Trans-SAC: Robust and Transferable Maximum Entropy Reinforcement Learning for Heat Pump Control

Qiong Huang*
Karlsruhe Institute of Technology
Karlsruhe, Germany
qiong.huang@kit.edu

Adrian Till Assmuth
Karlsruhe Institute of Technology
Karlsruhe, Germany
adrian.assmuth@kit.edu

Felix Langner
Karlsruhe Institute of Technology
Karlsruhe, Germany
felix.langner@kit.edu

Veit Hagenmeyer
Karlsruhe Institute of Technology
Karlsruhe, Germany
veit.hagenmeyer@kit.edu

Benjamin Schäfer
Karlsruhe Institute of Technology
Karlsruhe, Germany
benjamin.schaefer@kit.edu

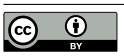
Abstract

Residential heating electrification through heat pumps is a cornerstone of building decarbonization, yet their potential for grid flexibility remains underutilized due to the difficulty of scaling control strategies across heterogeneous building stocks. Although reinforcement learning (RL) offers model-free adaptability, standard approaches often fail to scale, suffering from brittle policies that do not generalize across diverse thermal dynamics or discrete control actions that damage hardware. To address this, we introduce Trans-SAC (Transferable Soft Actor-Critic), a robust control framework designed to solve the "cold start" problem in city-scale deployments. By leveraging maximum entropy RL, Trans-SAC optimizes a dynamic trade-off between reward and entropy, ensuring continuous, hardware-safe actuation. Unlike prior work limited to fixed temperature bands, we target a challenging time-varying comfort objective under dynamic pricing and rigorously evaluate SAC against model predictive control (MPC) baselines, deep Q-networks (DQN) and proximal policy optimization (PPO), across a dataset of ten heterogeneous residential buildings calibrated with real-world weather and price data. Our experiments reveal that while the standard feature-freezing technique accelerates training, they remain suboptimal due to physical mismatches in building envelopes, and independent learning remains prohibitively sample-inefficient. In contrast, Trans-SAC's full-network adaptation strategy successfully bridges this gap, enabling robust generalization and reducing the training time and significantly outperforming other transfer baselines. This establishes Trans-SAC as a viable and data-efficient blueprint for aggregating large-scale heat pump fleets into smart buildings.

CCS Concepts

• **Computing methodologies** → **Reinforcement learning**; **Transfer learning**.

*Corresponding author



E-Energy '26, Banff, AB, Canada

ACM ISBN 979-8-4007-2011-6/26/06
https://doi.org/10.1145/3744255.3811723

Keywords

Heat Pump Control, Deep Reinforcement Learning, Transfer Learning

ACM Reference Format:
Qiong Huang, Adrian Till Assmuth, Felix Langner, Veit Hagenmeyer, and Benjamin Schäfer. 2026. Trans-SAC: Robust and Transferable Maximum Entropy Reinforcement Learning for Heat Pump Control. In *The 17th ACM International Conference on Future and Sustainable Energy Systems (E-Energy '26), June 22–25, 2026, Banff, AB, Canada.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3744255.3811723

1 Introduction

The electrification of residential heating via heat pumps is a critical strategy for decarbonizing the building sector. Beyond efficiency, heat pumps offer significant potential for demand response (DR) by leveraging building thermal inertia to shift electrical loads away from peak price periods [19]. However, realizing this potential at city-scale presents a formidable control challenge. Residential buildings are highly heterogeneous and vary widely in thermal capacitance, insulation levels, and usage patterns. While classical methods such as model predictive control (MPC) are effective, they rely on accurate physics-based models that are prohibitively expensive to develop and calibrate for millions of unique residential units [10].

Reinforcement learning (RL) has emerged as a promising model-free alternative, capable of learning optimal control policies directly from interaction data [16]. Despite its potential, two critical barriers prevent the widespread deployment of RL in real-world heating systems. First, many widely used RL methods operate on discrete action spaces, which is poorly matched to inverter-driven heat pumps and can induce high-frequency actuation that accelerates equipment wear [13, 24]. Second, policies trained on one building often transfer poorly to others, making building-by-building training prohibitively slow for large-scale adoption [28]. These difficulties become more severe when the controller must satisfy time-varying comfort bands under dynamic electricity prices rather than track a fixed setpoint [6].

In this work, we propose Trans-SAC (Transferable Soft Actor-Critic), a heat-pump control framework that combines maximum-entropy RL with cross-building transfer learning. We show that an end-to-end SAC transfer strategy can provide a robust and scalable

control pipeline for heterogeneous residential buildings. We evaluate the approach on ten buildings with distinct thermal dynamics and compare it against MPC, DQN, PPO, and alternative transfer strategies.

Our contributions are threefold. First, we formalize a continuous-control RL problem for heat-pump operation under dynamic pricing and time-varying comfort requirements. Second, we show empirically that SAC is substantially more robust than DQN and PPO in this setting, producing smooth, hardware-compatible actions and more reliable convergence across heterogeneous buildings. Third, we demonstrate that full-network transfer is consistently more effective than feature freezing and independent-learner baselines, reducing the cold-start burden for cross-building deployment.

The remainder of this paper is organized as follows: Section 2 reviews prior literature on building control, RL for HVAC, and transfer learning. Section 3 details the methodology, including the building thermal model, RL formulation, and transfer strategies. Section 4 presents the experimental setup and results. Finally, Section 6 concludes the paper with a summary of findings and future research directions. Our code repository is available on GitHub[1].

2 Related Work

Building control for heat pumps and demand response. Building energy management has long been dominated by rule-based and model-based strategies. Among these, MPC is widely regarded as a strong benchmark because it can explicitly optimize comfort and operating cost over a receding horizon [10, 15]. MPC has been applied successfully to flexible heating and thermal storage problems, including price-aware heat-pump scheduling [6]. Its main limitation is the modeling burden: each building requires an accurate and maintainable representation of its thermal dynamics, which is difficult to achieve for large residential fleets. This limitation motivates model-free or data-driven control approaches that can scale across heterogeneous buildings.

Reinforcement learning for HVAC and building energy control. RL has emerged as a promising alternative because it can optimize sequential decisions directly from data [14, 16, 25]. Prior studies have shown that RL can reduce operational cost while maintaining comfort in residential and commercial HVAC settings, but many evaluations remain confined to single buildings, simplified thermal models, or fixed setpoint tracking tasks [8]. For example, recent work in [4] demonstrates that PPO can learn effective control for a single building with time-varying setpoints. However, broader evidence suggests that algorithm choice matters substantially in building control: discrete-action methods such as DQN are not naturally aligned with continuously modulated heat-pump actuation [13, 24], while on-policy methods such as PPO can become conservative and sample-inefficient in long-horizon control tasks [20]. Maximum-entropy methods such as SAC [7] are therefore appealing because they combine continuous control with robust exploration, yet their role in transferable heat-pump control across heterogeneous buildings remains underexplored.

Transfer learning for RL-based building control. Transfer learning has been proposed to mitigate the cold-start problem by reusing knowledge from a source building when training on a new target building [21, 28]. In the building-control literature, prior work has explored policy reuse, action-guided learning, and parameter transfer to accelerate adaptation across environments [3, 9, 12, 17, 26]. These studies establish that transfer can reduce training effort, but they also show that transfer performance is sensitive to building mismatch, source-target similarity, and the chosen adaptation protocol. Most importantly, existing studies have not clearly established how a maximum-entropy continuous controller should be combined with cross-building transfer for heat pumps operating under dynamic comfort bands and price-responsive objectives. Our work addresses this gap by comparing zero-shot transfer, independent learners, frozen-feature transfer, and end-to-end fine-tuning within a unified SAC-based framework, and by evaluating these strategies across a diverse set of residential buildings. This literature positioning clarifies that the contribution of Trans-SAC is the integration and systematic validation of a scalable transfer pipeline for heterogeneous heat-pump control, rather than a new SAC variant in isolation.

3 Methodology

3.1 Problem Formulation

We formulate heat-pump control as a Markov decision process (MDP) in which the agent observes the current building condition and exogenous forecasts, selects a heat-pump modulation action, and receives a reward that balances comfort and operating cost.

State. At time step t, the state s_t contains the current thermal information of the building together with exogenous information required for anticipative control. In particular, the observation includes the current building temperatures, the active comfort target, and forecast information such as future electricity prices and weather-related disturbances over the control horizon. This formulation allows the agent to react not only to the present indoor condition but also to upcoming price spikes and ambient changes.

Action. The control action is the normalized heat-pump modulation level

$$a_t \in [0, 1], \tag{1}$$

where $a_t = 0$ corresponds to no heating and $a_t = 1$ corresponds to full heating power. For the continuous-control agents (PPO and SAC), this action is used directly. For the DQN baseline, the same control variable is discretized into five levels $\{0, 0.25, 0.50, 0.75, 1.0\}$ to enable a fair comparison under a shared control objective.

