypical scenarios; and a multinomial classifier g_φ provides online scenario probabilities p_t. We keep the hyper-parameters of the original module and choose K with the silhouette coefficient [24]. The action set defined below also mirrors the regenerative-braking-intensity, power-allocation and climate-control decisions of that framework, with power allocation interpreted as the division of thermal power between cabin and battery.

4.2. Multi-Objective Formulation

The state $s_t = [v, a, P_w, \mathrm{SOC}, T_b, T_{cab}, T_{amb}, c_{BMS}, t, p_t]$ has $9 + K$ normalised entries, where c_{BMS} is the admissible charge C-rate. The action $a_t = [a_{rb}, a_{cab}, a_{btm}] \in [-1, 1]^3$ sets $u_{rb} = (a_{rb} + 1)/2$ and two conditioning efforts $e = (a + 1)/2 \in [0, 1]$: the cabin receives e_{cab} times the maximum heating power when it is below the 22 °C set-point and times the maximum cooling power when it is above, and the BTMS conditions the pack towards 25 °C in the same way (Section IV-D). Each 1-s step incurs the cost vector

$$c_t = [U_{oc} I \Delta t, \quad \Delta q, \quad \max(|T_{cab} - 22°\mathrm{C}| - 0.5\ \mathrm{K}, 0)\Delta t]^{\mathrm{T}} \tag{7}$$

namely the chemical energy drawn from the pack (including resistive losses), the capacity fade and the degree-seconds outside a ±0.5 K comfort band. For a preference ω on the simplex Δ^2, the agent minimises $\mathrm{E}[\Sigma_t \gamma^t \omega^{\mathrm{T}} D^{-1} c_t]$, where D holds per-step cost scales measured under the production controller on the training set. Each trip is one episode.

4.3. Preference-Conditioned Soft Actor-Critic

PC-SAC extends SAC [15] in three ways. First, the actor $\pi_\theta(a|s, \omega)$ and two critics $Q_{\psi i}(s, a, \omega) \in \mathbb{R}^3$ receive ω as an input, and each critic has one output head per objective. Second, the clipped double-Q rule [25] is applied to the scalarised value. For a target action $a' \sim \pi_\theta(\cdot|s', \omega)$,

$$y = r + \gamma(1 - d)[\bar{Q}_{j*}(s', a', \omega) - \alpha \log \pi_\theta(a'|s', \omega) \cdot 1] \tag{8}$$

with $j* = \arg\min_j \omega^{\mathrm{T}} \bar{Q}_j(s', a', \omega)$ and $r = -D^{-1}\tilde{c}_t$ (Section IV-D). The critics minimise $\Sigma_i \|Q_{\psi i} - y\|^2$, the actor minimises $\mathrm{E}[\alpha \log \pi_\theta - \min_i \omega^{\mathrm{T}} Q_{\psi i}]$, and the entropy temperature α is tuned automatically [26]. Separate heads learn targets of homogeneous scale and let one replayed transition serve any preference. Third, every sampled transition is paired with a fresh preference (preference relabelling [18], [19]), drawn with probability 0.7 from Dirichlet(1, 1, 1) and otherwise from an edge of the simplex, so that two-objective trade-offs such as the energy–comfort modes are covered.

4.4. Re-Centred Costs and Effort Actions

Exogenous re-centring. Each critic head must resolve the effect of one action on a sum of about $1/(1-\gamma)$ future costs dominated by the uncontrollable traction demand. In a preliminary version without re-centring, probing the trained critic gave per-objective values near −250, while switching the BTMS from idle to full cooling changed the energy head by about 0.1, with the wrong sign; the actor consequently ignored ω. We therefore learn from $\tilde{c}_t = c_t - b_t$, where b_t is computed from the recorded trace alone (traction energy with unrestricted regeneration at nominal battery conditions, and the corresponding nominal-severity fade). Since b_t depends neither on actions nor on controllable states, it shifts every Q-function by an action-independent amount and leaves the optimal policy unchanged, in the spirit of input-dependent baselines [22], while removing most of the return variance.

Effort actions. With signed thermal powers as actions, an earlier version learned to exploit the model: in a −5.5 °C test trip its energy-oriented policy drove the cabin to −15 °C by cooling, and its health-oriented policy chilled the pack below ambient because cold discharge ages slowly in the cycle-life model. A production thermal system never conditions away from its target. We therefore parameterise both thermal actions as an effort towards the target temperature, so that "off" is always $a = -1$, every action value is physically meaningful, and no clipping creates flat regions in which the critic gradient vanishes. The production controllers are expressible in the same action space, so the baselines are unaffected.

