Environmental Science Advances

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Environ. Sci.: Adv.*, 2026, **5**, 1292

From one building to many: transferability of a deep reinforcement learning agent for optimizing pollutant exposure and energy consumption

Nishchaya Kumar Mishra [a] and Sameer Patel *[abc]

Minimizing indoor pollutant exposure while conserving energy is essential for protecting human health and the environment. Deep reinforcement learning (DRL) has emerged as a promising approach for optimizing residential ventilation and air conditioning systems. While DRL deployment is simpler than fully physics-driven strategies like dynamic optimization (DynOpt), its generalizability across diverse buildings and ambient conditions remains challenging. Although researchers have studied transfer and imitation learning techniques to address these challenges, they still require house characteristics and field measurements to adaptively train an agent. Therefore, the large-scale deployment of DRL agents can still be potentially challenging. This study assesses the performance of a trained DRL agent against the DynOpt (benchmark) when transferred to houses with varying characteristics and environmental conditions using digital twins. When varying house characteristics one at a time, the agent's performance remained comparable to DynOpt, with particulate matter (PM) exposure and energy ratios near unity (1.05 ± 0.03). Similarly, under simultaneous variations in house characteristics, the exposure (1.03 ± 0.07) and energy (1.09 ± 0.06) ratios remained close to one. However, the agent's performance declines in houses with high PM infiltration under high ambient parameters. The results indicate that the agent can still be integrated into different houses under varying ambient conditions by restricting the infiltration of PM, as evident by lower exposure and energy ratios in houses with lower infiltration. Moving forward, uncertainty quantification and benchmarking of the agent's performance are critical for enhancing confidence in predictions.

Received 27th November 2025
Accepted 13th March 2026

DOI: 10.1039/d5va00438a

rsc.li/esadvances

Environmental significance

Indoor environments govern occupants' health, comfort, and disease transmission, while also consuming a significant fraction of global energy, thereby necessitating optimization. Simultaneously, ensuring a healthy and comfortable indoor environment for the masses requires a robust, low-cost solution that can be deployed at scale. Therefore, it is critical to understand the capability of data-driven learning algorithms, such as reinforcement learning (RL), that could be a potential solution. This study aims to understand the scalability challenges associated with the adoption of a deep RL agent under varying household characteristics, different indoor pollutant emission scenarios, and diverse ambient weather conditions.

1. Introduction

The indoor environment of buildings significantly affects occupants' health and well-being while also impacting the regional and global environment through energy consumption and pollutant emissions.[1–5] The health impacts arise from various indoor pollutants originating from emission sources such as the building itself (*e.g.*, wood, plastics, paints), consumer products (personal care, cleaning, cooking, equipment, and office products), and human activities (metabolic, microbial, and occupant activities).[6–10] The interaction of the indoor and outdoor environment also affects occupants' health as pollutants of outdoor origin infiltrate the indoor climate, and recent studies have demonstrated that exposure to infiltrated pollutants is equal to or greater than that of pollutants of indoor origin.[11,12] Another critical challenge is the energy consumption of buildings, which amounts to more than one-third of global energy and has increased at an average of 1% annually over the past decade.[1] The heating, ventilation, and air conditioning (HVAC) system accounts for ~40% of the building's energy consumption and ~12% of the total energy consumption.[13]

Studies have demonstrated that reducing pollutant exposure inside buildings is associated with increased energy consumption of HVAC systems to ensure thermal comfort.[14,15] Therefore, a complex, interdependent relationship exists between thermal

