MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Scalable Physics-Informed Multi-Agent Reinforcement Learning for Building Energy System Control

Wang, Xuezheng; Ren, Zhaolin; Li, Na; Dong, Bing

TR2026-106 July 08, 2026

Abstract

Optimizing multi-zone building heating, ventilation, and air conditioning systems for energy efficiency while maintaining thermal comfort is a critical challenge, as buildings account for 40% of total energy consumption in the United States and approximately 30% globally. Although multi-agent reinforcement learning has emerged for building system control, existing studies typically involve fewer than 10 agents, rely on model-free methods despite having access to environment models, and lack real-building validation. This paper proposes a physics-informed model-based multi-agent reinforcement learning framework for scalable multi-zone control. First, we develop a physics-consistent graph neural network that combines group-shared multi-scale causal convolutions with a heat diffusion graph layer for inter-zone thermal coupling, enabling scalable multi-zone temperature prediction. Second, we introduce K-neighborhood truncation for the multi-agent soft actor-critic algorithm, where each agent's critic receives only states within its K-hop neighborhood, reducing critic input dimensionality by up to 72% with provable approximation guarantees. Third, we integrate the learned dynamics model as a differentiable world model for policy optimization, substantially improving sample efficiency over model-free alternatives. We validate the framework through a 114-day simulation study on 6 zones and a 42-day real-building deployment on 18 zones across 4 floors. The dynamics model achieves below 1.4C mean absolute error across all zones. Ablation studies confirm that model-based K-truncated training converges faster and to higher reward than model-free counterparts. The deployed controller achieves 15.7% and 35-70% energy savings in simulation and real-building studies, respectively, with modest thermal comfort degradation.

Advances in Applied Energy 2026

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Scalable Physics-Informed Multi-Agent Reinforcement Learning for Building Energy System Control

Abstract

Optimizing multi-zone building heating, ventilation, and air conditioning systems for energy efficiency while maintaining thermal comfort is a critical challenge, as buildings account for 40% of total energy consumption in the United States and approximately 30% globally. Although multi-agent reinforcement learning has emerged for building system control, existing studies typically involve fewer than 10 agents, rely on model-free methods despite having access to environment models, and lack real-building validation. This paper proposes a physics-informed model-based multi-agent reinforcement learning framework for scalable multi-zone control. First, we develop a physics-consistent graph neural network that combines group-shared multi-scale causal convolutions with a heat diffusion graph layer for inter-zone thermal coupling, enabling scalable multi-zone temperature prediction. Second, we introduce κ-neighborhood truncation for the multi-agent soft actor-critic algorithm, where each agent's critic receives only states within its κ-hop neighborhood, reducing critic input dimensionality by up to 72% with provable approximation guarantees. Third, we integrate the learned dynamics model as a differentiable world model for policy optimization, substantially improving sample efficiency over model-free alternatives. We validate the framework through a 114-day simulation study on 6 zones and a 42-day real-building deployment on 18 zones across 4 floors. The dynamics model achieves below 1.4°C mean absolute error across all zones. Ablation studies confirm that model-based κ-truncated training converges faster and to higher reward than model-free counterparts. The deployed controller achieves 15.7% and 35–70% energy savings in simulation and real-building studies, respectively, with modest thermal comfort degradation.

Keywords

Physics-informed neural network, multi-agent reinforcement learning, optimal building control, model-based learning, real-building deployment

1 Introduction

Buildings account for approximately 40% of total energy consumption and 36% of carbon dioxide emissions in the United States, with heating, ventilation, and air conditioning (HVAC) systems representing the single largest energy end-use [1]. As the urgency for decarbonization intensifies, optimizing HVAC control has emerged as one of the most impactful strategies for reducing building energy consumption while maintaining occupant thermal comfort. Traditional rule-based and proportional-integral-derivative (PID) controllers, though widely deployed, operate with fixed setpoints and limited adaptability, often leading to energy waste across different zones of a building [2] .

Model predictive control (MPC) has been extensively studied as an advanced alternative, leveraging physics-based or data-driven models to optimize control actions over a receding horizon. However, MPC faces significant practical barriers for multi-zone buildings: the computational burden of solving optimization problems at each control step grows substantially [3]. Moreover, convexity of the problem formulation is usually needed for global optima, which requires reformulation of the original problem into different hierarchies and could result in performance degradation due to information gaps [4]. Even though attempts have been made to

use neural networks to learn MPC control laws [5][6][7], RL still outperforms in terms of stable and optimal performance with low computational cost for online decision making, despite higher offline training requirements [4][8].

Single-agent RL has been extensively studied for building energy system control in recent years. Figure 1 (a) shows the number of zones and scale of some single-agent RL studies, with detailed characteristics and references listed in Table 7 in the Appendix. Simulation studies have demonstrated 10-50% energy savings across a range of HVAC configurations, with the majority controlling only a single zone or two zones. Experimental deployments have validated these savings on real buildings, with studies addressing up to 6 controlled zones. Despite the success of single-agent RL for building system control, its application is still limited to 6 zones at maximum. The biggest challenge is the curse of dimensionality: for a building with N zones, each with d_s state variables and d_a action variables, the joint state-action space scales as $O(d_s^N \times d_a^N)$. For example, a 20-zone building with 7 states and 1 action per zone yields a joint state-action space with dimensionality on the order of 10^{18}, a scale that requires a significant amount of samples to learn an optimal control policy, which increases offline training time significantly and therefore is challenging for critic and policy networks to learn state and action values and optimal control. Besides the challenges from increased dimensions, using single-agent RL to control multiple systems and zones creates a monolithic architecture in which the failure of any single component can compromise the entire system, thereby increasing deployment complexity and reducing operational robustness.

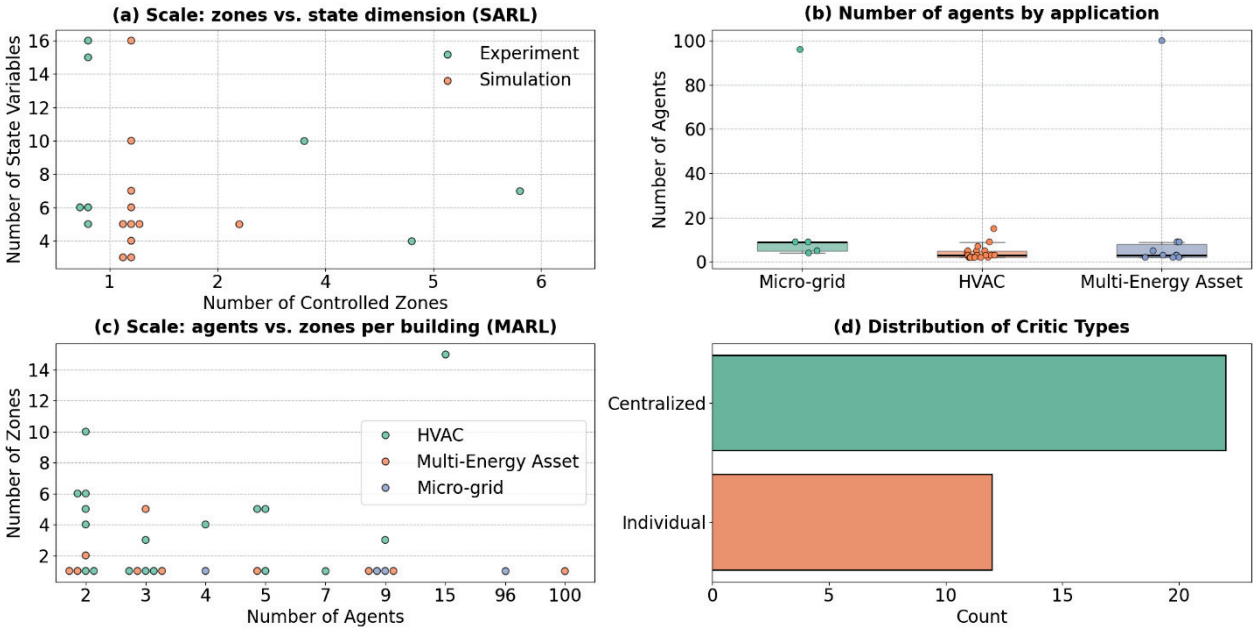

Figure 1. Landscape of reinforcement learning for building energy systems: (a) scale of single-agent studies, (b) number of agents in multi-agent studies, (c)number of agents and controlled zones in multi-agent studies, and (d) critic architecture.

Multi-agent reinforcement learning is naturally a good candidate for optimal building system control, where each zone or subsystem can be treated as an individual agent that learns a local control policy while coordinating with others. MARL has been studied extensively in recent years for the control of building HVAC systems, multi-energy assets, and microgrids. Figure 1 (b) – (c) summarizes the key characteristics of 34 recent studies based on Table 8 in the Appendix. MARL

has been demonstrated to achieve 3.95% to 41.7% energy reduction for HVAC systems, improved self-sufficiency for multi-energy assets, and near-optimal performance for microgrid control. Despite these promising results, four common limitations can be identified that prevent scaling to larger, more complex buildings.

Limited scale of thermally coupled zones: Most existing MARL studies involve only 2–5 agents representing distinct subsystems such as an AHU with VAV zones, or a chiller and a battery. Among the 34 studies reviewed in Table 10, 11 consider multi-zone indoor thermal dynamics, of which 9 model no more than 6 zones. Only two studies address 10 and 15 thermally connected zones [53][44], respectively. Two additional studies scale to 96 and 100 agents [37][56], but in both cases each agent controls a separate building treated as a single node rather than multiple thermally coupled zones within one building. A key factor limiting this scaling is the modeling effort required to capture multi-zone dynamics. All reviewed MARL studies rely on physics-based simulation environments, where developing a detailed thermal model for each building demands extensive effort due to differences in building layouts, envelope properties, and HVAC configurations. Although physics-informed machine learning has emerged as a promising and more scalable alternative [64], existing studies applying this approach to building thermal modeling have only been demonstrated for up to 6 zones [60][61][62][63]. Extending such models to buildings with tens of zones remains an open challenge, which in turn limits the scale at which MARL can be deployed for coordinated multi-zone control.

Scalability of the critic architecture: Even when the number of zones is modest, the critic design poses a scalability concern. Among the 34 studies, 22 employ a centralized critic that receives the concatenated states and actions of all agents. Usually, centralized training and decentralized execution (CDTE) are used. In this setting, the policy networks make control decisions based on local observations, while the global value is learned by critic networks using the concatenation of local observations. At the current scale of 2–5 agents with 3–15 state variables each for building applications, this concatenation yields critic inputs on the order of 20–50 dimensions, which is manageable for standard neural network function approximators. However, the input dimension grows with the number of agents. For a large building with hundreds of zones, the concatenated input to critic networks would reach thousands of dimensions. High-dimensional critic inputs degrade learning efficiency in two compounding ways: first, the number of samples required for accurate value function approximation grows exponentially with input dimension, as the critic must cover an increasingly high-dimensional space; second, the credit assignment problem becomes more severe, as the critic must disentangle each agent's contribution to the global return from a joint signal that mixes information from all zones, many of which have negligible influence on the agent in question. Together, these effects slow convergence and destabilize training, making centralized critics challenging for buildings with tens or hundreds of zones. The remaining 12 studies use independent per-agent critics that rely only on the local states of the agent itself. While this avoids the dimensionality problem, it sacrifices inter-agent coordination entirely: each agent optimizes its local objective without awareness of neighboring zones' states or actions. In thermally coupled buildings, this can lead to conflicting control decisions, for example, one zone increasing its heating load while an adjacent zone simultaneously cools, resulting in wasted energy and degraded comfort. In fact, the indoor dynamics of one zone is usually affected by its direct adjacencies, which can be leveraged to reduce the learning dimension of centralized critic networks. However, none of the existing studies exploit this property to design a critic whose input dimension scales with the local neighborhood size rather than the total number of zones.

Underutilization of dynamic models: Beyond the critic design, how the agents learn also presents an efficiency gap. A widely recognized advantage of model-free RL is that it requires no explicit dynamics model, learning optimal policies directly from interaction with the environment. However, in real building control practice, directly training agents through interaction with the physical system is cost-prohibitive due to occupancy comfort requirements, seasonal constraints, and the risk of equipment damage from exploratory actions. Consequently, nearly all existing studies, 31 of 34, train their model-free agents in simulation environments built upon physics-based models that encode thermodynamic equations such as energy balance, heat conduction, and convective heat transfer. This creates a paradox: the simulation environment already contains an explicit dynamics model, yet the model-free algorithm treats it as a black box that only returns states and rewards after each interaction. The structural knowledge embedded in the governing equations, the causal relationships between control actions, disturbances, and temperature responses, is entirely discarded by the learning algorithm, which must rediscover these relationships through extensive trial and error. A model-based approach, by contrast, can internalize a learned dynamics model to generate synthetic future states, enabling agents to anticipate the consequences of their actions without additional environment interactions. This substantially reduces the sample complexity of policy learning and accelerates convergence—advantages that become even more significant when fine-tuning policies on real buildings, where every interaction is costly.

Lack of real-building validation: Finally, 33 of the 34 studies evaluate their MARL algorithms exclusively in simulation. While simulation is essential for safe and reproducible experimentation, it cannot fully capture the disturbances, sensor noise, and occupant behavior encountered in practice. Real-building deployment of MARL for HVAC control remains rare, limiting the practical credibility of existing approaches.

The adoption of multi-agent reinforcement learning for multi-zone building control is motivated by three technical challenges that arise when scaling beyond a handful of zones. First, single-agent RL suffers from the curse of dimensionality: the joint state-action space grows exponentially with the number of zones, making sample-efficient learning intractable for buildings with tens or hundreds of zones. Second, the standard centralized training with decentralized execution for MARL introduces its own scalability bottleneck: the centralized critic must process the concatenated states of all agents, leading to degraded credit assignment and slower convergence as the number of zones increases. Fully independent critics avoid this dimensionality problem but sacrifice inter-zone coordination. Third, model-free MARL methods are sample-inefficient because they discard the structural knowledge embedded in the building's thermodynamic models, requiring extensive environment interactions to learn policies that a model-based approach could discover more directly. These three challenges motivate the framework proposed in this paper: κ-neighborhood truncation addresses the first two by restricting each agent to a provably sufficient local neighborhood, while model-based training with the learned physics-consistent graph neural network dynamics addresses the third.

This paper proposes an integrated framework for scalable, physics-consistent, model-based MARL for multi-zone building HVAC control and validates it through both simulation and real-building deployment. Here, scalability refers to the ability of the framework to maintain computational tractability and learning performance as the number of controlled zones increases, achieved through sparse graph-based modeling that scales sub-linearly with building size and a truncated critic architecture whose input dimension depends on local neighborhood size rather than total zone count. The specific contributions are as follows:

1. **A scalable physics-informed machine learning (PIML) model for multi-zone indoor dynamics**: Based on the same rationale of our previous studies [8] [60][65], we develop a scalable PIML model for multi-zone indoor dynamics. The proposed model combines multi-scale causal temporal convolutions with a heat diffusion graph neural network that captures inter-zone thermal coupling through learned graph Laplacian operators, enabling scalable and physically interpretable multi-zone temperature prediction.
2. **A scalable MARL with κ-neighborhood information**: A κ-neighborhood truncated critic design is introduced for the multi-agent soft actor-critic algorithm. Leveraging the exponential decay property of network Markov decision processes, where each agent's value function has diminishing sensitivity to distant agents, the critic for each zone receives only the states within its κ-hop neighborhood rather than the full building state. This reduces the input dimension of critic networks with provable approximation accuracy bounds, enabling the framework to scale to buildings with tens of zones.
3. **A model-based MARL training with PIML dynamic model**: Although model-based RL is well-established in the general literature, 31 of 34 reviewed MARL studies for building systems rely on model-free methods; this work bridges that gap by integrating the learned PCGNN as a differentiable world model for policy optimization. Rather than treating the simulation environment as a black box, the trained PIML model is used to generate synthetic rollouts for policy optimization, substantially improving sample efficiency compared to model-free alternatives.
4. **Multi-scale validation in both simulation and real-world implementation**: The proposed framework is validated at two scales: a small-scale simulation study on 6 zones of a commercial building, and a real-world deployment on an 18-zone, 4-floor medium-sized commercial building.

The remainder of paper is organized as follows. Section 2 presents the methodology, including the multi-zone dynamics model, the model-based multi-agent SAC algorithm with κ-truncated critics, and the training procedure. Section 3 describes the experimental setup for both simulation and real-building studies. Section 4 presents results. Section 5 discusses implications and limitations, and Section 6 concludes the paper.

2 Methodology

Figure 2 shows the overview of our proposed physics-informed model-based MARL. It consists of three parts: 1) physics-consistent graph neural network (PCGNN) for multi-zone temperature prediction, 2) model-based MARL with κ-neighborhood truncation for dimension reduction, and 3) real-building case studies for validation. The PCGNN combines two complementary layer types: causal 1-D convolutional layers that capture the temporal effect of zonal inputs (e.g., supply air condition, occupancy, and weather) on indoor temperature, and 2) heat diffusion graph layers that model inter-zone thermal coupling through learned non-negative heat transfer parameters on the building adjacency graph. In model-based MARL, PCGNN is not only used in the environment to generate states and rewards but also used by agents as a differentiable model to optimize their policy and critic networks. The κ-neighborhood truncation makes each agent's critic network receive only the local state from the zone and its κ-hop neighbors rather than the full building state. The proposed PCGNN and model-based MARL with κ-neighborhood truncation is validated in two case studies using different commercial buildings: a 6-zone case for simulation-based evaluation using a commercial building in Boston (building 1), and a 4-floor 18-zone case for real-

world implementation using a commercial building in Syracuse (building 2). The following subsections and Section 3 describe each part of Figure 2 in detail.