4.5. Training

Actor and critics have two hidden layers of 128 ReLU units and are trained with Adam (learning rate $3 \cdot 10^{-4}$), batch 256, $\gamma = 0.99$ and Polyak factor 0.005. Each run collects 280,000 transitions from 32 parallel training trips (episodes capped at 30 min), performing 0.25 gradient updates per transition after 20,000 uniformly random warm-up transitions, which takes about 4 min on a single CPU core with JAX. Algorithm 1 summarises the procedure.

Algorithm 1 PC-SAC training

1: Initialise θ, ψ_1, ψ_2, targets $\bar{\psi}_j$, replay buffer $\mathcal{B}$

2: **for** each parallel environment step **do**

3: sample ω_i for every environment that starts a new trip

4: $p_t \leftarrow g_\varphi(\text{encoder(60-s window)})$; form s_t

5: $a_t = \tanh(\mu_\theta(s_t, \omega_i) + \sigma_\theta(s_t, \omega_i)\varepsilon)$; map to u_{rb}, e_{cab}, e_{btm}

6: simulate; store $(s_t, a_t, -D^{-1}(c_t - b_t), s_{t+1}, d_t)$ in $\mathcal{B}$

7: sample a minibatch; draw a fresh ω per transition

8: update critics with (8), then actor and α; Polyak-update targets

9: **end for**

5. EXPERIMENTS

5.1. Setup

Baselines. RB-Std emulates a production controller: proportional cabin conditioning towards 22 °C (full effort at 1 K deviation), maximum admissible regeneration, BTMS heating below 5 °C and cooling above 35 °C. RB-Eco relaxes the comfort band to heating up to 20.5 °C and cooling from 23.5 °C. RB-Health derates regeneration linearly below 15 °C and keeps the pack between 10 and

30 °C. Mode-SAC denotes three SAC agents trained with the fixed energy–comfort weightings of the Eco (0.8/0.2), Balanced (0.5/0.5) and Comfort (0.2/0.8) modes of Wang *et al.* [11], with the same scenario-aware state, effort actions and re-centring as PC-SAC but without the battery objective. SAC-Health is a fixed-preference agent trained at ω = (0.25, 0.5, 0.25). DOE workplace-charging sessions [27] support the fleet-level projection of Section V-G. All learning agents share networks and training budget; PC-SAC is trained with three seeds and the other agents with one.

Metrics. On the 853 test trips we report energy (Wh/km), capacity fade (10^{-3} % per 1000 km), discomfort (degree-seconds per second outside the comfort band) and the mean $|\Delta T_{cab}|$. Objectives are normalised by RB-Std, which therefore scores 1 on each, and $J_\omega = \omega^T \hat{m}$ is the normalised scalarised cost at preference ω. The hypervolume (HV) [21] of the 21 policies obtained by evaluating PC-SAC on a simplex grid with step 0.2 is computed in the normalised space with reference point (1.25, 1.25, 1.25). Test trips start from deterministic conditions (T_b = OAT + 4 K; T_{cab} = OAT + 1.5 K, or + 5 K above 20 °C).

5.2. Driving-Scenario Recognition

The NT-Xent loss decreased from 1.84 to 0.62 over 25 epochs on 228,659 training windows, and the silhouette coefficient selected K = 4 (0.253). Ordered by mean speed, the scenarios are congested traffic (S1, 12 km/h, stationary 58 % of the time), urban stop-and-go (S2, 30 km/h), arterial roads (S3, 40 km/h) and expressways (S4, 62 km/h, stationary 1 %) (Fig. 3). The online classifier reproduces the prototype assignment of held-out validation drivers with 99.3 % accuracy. Table 2 reports the share of variance (η^2) of window descriptors explained by the cluster labels: the contrastive labels explain traction-energy intensity better than clusters of raw-window principal components or of hand-crafted statistics, whereas hand-crafted statistics, which contain speed explicitly, separate mean speed and regeneration intensity better. The learned scenarios therefore complement, rather than replace, the instantaneous kinematic state.