Reward. The general reward combines comfort preservation with price-aware operation:

$$\begin{aligned} r_t = &- \lambda_{\text{comfort}} \Big(\max\{0, \underline{T}_t - T_{in,t}\}^2 \\ &\qquad + \max\{0, T_{in,t} - \overline{T}_t\}^2 \Big) \\ &- \lambda_{\text{cost}}\, p_t\, a_t, \end{aligned} \tag{2}$$

where $T_{in,t}$ is the indoor air temperature, $[\underline{T}_t, \overline{T}_t]$ is the active comfort band, p_t is the electricity price, and λ_{comfort} and λ_{cost} weigh comfort and cost, respectively. For the pilot experiments with a fixed comfort target, this formulation reduces to a standard tracking reward by holding the comfort bounds constant; the detailed reward settings used in each experiment are given in Appendix A.4.

[1] https://github.com/Flywienix/rl-heat-pump-control

3.2 Building Thermal Dynamics

To ensure that the proposed controller is evaluated against realistic yet computationally scalable thermal behaviors, we utilize a dataset of linear state-space building models from Vallianos et al. [23]. The same modeling setup is also described in our previous work [8]. These models were obtained through system identification on large-scale smart thermostat data of 60,000 residential buildings and provide one building-specific tuple (A_i, B_i, C_i) for each residential unit. For building i, the thermodynamic evolution follows

$$x_{t+1} = A_i x_t + B_i u_t$$
$$o_t = C_i x_t$$

where $x_t \in \mathbb{R}^5$ is the latent thermal state, u_t contains the control action and exogenous disturbances, and o_t is the observed indoor temperature. Detailed state definitions and model components are provided in Appendix A.1.

Heterogeneity and Generalization. A key challenge in this work is the "sim-to-real" gap caused by the diversity of the building stock. The system matrices A_i and B_i encode building-specific thermal properties such as insulation quality and thermal capacitance. By randomly sampling ten distinct models (Buildings A–J) from this dataset, we create a challenging testbed that requires the RL agent to learn a robust control policy capable of generalizing across widely varying time constants and gain parameters.

3.3 Reinforcement Learning Framework

To enable continuous control while ensuring robustness against modeling errors, we adopt the SAC algorithm proposed in [7]. Unlike standard RL methods that seek only to maximize the expected sum of rewards $\sum r_t$, SAC optimizes a maximum entropy objective:

$$J(\pi) = \sum_{t=0}^{T} \mathbb{E}_{(s_t,a_t)\sim\rho_\pi}[r(s_t, a_t) + \alpha\mathcal{H}(\pi(\cdot|s_t))],$$

where $\mathcal{H}(\pi(\cdot|s_t))$ is the entropy of the policy and α is the temperature parameter determining the relative importance of exploration.

In the context of heterogeneous building control, this entropy term provides two critical advantages over deterministic baselines such as DDPG [11] or low-entropy methods such as PPO. The agent is encouraged to explore widely during the training phase. As shown in our results (see Section 4.3), this prevents the policy from collapsing into the "safety traps" (e.g., constant overheating) observed in PPO. SAC outputs a stochastic policy (typically a Gaussian distribution). By sampling actions from this distribution and squashing them via a tanh function, the resulting control signal is naturally smoother and less prone to high-frequency chattering that damages inverter compressors. We utilize the actor-critic architecture (Figure 1) where the actor network $\pi_\phi(a_t|s_t)$ outputs the mean and standard deviation of the compressor power. The critic network Q_θ estimates the soft Q-value to mitigate overestimation bias. The temperature α is automatically tuned to maintain a target entropy, ensuring consistent exploration throughout the learning process.

3.4 Scalability and Transfer Learning Strategies

The fundamental barrier to the mass adoption of RL in HVAC is training time. Training a high-performing agent from scratch takes millions of timesteps, which is equivalent to years of real-world

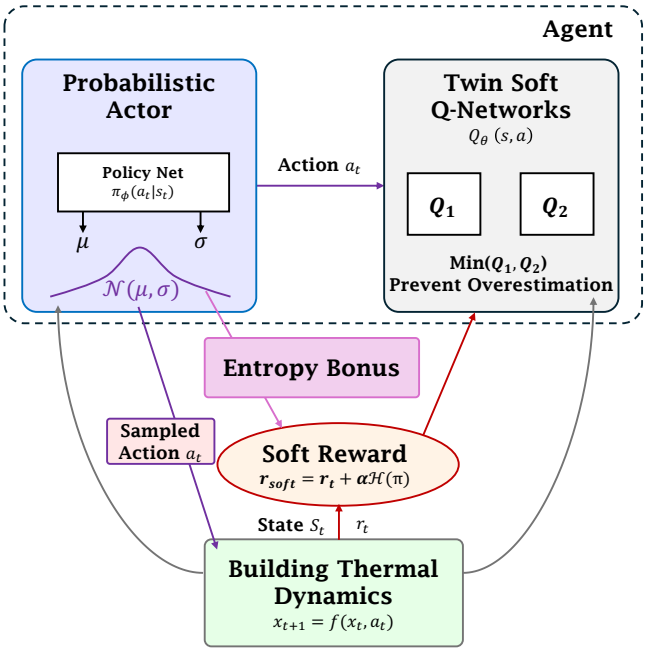


Figure 1: Schematic of the Soft Actor-Critic (SAC) Loop for Heat Pump Control. The Probabilistic Actor Network outputs a distribution over continuous compressor modulation actions. This stochasticity is reinforced by the Entropy Feedback loop, where the reward is augmented with an entropy bonus $\alpha\mathcal{H}(\pi)$. The Critic Network (Twin Soft Q-Networks) estimates the soft Q-value to guide policy updates.

operation. Transfer learning (TL) attempts to solve this by reusing the knowledge from a trained "source" agent to accelerate learning on a new "target" building.

Across all transfer experiments, the source SAC agent is first trained on one building until convergence. The resulting source-domain parameters are then used in different ways to construct a target-domain agent, which is subsequently evaluated or trained on a new building under the same state, action, and reward formulation. Importantly, the underlying SAC optimization remains unchanged during adaptation; transfer affects only the initialization of the target agent and which network components are allowed to update.

We evaluate four transfer regimes:

Zero-shot transfer. A converged source policy is deployed directly on the target building without any further parameter updates. This setting tests whether the control logic learned in the source domain is immediately usable under a new building's thermal dynamics.

Independent learners. Following [12], we also consider an action-guided learning baseline in which a new target agent is trained from scratch while receiving the action proposed by a pre-trained source agent as an additional input. In this regime, the source agent itself is fixed and only provides guidance; the target agent learns its own policy from random initialization. This baseline isolates the value of source-policy guidance without parameter transfer.

Frozen-Feature transfer. The target agent is initialized with the learned source parameters of both the actor and critic networks. The lower layers of these networks are then frozen, while the upper layers remain trainable and are fine-tuned on the target building [27]. This setting tests the hypothesis that low-level thermal-control features are transferable whereas higher-level control gains must adapt to the new environment.

Trans-SAC. The proposed Trans-SAC strategy performs full parameter transfer within the SAC architecture. After source training, the learned source parameters are used to initialize the target actor and twin critic networks, and the entire target network is then fine-tuned end-to-end on target-building data. Unlike frozen-feature transfer, no layers are locked. Thus, Trans-SAC preserves the source-domain control logic at initialization while allowing the agent to recalibrate its internal representation and control gains to the target building dynamics.

This formulation makes the interaction between RL and transfer learning explicit: transfer determines how source knowledge is injected into the SAC agent, while target-domain learning continues through standard SAC updates. The performance comparison among these regimes therefore reveals whether transfer should be used only for inference, only for guidance, or for full end-to-end adaptation. The TL architecture is illustrated in Figure 2.

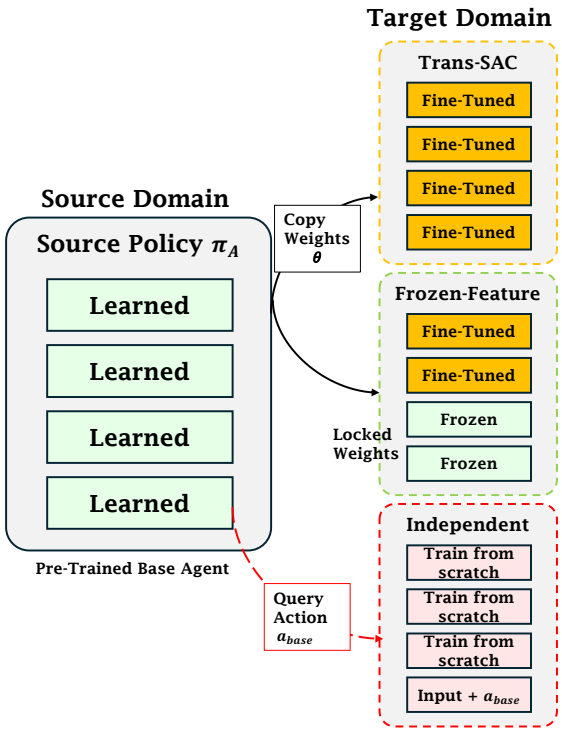


Figure 2: The Transfer Learning Pipeline. Trans-SAC: All neural network layers are "Transferred". Frozen-Feature transfer: Lower layers are "Frozen" (weights are locked), upper layers are "Fine-Tuned". Independent learners: A pre-trained base agent suggests actions to guide the learning of a new agent from scratch.