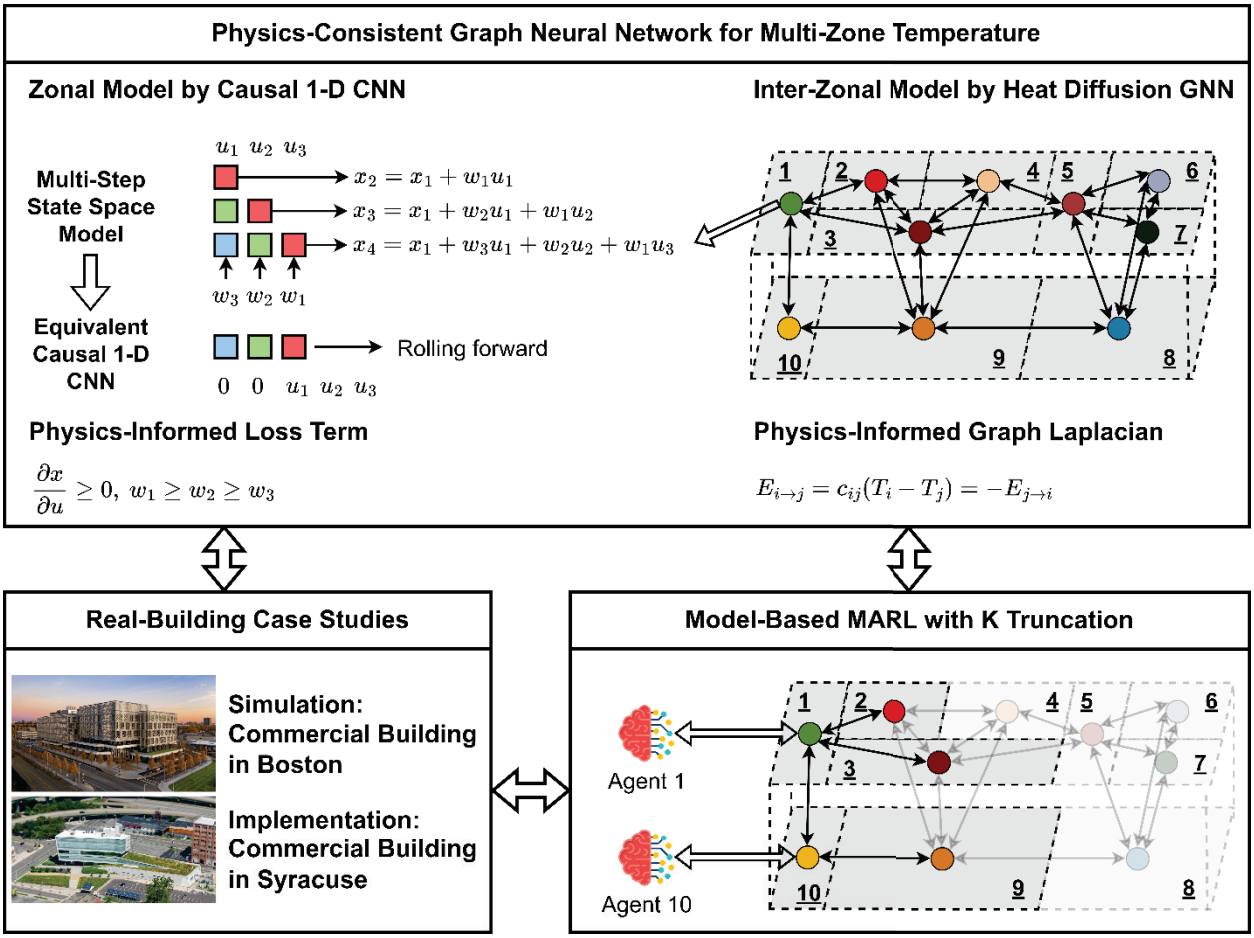

Figure 2. Overview of the proposed framework: physics-consistent graph neural network for multi-zone temperature prediction (top), model-based multi-agent soft actor-critic with κ-neighborhood truncation (bottom right), and case study validation (bottom left).

2.1 Physics-Consistent Graph Neural Network for Multi-Zone Temperature Prediction

In our previous work, we developed physics-consistent input convex neural networks (PCICNN) to model single-zone indoor temperature dynamics [8] [65]. PCICNN was inspired by the lower triangular block Toeplitz structure of coefficient matrices of linear state space model, as shown in multi-step state space model in Figure 2. In this structure, the coefficient matrix is lower triangular and shares the same weights along main and parallel diagonals. Such a structure was encoded into the weight matrices of multi-layer perceptron model, resulting in a partial connection of neurons between different layers and shared weights within the same layer. The extension to multi-zone cases was through the partial connection of neurons based on the adjacent information [60]. However, such development was highly customized and required significant manual effort to define the weights and neuron connections. Therefore, we propose a more standardized modeling approach that is more convenient to scale yet achieves equivalent effect as our previous work. As shown in Figure 2, the multi-step state-space model can be equivalent to causal 1-D convolutional layer. The inter-zonal interaction can be modeled by graph layers. As illustrated in the top panel

of Figure *3*, the PCGNN alternates the two types of layers: M convolutional layers and one graph layer. The two types of layers were alternated B_{blk} times to get the prediction of multi-zone dynamic predictions. The second panel of Figure *3* details the multi-scale temporal convolution within each causal convolutional layer, where a short- and long-branch kernels with different dilation rates are mixed via learned softmax weights to capture both short- and long-term input-output correlation. The third panel shows the group shared convolution strategy, where zones with similar dimensions and usage share the same convolutional weights to reduce model size. The fourth panel describes the physics-informed constraints: optional non-negative parameters or gradient penalties to enforce positive input-output correlation, and residual connections across convolutional layers. The bottom panel details the heat diffusion GNN layer, which models inter-zonal thermal coupling through conservation of energy and learned non-negative heat transfer coefficients on the building adjacency graph. The details of each component are described in the following subsections.

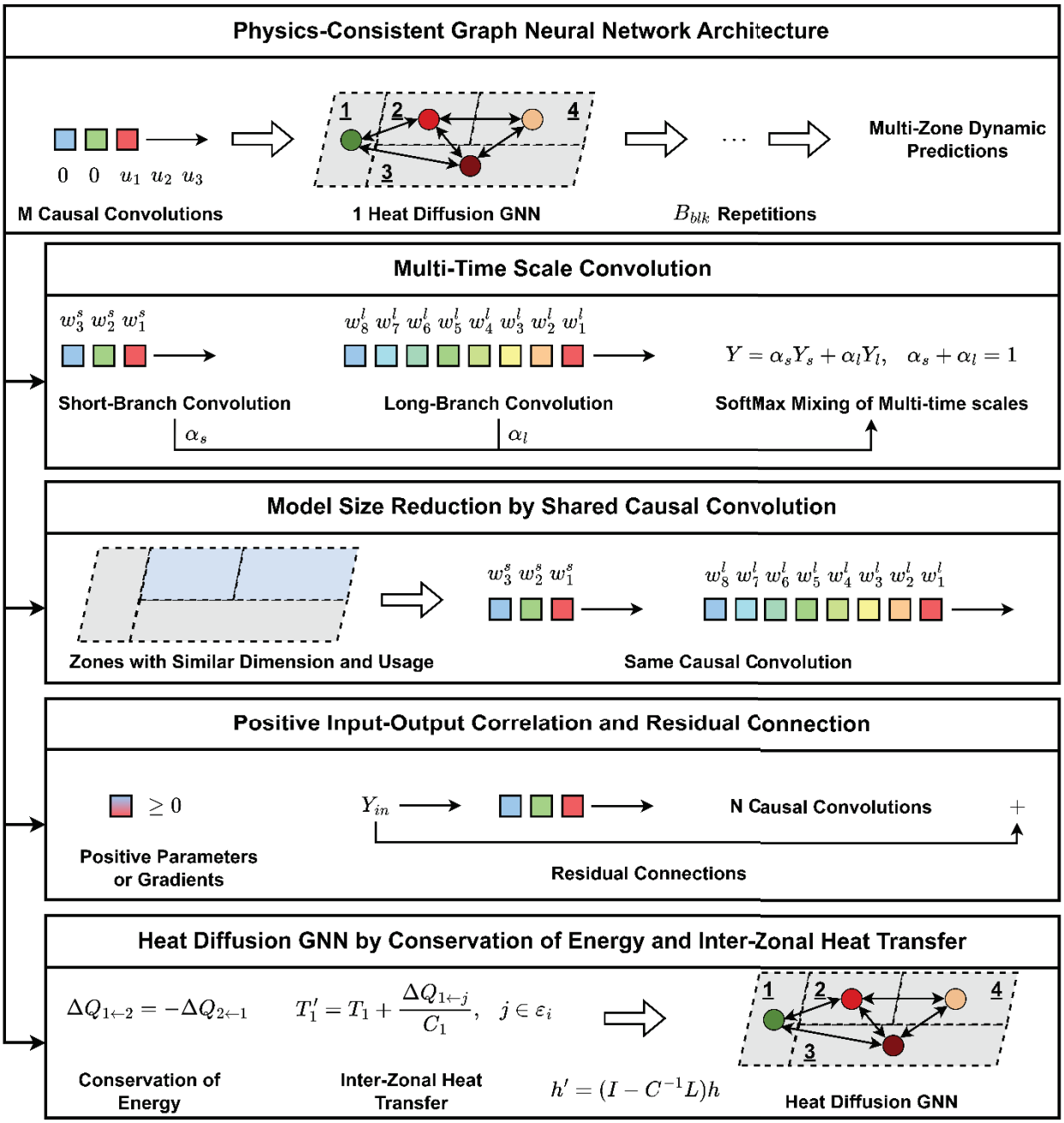

Figure 3. Detailed architecture of the proposed Physics-Consistent Graph Neural Network.

2.1.1 Group-Shared Multi-Scale Causal Convolutional Layer

Each causal convolutional layer of the PCGNN combines two parallel temporal convolutions with different receptive fields to capture short- and long-term effects of zonal inputs (e.g., supply air conditions, occupancy, weather) on indoor temperature. For a group $r \in R$, the short branch applies a 1-D causal convolution $Conv_{short}$ with kernel length k_s and dilation 1, while the long branch applies $Conv_{long}$ with kernel length k_l and dilation $d_l > 1$. Both branches use group-shared weights so that zones within the same group share parameters, reducing model size on

buildings with many similar zones. Causality is enforced by zero-padding past inputs ($H_{\{b,\tau,\cdot,\cdot\}} = 0, for\ \tau < 0$), so that predictions at time t depend only on current and past values.

The two branches are fused through learnable logits $\eta = (\eta_s, \eta_l)$ passed through a softmax to produce mixing weights (α_s, α_l) with $\alpha_s + \alpha_l = 1$; the layer output for group r is

$$\boldsymbol{Y}^{(r)} = \alpha_s \cdot Conv_{short}(\boldsymbol{H}^{(r)}; \boldsymbol{W}_{\boldsymbol{s}}^{(r)}) + \alpha_l \cdot Conv_{long}(\boldsymbol{H}^{(r)}; \boldsymbol{W}_{\boldsymbol{l}}^{(r)}) \quad (1)$$

followed by a shared bias, LeakyReLU activation, dropout, and a residual connection (with linear projection when input and output dimensions differ). An optional positivity constraint $\phi(W) = W \odot W$ applied to any kernel enforces non-negative correlation between specific input features (e.g., heating panel load, outdoor temperature) and predicted indoor temperature; when not enforced, the physical-consistency penalty in the loss function provides an alternative path. Full kernel formulations, dimensional notation, the positivity-constraint derivation, softmax mixing details, and the residual projection are given in Appendix 8.3.

2.1.2 Heat Diffusion GNN Layer

As shown in Figure 2, causal convolutional layer only processes feature inside the node (zone) over the prediction horizon. The inter-zonal effect is modeled by heat diffusion graph layer. It operates on each time slice independently, i.e., on each graph instance $g = (b, t)$. Let $\boldsymbol{h}_g \in \mathbb{R}^{Z \times D}$ be the zone embeddings for one (b, t). Let ε_u be the set of undirected edges (i, j) from the adjacency matrix for adjacency between zones i and zone j. The heat diffusion graph layer can be developed based on the heat transfer of adjacent zones.

Each undirected edge $(i, j) \in \varepsilon_u$ carries a learnable non-negative heat transfer coefficient $g_{ij} \geq 0$, enforced via the same positivity transform ϕ used in the convolutional layer, and each zone i carries a learnable strictly positive inverse heat capacity $c_i = \frac{1}{C_i}$. The layer update is written compactly in graph Laplacian form:

$$\boldsymbol{h}' = (\boldsymbol{I} - \boldsymbol{C}^{-1}\boldsymbol{L})\boldsymbol{h} \quad (2)$$

where $\boldsymbol{L}$ is the symmetric weighted graph Laplacian constructed from g_{ij} and $\boldsymbol{C}^{-1} = diag(c_1, \ldots, c_Z)$. Because $\boldsymbol{C}^{-1}$ is applied to the nodes, the effective transformation $\boldsymbol{C}^{-1}\boldsymbol{L}$ is asymmetric, reflecting the physical reality that the same heat flow across two adjacent zones produces different temperature changes when their thermal masses differ. Step-size capping for numerical stability, the per-edge symmetric heat flow definition, the per-node update form, and the explicit Laplacian construction are given in Appendix 8.4.

2.1.3 Full PCGNN

As aforementioned, the PCGNN alters the two types of layers in a way: M convolutional layers → one graph layer → M convolutional layers → one graph layer → … → M convolutional layers → one graph layer. Considering M convolutional layers and one graph layer as one block, the proposed PCGNN consists of B_{blk} to predict the multi-zone indoor dynamics. Let there be B_{blk} blocks, each containing M grouped multis-scale convolutional layers and one heat diffusion graph layer.

We initialize $\boldsymbol{H}^{(0)} = \boldsymbol{X}$, for block $\beta = 1, \ldots, B_{blk}$.

For grouped temporal stack, we have M convolutional layers for each zone group. The forward propagation is defined as follows:

$$\boldsymbol{H}^{(\beta,0)} = \boldsymbol{H}^{(\beta-1)} \quad (3)$$

$$\boldsymbol{H}^{(\beta,m)} = GroupedMSConv_{b,m}\left(\boldsymbol{H}^{(b,m-1)}\right), \qquad m = 1, \dots, M \quad (4)$$

Where each $GroupedMSConv$ is the group-assembled operator in section 2.1.1, which applies the same multi-time-scale convolution to the zones within the same groups.

After grouped temporal convolution stack, we have heat diffusion layer

$$\boldsymbol{H}^{(b)}_{b,t,:,:} = HeatDiffusion_b\left(\boldsymbol{H}^{(b,M)}_{b,t,:,:}\right), \qquad \forall(b,t) \quad (5)$$

After B_{blk} blocks of alternation between convolutional and graph layers, we apply a per-zone linear head:

$$\widehat{\boldsymbol{Y}}_{b,t,z,:} = W_{out} H^{(B_{blk})}_{b,t,z,:} + \boldsymbol{b}_{out} \quad (6)$$

Where $\hat{Y}$ gives us the prediction of all modeled zones over the prediction horizon.

Note that for RL implementation, we only need single-step ahead prediction of multiple zones. Such a single-step model can be easily built based on the trained multi-step model by using 1-step-ahead kernel from multi-scale convolutional layers.

2.1.4 Loss Function

The PCGNN is trained with a composite loss that combines prediction accuracy with physical-consistency penalties. For predictions $\widehat{\boldsymbol{Y}}$ and targets $\boldsymbol{Y}$ of shape (B, T, Z, C_{out}), the total training loss is

$$\mathcal{L} = \mathcal{L}_{mse} + \lambda_{\Delta}\mathcal{L}_{\Delta} + \lambda_{mono}\mathcal{L}_{mono} + \lambda_{grad}\mathcal{L}_{grad} + \lambda_{bias}\mathcal{L}_{bias} \quad (7)$$

where $\mathcal{L}_{mse}$ is per-zone mean-squared error; $\mathcal{L}_{\Delta}$ is a temporal-difference loss penalizing mispredicted step-to-step temperature changes; $\mathcal{L}_{mono}$ is a monotonicity regularization that encourages each convolutional kernel to weight near-term inputs more heavily than distant ones; $\mathcal{L}_{grad}$ is a gradient-monotonicity penalty enforcing non-negative gradients of predicted temperature with respect to selected input features (e.g., control variables) when the hard positivity constraint $\phi(W) = W \odot W$ is not applied; and $\mathcal{L}_{bias}$ is an optional penalty preventing predictions from collapsing onto layer biases when weights vanish. The $\lambda_{\cdot}$ are non-negative weighting coefficients. Full formulations of each loss term are given in Appendix 8.5.

The 1-D causal convolutional layer, heat diffusion graph layer, and loss terms for prediction accuracy and physical consistency consist of our proposed PCGNN. Next subsection will define the multi-zone control as a network MDP and describe model-based MARL with κ-neighborhood truncation.

2.2 Model-Based MARL with PCGNN Dynamics and κ-Hop Neighborhood Truncation

2.2.1 Problem Setup as a Network MDP

A network Markov decision process (network MDP) is defined on a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where each node $i \in \mathcal{V}$ represents an agent with its own local state $s_i \in S_i$ and action $a_i \in A_i$. The key

structural property that distinguished a network MDP from a standard MDP is the factorization of the transition dynamics: the next state of each agent depends on the states and actions of its immediate graph neighbors. In a traditional MDP, the transition probability is a monolithic function over the entire global state-action space with no assumed factorization structure, and a single policy maps the full state to an action. The network MDP exploits the locality of physical interactions to decompose the global control problem into coupled local problems. In multi-zone buildings, the temperature of one zone at any given step is directly influenced by its thermally adjacent zones. This locality gives rise to the exponential decay property [66][67]: each agent's Q-function has diminishing sensitivity to the states and actions of agents beyond its κ-hop neighborhood, with the approximation error decaying exponentially with graph distance. This property provides the theoretical foundation for κ-neighborhood truncation used in this paper.

We consider a multi-zone building control problem modeled as a network Markov decision process (MDP) defined on a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node $i \in \mathcal{V}$ represents a thermal zone (agent) and edges $(i, j) \in \mathcal{E}$ represent physical coupling or interaction. At a discrete time t, each agent observes a local zone state $s_t^{(i)} \in \mathbb{R}^{d_s}$ (e.g., zone temperature and related features), and selects a control action $a_t^{(i)} \in \mathbb{R}^{d_a}$ (e.g., valve/damper/setpoint or other actuator commands). The global state and action are

$$s_t = \left(s_t^{(1)}, \dots, s_t^{(N)}\right), \qquad a_t = \left(a_t^{(1)}, \dots, a_t^{(N)}\right)$$

With $N = |\mathcal{V}|$. The objective is to learn decentralized policies $\pi = \{\pi_i\}_{i=1}^{N}$ that maximize the expected discounted return:

$$\max_{\pi} \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]$$

Where $r(s_t, a_t)$ encodes thermal comfort, energy usage, and actuation smoothness constraints, and $\gamma \in (0,1)$ is the discount factor.

2.2.2 κ-Hop Truncation for Scalable Local Decision Making

A challenge in MARL is that the joint state-action space grows rapidly with the number of agents. For network-structured systems, theoretical results in network MDPs indicate that localized approximation can well represent global values if the influence of distant nodes on an agent's value can decay exponentially with graph distance [66][67][68]. The temperature of one zone at one step is usually impacted by itself and direct adjacency ($\kappa = 1$), which makes multi-zone temperature dynamics satisfy exponential decay property, and therefore the global state and action values can be well approximated using local information. Following this motivation, we apply κ-hop neighborhood truncation to reduce input dimensionality for scalable learning.