[a]*Department of Civil Engineering, India. E-mail: sameer.patel@iitgn.ac.in*
[b]*Department of Chemical Engineering, India*
[c]*Kiran C. Patel Centre of Sustainable Development, Indian Institute of Technology Gandhinagar, Palaj, Gandhinagar, Gujarat 382355, India*

comfort and pollutant exposure in indoor environments. Multiple studies have proposed physics-based optimization and artificial intelligence algorithms to balance the trade-off and optimize the operation of the HVAC systems to reduce exposure and energy consumption while ensuring thermal comfort.[14,16–22] For example, Mishra *et al.*[23] designed a deep reinforcement learning (DRL) agent for optimizing particulate matter (PM) exposure, energy consumption, and thermal comfort in a house. The same study compared the performance of the DRL agent with a dynamic optimization strategy and demonstrated that the DRL agent performed on par with it. Similar studies have shown the advantages of such agents over rule-based and traditional physics-driven algorithms in controlling indoor environments.[16,24,25] Moreover, reinforcement learning (RL) agents have the potential for wide-scale deployment owing to many advantages over their conventional counterparts, such as learning an optimal decision-making policy directly through environmental interaction, requiring no knowledge of the system's physics and building characteristics.[26–28] However, the dissemination of these agents at the community scale is limited owing to multiple challenges, such as transferability across buildings, non-intuitive performance, performance mismatch between simulation and real building, and datasets needed for training.[29–32]

RL agents are often trained in virtual indoor environments/digital twins of buildings.[27,29,33,34] Since the agent's training and testing are conducted offline using a digital twin, the performance of the trained agent in an actual building is susceptible to uncertainties and variations when deployed in the field.[35,36] On the contrary, online training (in real buildings) results in longer training times and discomfort for occupants during initial training phases when the agent is still learning.[37] Researchers have proposed various methods for HVAC control to overcome these challenges, where transfer learning,[31,38–41] imitation learning,[34,42,43] and multi-agent reinforcement learning[44,45] are some of the recently studied alternatives. Chen *et al.*[41] utilized transfer learning to predict indoor air temperature and relative humidity over a time horizon ranging from 10 minutes to 2 hours in a building. The same study demonstrated that the transferred model achieves higher accuracy in predicting indoor air temperature and RH with a mean square error lower than that of the benchmark model trained only on source or target data. Similarly, Deng *et al.*[40] transferred the behavior knowledge of an RL agent in different office buildings to control the set temperature and clothing level. The transferred model predicted occupant behavior with a high correlation (>0.8) and a mean square error of less than 1.1 °C. Further, Liu *et al.*[43] developed an imitation–interaction learning control method for multi-zone ventilation systems that accelerated RL training towards higher control performance and energy efficiency. Dey *et al.*[34] also proposed an RL-based building control method harnessing imitation learning, which reduced the