4 Experiments and Results

4.1 Simulation Environment

A detailed description of the building dynamics is included in Appendix A.

4.2 Reinforcement Learning Training Setup

We implement all agents using the Stable-Baselines3 library. We compare different algorithms including: DQN as standard discrete baseline; PPO as a standard continuous baseline; SAC as a max entropy architecture method; and we also use MPC for theoretical optimal benchmark. The state space dimension is fixed at 15 for all algorithms. For the network architecture, we employ multi-layer perceptrons (MLPs) with ReLU activations. Optimization tests with different parameters are conducted to find the best hyperparameters for each algorithm. The final hyperparameters used in the experiments are listed in Appendix B.1.

We consider two types of temperature comfort bands for the RL agents to track:

1. Fixed temperature band ([20, 22]°C) [8]
2. Time-varying temperature band

Both cases also include price-aware control terms. In the second configuration, an occupancy schedule is obtained from [22], which distinguishes occupant presence between "home", "away", and "sleep", and differentiates between weekends and weekdays. Figure 11 visualizes the time-varying temperature band.

4.3 Training Performances and Robustness Analysis

With the hyperparameters set, we first compare the training performance of SAC against the two baselines (PPO and DQN) on all buildings. Each algorithm is trained five times with different random seeds; Figure 3 reports the mean reward with the shaded area representing one standard deviation across seeds. Each subplot corresponds to one building, highlighting the final performance reached after training under identical experimental settings, including the same time-varying temperature comfort band and electricity price signal.

The SAC agent consistently outperforms both DQN and PPO agents in terms of cumulative reward during training in all buildings. The SAC (Blue) demonstrates superior sample efficiency, steadily converging to the optimal policy. DQN (Orange) suffers from discretization error, plateauing at a suboptimal local optimum significantly below SAC's performance. PPO (Green) fails to escape the suboptimal local optima, resulting in a significantly lower accumulated reward. Even in the single building case where DQN manages to learn a stable policy, it fails to match the asymptotic performance of SAC. This performance gap is a direct consequence of action discretization, which prevents the DQN agent from fine-tuning the heat pump modulation to the exact thermodynamic requirements of the building.

We also conduct a further hyperparameter sweep *n_steps* and *batch size* to identify configurations that might improve performance. Figure 12 presents the analysis of the PPO agent's failure mode on Building F. Despite extensive hyperparameter tuning and

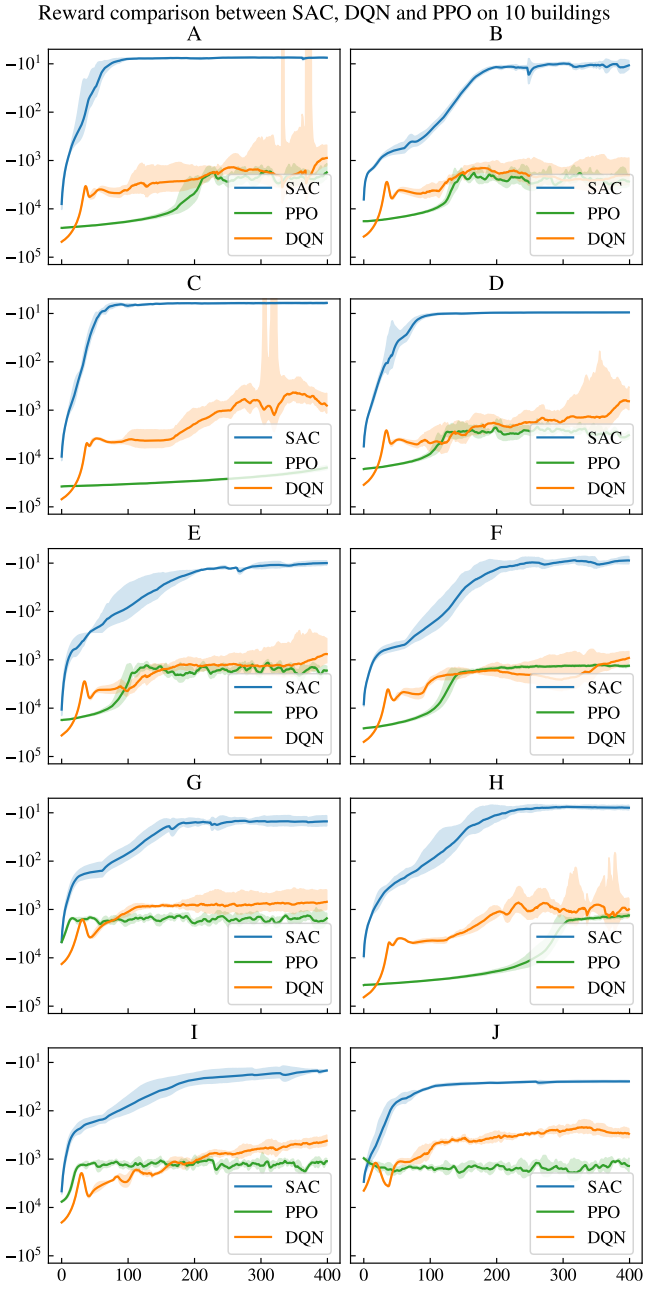


Figure 3: Generalization Performance: DQN, PPO and SAC across ten Buildings (A to J). Solid lines show the mean reward over five runs; shaded areas indicate one standard deviation.

prolonged training horizons, PPO consistently converges to suboptimal heating strategies, indicating that its on-policy optimization and clipped updates struggle with delayed comfort penalties and time-varying temperature constraints in this task. Figure 13 illustrates the behavior of the PPO and DQN agent when trained with a time-varying temperature comfort band and its performance over three days period. The upper panel shows the indoor air temperature together with the lower and upper bounds of the target comfort interval, as well as the outdoor temperature. The lower panel shows the corresponding electricity price signal and the heat pump activation level selected by the agent. Despite the presence of a dynamic comfort band, the PPO agent consistently drives the indoor temperature toward the upper bound and often exceeds it. The heating action remains persistently high and shows limited response to both the comfort constraints and the electricity price signal. This indicates that the agent fails to learn an effective trade-off between comfort and cost, instead converging to a degenerate policy that prioritizes continuous heating. The resulting behavior explains the low and stagnant reward values observed during training. Although the reward function penalizes deviations outside the comfort band, the PPO policy does not adapt its actions to correct these violations, even after extensive hyperparameter tuning and longer training horizons.

4.4 Control Quality Analysis

Across all ten buildings, the SAC agent consistently achieves higher rewards and converges to stable control policies. In contrast, the DQN agent fails to learn effective control behavior in most buildings, resulting in significantly lower reward values and high variability in performance. The consistent performance of the SAC agent can be attributed to its entropy-regularized objective and continuous action formulation, which promote stable exploration and smoother policy updates. These properties appear crucial when scaling training to multiple heterogeneous buildings with time-varying temperature targets. The results indicate that while the DQN agent performs well in a single-building setting with fixed comfort targets [8], it does not generalize robustly across diverse buildings. The SAC agent, by contrast, demonstrates strong robustness and scalability, making it better suited for multi-building heat pump control scenarios.

To validate the quality of the learned policy, we compare the SAC agent against a theoretical optimum derived from a Model Predictive Controller (MPC) with perfect foresight [15]. Figure 4 visualizes the temperature tracking and control actions for Building J over a representative period. The top panel shows the indoor temperature tracking, while the bottom panel displays the heat pump modulation alongside the dynamic electricity price. Unlike discrete baselines that suffer from high-frequency chattering, the SAC agent generates smooth, continuous modulation signals that closely follow the MPC trajectory. This validates that the Maximum Entropy framework successfully balances exploration with precise control, reducing mechanical stress on the inverter compressor. The agent exhibits intelligent load shifting without any explicit rule-based programming. As observed at time steps, e.g., $t \approx 50$ and $t \approx 120$, the agent anticipates upcoming price spikes (indicated by the price signal) and aggressively pre-heats the building during low-price intervals, storing thermal energy to coast through high-cost periods. This behavior mirrors the optimal MPC strategy, demonstrating that SAC effectively learns to exploit the building's thermal inertia for cost savings while maintaining comfort.