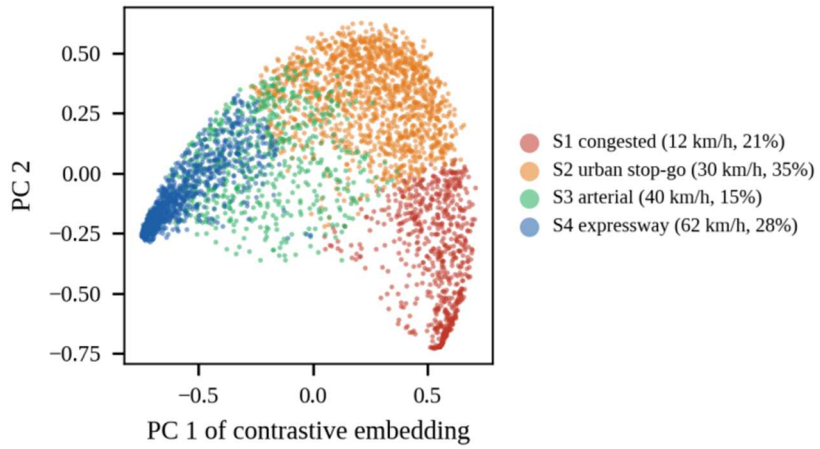


Fig. 3 Two-dimensional projection of the contrastive embeddings of 4,000 training windows, coloured by scenario (mean speed and share of windows in parentheses).

Table 2. Variance of window descriptors explained by K = 4 clusters

Representation	η^2 speed	η^2 traction	η^2 regen.
Contrastive (adopted)	0.576	0.078	0.078
PCA of raw windows	0.633	0.058	0.124
Hand-crafted statistics	0.650	0.070	0.155

5.3. Trade-Off Coverage and Operating Modes

On the test set, RB-Std consumes 181.6 Wh/km and loses $28.4 \cdot 10^{-3}$ % of capacity per 1000 km, with a mean cabin deviation of 2.8 K that is dominated by the warm-up of cold-soaked cabins. Table 3 lists the named operating points and Fig. 4 the normalised fronts. Because the agents optimise the re-centred costs of Section IV-D, the scalarised comparison ΔJ is expressed in the units of those training scales and relative to RB-Std (negative is better). Hierarchical reinforcement learning for integrated-circuit layout provides a complementary optimization example in which congestion-aware reward shaping helps represent competing design objectives [29].

Table 3. Test results at named preferences (PC-SAC: mean of 3 seeds)

Method	Wh/km	Fade[a]	Disc.[b]	$\lvert\Delta T\rvert$ K	ΔJ
RB-Std	181.6	28.4	2.52	2.8	0 (ref.)
RB-Eco	178.6	28.3	3.33	3.8	+0.022
RB-Health	200.6	24.7	2.52	2.8	+0.018
Mode-SAC Eco	146.5	31.6	10.87	11.3	+0.139
PC-SAC Eco	142.0	31.4	12.02	12.5	+0.165±0.006
Mode-SAC Balanced	175.8	28.0	3.16	3.5	+0.086
PC-SAC Balanced	174.9	28.9	2.99	3.4	+0.035±0.028
Mode-SAC Comfort	194.4	27.9	2.54	2.7	+0.070
PC-SAC Comfort	195.6	27.6	2.55	2.7	+0.080±0.020
SAC-Health	195.1	20.8	3.49	3.8	+0.004
PC-SAC Health	192.9	20.3	3.68	4.1	+0.001±0.031
PC-SAC Equal	180.1	25.0	3.20	3.6	+0.040±0.038

[a]10^{-3} % per 1000 km. [b]Degree-seconds per second outside ±0.5 K. Preferences (energy, fade, discomfort): Eco (0.8, 0, 0.2), Balanced (0.5, 0, 0.5), Comfort (0.2, 0, 0.8), Equal (1/3, 1/3, 1/3), Health (0.25, 0.5, 0.25). RB-Eco and RB-Health are scored at the Eco and Health preferences.