Let $\mathcal{N}_i^{\kappa} \subseteq \mathcal{V}$ denote the set of nodes within graph distance κ from agent i, including i. We define each agent's local observation as the concatenation of states from its neighborhood:

$$o_t^{(i)} = concat\left(\left\{s_t^{(j)} : j \in \mathcal{N}_i^{\kappa_\pi}\right\}\right)$$

Where κ_π is the truncation radius used by the policy. For value estimation we use $\kappa_\pi + 1$ for critic inputs:

$$\tilde{o}_t^{(i)} = concat\left(\left\{s_t^{(j)} : j \in \mathcal{N}_i^{\kappa_\pi + 1}\right\}\right)$$

This yields a centralized training and decentralized execution structure: each agent acts using only local information $o_t^{(i)}$ and value learning can incorporate a broader neighborhood to improve credit assignment and stability.

Note that the reference paper [66] introduces spectral representations as a scalable functional basis for continuous network MDPs. In this work, we adopt the same locality motivation but do not construct spectral embeddings. Instead, we rely on neural function approximation to represent policies and value functions.

2.2.3 Model-Based Soft Actor-Critic

Besides dimension reduction by κ-neighborhood truncation, we also use the learned PCGNN dynamics model as a differentiable world model to improve sample efficiency and stabilize learning. Let f_θ denote the learned one-step PCGNN, we can have state transition defined as follows:

$$s_{t+1} = f_\theta(s_t, a_t) \tag{8}$$

Where θ are model parameters trained separately as described in Section 2.1.

We employ a model-based variant of soft actor-critic (SAC) that leverages the differentiable PCGNN model to reduce the critic complexity. Rather than learning a state-action value function $Q(s,a)$ directly, we learn a state-value function $V(s)$ with a standard MLP and reconstruct the soft Q-value through a one-step model-based lookahead based on Bellman equation:

$$Q(s_t, a_t) \triangleq r(s_t, a_t) + \gamma V(s_{t+1}) = r(s_t, a_t) + \gamma V\big(f_\theta(s_t, a_t)\big) \tag{9}$$

Note that here we decompose the next state s_{t+1} to $f_\theta(s_t, a_t)$ based on the PCGNN. This decomposition shifts the learning burden from approximating $Q: \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ to approximating $V: \mathcal{S} \to \mathbb{R}$, while the transition mapping is captured by the pretrained PCGNN model.

To mitigate overestimation bias, we use two independent value networks $V_{\psi_1}(s)$ and $V_{\psi_2}(s)$ and define a clipped values estimate $V_{min}(s) = \min\Big(V_{\psi_1}(s), V_{\psi_2}(s)\Big)$, where ψ is the learnable parameter of the value networks.

For each sampled transition (s_t, a_t, r_t), we obtain the next state from the dynamics model $\tilde{s}_{t+1} = f_\theta(s_t, a_t)$. The soft target of critic is defined using target value networks and the policy entropy:

$$y_t = r_t + \gamma\big(V_{min,targ}(\tilde{s}_{t+1}) - \alpha \log \pi_\phi(a_{t+1}|\tilde{s}_{t+1})\big), \qquad a_{t+1} \sim \pi_\phi(\cdot\,|\tilde{s}_{t+1}) \tag{10}$$

Where $\alpha > 0$ is the temperature parameter.

Each value network is trained by minimizing the squared Bellman error:

$$\mathcal{L}_V(\psi_k) = \mathbb{E}\Big[\big(V_{\psi_k}(s_t) - y_t\big)^2\Big], \qquad k \in \{1,2\} \tag{11}$$

This trains $V(s_t)$ to match the entropy-regularized one-step return under the PCGNN dynamics.

The policy $\pi_\phi(a|s)$ is optimized to maximize expected soft return. Because the objective depends on the differentiable composition $V\big(f_\theta(s,a)\big)$, we can backpropagate through the dynamics model. The objective of policy network is defined as:

$$J(\phi) = \mathbb{E}\big[r(s_t, \tilde{a}_t) + \gamma V_{min}\big(f_\theta(s_t, \tilde{a}_t)\big) - \alpha \log \pi_\phi\big((\tilde{a}_t|s_t)\big)\big] \tag{12}$$

Where $\tilde{a}$ is the action sampled by reparameterization trick.

This explicitly encourages actions that drive the system, through f_θ into high-value next states while maintaining exploration through the entropy term.

Thereby, we have defined our PCGNN and model-based MARL with κ-neighborhood truncation. Next section will define our case studies.

3 Experiment Setup

3.1 Simulation Study: Building 1 in Boston

3.1.1 Overview of the Testbed

Building 1 is a commercial building in Boston used for simulation-based evaluation. Figure 4 shows the selected 6 zones: one corridor and five office rooms, together with the thermal adjacency graph. The corridor is conditioned by a dedicated outdoor air system (DOAS) with variable air volume boxes; the offices are served by both radiant ceiling panels and the DOAS. Two zones are connected in the adjacency graph if they share a physical wall or open boundary; only this binary structure is provided as prior knowledge, and all edge weights are learned from data. Full thermostat schedules, occupancy-mode logic, and seasonal heating/cooling switching thresholds are listed in Appendix 8.6.

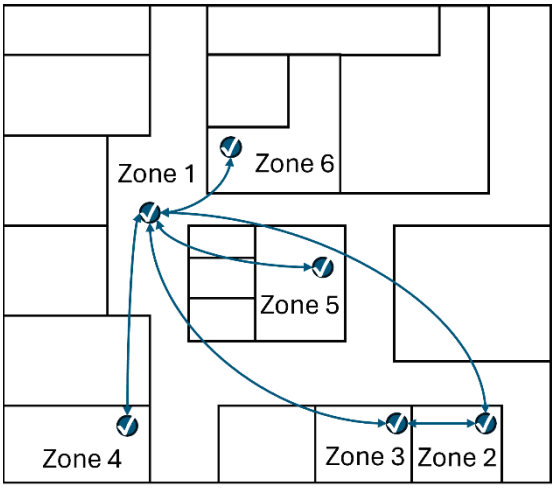

Figure 4. Floor plan of building 1 (Boston) with 6 modeled zones and thermal adjacency edges. Zone 1 is a corridor; zones 2–6 are offices served by radiant ceiling panels.

3.1.2 PCGNN for Building 1

The input features of the PCGNN for building 1 consist of the current indoor temperature at each modeled zone (a single time step) and sequences of the following variables over the prediction horizon: water-side radiant system load, air-side load of DOAS, outdoor temperature, solar radiation, and occupancy status. The dataset spans from April 15, 2024, to June 20, 2025, with 15-minute interval, and is split into training, validation, and testing sets with ratios of 0.70, 0.15, and 0.15, respectively. The hyperparameters of the PCGNN are tuned using the Optuna framework [69], and Table 9 lists the best configuration selected based on validation performance.

3.1.3 MARL for Building 1

The MARL agents control the radiant panel load in the five office zones; the corridor operates under default control, and the DOAS is excluded because it contributes only a small fraction of the total heating/cooling load and serves primarily for indoor air quality and condensation prevention. Each agent receives 23 local state variables (indoor temperature, DOAS load, outdoor temperature, solar radiation, occupancy, and current plus 8-step-ahead temperature bounds), and the reward penalizes temperature violations, action non-smoothness, and energy use while rewarding in-bound operation. A 1-hop truncation is used for the policy network and a 2-hop truncation for the critic; on the 6-zone graph the latter is equivalent to global concatenation since the graph diameter is 2. The full reward function, MARL hyperparameters, and training schedule are in Appendix 8.8.

3.2 Implementation Study: Building 2 in Syracuse

3.2.1 Overview of the Testbed

Building 2 is a real commercial building in Syracuse used for live deployment. Figure 5 shows the 18 selected zones spanning 4 floors including corridors, conference rooms, office rooms, and lobbies together with the thermal adjacency graph. All zones are conditioned by radiant ceiling panels and a DOAS with underfloor air distribution. The building is operated as a living laboratory with heating/cooling modes switchable throughout the year. Two zones are connected by an intra-floor edge if they share a wall or open boundary (including air walls), and by an inter-floor edge if they are vertically adjacent through a floor slab; as with building 1, only the binary adjacency structure is provided and all parameters are learned. Full setpoint schedules by room type are listed in Appendix 8.9.

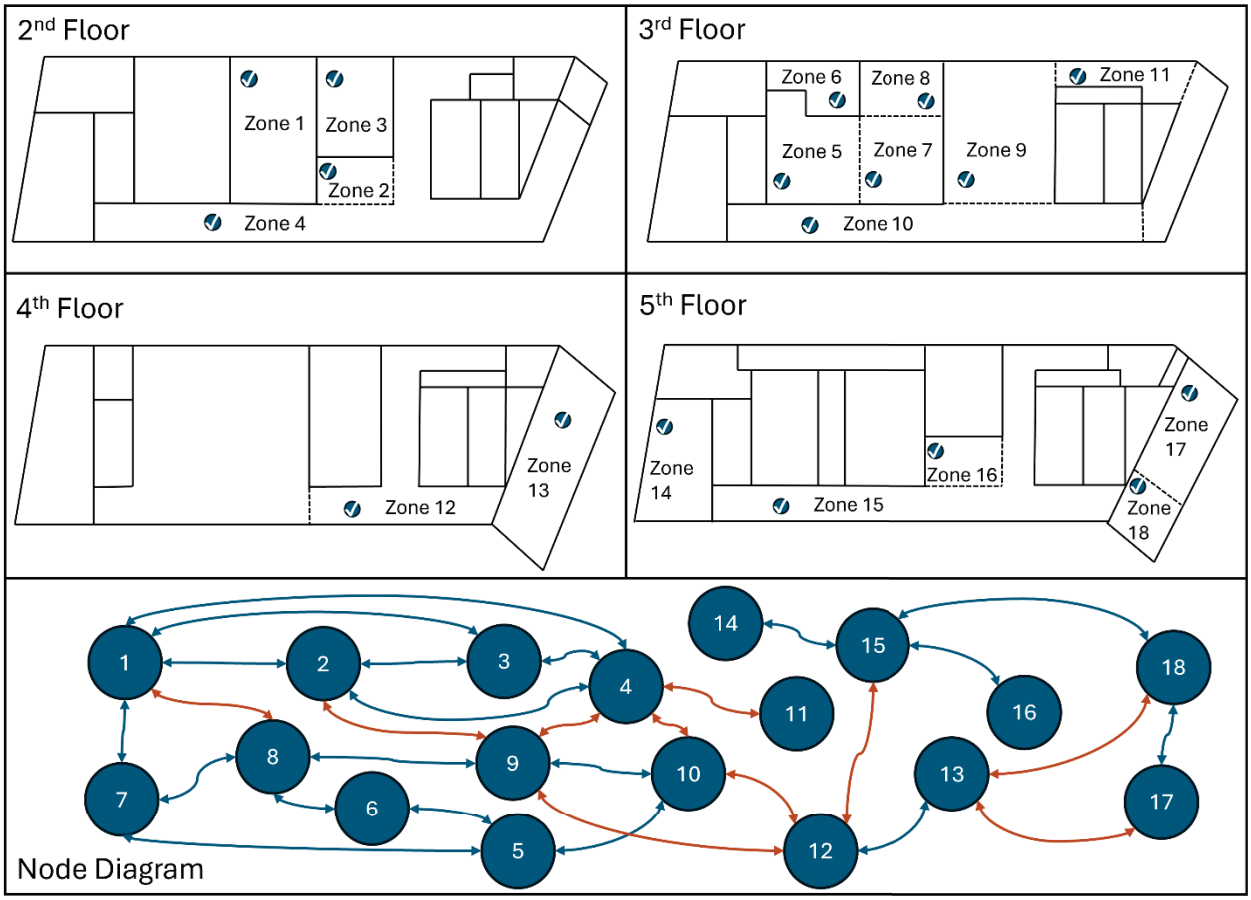

Figure 5. Floor plans of building 2 (Syracuse) showing 18 zones across 4 floors (top 4 diagrams) and the thermal adjacency graph (bottom). Dashed lines indicate air walls; blue and red edges represent intra-floor and inter-floor adjacency.

3.2.2 PCGNN for Building 2

The dataset spans April 15, 2024 to January 10, 2026 at 15-minute resolution and is split 0.70/0.15/0.15. Input features mirror those of building 1, with occupancy numbers (measured by people-counting sensors or estimated from daily schedules) replacing binary occupancy. Hyperparameters are tuned with Optuna; the selected configuration is listed in Appendix 8.10.

3.2.3 MARL for Building 2

As with building 1, the MARL agents control only the radiant ceiling panel systems for building 2, while the DOAS operates under its default control logic. The reward function and hyperparameters follow equation (36)**Error! Reference source not found.** and Table 10**Error! Reference source not found.**, respectively. Unlike building 1 case, the local state for each zone at building 2 does not include the temperature bounds at current and future time steps. Instead, sine and cosine encodings of the time of day are used, resulting in 7 local state variables per zone. After 1-hop neighborhood truncation, the maximum input dimension of the actor networks is 49, corresponding to zones with up to 7 neighbors (including themselves). For the critic networks with 2-hop truncation, the maximum and minimum input dimensions are 112 and 35, respectively. For

reference, the global state dimension without any truncation would be 126, demonstrating that the κ-truncation achieves up to a 72% reduction in critic input dimensionality.

The choice of κ is guided by the locality of building thermal dynamics: since each zone's temperature at a given step is directly influenced by its immediate adjacencies, $\kappa_\pi = 1$ captures the dominant thermal coupling for the policy. The critic uses $\kappa = \kappa_\pi + 1 = 2$ because the next-step state of the κ_π-hop neighborhood depends on one additional hop of neighbors, giving the critic the information needed to evaluate the policy's local decisions.

3.2.4 Implementation

Each MARL agent outputs a target radiant panel load, which is converted to a water flow rate using the ε-NTU method (full derivation and parameter definitions in Appendix 8.11) and mapped to a valve position based on the valve characteristics. The real-building deployment ran from January 12 to February 22, 2026, spanning 42 days, with MARL and the baseline controller alternating on a daily basis to enable direct comparison under similar weather conditions. During deployment, agents select actions deterministically using the policy mean rather than sampling, and all valve commands are clipped to 0–100% before being written to the SIEMENS Desigo building automation system via BACnet/IP at 15-minute intervals. No automated safety shield or rule-based backup controller was used; days affected by communication or execution errors were manually reverted to the baseline controller and excluded from the energy and comfort analysis. Table 1 summarizes the state space, action space, and key experimental parameters for both case studies.

Table 1. Summary of the two case studies.

	Building 1 (Boston)	**Building 2 (Syracuse)**
Number of Modeled Zones	6	18 (4 floors)
Number of Controlled Zones	5 (offices)	18
HVAC System	Radiant ceiling panels + DOAS	Radiant ceiling panels+DOAS
Control Variable (Action)	Radiant ceiling Panel Load	Radiant ceiling panel load
Action Dimension per Zone	1	1
Local State Variables	Indoor temp., DOAS load, outdoor temp., solar radiation, occupancy, indoor temp. bounds (current + 8 future steps)	Indoor temp., DOAS load, outdoor temp., solar radiation, occupancy, sin & cos time encoding
State Dimension Per Zone	23	7
Policy Input ($\kappa_\pi = 1$)	Up to 138	Up to 49
Critic Input ($\kappa = 2$)	138 (equivalent to global)	35-112
Control Interval	15 min	15 min
Study Type	Simulation (114 days)	Real-building deployment (42 days)

Evaluation Period	Apr 2024 – Jun 2025	Jan 12 – Feb 22, 2026
Safety Constraints	Soft constraints via reward penalty	Same as building 1 with valve positions clipped to 0-100% for deployment

4 Results

4.1 Simulation Study: Building 1

4.1.1 PCGNN Performance

The prediction performance of the PCGNN is first evaluated for building 1. Table 2 shows performance of PCGNN for the 6 modeled zones of building 1 on testing dataset (25 days). The predictions are generated in an autoregressive manner, where each predicted temperature serves as the initial condition for the next time step — the most operationally relevant evaluation mode, as it reflects how the model would be used during MARL rollouts. Table 2 shows that the PCGNN captures the temperature dynamic of all 6 zones accurately, with a mean absolute error (MAE) ranging from 0.31 to 1.18 ℃.

Table 2. Performance of PCGNN on testing dataset of building 1 for indoor temperature predictions.

Zone	**1**	**2**	**3**	**4**	**5**	**6**
MAE ℃	0.39	0.68	0.58	1.18	0.31	0.31

Figure 6 presents the temperature predictions under perturbed radiant panel loads, maximum cooling and maximum heating, compared to predictions using the measured load. In each row of subplots, only one zone's load is perturbed while all other zones retain their measured load profiles. The first row, for example, shows the response when the corridor (zone 1) is subjected to perturbed conditions. Because all modeled zones are adjacent to the corridor, every zone's prediction is affected, though with varying sensitivity — zone 4 exhibits the strongest response. When zone 2's load is perturbed, its direct neighbor zone 3 is significantly affected, while the remaining zones show only minor changes. Notably, zone 5 also produces a widespread impact across all modeled zones, as it serves as the core zone of the floor; its temperature change propagates through the corridor and subsequently influences other zones. Across all perturbation scenarios, the predicted temperature consistently increases under maximum heating load and decreases under maximum cooling load. The correct directional response to load perturbations, combined with the spatially decaying influence that follows the adjacency structure, confirms that the proposed PCGNN has learned both the zonal thermal dynamics and the inter-zone heat transfer physics.