Table 1 summarizes the aggregated average performance and standard deviation across Building A-J, comparing the proposed

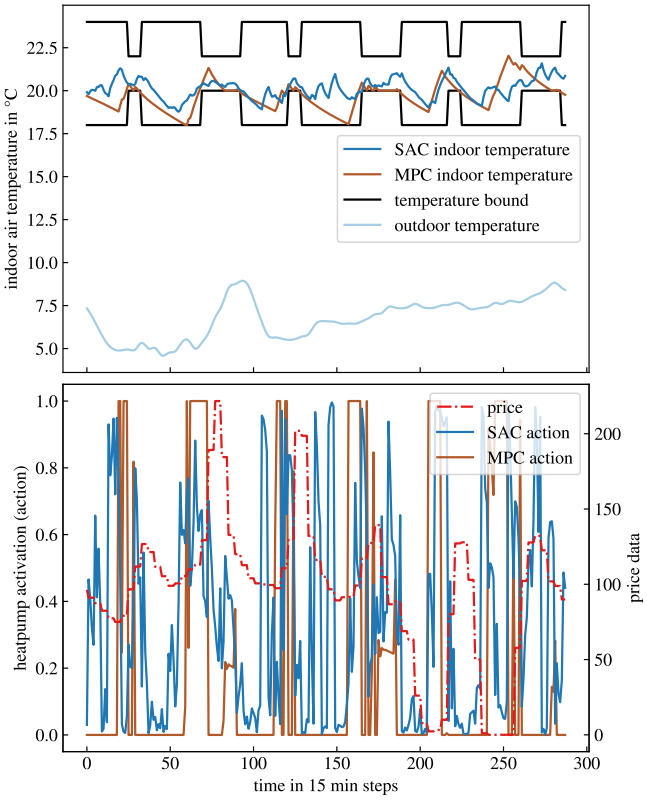


Figure 4: Indoor temperature and heat pump control actions of a SAC agent and MPC trained on Building J.

DRL agents against the MPC benchmark. As anticipated, the MPC establishes the theoretical performance upper bound, maintaining the strictest environmental comfort with a temperature bound deviation of 0.11 ± 0.15% and a negligible maximum temperature deviation of 0.25 ± 0.33°C. Among the learning-based methods, the SAC agent demonstrates the highest stability and sample efficiency. It achieves the best reward of -10.43 ± 4.68 and converges within 214 ± 109 episodes. Although the SAC agent incurs an operational cost gap of + 26.10 ± 7.42% compared to the MPC reference, it offers a favorable trade-off compared to the other model-free baselines. Conversely, the DQN agent exhibits a higher variance in comfort control (3.01 ± 1.95% deviation) and fails to converge. The PPO agent proves unstable in this environment, resulting in a significantly degraded policy with a temperature deviation rate of 72.25 ± 14.40% and an operation cost gap of +84.45 ± 27.49%, rendering it impractical for this specific building control formulation.

5 Transfer Learning

5.1 Pilot Validation

For each building, we have five different τ values: τ_e for envelope, τ_m for internal mass, τ_i for interior, τ_h for heater and τ_s for sensor. For each parameter $k \in \{e, m, i, h, s\}$, we normalize the time constant for each building b with $\hat{\tau}_k^{(b)} = \frac{\tau_k^{(b)} - \mu_k}{\sigma_k}$, where μ_k denotes the mean and σ_k is the standard deviation. We apply a **weighted euclidean distance** D metric to represent the similarity between source building and target buildings:

$$D(A, X) = \sqrt{\sum_{k \in \{e,m,i,h,s\}} w_k \left(\hat{\tau}_k^{(A)} - \hat{\tau}_k^{(X)}\right)^2},$$

where A is the source building, X is the target building, w_k are the importance weights assigned to each physical parameter. Small D of the building indicates physical similarity and large D indicates physical dissimilarity. We set the weights as $w_e = 0.90$, $w_m = 0.03$, $w_i = 0.07$, and $w_{h/s} = 0.00$, based on domain knowledge, prioritizing envelope and internal mass time constants due to their significant influence on thermal dynamics. To determine the baseline viability of transferring policies without re-training, we conduct a **zero-shot transfer** pilot study by selecting a diverse cohort of five buildings spanning four distinct Köppen-Geiger climate zones [1]. The source agent is trained on **Building P1**, located in a Humid subtropical zone (Cfa). The evaluation set is stratified to distinguish between physical and climate mismatches, which includes **Building P2** (also Cfa) to test intra-climate transfer, alongside three cross-climate targets: **P3** (Warm summer continental, Dfb), **P4** (Warm-summer Mediterranean, Csb), and **P5** (Hot-summer continental, Dfa).

We evaluate the transfer performance using the fixed comfort band of [20, 22]°C setpoint. As shown in Figure 14, our pilot zero-shot experiments reveal a crucial nuance that while the low physical distance (D) successfully predicts the transfer success for Building P2 from the same climate zone, it fails for Building P4 from different climate zones. Despite P4 having a minimal physical mismatch with the source, the zero-shot policy struggled to maintain comfort. This indicates that the learned policy is overfitted to the specific structural parameters and system sizing of the source domain. Although the weather input remains constant, the thermodynamic response of the target building differs significantly, leading to a control gain mismatch. Consequently, zero-shot transfer is brittle to climate shifts, even between physically similar buildings. This finding underscores the necessity of the following transfer methods with buildings in the same climate zone.

Figure 15 compares the behavior of the agent on the source building (familiar) versus the target Building P2 (alien). Although the zero-shot agent struggles to maintain strict comfort compliance in the target building and frequently breaches the lower 20°C bound compared to the tight control seen in the source domain, the control strategy remains structurally sound. The bottom panel reveals that the heating actions are well-aligned temporally, with both agents correctly pre-heating in response to price dips and coasting during peaks. This indicates that while the *magnitude* of the control action requires fine-tuning to match the new building's gain, the *logic* of the policy (Reacting to Price/Dynamics) is successfully transferred.

5.2 Large-Scale Transfer Performance

To determine the optimal mechanism for knowledge reuse, we evaluated three SAC-based transfer regimes on the target buildings (Building B to J) using the policy pre-trained on Source Building A, and additionally included DQN and PPO with and without transfer as algorithmic baselines. The reward curves of different transfer methods across ten buildings are shown in Figure 16. Figure 5

Table 1: Average performance comparison of the different methods on Buildings A-J

Method	Max reward / episode length	Episode of convergence	Time steps deviating from temperature bound	Maximum temperature deviation	Operational cost gap
MPC	-	-	0.11 ± 0.15%	0.25 ± 0.33°C	0% (Ref)
SAC	**-10.43 ± 4.68**	214 ± 109	0.43 ± 0.28%	0.65 ± 0.47°C	+26.10 ± 7.42%
DQN	-24.46 ± 12.89	-	3.01 ± 1.95%	2.02 ± 1.24°C	+28.27 ± 6.46%
PPO	-866.39 ± 87.01	-	72.25 ± 14.40%	12.06 ± 4.35°C	+84.45 ± 27.49%

presents the reward trajectories for these methods over three runs on Buildings C, E, and I, and Figure 6 quantifies the convergence speedup achieved by the SAC transfer strategies using the Time-to-Convergence (TTC) metric, in which we strictly qualify a run as 'successful' only if the reward converged. The bar graph compares the training episodes required to reach convergence across different initialization strategies.

Independent Learner: The independent learner serves as the baseline, illustrating the significant computational cost of learning thermal control policies from scratch. As shown in the bottom plot of Figure 16 and the Independent Learner column of Figure 5, these agents face a severe "cold start" problem, initializing with rewards as low as -10^5 due to random exploration. The learning curves are characterized by high variance and a slow, jagged ascent, indicating that the agents spend considerable time exploring unsafe or inefficient actions, such as overheating or underheating, before discovering a viable policy. Consequently, this approach requires an average of 350 episodes to converge, confirming that training without prior knowledge is computationally prohibitive for scalable deployment.

Frozen-Feature transfer: The frozen-feature transfer strategy attempts to mitigate this cost by reusing pre-trained feature extractors while fine-tuning only the control head. The hypothesis is that the agent can reuse the learned representations of building thermal dynamics in the early layers while relearning specific control gains to the new building. The performance trajectory confirms this hypothesis. Unlike the independent learner, there is no initial performance dip; the agent begins training with high rewards. This indicates that the transferred control logic is immediately effective. Although this method successfully avoids the initial random exploration phase, the middle plot of Figure 16 and the middle column of Figure 5 reveal its limitation: the assumption of universal thermal features is not strictly valid across diverse building envelopes. Because the lower layers are frozen, the agent cannot correct its internal physics model when the target building's thermal time constant (τ_e) differs significantly from the source. This leads to marked instability, where agents for physically distinct buildings (e.g., Building J) exhibit oscillating or degrading performance before recovering. As a result, the average convergence time is 145 episodes, and they have unpredictability and high variance.