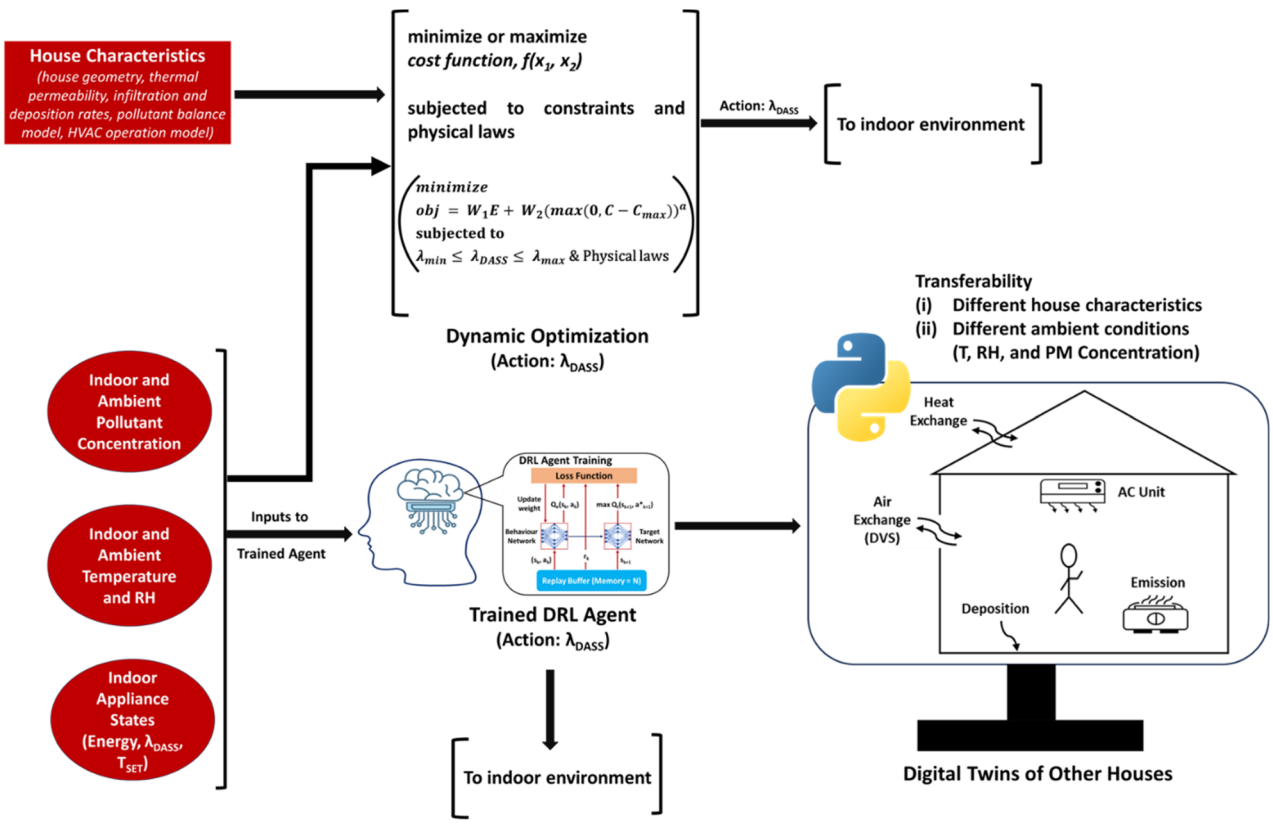

Fig. 1 The study framework outlines the input parameters needed for a dynamic optimization strategy and a deep reinforcement learning (DRL) agent, and evaluates the transferability of the DRL agent across different houses under varying ambient conditions. T: temperature, RH: relative humidity, PM: particulate matter, λ_{DASS}: indoor–outdoor air exchange rate, DVS: dedicated ventilation system.

training time while preventing unstable early exploration behavior and improving an accepted rule-based policy. However, techniques such as transfer learning and imitation learning still pose many challenges associated with their adoption. For imitation learning, the existence of an expert is crucial, as the agent learns the optimal policy by observing the expert's decisions. Therefore, learning an optimal policy *via* imitation learning is challenging when expert demonstrations are limited or the system is highly dynamic and complex. Similarly, in transfer learning, an optimal policy learned in a building is transferred to another building after retraining on a smaller dataset, which further requires data collection, monitoring, and an understanding of the target building's characteristics. These challenges restrict the wide-scale deployability of RL agents.

For the transferability of an RL agent, there are two key aspects to account for: (i) changes in household characteristics such as inner surface-to-volume ratio, thermal permeability, and pollutant penetration rate, and (ii) variations in climatic conditions and ambient pollutant concentration, since the performance of an RL agent could vary under these conditions. Therefore, for wide-scale deployment, it is imperative to evaluate the performance of these agents under varying climatic conditions across buildings with differing characteristics. Fig. 1 outlines the input and output parameters of a DRL agent and physics-based dynamic optimization strategy. Physical modeling of the house and HVAC systems is needed for dynamic optimization, in addition to sensor inputs (temperature, RH, and pollutant concentration). However, the trained DRL agent does not require house characteristics and physical models as inputs, and observations from low-cost monitors can be fed directly to the agent to obtain control actions. Based on this knowledge, the current study trains a DRL agent to optimize $PM_{2.5}$ (particles with a diameter of 2.5 microns or less; hereafter referred to as PM) exposure, thermal comfort, and energy consumption for a single house (the training house), which is then transferred to different houses (test houses). The agent transfer has been done under two conditions: (i) transferred to test houses (emulated through changing house characteristics) in the same neighborhood (same ambient conditions), and (ii) transferred to testing houses in different geographical locations (emulated by changing ambient conditions). Simulations have been performed to evaluate the performance of the transferred DRL agent compared to that of a dynamic optimization strategy (DynOpt) inside test houses under the defined conditions. Subsequently, alternatives are proposed to address the challenges associated with the transferability of the DRL agent.