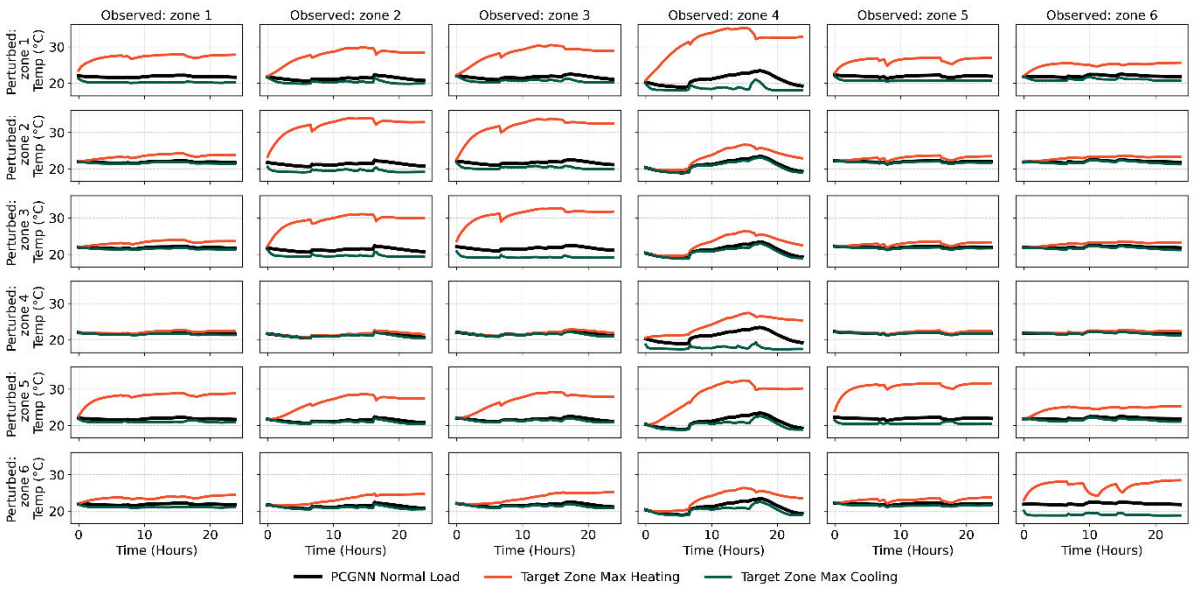

Figure 6. Cross-zone perturbation analysis of the PCGNN for building 1: each row perturbs one zone's load to maximum heating (red) or cooling (green) while others retain measured loads. Black lines show predictions under measured conditions.

4.1.2 MARL Performance

Figure 7 compares the indoor temperature under MARL control with the measured baseline (inherent control of the system) for building 1. Since occupants can adjust their thermostats, the temperature bounds vary across days. For clarity, one representative day is shown. Only five office rooms are presented, as the corridor (zone 1) is not equipped with radiant ceiling panels and is therefore not controlled by MARL.

Under the baseline, zones 2 and 3 exhibit temperature violations. The heating system was activated at 6:00 AM when zones became occupied, but due to the slow response of the hydronic system, room temperatures did not begin to rise until approximately 30 minutes later. In contrast, MARL provided proactive heating, as evidenced by an earlier rise in room temperature to meet the setpoints. For zone 4, MARL maintained the room temperature within bounds throughout the day, except for brief violations in the early morning, whereas the baseline failed to achieve thermal comfort until the afternoon. Similar performance differences were observed for zones 5 and 6, where MARL consistently maintained indoor temperatures within the comfort range, while the baseline intermittently failed to do so. Over the 114 simulated days, the mean cumulative temperature violation was reduced from 7.57°C under the baseline to 1.60°C under MARL,

representing a 79% reduction. Total energy consumption was 26,175 kWh and 22,053 kWh for the baseline and MARL, respectively, corresponding to an energy saving of 15.74%.

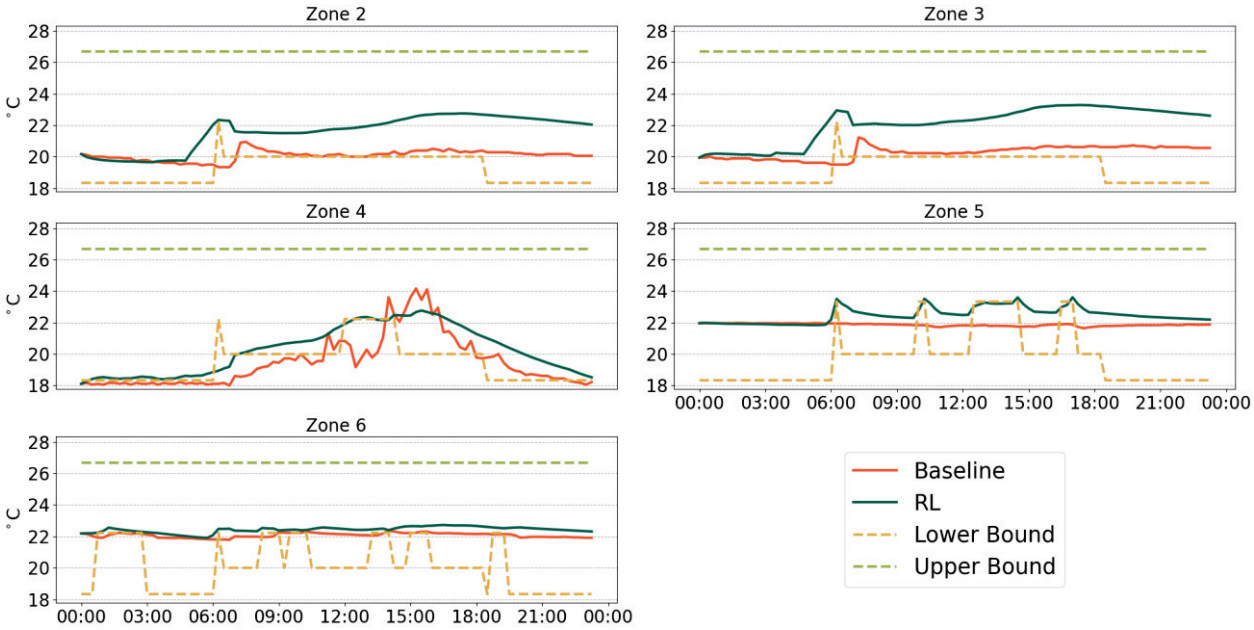

Figure 7. Indoor temperature comparison between MARL and baseline for building 1 on a representative day.

In summary, the simulation results on building 1 demonstrate that the proposed framework performs well at both the modeling and control stages. The PCGNN achieves below 1.18℃ MAE across all 6 zones in autoregressive prediction, and the perturbation analysis confirms that the model correctly captures both zonal thermal responses and inter-zone heat transfer through the building adjacency structure. Built upon the accurate and physically consistent dynamics model, the MARL agents reduce cumulative temperature violations by 79% compared to the baseline while achieving 15.74% energy savings. These results validate the framework in a controlled simulation setting and motivate its deployment into a larger, real-world building.

4.2 Implementation Study: Building 2

4.2.1 PCGNN Performance

Having validated the physical consistency of the PCGNN through the perturbation analysis on building 1, we now evaluate its prediction performance on building 2 to assess whether the model scales to a larger, more complex multi-zone system. Table 3 reports the autoregressive prediction accuracy on the testing dataset for all 18 modeled zones across 4 floors. Despite the threefold increase in the number of zones and the added complexity of inter-floor thermal coupling, the PCGNN maintains strong prediction accuracy: all zones achieve a MAE below 1.4℃, with the majority under 1.0℃. The lowest errors are observed in zones 5 (0.65℃) and 6 (0.86℃), while the highest are in zones 17 (1.40℃) and 18 (1.30℃), which are conference rooms where more dynamic occupancy profiles introduce additional variability. These results confirm that the proposed architecture, combining grouped causal convolutions with the heat diffusion graph layer, generalizes effectively from a 6-zone to an 18-zone building without degradation in prediction quality.

Table 3. Performance of PCGNN on testing dataset of building 2 for indoor temperature predictions.

Zone	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
MAE ℃	0.90	0.80	1.15	0.83	0.65	0.59	0.86	0.72	1.17
Zone	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
MAE ℃	1.30	0.75	0.88	1.10	0.69	1.00	0.86	1.40	1.30

4.2.2 MARL Performance

During the 42-day real-building deployment, the majority of zones are maintained within the prescribed temperature bounds under MARL most of the time, with the exceptions of rooms 3, 13, 17, and 18. For clarity, we show temperature comparison of representative zones between MARL and baseline in Figure 8 (good performance) and Figure 13 (underperformance). The underperformance in these four zones is not attributable to the PCGNN or the MARL algorithm, but rather to incorrectly commissioned radiant ceiling panels in zones 3 and 13, whose malfunction in turn affected the thermal conditions of neighboring zones 17 and 18. A detailed discussion of these commissioning issues is provided in the next section.

For the remaining zones, MARL exhibits clear predictive control behavior: temperatures begin rising ahead of the occupied period, earlier than under the baseline control, demonstrating that the agents have learned to pre-condition spaces in anticipation of upcoming comfort requirements. This is further confirmed by Figure 9 from the comparison of the radiant panel loads under both controllers for zones 6 and 9. A particularly noteworthy finding from Figure 8 is the control strategy learned for the third floor: MARL allocates a higher heating load to zone 6, the core zone of the floor, and leverages inter-zone heat transfer to condition its surrounding zones, rooms 4, 5, 7, 8 and 9, with less total energy than the baseline. This emergent strategy, which was not explicitly programmed but learned through multi-agent coordination, validates that the agents exploit the building's thermal network structure to achieve energy-efficient control.

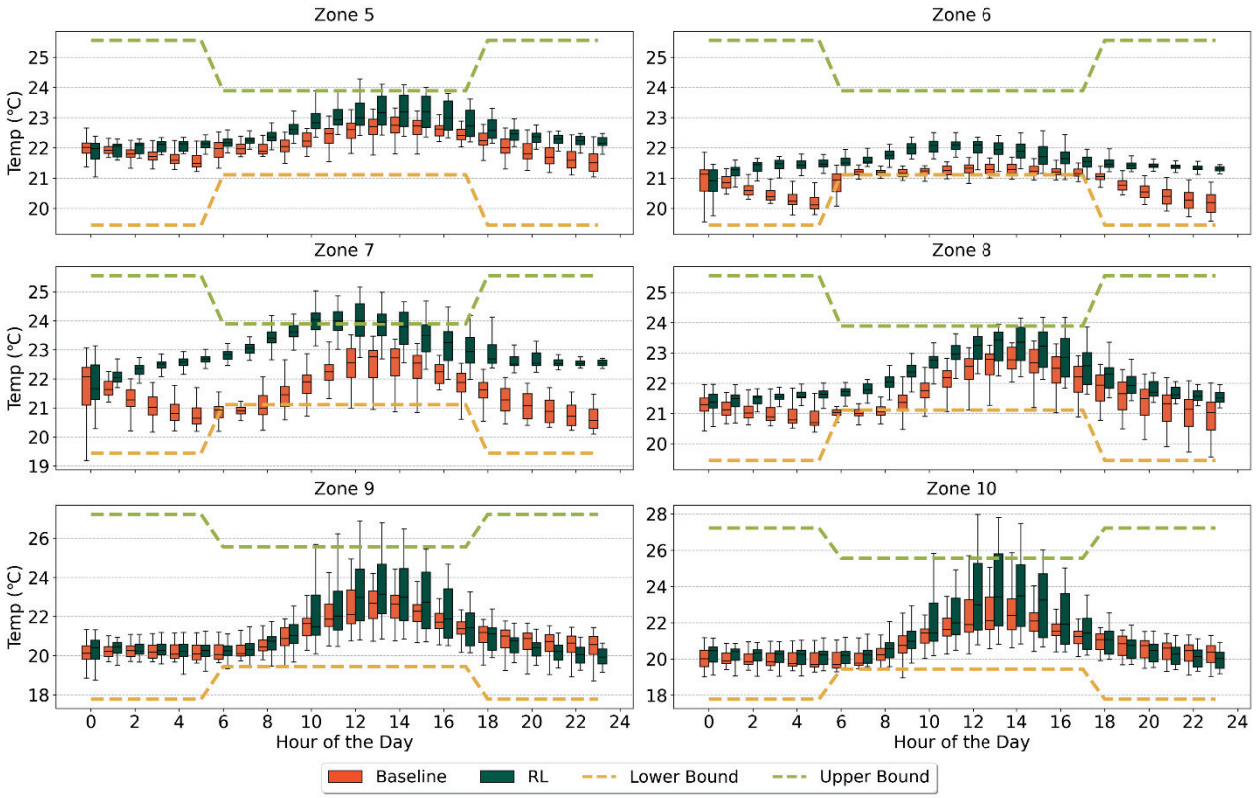

Figure 8. Indoor temperature of zones 4–9 during the 42-day deployment in building 2.

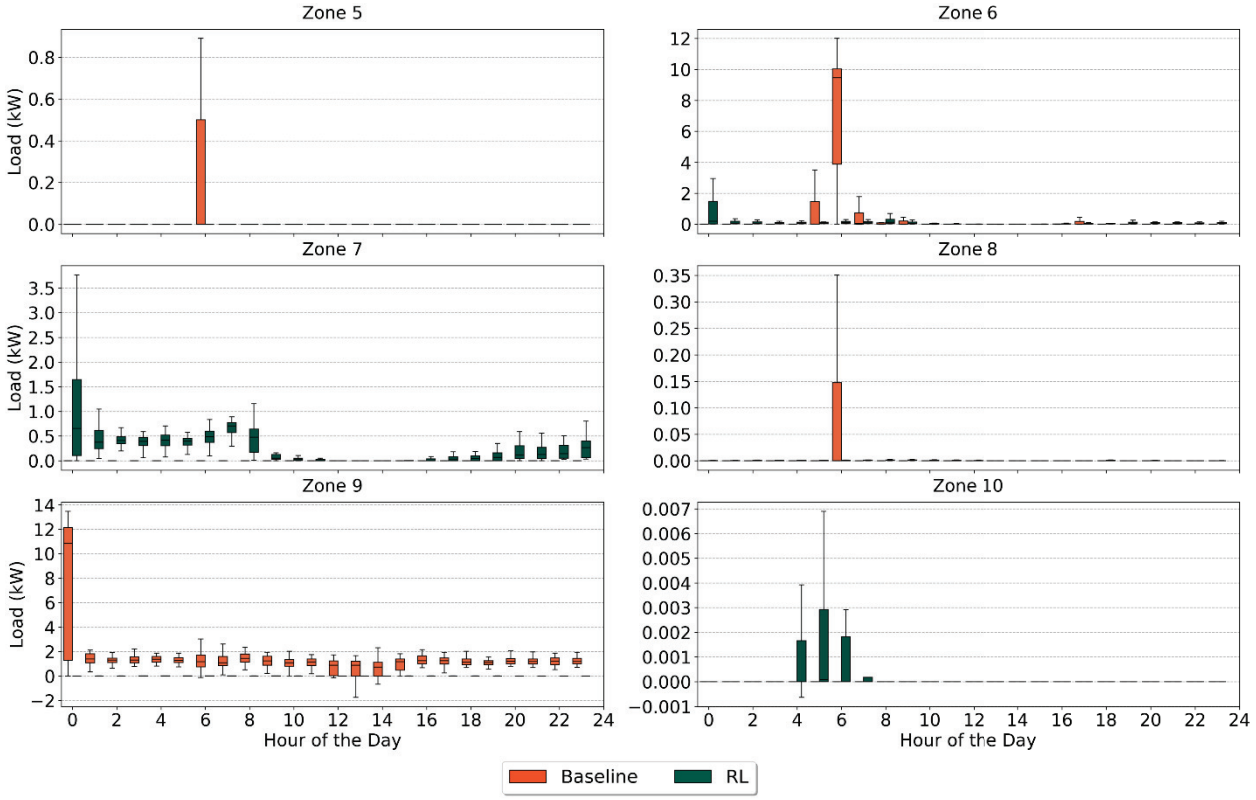

Figure 9. Radiant ceiling panel load comparison between baseline and MARL during implementation for zones 4-9.

The energy consumption and temperature violations under both controllers are quantified in Table 4. To account for varying weather conditions across the alternating deployment days, results are aggregated using a binning method: two bins are defined based on outdoor temperature and solar radiation with intervals of 5°C and 400 W/m², respectively, and the energy usage and temperature violation are averaged within each bin for each controller. The values in parentheses in the MARL load column indicate the percentage energy savings relative to the baseline.

Table 4 shows that the overall temperature violation under MARL is slightly higher than the baseline. This is attributable to two factors. First, as discussed above, the malfunctioning radiant panels in four zones inflate the MARL violation metric. Excluding these zones, the mean temperature violations are 0.32°C and 0.49°C for the baseline and MARL, respectively, indicating that MARL achieves substantial energy savings with modest comfort degradation. Second, the MARL formulation treats temperature bounds as soft constraints within the reward function. This trade-off is inherent to the soft constraint formulation: the penalty coefficients in the reward function (Equation (36)**Error! Reference source not found.**) allow the agents to balance energy reduction against comfort maintenance, and adjusting these coefficients provides a direct mechanism to shift this trade-off toward stricter comfort at the cost of reduced energy savings. This behavior is consistent with findings from our previous work [8], and enforcing hard constraints in RL remains an active research question in the community.

Despite the slightly higher temperature violations, Table 4 demonstrates that MARL achieves substantial energy savings ranging from 35.06% to 69.67% across weather bins. When the four malfunctioning zones are excluded, the savings increase to 40.89%–86.34%. These significant reductions are driven by two key mechanisms: the proactive pre-conditioning strategy that avoids reactive over-compensation, and the agents learned ability to leverage inter-zone heat transfer for more efficient load distribution across the thermal network.

In summary, the 42-day real-building deployment in building 2 demonstrates that the proposed framework scales effectively from simulation to practice. The PCGNN maintains below 1.4℃ MAE across all 18 zones, and the MARL agents achieve 35–70% energy savings (41–86% excluding malfunctioning zones) while maintaining comparable though slightly degraded thermal comfort relative to the baseline controller. Beyond raw performance metrics, the deployment reveals that the agents learn physically meaningful control strategies, notably proactive pre-conditioning and exploiting inter-zone heat transfer through core zones without being explicitly programmed to do so. These emergent behaviors confirm that the combination of a physics-consistent dynamics model and κ-truncated multi-agent learning enables the agents to discover and exploit the building's thermal network structure for energy-efficient control in real-world operating conditions.

Table 4. Energy usage and temperature violation of baseline and MARL control under different weather conditions.

Temp Bin (°C)	**Sol Bin (w/m²)**	**Load: Baseline (kW)**	**Load: MARL (kW)**	**Temp Vio: Baseline (°C)**	**Temp Vio: MARL (°C)**
(-17.5, -12.5]	**(0, 400]**	584.41	327.6 (43.95%)	0.38	0.81
(-12.5, -7.5]	**(0, 400]**	621.53	312.44 (49.73%)	0.395	0.82

	(400, 800]	554.36	359.98 (35.06%)	0.393	0.86
(-7.5, -2.5]	**(0, 400]**	507.88	318.02 (37.38%)	0.408	0.73
	(400, 800]	573.2	272.4 (52.48%)	0.395	0.877
(-2.5, 2.5]	**(0, 400]**	669.94	203.39 (69.67%)	0.803	0.71
(2.5, 7.5]	**(0, 400]**	281.33	135.49 (51.84%)	0.55	0.88

5 Discussion

Beyond the physically consistent model behavior and significant energy saving through real building implementation in the preceding section, this section discusses several aspects of the proposed framework: training efficiency of the PCGNN and MARL, observations on the measured data, practical considerations for RL-based control in both simulation and real-building settings, and limitations of the current study.