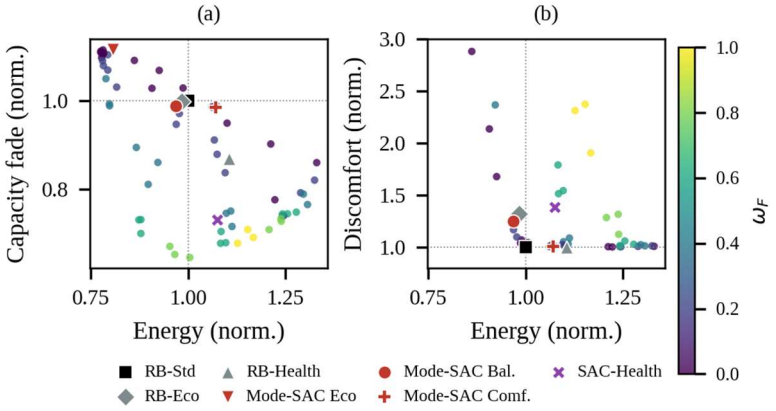


Fig. 4 Test-set objectives of PC-SAC on the 21-point preference grid (three seeds, colour = fade weight ω_F), the rule-based controllers and the fixed-preference agents, normalised by RB-Std (dotted lines). Discomfort is clipped at 3 for readability.

The single PC-SAC network moves along the front as the preference changes. Its Health point changes capacity fade by −28.5 % at +6.2 % energy and +46.2 % discomfort relative to RB-Std, whereas the hand-tuned RB-Health achieves −13.1 % fade at +10.5 % energy without changing comfort; PC-SAC thus obtains more than twice the fade reduction at a smaller energy cost. It does so mainly by lowering the regeneration intensity while the pack is cold (Section V-F), and it pays with comfort. At the Eco point, energy changes by −21.8 %, but the discomfort rises by +377.4 %: with 80 % of the weight on energy measured on the re-centred scale, the agent largely forgoes cabin conditioning. The Balanced, Comfort and Equal points lie between these extremes (Table 3).

PC-SAC attains a lower ΔJ than the corresponding specialist at two of four preferences (Balanced, Health). No learned agent improves the scalarised cost of RB-Std at the energy- and comfort-weighted preferences; the production controller, which holds the cabin at its set-point with maximum regeneration, is a strong baseline for these objectives, and at the Health preference PC-SAC is on par with it ($\Delta J = +0.001 \pm 0.031$) while offering a different trade-off. The hypervolume of the PC-SAC grid is 0.0235 ± 0.0015 (three seeds), compared with 0.0114 for the three Mode-SAC agents, 0.0204 for the three rule-based controllers and 0.0205 for all fixed agents and rule-based controllers together. Mode-SAC, which has no battery objective, does not enter the low-fade region of the front at any of its modes (Fig. 4a).

5.4. Seasonal Analysis

Fig. 5 splits the changes by the median outside temperature of each trip. In cold trips the Health point changes fade by −70.1 % and energy by +2.0 %; in warm trips fade changes by −6.3 % at +3.4 % energy. Warm trips dominate absolute ageing: RB-Std loses $45.1 \cdot 10^{-3}$ % per 1000 km in them versus $19.1 \cdot 10^{-3}$ % in cold trips. The Eco point changes energy by −36.0 % in cold trips, where cabin heating is the largest controllable load, and by −9.3 % in warm trips.

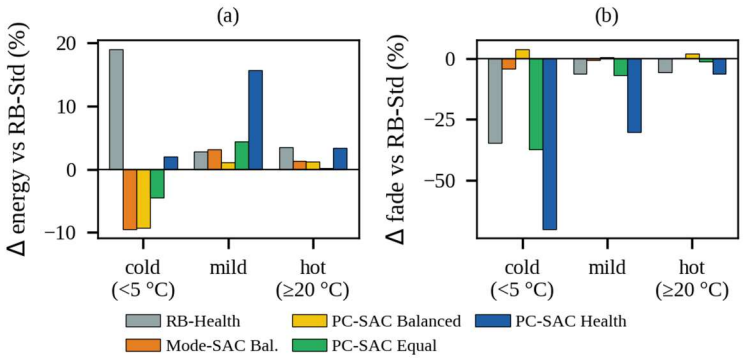


Fig. 5 Change of (a) energy and (b) capacity fade relative to RB-Std in cold (median outside temperature below 5 °C), mild and warm (20 °C and above) test trips; PC-SAC values are averaged over three seeds.

5.5. Ablation Study

Table 4 removes one component at a time with seed 0 and the same budget. Training each transition only under its behaviour preference gives a hypervolume of 0.0141, a scalar critic of the scalarised reward gives 0.0190, and removing the scenario probabilities gives 0.0222, against 0.0234 for the full method with the same seed. Removing the decomposed critics and preference relabelling therefore lowers the hypervolume by more than the seed spread (±0.0015). For the scenario features, the difference does not exceed the seed spread in the unfavourable direction, so this study does not demonstrate a benefit; the scenario probabilities carry little information beyond the speed, acceleration and power already in the state.