Trans-SAC: Trans-SAC method demonstrates superior sample efficiency and robustness in all target environments. By initializing with source weights but allowing the entire network to fine-tune, this approach effectively bypasses the cold start phase, beginning with rewards near -10^1 immediately upon deployment. The top plot shows a tight clustering of learning curves, indicating that the agent can rapidly adapt to both minor and major physical mismatches (such as the extreme envelope lag in Building G in Figure 16). This flexibility allows the agent to realign its internal physics representation with the new environment's dynamics, achieving stable convergence very quickly.

While the independent learner requires 350 episodes on average to learn an effective policy, the Trans-SAC approach drastically reduces this computational burden, achieving stable convergence in just ≈ 84 episodes on average, a 2.7 times speedup (63% reduction) over the baseline learner without knowledge transfer (no transfer), confirming that initializing the agent with source-domain knowledge effectively solves the "Cold Start" problem. Furthermore, compared with the frozen-feature transfer method (transferring only low-level logic), which also shows a strong improvement, the learned control features are highly robust.

DQN and PPO baselines (with and without transfer): Beyond the primary Trans-SAC results, we evaluate the baseline performance of DQN and PPO to contextualize the advantages of the maximum entropy framework. As shown in the Independent Learner column of Figure 5, both DQN and PPO trained from scratch struggle with the cold start problem, often initializing with rewards as low as -10^5 due to random exploration of inefficient heating strategies. While the transfer protocol (solid lines) provides an initial performance boost for these baselines, they consistently converge to lower reward plateaus compared to SAC. Specifically, the DQN agent suffers from discretization errors that prevent fine-tuned modulation, leading to chattering that can degrade hardware. Meanwhile, PPO frequently collapses into safety traps, such as constant overheating, to avoid comfort penalties, failing to learn the sophisticated cost-comfort trade-offs achieved by SAC's entropy-regularized objective. These results demonstrate that while transfer learning accelerates the initial learning phase for all algorithms, the choice of the underlying RL engine remains critical for asymptotic control quality.

Figure 7 illustrates the operational behavior of the SAC agent in the Target Building E following the Trans-SAC process. The top panel confirms that the agent successfully maintains the indoor temperature within the user's comfort bounds, exhibiting stable control despite the distinct thermal dynamics of the new environment. The lower panels reveal the agent's economic intelligence where the heat pump activation is strongly anti-correlated with the electricity price signal. The agent strategically performs "pre-heating" during off-peak hours when electricity is cheap, storing thermal energy in the building's mass. The heat pump then idles

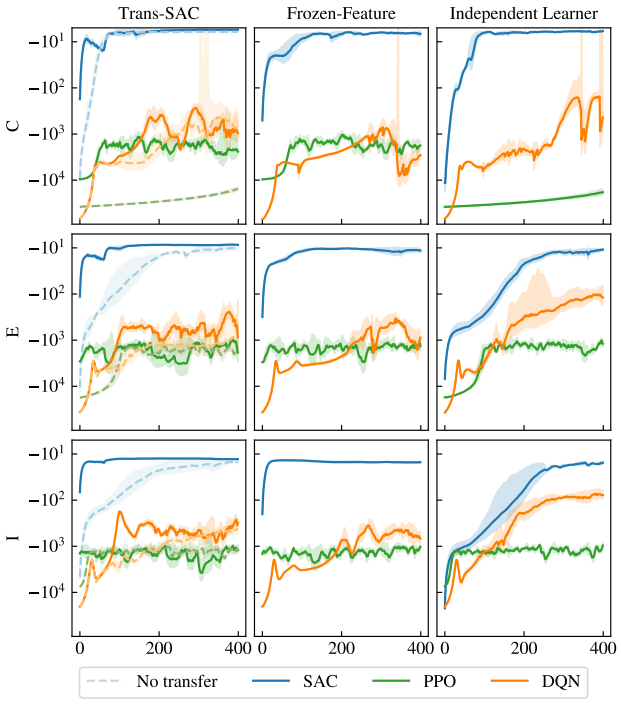


Figure 5: Comparative reward curves across Buildings C, E, and I for SAC, DQN, and PPO with and without transfer. Solid lines represent agents initialized with source-domain weights (transfer), and dashed lines represent the corresponding no-transfer baselines. No-transfer baselines are shown only in the Trans-SAC column. They are identical across all columns.

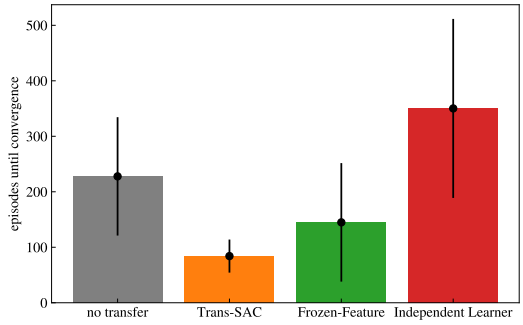


Figure 6: Convergence Speedup: Time-to-convergence comparison among No transfer, Trans-SAC, Frozen-Feature transfer learning and Independent Learner.

during price spikes (peak hours), allowing the temperature to drift slowly within the comfort zone. This behavior demonstrates that the Trans-SAC method successfully preserves the high-level planning logic, specifically the economic load shift, learned from the source domain, effectively applying it to the target building without the need to relearn the fundamentals of energy efficiency.

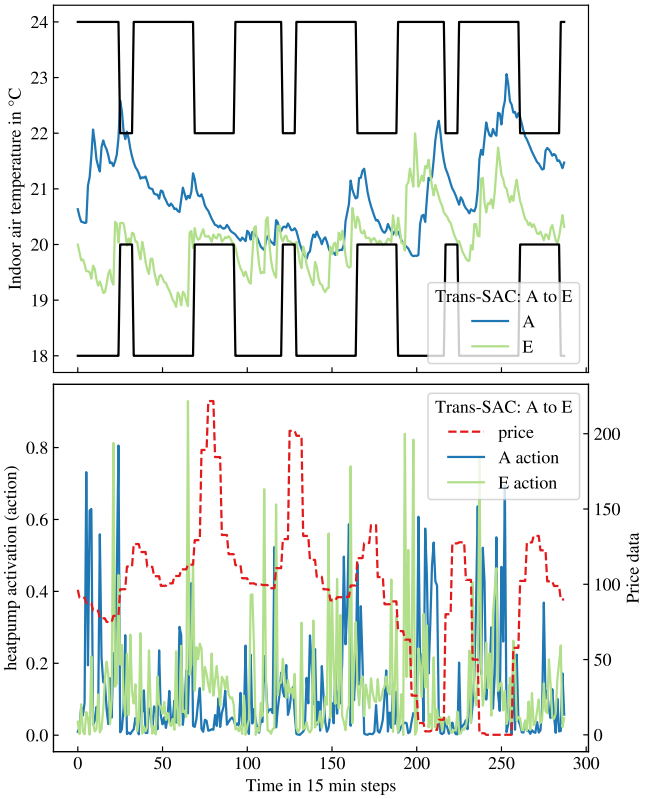


Figure 7: Operational performance of the SAC agent on Target Building E following Trans-SAC from Source Building A.

To understand the determinants of transfer efficiency, we analyzed the relationship between the physical similarity of the buildings and the computational cost of adaptation. Figure 8 shows the episodes to convergence against the proposed weighted physical distance metric D. The analysis reveals a strong positive correlation ($R \approx 0.72$) in Tran-SAC, confirming that training overhead is linearly dependent on the kinematic disparity between the source and target domains. Buildings with low D (e.g., B, C, D, E) converge in ≈ 70 episodes. The outliers G ($D = 1.89 \rightarrow 160$ episodes) and F ($D = 1.48 \rightarrow 110$ episodes) are perfectly predicted by their massive τ_e mismatch. There are also anomaly Building J ($D = 1.39$) converged fast (66 episodes). This suggests that transferring from a slow building (Source A) to a fast building (Target J) is easier than the reverse (Source A $\rightarrow$ Target G). Meanwhile, frozen-feature transfer ($R \approx 0.52$) shows a moderate correlation. It follows the trend, but is "noisier". Since the feature layers are frozen, the agent cannot fully adapt to the τ_e mismatch, leading to erratic convergence (e.g., Building G takes 594 episodes). In addition, the independent learner ($R \approx 0.36$) shows weak/no correlation, and the difficulty depends on the exploration complexity, not the similarity to Building A.