Succinctly, this work contrasts with prior studies that validate RL-based controllers within a fixed building configuration or climatic setting; it rigorously examines the cross-building and cross-climate transferability of a DRL agent trained on a single house. Rather than limiting evaluation to isolated parametric perturbations, the current work systematically analyzes multidimensional variations in house characteristics and shifts in ambient conditions to identify robustness boundaries relative to a physics-based dynamic optimization benchmark. The study further quantifies conjugate interaction effects that emerge under extreme configurations, an aspect that is underexplored in the existing literature. By doing so, this work advances DRL-based indoor environmental control from case-specific demonstrations toward scalable, generalizable real-world deployment.

2. Data and methods

2.1. Framework of the study

The DRL agent is trained on a digital twin of a house, which is validated against field-measurement data from a test house. The details of the experiments and the test house characteristics are presented in previous studies[8,46,47] and are briefly discussed in Section S1 of the SI. The DRL agent takes indoor and outdoor parameters, including temperature, relative humidity (RH), PM concentration, and HVAC energy consumption as input, and outputs the indoor–outdoor air exchange rate (AER) in real time. The inputs to the agent could be obtained through low-cost monitors. Fig. 1 demonstrates the general framework for constructing the digital twin, and training and testing of the DRL agent. The following sections discuss the various elements required for the development of the DRL agent, such as house characteristics, PM emission profiles, dataset availability, and performance evaluation of the agent under different operating conditions.

2.2. Creation of digital twins and training data

The monitored and derived parameters from a field study have been fed into the aerosol dynamics and energy balance models to create a digital twin of a house, which has been utilized for offline training of the DRL agent. The creation of the digital twin is discussed in detail in Section S2 of the SI and is similar to the previous study by Mishra *et al.*[23] Briefly, the indoor PM concentration measured during the field investigation[47] has been used to develop a pollutant balance model, and an energy balance is incorporated to model HVAC operation and imitate the indoor thermal environment of the house. During the field campaign, the size-resolved PM distribution showed that sub-500 nm particles accounted for the majority of $PM_{2.5}$;[8,14,46] therefore, although the data used in this study were obtained from a scanning mobility particle sizer, the defined PM may be considered as an estimate of $PM_{2.5}$.

The measurements from this digital twin are fed to the DRL agent for training. The agent takes simulated indoor parameters (temperature, RH, PM concentration, and energy consumption), measured outdoor parameters (temperature and RH), and hourly ambient PM concentration, obtained from[48] for the test house location, downscaled to 1-minute resolution as inputs to the DRL agent to predict the indoor–outdoor AER. Multiple parametric combinations have been utilized to vary the house characteristics and evaluate the performance of the trained DRL agent when transferred to different houses.

The variations in the volume, PM deposition rate, and thermal permeability of the tested houses are in the range of $\pm 60\%$ of the training house. The PM penetration factor controls

the infiltration rate of ambient PM into buildings and ranges from 0 to 1, where 0 represents 100% ambient air filtration, and 1 means no filtration. The penetration is governed by multiple factors, including house construction, cracks, gaps, openings, and transport through the ventilation system into the building envelope. Naturally, all houses allow penetration of a certain fraction of ambient PM. However, transport through the ventilation systems can be controlled by installing an air filter in the dedicated air supply system (DASS), which controls the indoor–outdoor AER. In this work, the penetration factor has been varied between 0.1 and 0.9, representing different PM infiltration scenarios.

The details of the training house and transferred house characteristics are shown in Table 1. Five cases (Case ID C1 to Case ID C5) have been defined where the impacts of different house characteristics on the DRL agent's performance have been assessed by varying one characteristic at a time. In other words, Case IDs C1–C5 were constructed as controlled univariate sensitivity analyses under identical ambient conditions. For each case, one key house characteristic (*e.g.*, volume, PM deposition rate, thermal permeability of the building envelope, and PM penetration factor) was varied from its minimum to maximum bound while keeping all other parameters fixed at baseline values. The selection of the minimum and maximum bounds was performed heuristically. These heuristic bounds were used to test robustness across incremental variability rather than to identify extreme cases.