Table 4. Ablation results (HV: hypervolume; ΔJ at three preferences)

Variant	HV	ΔJ Eco	ΔJ Equal	ΔJ Health
PC-SAC (3 seeds)	0.0235±0.0015	+0.165	+0.040	+0.001
PC-SAC (seed 0)	0.0234	–	–	–
w/o decomposed critic	0.0190	+0.160	+0.180	+0.063
w/o pref. relabelling	0.0141	+0.090	+0.186	+0.158
w/o scenario features	0.0222	+0.132	+0.043	−0.017

5.6. Learned Behaviour

Fig. 6 replays a cold test trip (-5 °C, 12 min) under RB-Std and under two preferences of the same PC-SAC network. RB-Std keeps the regeneration intensity at its maximum (mean 1.00); the Health point lowers it to a mean of 0.16 while the pack is cold, where the plating branch of the severity model makes charge throughput most damaging, and the Eco point keeps a mean of 0.86. At the end of the trip the cabin is at 21.5 °C under RB-Std, 21.9 °C under Health and -2.9 °C under Eco, and the pack at 5.0, -1.2 and -1.4 °C, respectively. The preference thus acts as an interpretable run-time knob.

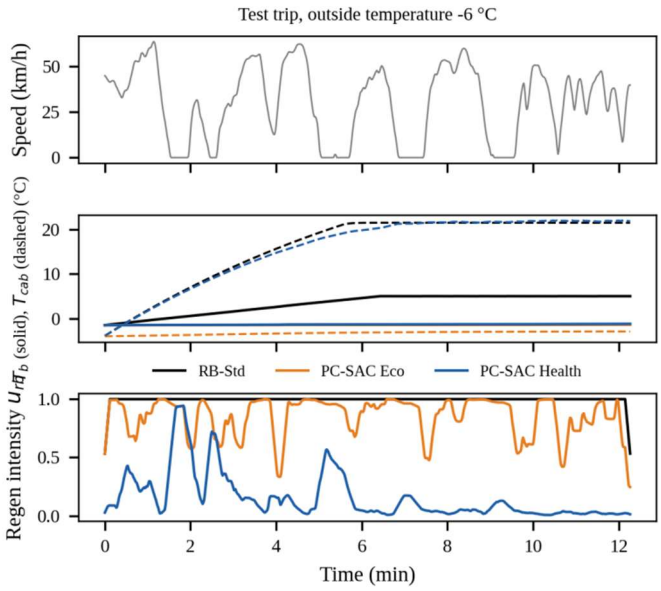


Fig. 6 A cold test trip replayed with RB-Std and with the Eco and Health preferences of one PC-SAC network: speed, battery (solid) and cabin (dashed) temperatures, and regeneration intensity (15-s moving average).

5.7. Fleet-Level Projection

To translate per-kilometre results into annual quantities, we annualised the energy that each of the 85 drivers drew from the workplace chargers (3,395 sessions over 319 days; median 114 kWh per year, interquartile range 33–437 kWh) and converted it into distance with the RB-Std consumption, a median of 627 km per driver and year supplied by workplace charging alone. Applying the test-set ratios, the Health point changes the annual energy of the median driver by +7 kWh and the annual capacity fade by −0.0051 percentage points, and the Balanced point changes the energy by −4 kWh. The projection inherits all model assumptions and ignores driving not charged at work; it indicates orders of magnitude only.

6. DISCUSSION AND LIMITATIONS

Three observations stand out. First, most of the energy is uncontrollable traction, so thermal and regeneration decisions have small per-step effects; without exogenous re-centring the critics could not resolve them, and with unconstrained thermal actions the agent exploited the model. These two design choices, rather than network size, decided whether learning worked at all. Second, the preference weights act on the training cost scales, on which the controllable energy looms large; energy-heavy preferences therefore trade comfort aggressively, and a deployment would combine the preference with hard comfort limits or rescaled weights. Third, the production rule set is hard to beat on energy and comfort, whereas the battery-health region of the front is where a learned, preference-conditioned policy adds options that the rules do not provide; the price of these options is energy and comfort, and it varies with the season.