6 Conclusion and Outlook

The present paper addresses the dual challenges of algorithmic robustness and deployment scalability in residential Demand Response. Our extensive comparative analysis across heterogeneous

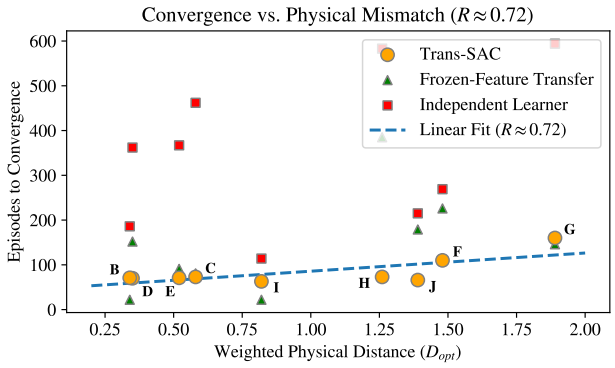


Figure 8: Correlation between physical similarity (D) and transfer efficiency (Time-to-Convergence).

building models establishes maximum entropy reinforcement learning (SAC) as the robust engine of modern thermal control. By optimizing a trade-off between reward and entropy, SAC successfully navigates the complex, continuous action spaces where discrete baselines (DQN) fail and on-policy methods (PPO) stagnate. We demonstrate that this stochastic policy prevents the deterministic collapse often seen in rule-based approximations, delivering hardware-safe modulation that respects user comfort constraints.

However, a robust engine alone is insufficient for city-scale adoption due to the prohibitive cost of training agents from scratch. To this end, we validate that Transfer Learning acts as the essential fuel for scale. Our results confirm that Trans-SAC reduces the average convergence time from 350 episodes in the independent learner to just 84 episodes, representing a 4.2 times speedup (76% reduction in data requirements). Crucially, we show that the frozen-feature transfer approach by freezing feature layers also speeds up training (average 145 episodes) but is slower than Trans-SAC, as it prevents the agent from correcting its internal physics representation when facing diverse thermal envelopes. Thus, unfreezing the entire network is necessary to bridge the sim-to-real gap effectively.

Our results confirm that even within a single climate zone, the diversity of the building stock poses a significant barrier to scalability. Future work will extend this framework to cross-climatic transfer, but our current findings establish that physics-aware fine-tuning is the prerequisite for aggregating heterogeneous fleets, regardless of their geographic proximity. Our analysis from Trans-SAC further reveals a strong linear correlation ($R \approx 0.72$) between the transfer cost and the weighted physical distance (D) of the buildings. Specifically, we identify that the envelope time constant (τ_e) is the dominant predictor of transfer difficulty; mismatches in thermal insulation require linearly more data to correct than mismatches in heating systems. A notable exception is the asymmetry observed when transferring from slow to fast buildings versus the reverse, suggesting that future work should explore curriculum learning strategies that sequence transfers from fast to slow dynamics. In addition, instead of a single universal source policy, future deployments should utilize a library of source agents trained on distinct clusters of τ_e (e.g., "Leaky/Low-Inertia" vs. "Sealed/High-Inertia"). By matching a target building with its nearest physical neighbor (D_{min}), we could theoretically reduce the convergence time to the robustness plateau (< 70 episodes) for every installation.

To further minimize the initialization gap, we propose using a hierarchical clustering framework for source policy selection. This two-tier strategy would first segment the building stock by Climate Zone to align external disturbance profiles, and subsequently cluster within these groups based on internal physical parameters, specifically the envelope time constant τ_e. By matching target buildings to source agents that share both climate and physical characteristics, this approach ensures that the transferred policy is pre-conditioned on both relevant environmental scenarios and the specific thermal dynamics of the target system. A key limitation of the present study is the reliance on identified Linear Time-Invariant (LTI) building models. While these provide a tractable framework for large-scale benchmarking across 60,000 homes, they may not fully capture the nonlinear thermodynamic and operational complexities of real-world structures. Future work should therefore evaluate Trans-SAC on non-LTI building dynamics, where time-varying parameters, nonlinear heat-transfer effects, and equipment nonlinearities may alter both control performance and transferability. Furthermore, because the identified state-space models used here do not provide the granular geometric or material metadata required for high-fidelity reconstruction (e.g., specific wall assemblies or orientation), a direct transition to more complex simulators was not feasible for this building set. Future research will focus on validating Trans-SAC in standard high-fidelity benchmark environments, such as EnergyPlus or Modelica [18], to ensure that the observed transfer-performance gains and physics-aware fine-tuning benefits persist in non-linear, high-dimensional settings.

Acknowledgments

The authors gratefully acknowledge funding from the Helmholtz Association under grant No. VH-NG-1727 and the Networking Fund through Helmholtz AI and within the framework of the Program-Oriented Funding POF IV in the program Energy Systems Design (ESD, project number 37.12.01). The authors acknowledge support by the state of Baden-Württemberg through bwHPC. We also thank Gökhan Demirel for his valuable discussions.

References

[1] Hylke E Beck, Niklaus E Zimmermann, Tim R McVicar, Noemi Vergopolan, Alexis Berg, and Eric F Wood. 2018. Present and future Köppen-Geiger climate classification maps at 1-km resolution. *Scientific data* 5, 1 (2018), 1–12. doi:10.1038/sdata.2018.214

[2] Bundesnetzagentur. 2025. SMARD. https://www.smard.de/home/downloadcenter/download-marktdaten/.

[3] Davide Coraci, Silvio Brandi, Tianzhen Hong, and Alfonso Capozzoli. 2023. Online transfer learning strategy for enhancing the scalability and deployment of deep reinforcement learning control in smart buildings. *Applied Energy* 333 (2023), 120598. doi:10.1016/j.apenergy.2022.120598

[4] Gökhan Demirel, Ömer Ekin, Jianlei Liu, Luigi Spatafora, Kevin Förderer, and Veit Hagenmeyer. 2025. Advanced Deep Reinforcement Learning for Heat Pump Control in Residential Buildings. In *2025 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT Europe)*. IEEE, Malta, 1–5. doi:10.1109/ISGTEurope64741.2025.11305332

[5] Joint Research Centre European Commission. 2025. Photovolatic Geographical Information System. https://re.jrc.ec.europa.eu/pvg_tools/de/tools.html.

[6] David Fischer and Hatef Madani. 2017. On heat pumps in smart grids: A review. *Renewable and Sustainable Energy Reviews* 70 (2017), 342–357. doi:10.1016/j.rser.2016.11.182

[7] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*. PMLR, Stockholm, Sweden, 1861–1870.
[8] Qiong Huang, Adrian Till Assmuth, Felix Langner, Benjamin Schäfer, and Veit Hagenmeyer. 2026. Deep Reinforcement Learning for Price-Aware Building Heating Control. *KI - Künstliche Intelligenz* (2026), 1–9. doi:10.1007/s13218-026-00908-0
[9] Kevlyn Kadamala, Des Chambers, and Enda Barrett. 2024. Enhancing HVAC control systems through transfer learning with deep reinforcement learning agents. *Smart Energy* 13 (2024), 100131. doi:10.1016/j.segy.2024.100131
[10] Michaela Killian and Martin Kozek. 2016. Ten questions concerning model predictive control for energy efficient buildings. *Building and Environment* 105 (2016), 403–412. doi:10.1016/j.buildenv.2016.05.034
[11] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2016. Continuous control with deep reinforcement learning. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). Open-access, Puerto Rico. http://arxiv.org/abs/1509.02971
[12] Paulo Lissa, Michael Schukat, Marcus Keane, and Enda Barrett. 2021. Transfer learning applied to DRL-Based heat pump control to leverage microgrid energy efficiency. *Smart Energy* 3 (2021), 100044. doi:10.1016/j.segy.2021.100044
[13] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *nature* 518, 7540 (2015), 529–533. doi:10.1038/nature14236
[14] Zoltan Nagy, Gregor Henze, Sourav Dey, Javier Arroyo, Lieve Helsen, Xiangyu Zhang, Bingqing Chen, Kadir Amasyali, Kuldeep Kurte, Ahmed Zamzam, et al. 2023. Ten questions concerning reinforcement learning for building energy management. *Building and Environment* 241 (2023), 110435. doi:10.1016/j.buildenv.2023.110435
[15] Frauke Oldewurtel, Alessandra Parisio, Colin N Jones, Dimitrios Gyalistras, Markus Gwerder, Vanessa Stauch, Beat Lehmann, and Manfred Morari. 2012. Use of model predictive control and weather forecasts for energy efficient building climate control. *Energy and buildings* 45 (2012), 15–27. doi:10.1016/j.enbuild.2011.09.022
[16] Thijs Peirelinck, Frederik Ruelens, and Geert Deconinck. 2018. Using reinforcement learning for optimizing heat pump control in a building model in Modelica. In *2018 IEEE International Energy Conference (ENERGYCON)*. IEEE, Cyprus, 1–6. doi:10.1109/ENERGYCON.2018.8398832
[17] Giuseppe Pinto, Zhe Wang, Abhishek Roy, Tianzhen Hong, and Alfonso Capozzoli. 2022. Transfer learning for smart buildings: A critical review of algorithms, applications, and future perspectives. *Advances in Applied Energy* 5 (2022), 100084. doi:10.1016/j.adapen.2022.100084
[18] Fabian Raisch, Thomas Krug, Christoph Goebel, and Benjamin Tischler. 2025. GenTL: A General Transfer Learning Model for Building Thermal Dynamics. In *Proceedings of the 16th ACM International Conference on Future and Sustainable Energy Systems*. ACM, Netherlands, 322–333. doi:10.1145/3679240.3734589
[19] Tobias Rohrer, Lilli Frison, Lukas Kaupenjohann, Katrin Scharf, and Elke Hergenröther. 2023. Deep reinforcement learning for heat pump control. In *Science and Information Conference*. Springer, London, 459–471. doi:10.1007/978-3-031-37717-4_29
[20] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).
[21] Matthew E. Taylor and Peter Stone. 2009. Transfer Learning for Reinforcement Learning Domains: A Survey. *J. Mach. Learn. Res.* 10 (Dec. 2009), 1633–1685.
[22] Tsuyoshi Ueno and Alan Meier. 2020. A method to generate heating and cooling schedules based on data from connected thermostats. *Energy and Buildings* 228 (2020), 110423. doi:10.1016/j.enbuild.2020.110423
[23] Charalampos Vallianos, José Candanedo, and Andreas Athienitis. 2024. Thermal modeling for control applications of 60,000 homes in North America using smart thermostat data. *Energy and Buildings* 303 (2024), 113811. doi:10.1016/j.enbuild.2023.113811
[24] José R Vázquez-Canteli and Zoltán Nagy. 2019. Reinforcement learning for demand response: A review of algorithms and modeling techniques. *Applied energy* 235 (2019), 1072–1089. doi:10.1016/j.apenergy.2018.11.002
[25] Zhe Wang and Tianzhen Hong. 2020. Reinforcement learning for building controls: The opportunities and challenges. *Applied Energy* 269 (2020), 115036. doi:10.1016/j.apenergy.2020.115036
[26] Shichao Xu, Yixuan Wang, Yanzhi Wang, Zheng O'Neill, and Qi Zhu. 2020. One for many: Transfer learning for building hvac control. In *Proceedings of the 7th ACM international conference on systems for energy-efficient buildings, cities, and transportation*. ACM, Virtual Event Japan, 230–239. doi:10.1145/3408308.3427617
[27] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks?. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2* (Montreal, Canada) *(NIPS'14)*. MIT Press, Cambridge, MA, USA, 3320–3328.
[28] Liang Yu, Shuqi Qin, Meng Zhang, Chao Shen, Tao Jiang, and Xiaohong Guan. 2021. A review of deep reinforcement learning for smart building energy management. *IEEE Internet of Things Journal* 8, 15 (2021), 12046–12063. doi:10.1109/JIOT.2021.3078462