5.1 Effect of Physics-Informed Constraints

To evaluate whether the physics-informed components of the PCGNN are necessary, we compared the proposed model against an ablated variant on building 1. The ablated model replaces the heat diffusion GNN layer with a standard message-passing GNN that uses unconstrained learned edge transformation, removes the energy conservation structure, and eliminates the non-negative heat transfer coefficients and learned thermal capacity. The causal CNN backbone, hyperparameters, training data, and training procedure remain identical between the two models. The loss function is also modified to remove physical-informed terms, which only contains MSE loss.

Table 5 reports the MAE of the ablated model on the testing dataset. Compared with PCGNN results in Table 2, the ablated model shows degraded prediction accuracy across all 6 zones. the largest degradations occur in zone 1 (0.39℃ to 0.70℃), zone 3 (0.58℃ to 1.14℃), and zone 5 (0.31℃ to 0.68℃)

Table 5. Performance of the ablated model on testing dataset of building 1.

Zone	**1**	**2**	**3**	**4**	**5**	**6**
MAE (℃)	0.70	1.23	1.14	1.26	0.68	0.45

Beyond prediction accuracy, we conduct the same perturbation analysis described in Section 4.1.1 on the ablated model. Figure 10 shows the cross-zone perturbation matrix for the ablated model. For the PCGNN (Figure 6), the predicted temperature consistently increases under maximum heating and decreases under maximum cooling across all zones, and the spatial influence decays with graph distance following the building adjacency structure. In contrast, the ablated model exhibits physically inconsistent behavior: in several zones, increasing the cooling load increases the predicted temperature rather than decreasing it. This inverted directional response indicates

that without the non-negative heat transfer coefficients, the symmetric diffusion update that enforces energy conservation, and physics-informed loss terms, the standard GNN fits the training data through spurious correlations that violate fundamental thermodynamic principles.

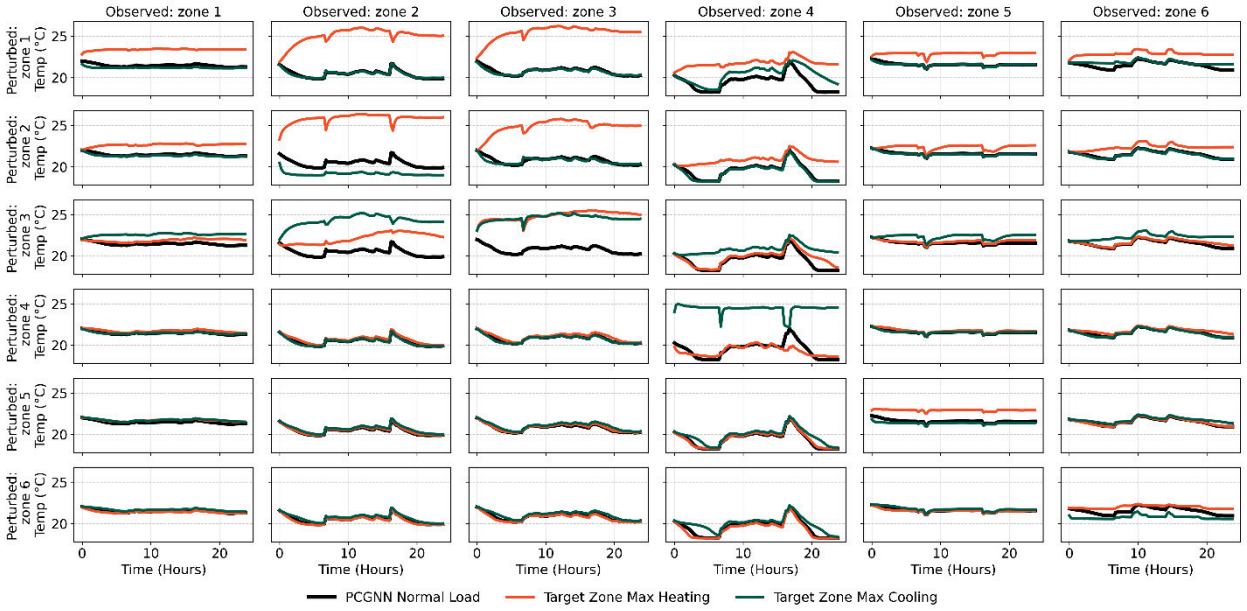

Figure 10. Cross-zone perturbation analysis of the ablated model (causal CNN + vanilla GNN without physics constraints) for building 1.

This distinction has direct implications for downstream control. In model-based MARL, the agents backpropagate policy gradients through the learned dynamics model to optimize control actions. A dynamics model that produces reasonable average prediction error but learns inverted causal relationships between control inputs and temperature responses would generate gradients that point in the wrong direction, causing the agents to learn counterproductive control strategies. The physics-informed constraints of the PCGNN are therefore not merely accuracy improvements but structural guarantees that the learned model respects the directionality of the underlying thermodynamics, which is essential for reliable policy optimization. Training MARL on the ablated model is therefore not expected to yield meaningful control behavior, as the reversed gradient signs would direct agents to increase cooling when heating is needed and vice versa.

5.2 Training Efficiency

5.2.1 PCGNN

As discussed earlier, a key architectural improvement of the proposed PCGNN over our previous work is the use of native PyTorch layers, causal 1-D convolutions and graph-based diffusion operations, to achieve mathematical equivalence to the manually connected neuron structure in [60]. This substitution enables significant acceleration of the training process. In [60], modeling 6 zones (zones 4 to 9) with a 6-hour prediction horizon typically required over 3 hours of training time. It was infeasible to train the model with more zones and longer prediction horizons. Table 6 compares the training time of the proposed PCGNN across different model scales with our previous model. For consistency, the same zones from [60] are used for the 3-zone and 6-zone configurations: in the 3-zone case, rooms 302, 303, and 303A are aggregated into one zone and rooms 304 and 399C into another, while the 6-zone case models each room individually. The 3-

zone and 6-zone models both converge in approximately 30 minutes regardless of the prediction horizon, representing a 6× speedup over the previous approach. The 18-zone model, which was computationally infeasible with the manual connection architecture, requires only 80 minutes for a 6-hour prediction horizon and 96 minutes for a 24-hour horizon. This training efficiency, scaling sub-linearly with the number of zones, is what makes it practical to extend physics-informed multi-zone modeling from a handful of zones to an entire building.

Table 6. Training time of PCGNN and previous model for building 2 with different number of zones and prediction horizons.

Number of Zones	Prediction Horizon	Training Time (PCGNN)	Training Time (Previous Model)
3 zones	**6 hours**	32 minutes	> 120 minutes
	24 hours	29 minutes	Infeasible
6 zones	**6 hours**	38 minutes	> 180 minutes
	24 hours	32 minutes	Infeasible
18 zones	**6 hours**	80 minutes	Infeasible
	24 hours	96 minutes	Infeasible

5.2.2 κ-Neighbor Truncated Model-Based MARL

To demonstrate the effectiveness of the proposed κ-truncated model-based MARL, we compare four configurations on the building 2 case: (1) model-free MARL without truncation (SAC: vanilla, each zone has an agent and value network with global states and actions), (2) model-free MARL with κ-hop truncation (SAC: κ-hop only), (3) model-based MARL without truncation (SAC: dynamic only), and (4) model-based MARL with both truncation and the dynamics model (SAC: dynamic & κ-hop).

Figure 11 shows the training curves for each individual objective and the total reward under the four MARL configurations, averaged over 5 random seeds with ±1 standard deviation shown as shaded bands. All four configurations use the same computational budget: identical number of gradient steps (40,000), batch size (2,048), replay buffer size (1,000,000), and network architecture, with evaluation performed every 100 training steps. The model-based κ-truncated configuration (dynamic & κ-hop) consistently converges faster and achieves the highest total reward with the smallest variance across all seeds. The two model-free configurations (κ-hop only and vanilla) exhibit larger variance, lower final reward, and instability during early training, particularly in the non-smoothness and IEQ objectives. For the two model-free approaches, the agents initially achieve similar total rewards regardless of whether truncation is applied, indicating that the truncated critic can approximate the true value function well even with reduced information. This is consistent with the theoretical result in [66], which proves that network MDPs with exponential decay properties admit accurate local approximations of the Q-function. However, neither model-free approach reaches the reward level of the model-based configurations.

The decomposition of the total reward into individual objectives is also shown in Figure 11. The model-free approaches struggle to learn a stable balance: improvements in energy reduction come at the cost of deteriorating IEQ performance in the later stages of training, and agents struggle to make smooth action. By contrast, model-based approaches improve both objectives simultaneously, suggesting that access to the dynamics model enables the agents to discover control strategies that reduce energy consumption without sacrificing comfort and action

smoothness. With κ-truncation further reducing the critic input dimensionality, and consequently the learning complexity, the proposed combined approach achieves the best overall performance among the four configurations.

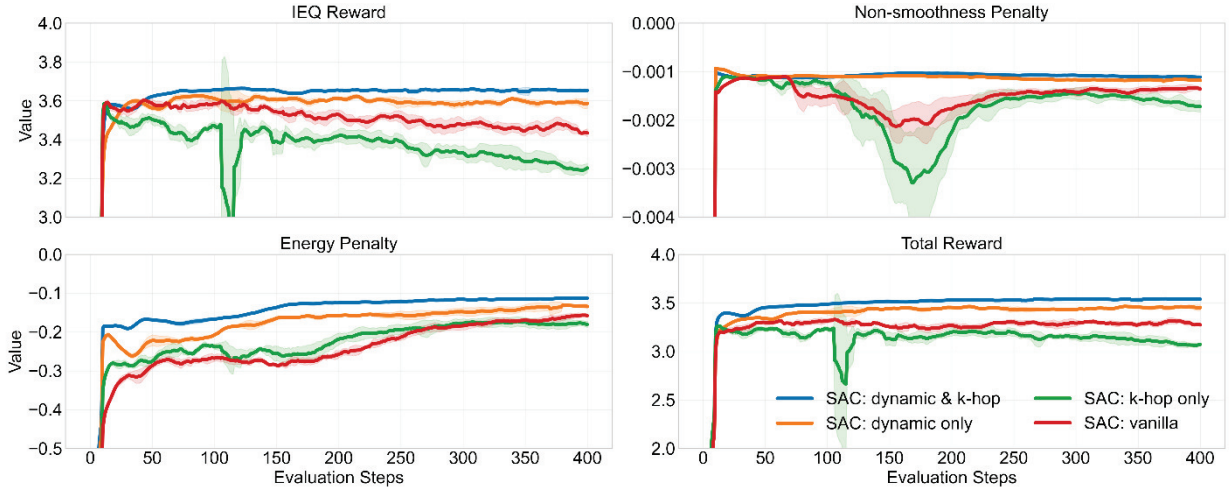

Figure 11. Training curves of four MARL configurations for building 2, averaged over 5 random seeds with ±1 standard deviation (shaded bands): IEQ reward (top left), non-smoothness penalty (top right), energy penalty (bottom left), and total reward (bottom right).

5.3 Observations on Measured Data and Its Impact of MARL Performance

5.3.1 Building 1

As described in section 3.1.1, building 1 determines the effective setpoint and operating mode based on occupancy status and outdoor temperature. Notably, the threshold for switching between heating and cooling modes has only a 0.5°C difference. This narrow deadband is problematic in practice because outdoor temperature sensors typically have an accuracy range of ±1°C, meaning sensor noise alone can trigger unintended mode switching. As shown in plot (a) of Figure 12, the temperature bounds derived from the effective setpoint and load sign reveal that after 12 PM, the system frequently alternates between heating and cooling modes at the same effective setpoint, resulting in simultaneous heating and cooling throughout the afternoon — a significant source of energy waste.

This issue also has direct implications for MARL training. If the lower and upper temperature bounds are taken directly from the raw measurement data, the agents are trained against inconsistent and physically unrealistic comfort expectations that reflect control system faults rather than genuine occupant requirements. Under such conditions, the agents fail to learn a meaningful balance among competing objectives. Diagnosing and cleaning the temperature bounds, as shown in plot (b) of Figure 12, is therefore a necessary preprocessing step. With the corrected bounds, the proposed MARL achieves significant improvements in both energy efficiency and indoor temperature regulation, underscoring the importance of data quality for RL-based building control.

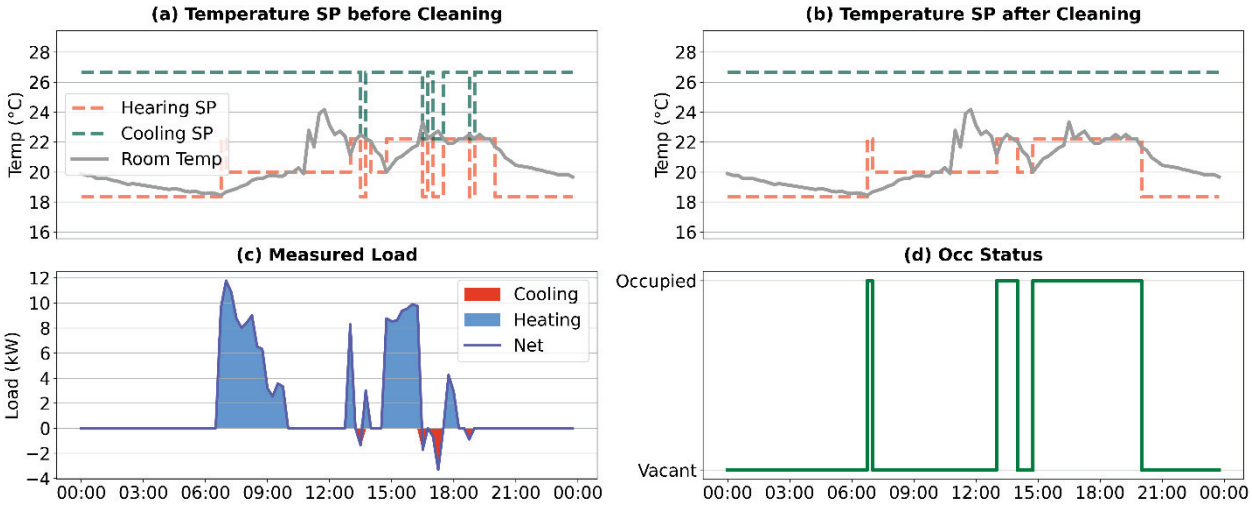

Figure 12. Data quality in building 1: (a) raw temperature bounds showing simultaneous heating and cooling, (b) corrected bounds, (c) measured load, (d) valve position, and (e) occupancy.

5.3.2 Building 2

As noted in Section 4.2.2, four zones, zones 3, 13, 17, and 18, exhibited poor temperature performance under MARL and baseline control (Figure 13). Investigation revealed that this underperformance is attributable to hardware malfunction rather than deficiencies in the PCGNN or MARL algorithm. As shown in Figure 14, the radiant ceiling panels in zones 3 and 13 were commanded to provide heating, indicated by positive valve positions. However, the load feedback shows fluctuations between cooling and heating even at 100% heating valve opening, suggesting hydraulic short circuits or leakage through the radiant ceiling panels. This malfunction results in an unmet heating demand, explaining the low indoor temperatures observed for zones 3 and 13 in Figure 13.

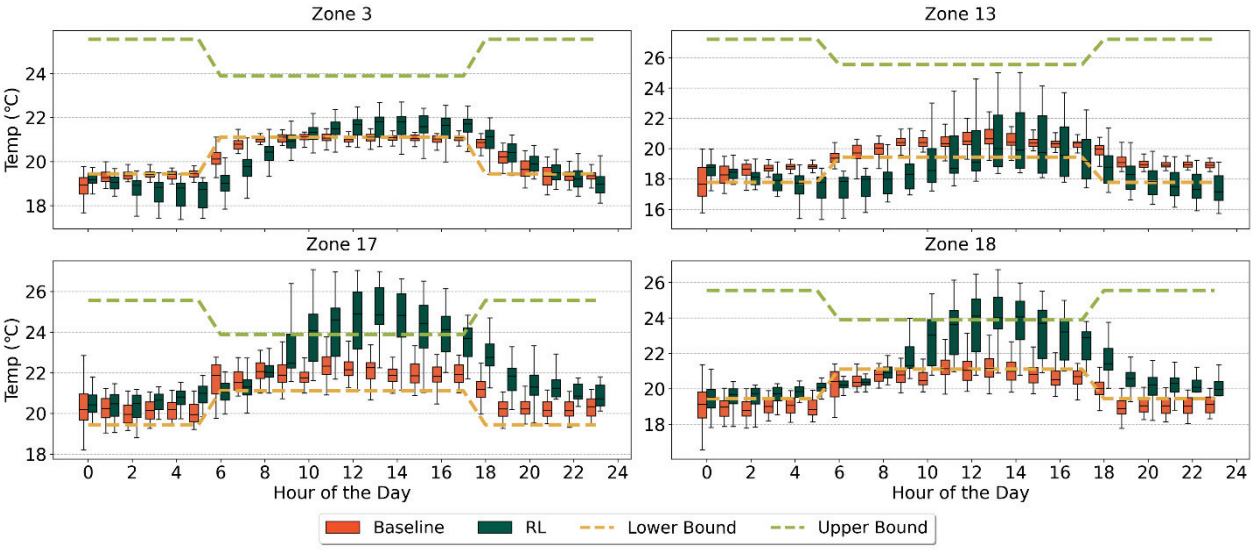

Figure 13. Temperature comparison of zones 3, 13, 17, and 18 between baseline and MARL during implementation.

The consequences propagate through the building's thermal network. Zones 17 and 18 are located directly above zone 13, and the MARL agents, having learned the inter-floor thermal coupling, attempted to compensate for the heating deficit in zone 13 by allocating additional heating to the

fifth-floor zones. This compensatory strategy, while physically logical, led to overheating in zones 17 and 18. This finding illustrates both a strength and a limitation of the framework: the agents successfully learned to exploit cross-floor heat transfer, but they cannot fully overcome persistent hardware faults that violate the assumptions of the learned dynamics model.