The study has limitations. All results are simulated: the traction model is calibrated on real measurements, but road grade is missing and a driver-specific bias of up to about 24 % remains. Thermal parameters, BTMS and BMS characteristics and the heat-pump map are representative values rather than measurements of the VED vehicles. The ageing model was identified on LFP cells, whereas the calibrated vehicle uses a manganese-based chemistry, and the mirrored activation energy of the plating branch is an assumption; absolute fade values are indicative, and the relative comparisons are the main result. Episodes are single trips, so calendar ageing and the thermal history between trips are not modelled. PC-SAC was trained with three seeds and the baselines and ablations with one, on a single CPU core.

7. CONCLUSION

We presented a battery-health-aware, preference-conditioned soft actor-critic that jointly manages regenerative braking, cabin climate and battery thermal management of an EV and exposes the energy–health–comfort trade-off as a run-time preference. Trained and tested on real-world VED trips with a calibrated powertrain model, one network spanned capacity-fade changes from +10.8 % to −28.5 % and energy changes from −21.8 % to +7.7 % relative to a production-style controller, and reached battery-health operating points that neither the rule set nor mode-specific agents without a battery objective offer. Exogenous cost re-centring and effort-type thermal actions were decisive for trainability, and the decomposed critics and preference relabelling improved the front in ablations. Future work will add calendar ageing and inter-trip thermal history, comfort constraints, preview information from navigation, chemistry-specific plating data and hardware-in-the-loop validation.

REFERENCES

[1] Yuksel, T., & Michalek, J. J. (2015). Effects of regional temperature on electric vehicle efficiency, range, and emissions in the United States. Environmental Science & Technology, 49(6), 3974–3980. https://doi.org/10.1021/es505621s

[2] Al-Wreikat, Y., Serrano, C., & Sodré, J. R. (2021). Driving behaviour and trip condition effects on the energy consumption of an electric vehicle under real-world driving. Applied Energy, 297, Article 117096. https://doi.org/10.1016/j.apenergy.2021.117096

[3] Fiori, C., Ahn, K., & Rakha, H. A. (2016). Power-based electric vehicle energy consumption model: Model development and validation. Applied Energy, 168, 257–268. https://doi.org/10.1016/j.apenergy.2016.01.097

[4] Zhou, F. (2026). Moment-constrained side-force allocation for wind-assisted merchant ships: Hydrodynamic performance of retractable anti-leeway fins and flap rudders. Innovation and Technology Studies, 3(1), 27–41.

[5] Waldmann, T., Wilka, M., Kasper, M., Fleischhammer, M., & Wohlfahrt-Mehrens, M. (2014). Temperature dependent ageing mechanisms in lithium-ion batteries-A post-mortem study. Journal of Power Sources, 262, 129–135. https://doi.org/10.1016/j.jpowsour.2014.03.112

[6] Ebbesen, S., Elbert, P., & Guzzella, L. (2012). Battery state-of-health perceptive energy management for hybrid electric vehicles. IEEE Transactions on Vehicular Technology, 61(7), 2893–2900. https://doi.org/10.1109/TVT.2012.2203836

[7] Liu, T., Zou, Y., Liu, D., & Sun, F. (2015). Reinforcement learning of adaptive energy management with transition probability for a hybrid electric tracked vehicle. IEEE Transactions on Industrial Electronics, 62(12), 7837–7846. https://doi.org/10.1109/TIE.2015.2475419

[8] Hu, X., Liu, T., Qi, X., & Barth, M. (2019). Reinforcement learning for hybrid and plug-in hybrid electric vehicle energy management: Recent advances and prospects. IEEE Industrial Electronics Magazine, 13(3), 16–25. https://doi.org/10.1109/MIE.2019.2913015

[9] Wang, J., Liu, J., Ge, Y., & Jiang, B. (2023). Risk-aware ensemble learning for credit default prediction under economic uncertainty. World Journal of Economics and Business Research, 1(1), 18–32.

[10] Qi, C., Zhu, Y., Song, C., Yan, G., Xiao, F., Zhang, X., Cao, J., & Song, S. (2022). Hierarchical reinforcement learning based energy management strategy for hybrid electric vehicle. Energy, 238, Article 121703. https://doi.org/10.1016/j.energy.2021.121703

[11] Wang, J., Zhang, H., Wu, B., & Liu, W. (2025). Symmetry-guided electric vehicles energy consumption optimization based on driver behavior and environmental factors: A reinforcement learning approach. Symmetry, 17(6), Article 930. https://doi.org/10.3390/sym17060930