For Case ID C6, 16 parametric combinations have been simulated, with the minimum and maximum variations for each house characteristic adopted, and each combination assigned a House ID (C6_1 to C6_16; see Table 1). These configurations were created to assess the conjugate (interaction) effects among parameters rather than from individual thresholds. It involved two methodological approaches: stress-testing the DRL agent at the edges of different house characteristics, and analyzing cross-factor interactions that might not be observed under univariate perturbation. Instead of sampling at intermediate levels of house characteristics, this case investigates the extreme envelope of joint house characteristics, where generalization limits can be tested. It is acknowledged that these different combinations may not describe all existing house characteristics in a community. Nevertheless, they provide critical insights into assessing the transferability of the DRL agent for indoor PM control and energy optimization.

2.3. Deep reinforcement learning (DRL) agent

The DRL agent trained in this work is a deep Q-network (DQN) that controls the indoor–outdoor AER while interacting with the environment and is identical to the agent proposed by Mishra *et al.*[23] The agent aims to optimize the PM exposure and energy consumption in an indoor environment while ensuring the thermal comfort of occupants. It learns to make optimal decisions through an iterative interaction with the environment

Table 1 Details of the house characteristics for training and testing of the deep reinforcement learning agent[a]

Purpose	Case ID	Volume factor ($V_{testing}/V_{training}$) ($V_{training}$ = 250 m^3)	PM deposition factor ($\lambda_{testing}/\lambda_{training}$) ($\lambda_{training}$ = 1.6 h^{-1})	Thermal permeability factor ($\alpha_{testing}/\alpha_{training}$) ($\alpha_{training}$ = 0.068 kJ s^{-1} C^{-1})	PM penetration factor (p_{DASS})
Training	C1	1	1	1	0.5
Testing	C2	[0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6]	1	1	0.5
	C3	1	[0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6]	1	0.5
	C4	1	1	[0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6]	0.5
	C5	1	1	1	[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]
	C6_1	0.4	0.4	0.4	0.1
	C6_2	0.4	0.4	0.4	0.9
	C6_3	0.4	0.4	1.6	0.1
	C6_4	0.4	0.4	1.6	0.9
	C6_5	0.4	1.6	0.4	0.1
	C6_6	0.4	1.6	0.4	0.9
	C6_7	0.4	1.6	1.6	0.1
	C6_8	0.4	1.6	1.6	0.9
	C6_9	1.6	0.4	0.4	0.1
	C6_10	1.6	0.4	0.4	0.9
	C6_11	1.6	0.4	1.6	0.1
	C6_12	1.6	0.4	1.6	0.9
	C6_13	1.6	1.6	0.4	0.1
	C6_14	1.6	1.6	0.4	0.9
	C6_15	1.6	1.6	1.6	0.1
	C6_16	1.6	1.6	1.6	0.9

[a] V: volume of a house, λ: PM deposition rate, α: thermal permeability of a house, p_{DASS}: PM penetration factor of dedicated air supply system (DASS). 'Testing' and 'training' refer to the houses where the agent is being tested and trained.

during offline training. The DRL agent consists of three components: the state (s_t), the action (a_t), and the reward (r_t), and it observes the state s_t in real time, takes action a_t based on a policy determined using a value function that changes the state to s_{t+1}, and in return receives a reward r_t (defined in eqn (1)).

$$r_t = -W_1 E - W_2(\max(0,\, C - C_{\max}))^a \quad (1)$$

The first term in eqn (1) corresponds to energy consumption, and the second is a proxy for exposure. W_1 and W_2 are the user-assigned weightage to energy (E) and exposure terms. $C_{\max}$ is the threshold PM concentration whose exceedances are penalized, and a represents the order of penalty when the indoor PM concentration (C) exceeds $C_{\max}$. For all simulations and evaluation purposes, W_1 is 1, W_2 is 10, a is 2, and $C_{\max}$ is 10 $\mu g\ m^{-3}$. A detailed discussion on the selection of W_1, W_2, and a is presented in Mishra *et al.*,[14] and discussed briefly in Section S4 of the SI.