A Building Information

A.1 Building Model Dynamics

State Vector (x_t): The state vector encapsulates the complex thermal inertia of the building components. It consists of five variables:

(1) Indoor air temperature (T_{in})
(2) Interior thermal mass temperature (T_{int_mass})
(3) Building envelope thermal mass temperature (T_{env_mass})
(4) Heater component temperature (T_{heater})
(5) Sensor casing temperature (T_{sensor})

Input Vector (u_t): The input vector $u_t \in \mathbb{R}^3$ combines controllable actuation and uncontrollable environmental disturbances:

$$u_t = [a_t, T_{amb,t}, I_{solar,t}]^\top,$$

where $a_t \in [0, 1]$ is the continuous heating modulation action (0% to 100% capacity), T_{amb} is the ambient outdoor temperature, and I_{solar} is the solar irradiance.

Output Matrix (C_i): The agent observes the indoor air temperature. Consequently, the output matrix is defined as a selector vector $C_i = [1, 0, 0, 0, 0]$, such that $o_t = T_{in,t}$.

A.2 Environmental Data

Training utilizes Typical Meteorological Year (TMY) data sampled at 15-minute intervals, specifically focusing on heating seasons in climates such as Csb (warm-summer Mediterranean, e.g., Seattle) to ensure relevant load profiles [5]. The time series for the data consist of 16,263 timesteps, corresponding to 169.4 days (or 24.2 weeks). Figure 9 shows the data. The simulation takes as input the ambient air temperature and solar radiation.

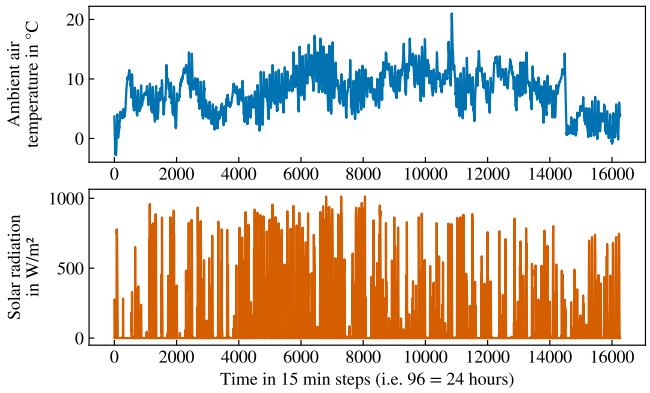


Figure 9: Weather data for a typical meteorological year in Cbs climate during heating season.

A.3 Pricing Data

To test economic optimization, agents are fed real-world electricity price signals, such as day-ahead market prices from the German

energy market [2], which reflects a realistic price forecast scenario. It is used in experiments where the agent considers cost in its control strategy. The price trajectory is visualized in Figure 10. These prices exhibit significant volatility, providing the signal necessary for the agent to learn arbitrage.

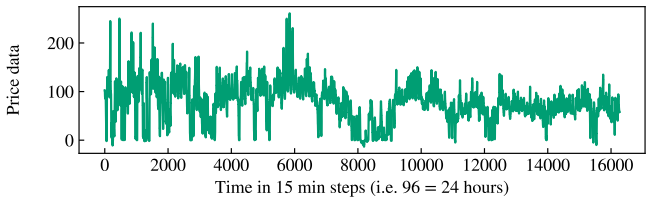


Figure 10: Price data in €/MWh during heating seasons).

A.4 Temperature Comfort Zone

The target temperature band is illustrated in Figure 11. For time-varying temperature, while the occupancy status is "home", the temperature band is tightened to ensure high comfort, while it is relaxed during "away" and "sleep" to increase flexibility and reduce costs. On weekdays, "home" spans 06:00–08:00 and 18:00–23:00, while on weekends, it is 06:00–23:00. "Sleep" is always 23:00–06:00, and "away" applies only on weekdays from 08:00–18:00.

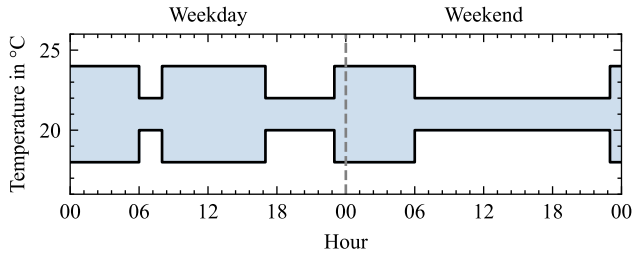


Figure 11: The target temperature band for weekdays and weekends.

B Additional Experimental Settings and Results

This appendix section provides supplementary material for the main experiments. It summarizes implementation details, hyperparameter settings, and additional results that support the analyses reported in the main text.

B.1 Hyperparameter Settings

DQN: For the discrete action space (5 actions), the DQN agent used two hidden layers with 64 neurons each. Exploration was managed via an ϵ-greedy strategy, decaying linearly from 1.0 to 0.05 over the first 10% of training.

PPO: The PPO agent (two hidden layers, 64 neurons) was specifically tuned to capture the temporal periodicity of the building dynamics. We set the rollout buffer length (n_steps) and batch size to 672 steps, corresponding to exactly one week of simulation time (assuming 15-minute intervals), to ensure the policy updates based on a full weekly cycle.

SAC: As an off-policy maximum entropy algorithm, SAC was configured with a larger capacity network consisting of two hidden layers with 256 neurons each. We utilized automatic entropy coefficient tuning (ent_{coef} ='auto') to balance exploration and exploitation dynamically.

B.2 PPO Hyperparameter Tuning

We conducted an extensive hyperparameter sweep for the PPO agent to identify configurations that could potentially enhance its performance in the building control task. Figure 12 summarizes the results of this sweep, illustrating the impact of various hyperparameter settings on the cumulative reward achieved by the PPO agent on Building F.

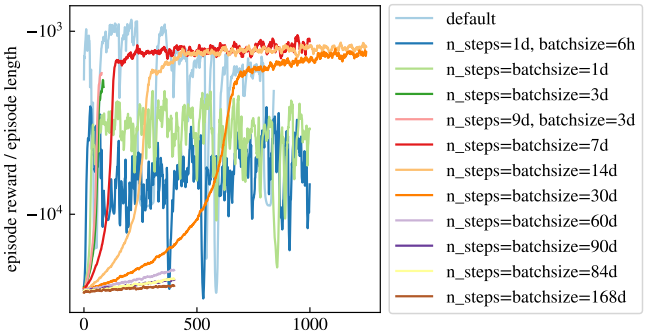


Figure 12: PPO Hyperparameter Sweep on Building F.