[12] Oh, G., LeBlanc, D. J., & Peng, H. (2022). Vehicle Energy Dataset (VED), a large-scale dataset for vehicle energy consumption research. IEEE Transactions on Intelligent Transportation Systems, 23(4), 3302–3312. https://doi.org/10.1109/TITS.2020.3035596

[13] Wang, J., Liu, P., Hicks-Garner, J., Sherman, E., Soukiazian, S., Verbrugge, M., Tataria, H., Musser, J., & Finamore, P. (2011). Cycle-life model for graphite-LiFePO4 cells. Journal of Power Sources, 196(8), 3942–3948. https://doi.org/10.1016/j.jpowsour.2010.11.134

[14] Hu, X., Li, S., & Peng, H. (2012). A comparative study of equivalent circuit models for Li-ion batteries. Journal of Power Sources, 198, 359–367. https://doi.org/10.1016/j.jpowsour.2011.10.013

[15] Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Proceedings of the 35th International Conference on Machine Learning (ICML) (pp. 1861–1870).

[16] Roijers, D. M., Vamplew, P., Whiteson, S., & Dazeley, R. (2013). A survey of multi-objective sequential decision-making. Journal of Artificial Intelligence Research, 48, 67–113. https://doi.org/10.1613/jair.3987

[17] Hayes, C. F., Rădulescu, R., Bargiacchi, E., Källström, J., Macfarlane, M., Reymond, M., Verstraeten, T., Zintgraf, L. M., Dazeley, R., Heintz, F., Howley, E., Irrthum, A., Mannion, P., & Roijers, D. M. (2022). A practical guide to multi-objective reinforcement learning and planning. Autonomous Agents and Multi-Agent Systems, 36(1), Article 26. https://doi.org/10.1007/s10458-022-09552-y

[18] Abels, A., Roijers, D. M., Lenaerts, T., Nowé, A., & Steckelmacher, D. (2019). Dynamic weights in multi-objective deep reinforcement learning. In Proceedings of the 36th International Conference on Machine Learning (ICML) (pp. 11–20).

[19] Yang, R., Sun, X., & Narasimhan, K. (2019). A generalized algorithm for multi-objective reinforcement learning and policy adaptation. In Advances in Neural Information Processing Systems (Vol. 32).

[20] Basaklar, T., Gumussoy, S., & Ogras, U. Y. (2023). PD-MORL: Preference-driven multi-objective reinforcement learning algorithm. In Proceedings of the International Conference on Learning Representations (ICLR).

[21] Yue, X., & Zhu, R. (2024). An explainable hybrid machine learning framework for cash-flow-conditioned multi-horizon liquidity risk early warning: An SME-oriented benchmark study. World Journal of Management Science, 2(1), 63–72.

[22] Mao, H., Venkatakrishnan, S. B., Schwarzkopf, M., & Alizadeh, M. (2019). Variance reduction for reinforcement learning in input-driven environments. In Proceedings of the International Conference on Learning Representations (ICLR).

[23] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In Proceedings of the 37th International Conference on Machine Learning (ICML) (pp. 1597–1607).

[24] Chen, J., Han, X., Zhang, X., & Yang, Y. (2023). Exact concept-level explanations for early detection of financial misstatements. Eurasia Journal of Science and Technology, 5(2), 34–48.

[25] Fujimoto, S., van Hoof, H., & Meger, D. (2018). Addressing function approximation error in actor-critic methods. In Proceedings of the 35th International Conference on Machine Learning (ICML) (pp. 1587–1596).

[26] Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., Kumar, V., Zhu, H., Gupta, A., Abbeel, P., & Levine, S. (2018). Soft actor-critic algorithms and applications. arXiv preprint arXiv:1812.05905.

[27] Asensio, O. I., Lawson, M. C., & Apablaza, C. Z. (2021). Electric vehicle charging stations in the workplace with high-resolution data from casual and habitual users. Scientific Data, 8(1). https://doi.org/10.1038/s41597-021-00956-1

[28] Li, P., Bian, M., & Lin, Y. (2026). Carbon-Aware Offline Reinforcement Learning for Joint Inventory Replenishment and Transportation Scheduling in Multi-Echelon Supply Chains. Highlights in Business, Economics and Management, 69, 251–262.

[29] Zhang, H., Li, Y., Wang, X., Liu, Z., & Yu, B. (2025). Hierarchical Deep Reinforcement Learning for Multi-Objective Integrated Circuit Physical Layout Optimization With Congestion-Aware Reward Shaping. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3610615