The decision-making ability of the agent to take action is termed policy, denoted by $\pi_t(s_t|a_t)$, and the rewards of an action at a given state are determined using the values function (Q_π (s, a)), as shown in eqn (2).[49]

$$Q_\pi(s,a) = E\left[r + \gamma \max_{a'} Q_\pi(s,a)\right],\ \forall s \varepsilon S,\ \forall a \varepsilon A \quad (2)$$

where $E[\,]$ is the expectation of the expression inside the bracket under a given policy $\pi_t(s_t|a_t)$, S is the possible set of state space, and A is the feasible set of actions. The reward received by the agent depends on the evaluation of the action. After accounting for future rewards through a discount factor γ, the policy is updated to maximize the total reward, R_t, as expressed in eqn (3).

$$R_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots = \sum_{k=t}^{\infty} \gamma^{k-t} r_k \quad (3)$$

During training, the agent comprises of two fully connected neural networks-(a) behavior network, with weights w_b, and (b) target network, with weights w_t. The behavior network makes the decision and communicates with the environment, and the target network is used to update the behavior network.[49–51] A replay buffer is defined to store the agent's experience while interacting with the environment, which allows self-learning through past experiences.[49,51] A random batch sampling performed over the experiences stored in the replay buffer ensures learning from past mistakes, avoids overfitting, and eliminates the correlation between the input data in each batch.[52,53] The state, s_t, stores various indoor and outdoor parameters and serves as input to the behavior network at each time step (as shown in Eq. (4)).

$$s_t = [C_{out}(t),\ T_{out}(t),\ RH_{out}(t),\ C_{in}(t),\ T_{in}(t),\ RH_{in}(t),\ \lambda_{DASS}(t),\ E(t)] \quad (4)$$

where C is $PM_{2.5}$ concentration, T is air temperature, RH is relative humidity, λ_{DASS} is air changes per hour (ACH), E is the total energy consumption of HVAC and DASS combined, and the subscripts out and in mean outdoor and indoor, respectively.

The agent takes action, a_t, through the behavior network ($Q_b(s_t, a_t^*)$) and the epsilon-greedy strategy. The epsilon-greedy policy refers to taking the action corresponding to the maximum value of $Q_b(s_t, a_t^*)$ with a probability of epsilon (ε) and at random with a probability of $1 - \varepsilon$. The action space, A, consists of the indoor–outdoor AER (λ_{DASS}), a step function ranging from 0.5 ACH to 10 ACH at a step size of 0.1 ACH, meaning a total of 96 actions are possible. The state, s_t, the action, a_t, the reward, r_t, and the next state, s_{t+1}, are stored in the replay buffer, and a batch of samples (s_t, a_t, r_t, s_{t+1}) is randomly selected to update the behavior network. The behavior network takes (s_k, a_k) as input and outputs $Q_b(s_k, a_k)$, while the target network takes (r_k, a_{k+1}) as input to give an output of maximum $Q_t(s_{k+1}, a_{k+1}^*)$. The behavior network is then updated based on the loss function, L_k, estimated as shown in eqn (5).[49] The parameters of the target network are updated every m timesteps by replacing them with the behavior network. This update frequency, m, is a trainable parameter whose selection has been discussed later.

$$L_k = \left(r_k + \gamma_{\max} Q_t\left(s_{k+1}, a_{k+1}^*\right) - Q_b(s_k,\ a_k)\right)^2 \quad (5)$$

The agent aims to learn the policy that maximizes the total reward, where the reward function consists of energy and exposure terms, and the indoor set temperature is kept at 25 °C at all times to ensure the thermal comfort of the occupants.

2.4. Physics-based dynamic optimization strategy (DynOpt)

A fully physics-driven dynamic optimization (DynOpt) strategy has been used as a benchmark to evaluate the performance of the DRL agent when transferred to houses with characteristics different from those of the digital twin used for training. The DynOpt relies on the knowledge of the physics of the house, defined using aerosol dynamics, house characteristics, and HVAC modeling. On the other hand, the DRL agent is a data-driven model trained on observed parameters within a digital twin of the house and does not require any household characteristics or knowledge of aerosol physics or HVAC models. Hence, the trained DRL agent faces limited challenges in terms of deployability. However, benchmarking the performance of the DRL agent is needed to increase confidence in its adoption and transfer across different houses. Therefore, a comparative performance assessment of the trained DRL agent with DynOpt has been conducted.