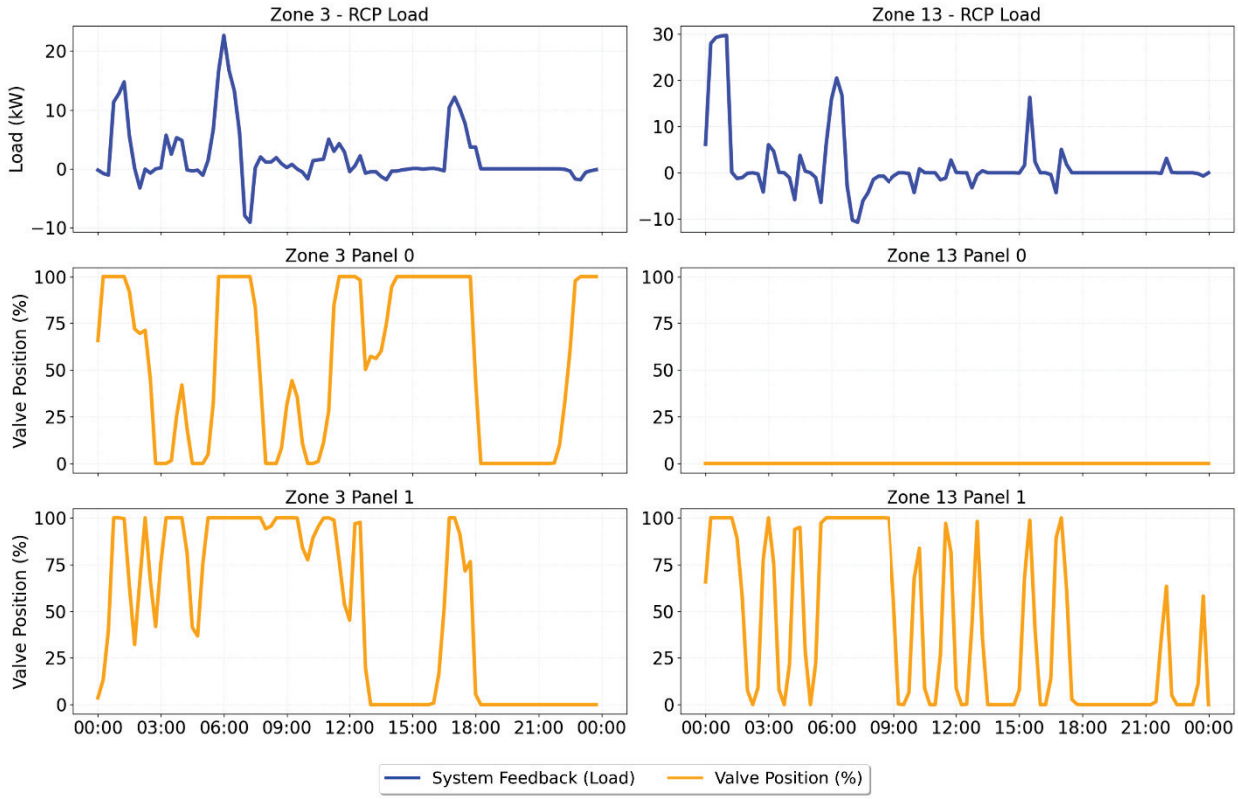

Figure 14. Feedback of radiant panel load and valve positions of zones 3 and 13 of building 2.

5.4 Practical Considerations of RL Control

Building 1 is equipped with interactable thermostats, allowing occupants to adjust temperature settings based on their preferences. This dynamic setpoint introduces additional complexity for RL training, as the agents must respond to comfort requirements that change unpredictably throughout the day.

To investigate the impact of setpoint information on agent performance, we compare three state configurations in Figure 15: (1) current and next 8-step setpoints included as states (Current & Future SP), (2) current and previous 8-step setpoints included as states (Current & Previous SP), and (3) current setpoints only (Current SP Only). The results show that including only the current setpoint — the configuration that strictly satisfies the Markov property assumption — leads to poor performance, with the agents unable to learn strategies that adequately meet the temperature requirements. Including future setpoint information yields a significant improvement, with substantially fewer temperature violations. The configuration with past setpoint information performs better than the current-only case but noticeably worse than the future-information case: the agents exhibit delayed reactions to setpoint changes, confirming that historical setpoints serve as a partial but insufficient proxy for anticipating upcoming comfort demands.

These findings have practical implications for real-building deployment. Achieving the best RL performance in building 1 would require forecast models for both occupant thermostat interactions and weather conditions to supply the agents with anticipated future setpoints. This adds implementation complexity and introduces an additional source of error, as forecast inaccuracies could degrade the RL control performance. By contrast, building 2 uses schedule-based setpoints that are known in advance, avoiding this challenge entirely. The comparison highlights that the complexity of RL deployment depends not only on the algorithm design but also on the controllability and predictability of the building's operational context.

Beyond the simulation findings, the real-building deployment in building 2 revealed that the majority of control failures were not caused by the MARL algorithm itself but by two infrastructure-level issues: sensor failures and BACnet communication errors. Sensor failures occurred due to power outages at individual sensors or disruptions to the cloud service that relays sensor data, both of which prevent the agents from receiving valid state observations and therefore lead to erroneous control decisions. A stable and reliable sensing infrastructure is thus a prerequisite for deploying any advanced control strategy in real buildings. BACnet communication failures arose because the protocol is designed for sequential access; when multiple processes attempt to read or write simultaneously, communication congestion causes errors that interrupt the control loop. Implementing asynchronous access would mitigate this issue in future deployments. These practical challenges underscore that the gap between simulation success and real-building performance is often determined not by the control algorithm but by the reliability of the underlying building automation infrastructure.

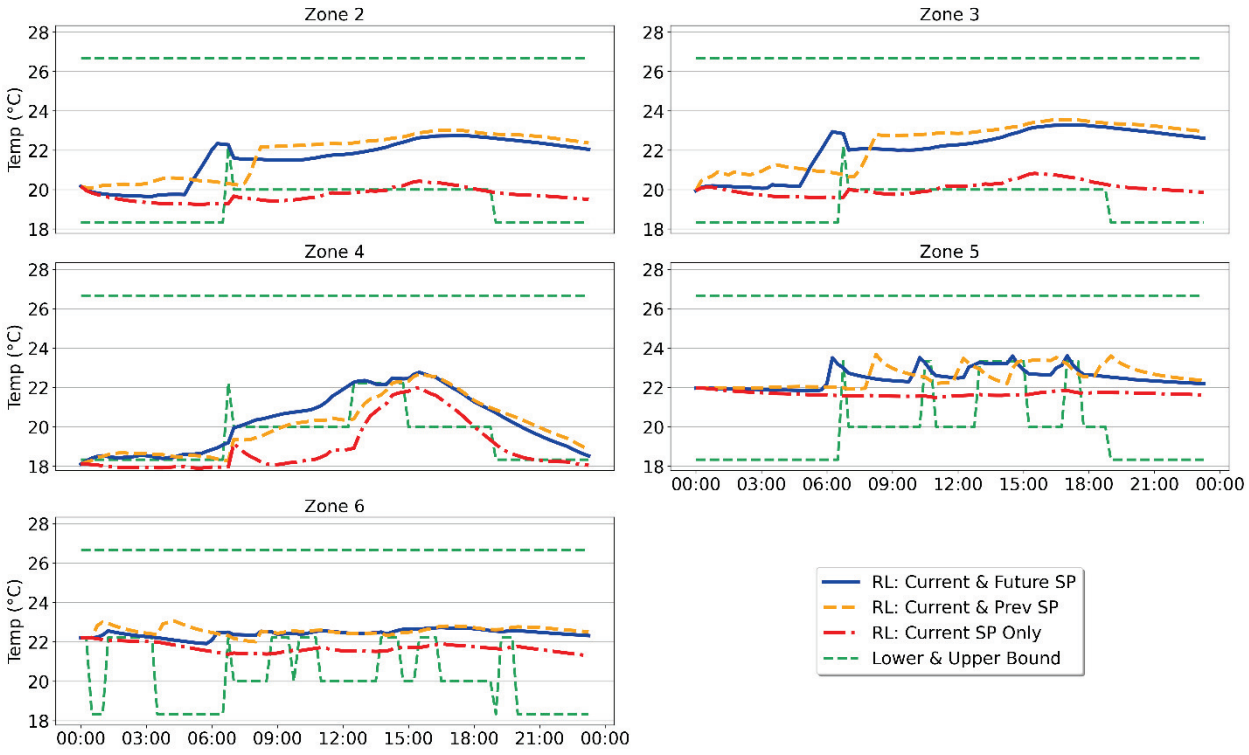

Figure 15. Effect of setpoint information on MARL performance for building 1: comparison of agents with current+future, current+past, and current-only setpoint states..

5.5 Applicability of κ-Truncation

The κ-truction relies on the exponential decay property of network MDPs, where each agent's value function has diminishing sensitivity to agents beyond its κ-hop neighborhood. This property holds for localized policies with decay rate $\rho = \gamma$ regardless of the control interval. A higher control frequency reduces the per-step temperature change between adjacent zones, which strengthens the locality of thermal coupling at each step and makes the κ-truncation approximation tighter rather than looser.

However, κ-truncation assumes that inter-zone coupling is mediated through the building's thermal adjacency graph, where influence propagates hop by hop. This assumption can be violated when centralized systems such as shared thermal storage or centralized chillers create direct coupling between physically distant zones that bypass the adjacency structure. In such cases, the exponential decay property may not hold for the subsystem served by the shared resource. A natural extension is hierarchical MARL, where a higher-level agent controls the shared system (e.g., thermal storage system or central plant equipment) and lower-level zone agents continue to use κ-truncated MARL for local radiant or air-side control. The higher-level agent would receive aggregated building-wide information to coordinate distant zones, while the lower-level agents retain the scalability advantages of local neighborhood truncation. Exploring this hierarchical architecture is left for future work.

5.6 Limitations of This Work

Despite promising results demonstrated in proceeding sections, several limitations of the current study should be acknowledged.

1. **Data requirements and transferability**. The PCGNN models for buildings 1 and 2 were trained on approximately 14 and 21 months of historical data, respectively. The minimum amount of data required for effective PCGNN training was not systematically investigated in this study. Additionally, while the same PCGNN architecture is applied to both buildings without structural changes, the model parameters are trained independently for each building. Exploring the minimum data requirement for reliable PCGNN training and the potential for transfer learning across buildings with similar HVAC configurations are directions for future work.
2. **More real-building implementations**. While the simulation study on building 1 validates the framework in a controlled setting, deploying the MARL agents on the physical system of building 1 would provide stronger evidence of generalizability across different building types and HVAC configurations. Future work will pursue this deployment while addressing the practical challenges identified in this study, including the need for occupant behavior forecasting and robust sensor infrastructure.
3. **Long-term implementation in building 2**. The 42-day implementation period in building 2, while sufficient to demonstrate the framework's viability, covers only winter heating conditions. A longer deployment spanning multiple seasons would capture a wider range of operating scenarios — including cooling, swing seasons, and varying occupancy patterns — and provide a more comprehensive evaluation of the agents' adaptability and long-term performance stability.
4. **Scaling up to 100+ zones**. Although the κ-truncation strategy is designed to enable scalability, the largest building tested in this study contains 18 zones. Testing the framework on buildings with 100 or more zones would more rigorously evaluate whether

the computational and learning efficiency gains from κ-neighborhood truncation hold at significantly larger scales, and whether the PCGNN's grouped convolution and heat diffusion architecture remains effective as the building graph grows in size and complexity.

6 Conclusion

This paper presented a physics-informed model-based multi-agent reinforcement learning framework for scalable multi-zone building HVAC control. The framework integrates three components: a physics-consistent graph neural network (PCGNN) for multi-zone temperature prediction, a κ-neighborhood truncated multi-agent soft actor-critic algorithm for scalable policy learning, and a model-based training procedure that leverages the learned dynamics for optimal policy learning.

The PCGNN combines group-shared multi-scale causal convolutions with heat diffusion graph neural network layers, replacing the manually connected neuron architecture of our previous work with standard PyTorch operations. This architectural redesign achieves mathematical equivalence while reducing training time from over 3 hours to approximately 30 minutes for 6-zone models and enables scaling to 18 zones in under 100 minutes, a configuration that was computationally infeasible with the prior approach. The model achieves below 1.18 and 1.4℃ MAE across all zones of the 6-zone building 1 and the 18-zone building 2, respectively. Perturbation analysis confirms that the model learns both the correct directional response to control inputs and the spatially decaying inter-zone influence dictated by the building adjacency structure.

The κ-truncated model-based multi-agent reinforcement learning (MARL) demonstrated clear advantages over model-free alternatives. Ablation studies showed that the combination of κ-truncation and the dynamics model converges faster and to higher reward than model-free approaches, which struggle to balance energy efficiency and thermal comfort objectives simultaneously. The κ-truncation reduces critic input dimensionality by up to 72% for the 18-zone building while preserving learning performance.

In the 42-day real-building deployment in building 2, MARL achieved energy savings of 35–70% across weather conditions while maintaining comparable though slightly degraded thermal comfort relative to the baseline controller. The agents learned optimal control strategies, including proactive pre-conditioning and leveraging inter-zone heat transfer through core zones. The deployment also revealed that practical challenges such as sensor reliability, BACnet communication, and hardware commissioning are often the binding constraints on real-building performance, rather than the control algorithm itself.

Future work will extend the real-building deployment to additional buildings including the building 1, conduct longer-term implementation in building 2 spanning multiple seasons, and test the framework's scalability on buildings with 100 or more zones to further evaluate and improve the limits of κ-neighborhood truncation and the PCGNN architecture.

7 Acknowledgement

Xuezheng Wang and Bing Dong’s work is supported by the U.S. National Science Foundation (Award No. 1949372). Zhaolin Ren and Na Li’s work is supported by the U.S. National Science Foundation (Award No. 2328241, No. 2112085, and No. 2401390).

8 Appendix

8.1 Literature Review Tables

Table 7. Representative studies of single-agent RL for building HVAC control.

Ref.	Simulation/Experiment	No. of Zones
[8]	Experiment	1
[9]	Simulation	1
[10]	Simulation	1
[11]	Simulation	1
[12]	Simulation	1
[13]	Simulation	1
[14]	Simulation	1
[15]	Simulation	1
[16]	Simulation	2
[17]	Simulation	1
[18]	Simulation	1
[19]	Experiment	5
[20]	Experiment	1
[21]	Experiment	1
[22]	Experiment	1
[23]	Experiment	4
[24]	Simulation	1
[61]	Experiment	1
[62]	Experiment	6

Table 8. Representative studies of MARL for building systems from 2018 to 2025. S: Simulation. E: Experiment. MF: model-free. MB: model based. C: Centralized. I: Individual

Ref.	Application	Simulation /Experiment	No. of Building	No. of Zone	No. of Agent	Critic	States (Max)	Actions (Max)	M
[25]	Microgrid	S	NA	NA	5	C	13	7	MF
[26]	Microgrid	S	4	1	4	I	3	1	MF
[27]	HVAC/Thermal	S	1	1	2	I	9	1	MF
[28]	HVAC/Thermal	S	1	4	2	C	9	1	MF
[30]	HVAC/Thermal	S	1	5	3	C	15	1	MF
[32]	HVAC/Thermal	S	1	5	5	I	2	1	MF
[31]	HVAC/Thermal	S	1	1	3	C	3	1	MB
[29]	HVAC + Storage	S	1	1	5	C	10	3	MF
[33]	HVAC/Thermal	S	1	4	4	I	6	2	MF

[34]	HVAC + Storage	S	1	1	3	C	9	1	MF
[35]	HVAC/Thermal	S	1	5	5	I	4	1	MF
[36]	HVAC + Storage	S	9	1	9	C	27	3	MF
[37]	HVAC + Storage	S	96	1	96	C	18	1	MF
[38]	Energy Storage	S	1	1	2	C	7	2	MF
[39]	HVAC + Storage	S	1	2	2	C	5	1	MF
[40]	Energy Storage	S	9	1	9	C	14	1	MF
[41]	HVAC/Thermal	E	1		3	C	10	1	MF
[42]	HVAC/Thermal	S	2	1	7	C	3	1	MB
[43]	Energy Storage	S	9	1	9	C	N/A	3	MF
[44]	HVAC/Thermal	S	3	15	15	C	6	1	MF
[45]	HVAC + Storage	S	1	1	3	C	6	1	MF
[46]	HVAC/Thermal	S	1	1	3	I	6	1	MF
[47]	HVAC + Storage	S	1	1	3	C	6	1	MF
[48]	HVAC + Storage	S	9	1	9	I	11	5	MF
[49]	HVAC/Thermal	S	1	6	2	I	8	1	MF
[50]	HVAC/Thermal	S	1	3	3	I	3	1	MF
[51]	HVAC/Thermal	S	1	1	5	I	4	1	MF
[52]	HVAC/Thermal	S	1	3	9	I	7	1	MF
[53]	HVAC/Thermal	S	1	10	2	C	6	1	MF
[54]	HVAC/Thermal	S	1	6	2	C	7	1	MF
[55]	Energy Storage	S	1	5	2	I	6	1	MF
[56]	Energy Storage	S	100	1	100	C	13	2	MF
[57]	HVAC/Thermal	S	1	1	2	C	3	1	MB
[58]	HVAC/Thermal	S	1	1	2	C	22	2	MF

8.2 Notation and Data Layout

To model multi-zone temperature dynamics, we define batch size B, time length (prediction horizon) T, number of zones (nodes) Z and number of features C_{in}. Then the input sequence has the dimension of:

$$\boldsymbol{X} \in \mathbb{R}^{B\times T\times Z\times C_{in}}, \qquad \boldsymbol{X}_{b,t,z} \in \mathbb{R}^{C_{in}}$$

Where $\boldsymbol{X}$ is model input tensor and $\boldsymbol{X}_{b,t,z}$ is the feature vector of zone z at time t in batch b.

To reduce model parameters, same convolutional layer can be applied to zones that have similar dimensions and functions. We can partition zones into manual groups $\mathcal{R}$, where each group $r \in \mathcal{R}$ contains zone indices $S_r \subset \{1, \dots, Z\}$, and $\cup_r S_r = \{1, \dots, Z\}$, $S_r \cap S_{r'} = \emptyset$. That being said, for multiple zones, we can manually categorize them into different groups based on their size and intended use, and each zone only belongs to one type.

We denote the hidden representation after some layer as:

$$\boldsymbol{H} \in \mathbb{R}^{B\times T\times Z\times D}$$

Where D is the hidden feature dimension inside the network.

8.3 Derivation of Group-Shared Multi-Scale Causal Convolutional Layer

Causal Convolution Branches

For a group r, take the group slice:

$$\boldsymbol{H}^{(r)}_{b,t,z} = \boldsymbol{H}_{b,t,z}, \qquad z \in S_r, \qquad \boldsymbol{H}^{(r)} \in \mathbb{R}^{B\times T\times |S_r|\times D_{in}}$$

Then we can define two temporal convolutions (short and long) that map D_{in} to D_{out} to capture the short-term and long-term impact of input futures on the zone temperature.