Figure 13 illustrates the control behavior of the PPO and DQN agents over a three-day period on Building F. The top panel displays the indoor temperature trajectories relative to the dynamic comfort bounds, while the bottom panel shows the corresponding heat pump modulation actions. The DQN and PPO agents exhibit severe control instability. Rather than converging to a safe and conservative policy, the PPO agent fails to learn the system dynamics effectively, resulting in a highly oscillatory temperature trajectory that frequently violates the comfort bound. This behavior is reflected in the agent's control signal, which shows erratic, high-frequency switching (chattering) rather than smooth modulation. This inability to stabilize the system explains the poor aggregate metrics observed in Table 1, where PPO records a 72.25% violation rate and an operational cost gap of +84.45%. These results suggest that the on-policy nature of PPO is ill-suited for this specific building control formulation, likely due to high variance in gradient estimates preventing the policy from settling into a viable control range.

B.3 Zero-shot Transfer

Figure 14 shows the zero-shot transfer performance on the four pilot target buildings (P2 to P5) from different climate zones with diverse thermal distance (D).

Figure 15 compares the behavior of the agent on the original source building P1 (familiar) versus the target Building P2 (alien). The *logic* of the policy (Reacting to Price/Dynamics) is successfully transferred.

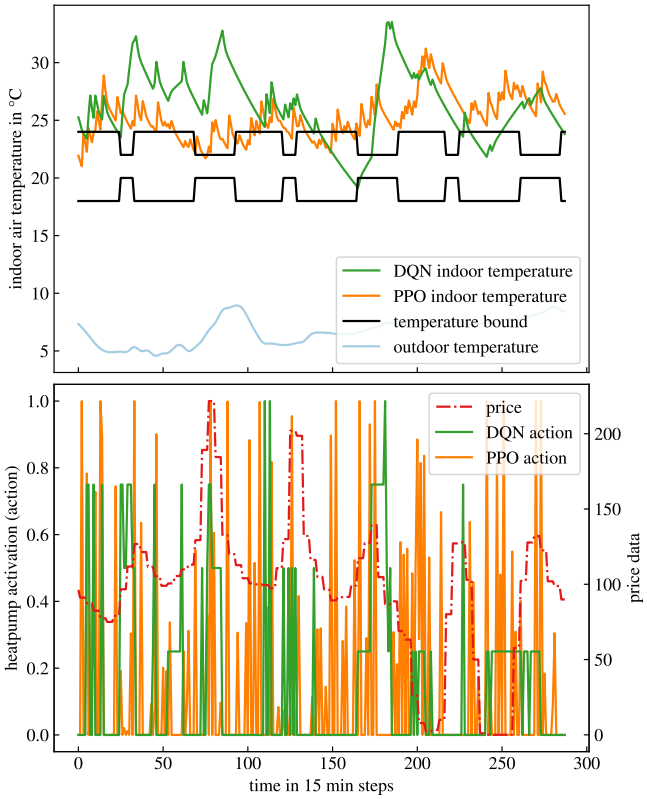


Figure 13: Indoor temperature trajectory and heat pump control actions of PPO and DQN agent trained on Building F.

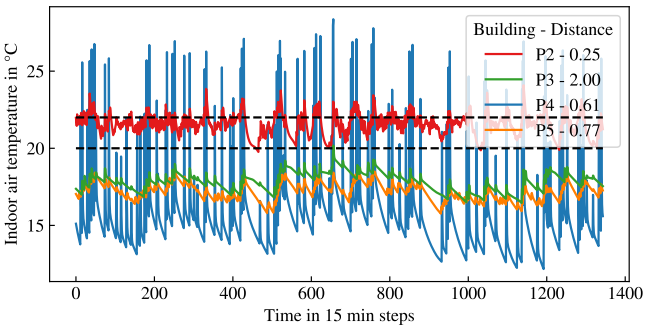


Figure 14: Zero-shot transfer performance on Pilot Targets.

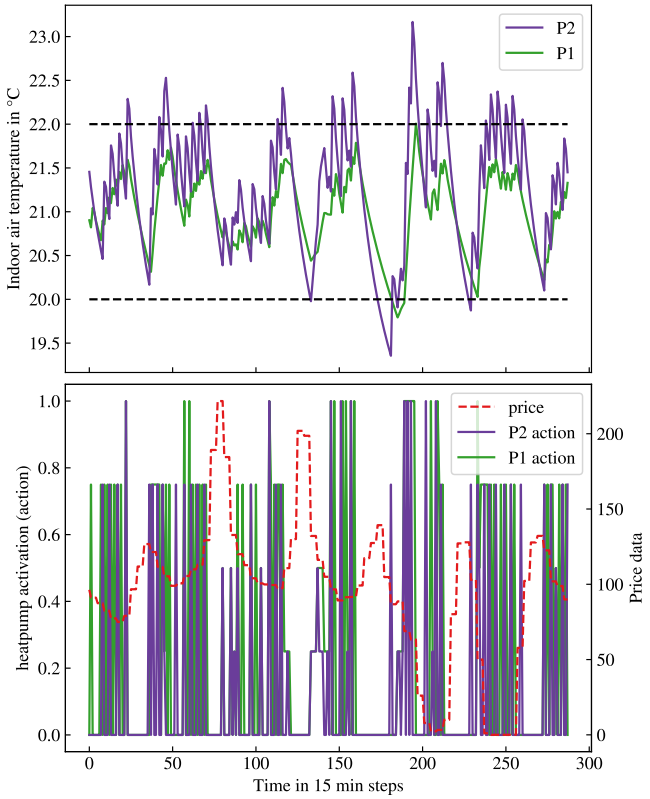


Figure 15: Comparison of indoor air temperature, agent actions, and energy price over the course of three days for a DQN agent trained in the original (Building P1) and transferred building (Building P2) with zero-shot transfer.

C Additional Transfer-Learning Results across Ten Buildings

Figure 16 reports supplementary transfer-learning results for the ten-building study. It provides the full cross-building reward comparison that complements the summarized transfer-learning analysis in the main text.

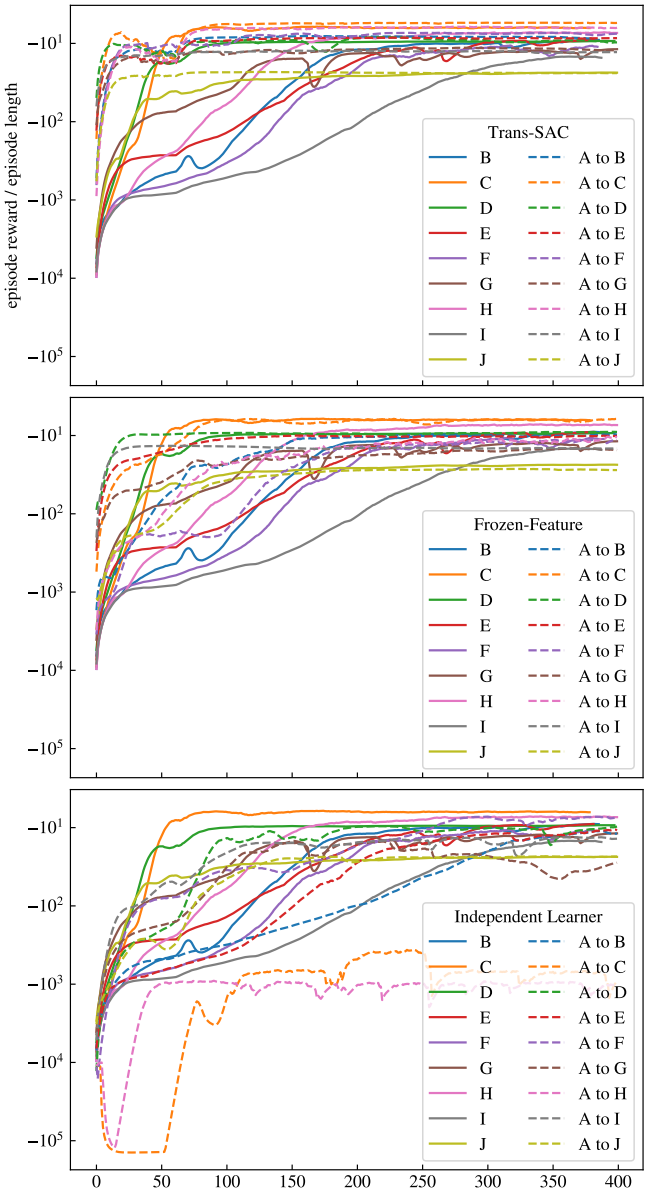


Figure 16: Reward curve of Trans-SAC, Frozen-Feature, and Independent Learner from Building A to other Buildings.