For the short-branch temporal convolution, we have kernel length k_s with dilation 1:

$$\boldsymbol{Y}^{(r)}_s(b,t,z,c) = \sum_{i=0}^{k_s-1}\sum_{m=1}^{D_{in}} W^{(r)}_s(c,m,i)\boldsymbol{H}^{(r)}(b,t-i,z,m) \tag{13}$$

Where $\boldsymbol{Y}^{(r)}_s$ is short-branch convolution output for group r, $W^{(r)}_s \in \mathbb{R}^{D_{out}\times D_{in}\times k_s}$ is short-branch kernel for group g. For the causality between input and output, we use the causal convention $\boldsymbol{H}(b,\tau,\cdot,\cdot) = 0, \forall \tau < 0$, i.e., zero padding as shown in Figure 2. By zero padding, the prediction of time step only uses the current and past information, i.e., inputs in the future steps have no impact on the prediction from the current step. This achieves the same effect as partial connection of MLP's neurons in our previous study but in a more elegant way.

For the long-branch temporal convolution, we have kernel length k_l with dilation d_l:

$$\boldsymbol{Y}^{(r)}_l(b,t,z,c) = \sum_{i=0}^{k_l-1}\sum_{m=1}^{D_{in}} W^{(r)}_l(c,m,i)\boldsymbol{H}^{(r)}(b,t-d_l i,z,m) \tag{14}$$

Where $\boldsymbol{Y}^{(r)}_s$ is long-branch convolution output for group r, $W^{(r)}_l \in \mathbb{R}^{D_{out}\times D_{in}\times k_l}$ is long-branch kernel for group r.

The rationale of short- and long-branch convolutions is to capture short- and long-term temporal relationship between input and output. However, implementing the two branches of convolutions

in their naïve form lacks the guarantee of physical consistency. In our previous study, physical consistency was enforced by non-negative weights. Here we define the same mechanism as an option in the following.

Optional Positivity Constraint on Kernels

The indoor temperature should be non-negatively related to its input features, such as the radiant panel load (positive for heating, negative for cooling), outdoor temperature, and solar radiation, and such a correlation was guaranteed by enforcing non-negative weights of the neural network in our previous study [8][60][65]. However, non-negative weights are very restrictive and sometimes can limit the expressiveness of the model. Therefore, here we use an optional positivity constraint on kernels:

$$\phi(W) = \begin{cases} W \odot W, & if\ enforce\ positive \\ W, & otherwise \end{cases} \tag{15}$$

$\odot$ is Hadamard product where we take elementwise square of the weight matrix. In the implementation, we will have:

$$W \leftarrow \phi(W)$$

Note that the $\phi(W)$ can be applied to any weights in the PCGNN, not only limited to convolutional layer.

If we enforce non-negative weights, then the input and output are guaranteed to be nonnegatively correlated. When such enforcement is not applied, we will need a loss term to penalize negative gradients of output with respect to input features as described in equation (34) **Error! Reference source not found.** for physical consistency. The choice of the two mechanisms depends on the tradeoff between model expressiveness and the extent of physical consistency required.

Softmax Mixing of the Two Time Scales

As aforementioned, we have two branches of convolution to capture short- and long-term impact of input features on the output. The contribution of the two branches is learned through two logits for each convolutional layer. For every convolutional layer, two learnable logits $\boldsymbol{\eta} = [\eta_s, \eta_l]$ are defined to calculate the mixing weights:

$$\alpha_s, \alpha_l = softmax(\boldsymbol{\eta}), \qquad \alpha_s + \alpha_l = 1, \qquad \alpha_s, \alpha_l \geq 0 \tag{16}$$

Then the mixed output is weighted sum of short and long branches output:

$$\boldsymbol{Y}^{(r)} = \alpha_s \boldsymbol{Y}_s^{(r)} + \alpha_l \boldsymbol{Y}_l^{(r)} \tag{17}$$

Adding a shared bias $\boldsymbol{b} \in \mathbb{R}^{D_{out}}$, we have:

$$\widetilde{\boldsymbol{Y}}^{(r)}(b,t,z,:) = \boldsymbol{Y}^{(r)}(b,t,z,:) + \boldsymbol{b} \tag{18}$$

Applying activation (LeakyReLU) and dropout, we have:

$$\boldsymbol{Z}^r = Dropout\left(\sigma\left(\widetilde{\boldsymbol{Y}}^{(r)}\right)\right), \qquad \sigma(\cdot) = LeakyReLU(\cdot) \tag{19}$$

Residual Connection

Multiple convolution layers could be involved in PCGNN, making it a deep network. Therefore, we apply residual connections to the layers of PCGNN to improve model training.

For a given convolutional layer, if the dimension of input is different from the output, i.e., $D_{in} \neq D_{out}$, we will use linear projection:

$$\boldsymbol{R}^{(r)}(b,t,z,:) = P\boldsymbol{H}^{(r)}(b,t,z,:) + \boldsymbol{p}, \quad P \in \mathbb{R}^{D_{out} \times D_{in}} \tag{20}$$

Otherwise, we will have $\boldsymbol{R}^{(r)} = \boldsymbol{H}^{(r)}$.

The final output for group g is:

$$\boldsymbol{H'}^{(r)} = \boldsymbol{Z}^{(r)} + \boldsymbol{R}^{(r)} \tag{21}$$

Lastly, we can define the full output $\boldsymbol{H'} \in \mathbb{R}^{B \times T \times Z \times D_{out}}$ by writing each group result into its zone indices:

$$\boldsymbol{H'}_{:,:,S_r:} \leftarrow \boldsymbol{H'}^{r}, \qquad \forall r \in \mathcal{R}$$

The short- and long-branches of convolution, positivity constraints of kernels, learnable mixing of two scales, and residual connections define the 1-D causal convolutional layers of our proposed PCGNN. As shown in Figure 2, convolutions are applied within the node (zone), which aims to capture the temporal impact of features within the zone. To model spatial inter-zonal impact, we use heat diffusion graph layer defined in the following subsection.

8.4 Derivation of Heat Diffusion GNN Layer

Nonnegative Heat Transfer Coefficient

Here we assume that the hidden representations denote the temperature of each zone. Each undirected edge (i,j) has a learnable scalar weight $w_{i,j}$ that represents heat transfer coefficient between zones i and j. The heat transfer coefficients are non-negative values. By enforcing nonnegativity using equation (15), we have:

$$\tilde{g}_{i,j} = \phi(w_{ij}) \geq 0$$

To improve training stability, for each graph instance, we define base degree

$$d_i = \sum_{j:(i,j)\in\varepsilon_u} \tilde{g}_{i,j} \tag{22}$$

Let $d_{max} = \max_i d_i$. We cap the step size:

$$\alpha_{cap} = \frac{0.9}{\max(d_{max}, \epsilon)} \tag{23}$$

Where ϵ is a small constant to avoid division by zero, e.g. 10^{-6}.

Then the effective step is:

$$\alpha_{eff} = \delta \cdot \min(\Gamma, \alpha_{cap}) \tag{24}$$

Where a learnable global scale Γ and a gate $\delta \in [0,1]$ are used:

$$\Gamma = \phi(global\ scale) \geq 0, \qquad \delta = \sigma(gate\ logit) \in (0,1)$$

Finally, the effective heat transfer coefficient is:

$$g_{ij} = \alpha_{eff}\tilde{g}_{ij} \geq 0 \quad (25)$$

Diffusion Update

Having the effective heat transfer coefficient, for each undirected edge (i, j), we define the symmetric heat flow between zones:

$$\Delta Q_{i\leftarrow j} = g_{ij}(\boldsymbol{h}_j - \boldsymbol{h}_i), \qquad \Delta Q_{j\leftarrow i} = g_{ij}(\boldsymbol{h}_i - \boldsymbol{h}_j) \quad (26)$$

$\Delta Q_{i\leftarrow j}$ represents heat gain of zone i from its adjacencies. Note that $\Delta Q_{i\leftarrow j} = -\Delta Q_{j\leftarrow i}$. It can be interpreted as the heat balance between zone i and zone j, i.e., the heat loss from zone i contributes to the heat gain of zone j, and vice versa.

To translate this heat flow into a temperature change, we must account for the distinct thermal mass of each zone. We introduce a learnable, strictly positive inverse heat parameter $c_i = 1/C_i$ for each node i. Considering all adjacencies of zone i, the node i update is:

$$\boldsymbol{h}_i' = \boldsymbol{h}_i + c_i \sum_{j:(i,j)\in\varepsilon_u} g_{ij}(h_j - h_i) \quad (27)$$

This formulation ensures that while the heat exchange is symmetric, the resulting temperature change is physically asymmetric, reflecting the varying sizes, air volumes, and thermal capacities of different building zones.

Laplacian Form

We can denote diffusion updates in a compact way with Laplacian form. Let $\boldsymbol{L}$ be the symmetric weighted graph Laplacian constructed from the heat transfer coefficients g_{ij}:

$$\boldsymbol{L}_{ii} = \sum_{j:(i,j)\in\varepsilon_u} g_{ij}\,, \qquad \boldsymbol{L}_{ij} = \begin{cases} -g_{ij}, & (i,j) \in \varepsilon_u \\ 0, & otherwise \end{cases} \quad (28)$$

Let $\boldsymbol{C}^{-1} = diag(c_1, c_2, \dots, c_Z)$ be the diagonal matrix of the learned inverse heat capacities for all Z zones. The full graph layer update becomes:

$$\boldsymbol{h}' = (\boldsymbol{I} - \boldsymbol{C}^{-1}\boldsymbol{L})\boldsymbol{h} \quad (29)$$

Because $\boldsymbol{C}^{-1}$ is applied directly to the nodes, the effective transformation matrix $\boldsymbol{C}^{-1}\boldsymbol{L}$ is asymmetric. At this point, we defined heat diffusion graph layer that obeys heat transfer and energy balance among the zone and its adjacencies.

8.5 Loss Function of PCGNN

Per-Zone MSE

$$\mathcal{L}_{mse} = \frac{1}{Z}\sum_{z=1}^{Z} MSE_z\,, \qquad MSE_z = \frac{1}{BTC_{out}}\sum_{b,t,c}\left(\hat{Y}_{b,t,z,c} - Y_{b,t,z,c}\right)^2 \quad (30)$$

Temporal Difference Loss

$$\mathcal{L}_{\Delta} = \frac{1}{B(T-1)ZC_{out}} \sum_{b,t=1}^{T-1} \sum_{z,c} \left[\left(\hat{Y}_{b,t+1,z,c} - \hat{Y}_{b,t,z,c} \right) - \left(Y_{b,t+1,z,c} - Y_{b,t,z,c} \right) \right]^2 \quad (31)$$

This term is to help PCGNN to learn the temperature change between time steps.

Monotonicity Regularization

For a kernel tensor $W \in \mathbb{R}^{D_{out} \times D_{in} \times K}$, we define tap magnitudes by Frobenius norm

$$s_i = \| \phi(W[:,:,i]) \|_F, \qquad i = 0, \ldots, K-1 \quad (32)$$

We can regularize the convolutional layer to be near-time-step focus by the loss:

$$\mathcal{L}_{mono} = \sum_{i=0}^{K-2} [\max(0, s_i - s_{i+1})]^2 \quad (33)$$

The rationale of $\mathcal{L}_{mono}$ is that the prediction of a certain step should be largely impacted by its near-term historical inputs rather than distant ones.

Gradient Monotonicity

When non-negative weights are not enforced for all weights, i.e., $\phi(W) = W$ in equation (15), we can guarantee a positive gradient of the model output $\hat{\boldsymbol{Y}}$ with respect to the input $\boldsymbol{X}$.

$$\mathcal{L}_{grad} = \sum \max(0, -G), \qquad G = \frac{\partial \bar{y}}{\partial \boldsymbol{X}}, \qquad \bar{y} = \frac{1}{BTZC_{out}} \sum_{b,t,z,c} \hat{Y}_{b,t,z,c} \quad (34)$$

Note that it is not necessary to have nonnegative gradients for all input features. For example, we can enforce nonnegative gradients only with respect to control variables.

Bias Penalty

To avoid the prediction heavily relying on bias with vanishing weights, we can also have optional bias penalty.

$$\mathcal{L}_{bias} = \sum_{all\ conv\ layers} \|b\|_1 \quad (35)$$

8.6 System Description of Building 1

The simulation study includes 6 zones: 1 corridor and 5 office rooms. The corridor is conditioned by a dedicated outdoor air system (DOAS) with variable air volume (VAV) boxes, while the offices are served by both radiant ceiling panel system and DOAS. Each office is equipped with an occupancy presence sensor. The occupancy status determines the operating mode of the radiant panel system: (1) unoccupied mode before 7 AM and after 7 PM, (2) occupied mode when occupancy is detected, and (3) standby mode when the room is unoccupied between 7 AM and 7 PM. The radiant system switches between heating and cooling based on the outdoor air temperature: heating mode when the outdoor temperature is below 12.8 °C and cooling mode when

it exceeds 13.3 °C. Occupants can also adjust the thermostat setpoint within a predefined range: 22.2-26.7 °C in cooling mode and 18.3-23.3 °C in heating mode.

8.7 Hyperparameters of PCGNN for Building 1

Table 9. Hyperparameters of PCGNN for building 1.

Hyperparameters	**Values**
Hidden size	32
Number of blocks B_{blk}	3
Grouped MS conv layers M (each block)	3
Long-branch kernel size	0
Short-branch kernel size	24
Enforce positivity for MS conv	False
Enforce positivity for heat diffusion layer	True
Conv dropout	0.1
Learning rate	1e-3
Batch size	512
Number of epochs	200
Grad clip	0.5
Coefficient λ_{Δ}	1.0
Coefficient λ_{mono}	0
Coefficient λ_{bias}	0
Coefficient λ_{grad}	1.5
Gradient monotonic features	Indoor temperature and radiant load
Early stop patience	20
Shrink factor of learning rate	0.5
Learning rate patience	10
Minimal learning rate	1e-5
Prediction horizon	6 hours

8.8 Detailed MARL Settings of Building 1

The reward function is defined as:

$$r(s_t, a_t, a_{t-1}) = r_1 s_t^{vio} + r_2|a_t - a_{t-1}| + r_3|a_t| + r_4 s_t^{good} \quad (36)$$

Where s_t^{vio} is temperature violation calculated by the deviation from lower and upper bounds; s_t^{good} is a binary indicator of whether the zone is within the comfort bounds (i.e., no temperature violation); and r_1 to r_4 are penalty and reward coefficients for different objectives.

Error! Reference source not found. lists the MARL hyperparameters for building 1. A 1-hop truncation is used for the policy network, resulting in a maximum of 138 input states for agents. A

2-hop truncation is used for the critic network, which in this case is equivalent to global state concatenation since the 6-zone network has a diameter of 2.

Table 10. Hyperparameters of MARL for building 1.

Hyperparameters	**Values**
κ-hop truncation for policy network	1
κ-hop truncation for critic network	2
Number of episodes	4e4
Batch size	2048
Hidden dimension	64
Discount factor γ	0.93
Learning rate	1e-4
Target network update rate	0.05
Number of layers of neural networks	10
Hidden dimension of neural networks	512
Penalty for temperature violation r_1	-2
Penalty for non-smoothness of action r_2	1e-3
Penalty for energy usage r_3	0.05
Reward for good IEQ r_4	0.04

8.9 System Description of Building 2

Building 2 is designed as a living laboratory, allowing the heating and cooling modes to be switched throughout the year. The operating mode is determined by the relationship between indoor temperature and setpoints: heating is provided when the temperature falls below the heating setpoint, and cooling when it exceeds the cooling setpoint. The setpoints are defined based on a daily schedule and room usage type. The building follows an occupied schedule from 6 AM to 6 PM. For frequently occupied rooms such as offices, the heating and cooling setpoints are 21.1 °C and 23.8 °C during occupied hours, relaxing to 17.8 °C and 27.2 °C otherwise. For less frequently occupied spaces such as corridors and lobbies, the daytime setpoints are 19.4 °C and 25.6 °C, and the nighttime setpoints are 17.8 °C and 27.2 °C.

8.10 Hyperparameters of PCGNN for Building 2

Table 11. Hyperparameters of PCGNN for building 2.

Hyperparameters	**Values**
Hidden size	16
Number of blocks B_{blk}	2
Grouped MS conv layers M (each block)	3
Long-branch kernel size	0

Short-branch kernel size	24
Enforce positivity for MS conv	False
Enforce positivity for heat diffusion layer	True
Conv dropout	0.1
Learning rate	1e-3
Batch size	512
Number of epochs	200
Grad clip	0.5
Coefficient λ_{Δ}	0
Coefficient λ_{mono}	0
Coefficient λ_{bias}	0
Coefficient λ_{grad}	2
Gradient monotonic features	Indoor temperature and radiant load
Early stop patience	20
Shrink factor of learning rate	0.5
Learning rate patience	10
Minimal learning rate	1e-5
Prediction horizon	6 hours

8.11 ϵ-NTU for Radiant Ceiling Panels

Following our previous work [60], we use ϵ-NTU method:

$$\epsilon = \frac{T_s - T_r}{T_s - T_z} = 1 - \exp\left(-\frac{A(G^B + C)}{G}\right) \quad (37)$$

$$Q = \rho \cdot c_p \cdot G \cdot (T_s - T_z)\left(1 - \exp\left(-\frac{A(G^B + C)}{G}\right)\right) \quad (38)$$

Where T represents temperature; the subscripts s, r and z indicate supply water, return water, and zone air. A, B, and C are tunable parameters indicating the heat transfer properties of the radiant ceiling panels; G is the water flow rate; Q is the radiant panel load; ρ and c_p are density and specific heat of water.

Using equation (38)**Error! Reference source not found.**, we can convert sampled action Q to the water flow rate G. The flow rate is then mapped to a valve position based on the valve characteristics, enabling direct implementation of the agent's action through the building automation system.

9 Reference

[1]. Langevin, J., Harris, C., Satre-Meloy, A., Putra, H. C., Speake, A., Present, E., Adhikari, R., Wilson, E., & Satchwell, A. (2021). US building energy efficiency and flexibility as an electric grid resource. Joule, 5(8), 2102–2128. https://doi.org/10.1016/j.joule.2021.06.002

[2]. Wang, X., Dong, B. & Zhang, J.J. Nationwide evaluation of energy and indoor air quality predictive control and impact on infection risk for cooling season. Build. Simul. 16, 205–223 (2023). https://doi.org/10.1007/s12273-022-0936-6
[3]. Deng, Z., Wang, X., & Dong, B. (2023). Quantum computing for future real-time building HVAC controls. Applied Energy, 334, 120621. https://doi.org/10.1016/j.apenergy.2022.120621
[4]. Song, Y., Romero, A., Müller, M., Koltun, V., & Scaramuzza, D. (2023). Reaching the limit in autonomous racing: Optimal control versus reinforcement learning. Science Robotics, 8(82). https://doi.org/10.1126/scirobotics.adg1462
[5]. Drgoňa, J., Picard, D., Kvasnica, M., & Helsen, L. (2018). Approximate model predictive building control via machine learning. Applied Energy, 218, 199–216. https://doi.org/10.1016/j.apenergy.2018.02.156
[6]. Drgoňa, J., Tuor, A., Skomski, E., Vasisht, S., & Vrabie, D. (2021). Deep learning Explicit differentiable predictive control laws for buildings. IFAC-PapersOnLine, 54(6), 14–19. https://doi.org/10.1016/j.ifacol.2021.08.518
[7]. Drgoňa, J., Kiš, K., Tuor, A., Vrabie, D., & Klaučo, M. (2022). Differentiable predictive control: Deep learning alternative to explicit model predictive control for unknown nonlinear systems. Journal of Process Control, 116, 80–92. https://doi.org/10.1016/j.jprocont.2022.06.001
[8]. Wang, X., & Dong, B. (2024). Long-term experimental evaluation and comparison of advanced controls for HVAC systems. Applied Energy, 371, 123706. https://doi.org/10.1016/j.apenergy.2024.123706
[9]. Anderson, C. W., Hittle, D. C., Katz, A. D., & Kretchmar, R. M. (1997). Synthesis of reinforcement learning, neural networks and PI control applied to a simulated heating coil. Artificial Intelligence in Engineering, 11(4), 421–429. https://doi.org/10.1016/s0954-1810(97)00004-6
[10]. Dalamagkidis, K., Κολοκότσα, Δ., Kalaitzakis, K., & Stavrakakis, G. (2007). Reinforcement learning for energy conservation and comfort in buildings. Building and Environment, 42(7), 2686–2698. https://doi.org/10.1016/j.buildenv.2006.07.010
[11]. Du, D., & Fei, M. (2008). A two-layer networked learning control system using actor–critic neural network. Applied Mathematics and Computation, 205(1), 26–36. https://doi.org/10.1016/j.amc.2008.05.062
[12]. Yu, Z., & Dexter, A. (2010). Online tuning of a supervisory fuzzy controller for low-energy building system using reinforcement learning. Control Engineering Practice, 18(5), 532–539. https://doi.org/10.1016/j.conengprac.2010.01.018
[13]. Ruelens, F., Iacovella, S., Claessens, B., & Belmans, R. (2015). Learning Agent for a Heat-Pump Thermostat with a Set-Back Strategy Using Model-Free Reinforcement Learning. Energies, 8(8), 8300–8318. https://doi.org/10.3390/en8088300
[14]. Barrett, E., & Linder, S. (2015). Autonomous HVAC control, a reinforcement learning approach. In Lecture Notes in Computer Science (pp. 3–19). https://doi.org/10.1007/978-3-319-23461-8_1
[15]. Yang, L., Nagy, Z., Goffin, P., & Schlueter, A. (2015). Reinforcement learning for optimal control of low exergy buildings. Applied Energy, 156, 577–586. https://doi.org/10.1016/j.apenergy.2015.07.050

[16]. Biemann, M., Scheller, F., Liu, X., & Huang, L. (2021). Experimental evaluation of model-free reinforcement learning algorithms for continuous HVAC control. Applied Energy, 298, 117164. https://doi.org/10.1016/j.apenergy.2021.117164
[17]. Han, X., & Malkawi, A. (2023). Model-Free reinforcement Learning-Based control for radiant floor heating systems. In Environmental science and engineering (pp. 1447–1455). https://doi.org/10.1007/978-981-19-9822-5_150
[18]. Gokhale, G., Claessens, B., & Develder, C. (2022, November 21). PhysQ: A Physics Informed Reinforcement learning Framework for building Control. arXiv.org. https://arxiv.org/abs/2211.11830
[19]. Park, J. Y., Dougherty, T., Fritz, H., & Nagy, Z. (2019). LightLearn: An adaptive and occupant centered controller for lighting based on reinforcement learning. Building and Environment, 147, 397–414. https://doi.org/10.1016/j.buildenv.2018.10.028
[20]. Zhang, Z., Chong, A., Pan, Y., Zhang, C., & Lam, K. P. (2019). Whole building energy model for HVAC optimal control: A practical framework based on deep reinforcement learning. Energy and Buildings, 199, 472–490. https://doi.org/10.1016/j.enbuild.2019.07.029
[21]. Chen, B., Cai, Z., & Bergés, M. (2020). GNU-RL: a practical and scalable reinforcement learning solution for building HVAC control using a differentiable MPC policy. Frontiers in Built Environment, 6. https://doi.org/10.3389/fbuil.2020.562239
[22]. Zou, Z., Yu, X., & Ergan, S. (2020). Towards optimal control of air handling units using deep reinforcement learning and recurrent neural network. Building and Environment, 168, 106535. https://doi.org/10.1016/j.buildenv.2019.106535
[23]. Lei, Y., Song, Z., Ono, E., Peng, Y., Zhang, Z., Hasama, T., & Chong, A. (2022). A practical deep reinforcement learning framework for multivariate occupant-centric control in buildings. Applied Energy, 324, 119742. https://doi.org/10.1016/j.apenergy.2022.119742
[24]. Wang, X., Kang, X., An, J., Chen, H., & Yan, D. (2023). Reinforcement learning approach for optimal control of ice-based thermal energy storage (TES) systems in commercial buildings. Energy and Buildings, 301, 113696. https://doi.org/10.1016/j.enbuild.2023.113696
[25]. Kofinas, P., Dounis, A., & Vouros, G. (2018). Fuzzy Q-Learning for multi-agent decentralized energy management in microgrids. Applied Energy, 219, 53–67. https://doi.org/10.1016/j.apenergy.2018.03.017
[26]. Prasad, A., & Dusparic, I. (2019). Multi-agent deep reinforcement learning for zero energy communities. In 2019 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe) (pp. 1–5). https://doi.org/10.1109/isgteurope.2019.8905628
[27]. Nagarathinam, S., Menon, V., Vasan, A., & Sivasubramaniam, A. (2020). MARCO - Multi-Agent Reinforcement Learning based COntrol of building HVAC systems. In the Eleventh ACM International Conference on Future Energy Systems (e-Energy '20) (pp. 57–67). https://doi.org/10.1145/3396851.3397694
[28]. Yu, L., Sun, Y., Xu, Z., Shen, C., Yue, D., Jiang, T., & Guan, X. (2020). Multi-Agent deep reinforcement learning for HVAC control in commercial buildings. IEEE Transactions on Smart Grid, 12(1), 407–419. https://doi.org/10.1109/tsg.2020.3011739
[29]. Zhang, B., Hu, W., Ghias, A. M., Xu, X., & Chen, Z. (2022). Multi-agent deep reinforcement learning-based coordination control for grid-aware multi-buildings. Applied Energy, 328, 120215. https://doi.org/10.1016/j.apenergy.2022.120215
[30]. Deng, X., Zhang, Y., Zhang, Y., Zhang, Y., Zhang, Y., & Qi, H. (2022). Towards optimal HVAC control in non-stationary building environments combining active change detection and

deep reinforcement learning. Building and Environment, 211, 108680. https://doi.org/10.1016/j.buildenv.2021.108680
[31]. Homod, R. Z., Togun, H., Hussein, A. K., Al-Mousawi, F. N., Yaseen, Z. M., Al-Kouz, W., Abd, H. J., Alawi, O. A., Goodarzi, M., & Hussein, O. A. (2022). Dynamics analysis of a novel hybrid deep clustering for unsupervised learning by reinforcement of multi-agent to energy saving in intelligent buildings. Applied Energy, 313, 118863. https://doi.org/10.1016/j.apenergy.2022.118863
[32]. Fu, Q., Chen, X., Ma, S., Fang, N., Xing, B., & Chen, J. (2022). Optimal control method of HVAC based on multi-agent deep reinforcement learning. Energy and Buildings, 270, 112284. https://doi.org/10.1016/j.enbuild.2022.112284
[33]. Blad, C., Bøgh, S., & Kallesøe, C. S. (2022). Data-driven offline reinforcement learning for HVAC-systems. Energy, 261, 125290. https://doi.org/10.1016/j.energy.2022.125290
[34]. Shen, R., Zhong, S., Wen, X., An, Q., Zheng, R., Li, Y., & Zhao, J. (2022). Multi-agent deep reinforcement learning optimization framework for building energy system with renewable energy. Applied Energy, 312, 118724. https://doi.org/10.1016/j.apenergy.2022.118724
[35]. Yu, L., Xu, Z., Zhang, T., Guan, X., & Yue, D. (2022). Energy-efficient personalized thermal comfort control in office buildings based on multi-agent deep reinforcement learning. Building and Environment, 223, 109458. https://doi.org/10.1016/j.buildenv.2022.109458
[36]. Nweye, K., Liu, B., Stone, P., & Nagy, Z. (2022). Real-world challenges for multi-agent reinforcement learning in grid-interactive buildings. Energy and AI, 10, 100202. https://doi.org/10.1016/j.egyai.2022.100202
[37]. Pigott, A., Crozier, C., Baker, K., & Nagy, Z. (2022). GridLearn: Multiagent reinforcement learning for grid-aware building energy management. Electric Power Systems Research, 213, 108521. https://doi.org/10.1016/j.epsr.2022.108521
[38]. Gao, Y., Matsunami, Y., Miyata, S., & Akashi, Y. (2022). Multi-agent reinforcement learning dealing with hybrid action spaces: A case study for off-grid oriented renewable building energy system. Applied Energy, 326, 120021. https://doi.org/10.1016/j.apenergy.2022.120021
[39]. Shen, R., Zhong, S., Zheng, R., Yang, D., Xu, B., Li, Y., & Zhao, J. (2023). Advanced control framework of regenerative electric heating with renewable energy based on multi-agent cooperation. Energy and Buildings, 281, 112779. https://doi.org/10.1016/j.enbuild.2023.112779
[40]. Nweye, K., Sankaranarayanan, S., & Nagy, Z. (2023). MERLIN: Multi-agent offline and transfer learning for occupant-centric operation of grid-interactive communities. Applied Energy, 346, 121323. https://doi.org/10.1016/j.apenergy.2023.121323
[41]. Blad, C., Bøgh, S., Kallesøe, C., & Raftery, P. (2023). A laboratory test of an Offline-trained Multi-Agent Reinforcement Learning Algorithm for Heating Systems. Applied Energy, 337, 120807. https://doi.org/10.1016/j.apenergy.2023.120807
[42]. Homod, R. Z., Mohammed, H. I., Abderrahmane, A., Alawi, O. A., Khalaf, O. I., Mahdi, J. M., Guedri, K., Dhaidan, N. S., Albahri, A., Sadeq, A. M., & Yaseen, Z. M. (2023). Deep clustering of Lagrangian trajectory for multi-task learning to energy saving in intelligent buildings using cooperative multi-agent. Applied Energy, 351, 121843. https://doi.org/10.1016/j.apenergy.2023.121843

[43]. Xie, J., Ajagekar, A., & You, F. (2023). Multi-Agent attention-based deep reinforcement learning for demand response in grid-responsive buildings. Applied Energy, 342, 121162. https://doi.org/10.1016/j.apenergy.2023.121162
[44]. GS, A. K., Zhang, T., Ardakanian, O., & Taylor, M. E. (2023). Mitigating an adoption barrier of reinforcement learning-based control strategies in buildings. Energy and Buildings, 285, 112878. https://doi.org/10.1016/j.enbuild.2023.112878
[45]. Yang, D., Wang, X., Shen, R., Li, Y., Gu, L., Zheng, R., Zhao, J., & Tian, X. (2024). Global optimization strategy of prosumer data center system operation based on multi-agent deep reinforcement learning. Journal of Building Engineering, 91, 109519. https://doi.org/10.1016/j.jobe.2024.109519
[46]. Liu, S., Liu, X., Zhang, T., Wang, C., & Liu, W. (2024). Joint optimization for temperature and humidity independent control system based on multi-agent reinforcement learning with cooperative mechanisms. Applied Energy, 375, 123968. https://doi.org/10.1016/j.apenergy.2024.123968
[47]. Wang, Z., Xiao, F., Ran, Y., Li, Y., & Xu, Y. (2024). Scalable energy management approach of residential hybrid energy system using multi-agent deep reinforcement learning. Applied Energy, 367, 123414. https://doi.org/10.1016/j.apenergy.2024.123414
[48]. Wu, H., Qiu, D., Zhang, L., & Sun, M. (2024). Adaptive multi-agent reinforcement learning for flexible resource management in a virtual power plant with dynamic participating multi-energy buildings. Applied Energy, 374, 123998. https://doi.org/10.1016/j.apenergy.2024.123998
[49]. Chen, Z., Xing, T., Wang, Y., Zhuang, Y., Zheng, M., Zhao, Q., & Jia, Q. (2025). Coupling time-scale reinforcement learning methods for building operational optimization with waste heat. Applied Energy, 391, 125851. https://doi.org/10.1016/j.apenergy.2025.125851
[50]. Xue, W., Jia, N., & Zhao, M. (2025). Multi-agent deep reinforcement learning based HVAC control for multi-zone buildings considering zone-energy-allocation optimization. Energy and Buildings, 329, 115241. https://doi.org/10.1016/j.enbuild.2024.115241
[51]. Liu, J., Dou, W., Meng, X., Wu, J., & Ma, Z. (2025). Multi-agent deep reinforcement learning-based hierarchical energy management for better indoor air quality and energy-savings in building energy systems. Energy Conversion and Management, 342, 120103. https://doi.org/10.1016/j.enconman.2025.120103
[52]. Tariq, S., Ali, U., Kim, S., & Yoo, C. (2025). Multi-agent distributed reinforcement learning for energy-efficient thermal comfort control in multi-zone buildings with diverse occupancy patterns. Energy, 332, 137082. https://doi.org/10.1016/j.energy.2025.137082
[53]. Wu, Y., Cong, M., Lu, Q., Zhou, Z., Miao, Y., Liu, J., & Yang, D. (2025). Hierarchical control strategy for heating systems using multi-agent deep reinforcement learning. Journal of Building Engineering, 107, 112699. https://doi.org/10.1016/j.jobe.2025.112699
[54]. Zhang, Y., Zhao, Y., Zhang, C., & Feng, C. (2025). Multi-agent reinforcement learning-based method for demand response of building HVAC systems. Journal of Building Engineering, 108, 112734. https://doi.org/10.1016/j.jobe.2025.112734
[55]. Liu, Y., Song, Y., & Cui, C. (2024). Towards smart control and energy efficiency for multi-zone ventilation systems via an imitation-interaction learning method in energy-aware buildings. Energy, 314, 134220. https://doi.org/10.1016/j.energy.2024.134220
[56]. Savino, S., Minella, T., Nagy, Z., & Capozzoli, A. (2025). A scalable demand-side energy management control strategy for large residential districts based on an attention-driven multi-

agent DRL approach. Applied Energy, 393, 125993. https://doi.org/10.1016/j.apenergy.2025.125993

[57]. Homod, R. Z., Mohammed, H. I., Sadeq, A. M., Alhasnawi, B. N., Al-Fatlawi, A. W., Al-Manea, A., Alawi, O. A., Alahmer, A., Mahdi, J. M., Al-Kouz, W., & Yaseen, Z. M. (2024). Massive energy reduction and storage capacity relative to PCM physical size by integrating deep RL clustering and multi-stage strategies into smart buildings to grid reliability. Journal of Energy Storage, 109, 115058. https://doi.org/10.1016/j.est.2024.115058

[58]. Liao, C., Miyata, S., Qu, M., & Akashi, Y. (2025). Year-round operational optimization of HVAC systems using hierarchical deep reinforcement learning for enhancing indoor air quality and reducing energy consumption. Applied Energy, 390, 125816. https://doi.org/10.1016/j.apenergy.2025.125816

[59]. Jiang, Z., Wang, X., & Dong, B. (2025). Physics-informed modularized neural network for advanced building control by deep reinforcement learning. Advances in Applied Energy, 19, 100237. https://doi.org/10.1016/j.adapen.2025.100237

[60]. Wang, X., Wang, X., Kang, X., Dong, B., & Yan, D. (2024). Physics-consistent input convex neural network-driven reinforcement learning control for multi-zone radiant ceiling heating and cooling systems: An experimental study. Energy and Buildings, 327, 115105. https://doi.org/10.1016/j.enbuild.2024.115105

[61]. Liang, W., Li, H., Zhan, S., Chong, A., & Hong, T. (2024). Energy flexibility quantification of a tropical net-zero office building using physically consistent neural network-based model predictive control. Advances in Applied Energy, 14, 100167. https://doi.org/10.1016/j.adapen.2024.100167

[62]. Jiang, Z., & Dong, B. (2024). Modularized neural network incorporating physical priors for future building energy modeling. Patterns, 5(8), 101029. https://doi.org/10.1016/j.patter.2024.101029

[63]. Di Natale, L., Svetozarevic, B., Heer, P., & Jones, C. (2023). Towards scalable physically consistent neural networks: An application to data-driven multi-zone thermal building models. Applied Energy, 340, 121071. https://doi.org/10.1016/j.apenergy.2023.121071

[64]. Jiang, Z., Wang, X., Li, H., Hong, T., You, F., Drgoňa, J., Vrabie, D., & Dong, B. (2025). Physics-informed machine learning for building performance simulation-A review of a nascent field. Advances in Applied Energy, 18, 100223. https://doi.org/10.1016/j.adapen.2025.100223

[65]. Wang, X., & Dong, B. (2023). Physics-informed hierarchical data-driven predictive control for building HVAC systems to achieve energy and health nexus. Energy and Buildings, 291, 113088. https://doi.org/10.1016/j.enbuild.2023.113088

[66]. Ren, Z., Zhang, R., Dai, B., & Li, N. (2024, October 22). Scalable spectral representations for multi-agent reinforcement learning in network MDPs. arXiv.org. https://arxiv.org/abs/2410.17221

[67]. Qu, G., Wierman, A., & Li, N. (2019, December 5). Scalable reinforcement learning for Multi-Agent networked systems. arXiv.org. https://arxiv.org/abs/1912.02906

[68]. Qu, G., Lin, Y., Wierman, A., & Li, N. (2020, June 11). Scalable Multi-Agent Reinforcement Learning for Networked Systems with Average Reward. arXiv.org. https://arxiv.org/abs/2006.06626

[69]. Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. In KDD.