The DynOpt strategy has been formulated using a cost function defined in eqn (6), with the indoor–outdoor AER and physics-based knowledge of the indoor environment dynamics as constraints. The cost function in eqn (6) is identical to the reward (shown in eqn (1)) and is a weighted combination of two terms: the first term corresponds to energy consumption, and the second is a measure of pollutant exposure.

Minimize

$$\text{obj} = W_1E + W_2(\max(0, C - C_{\max}))^a \text{ subjected to } \lambda_{\min} \leq \lambda_{\text{DASS}} \leq \lambda_{\max} \quad (6)$$

where $\lambda_{\min}$ (0.5 h^{-1}) and $\lambda_{\max}$ (10 h^{-1}) are the lower and upper bounds of the indoor–outdoor AER. Various terms of the cost function have been discussed previously in Section 2.3 while defining the reward function. The above-formulated optimization problem aims to achieve optimal operation of DASS that concurrently reduces exposure and energy consumption while ensuring thermal comfort inside the house. A detailed discussion of DynOpt, including different parameters of the cost function, is provided by Mishra *et al.*[14]

3. Results and discussion

3.1 Agent training and hyperparameter tuning

The DRL agent is trained using a digital twin of the house, and the trained agent is similar to that proposed by Mishra *et al.*[23] The augmented dataset, synthesized from monitored and derived parameters from six experimental days of the field study, has been used for training. The agent training is performed over 1500 episodes, with a grid search over all possible hyperparameter combinations (Table 2), including learning rate, batch size, and the number of hidden layers and nodes, to obtain the optimal model parameters. Based on the performance of the DRL agent, the hyperparameters with the highest total reward across all combinations and episodes have been adopted for all subsequent simulations in this study.

The state space (s_t) serves as the input layer, and the action space (a_t) serves as the output layer, with a learning rate of 0.001, batch size of 512, and target network update frequency of 200 timesteps. The selected network architecture of the DRL agent is a fully connected newyrok with dimensions $s_t \times 128 \times 256 \times 128 \times a_t$. The replay memory is set to store 20 000 to allow the agent to learn from past experiences.

The subsequent section first demonstrates the agent's performance in houses with different characteristics (volume, deposition rate, penetration factor, and thermal permeability as outlined in Table 1) for moderate variations in ambient temperature (25 °C to 33 °C), RH (40% to 73%), and PM (<20 μg m^{-3}) similar to that in the training dataset. Then, the performance of the agent is assessed across varying house characteristics with higher variations in ambient temperature (28 °C to 44 °C), RH (40% to 80%), and PM (up to 110 μg m^{-3}). The indoor PM emission periods for three days with low and high emission activities have been adopted and kept the same for all houses because the emission rate for the same type of activity is independent of house characteristics.

3.2 Variability in agent's performance with house characteristics

The trained agent's performance, quantified in terms of exposure and energy consumption, is compared to DynOpt. Under all emission and control scenarios, the DASS maintains a minimum indoor–outdoor AER of 0.5 h^{-1}, and the air conditioning unit operates to maintain an indoor temperature of 25 °C. Moreover, the same ambient conditions were used for simulations across all houses, with outdoor temperatures ranging from 25 °C to 33 °C, RH from 40% to 73%, and PM up to 18 μg m^{-3}.

Cumulative exposure (Exp) and energy consumption (Enr) ratios $\left(\frac{\text{Exp}_{\text{DRL}}}{\text{Exp}_{\text{DynOpt}}} \text{ and } \frac{\text{Enr}_{\text{DRL}}}{\text{Enr}_{\text{DynOpt}}}\right)$ between the two control strategies (DRL agent and DynOpt) for the same house have been used to assess the performance of the DRL agent. A ratio of one for any parameter (exposure or energy) indicates equal values of that parameter in both control scenarios, and lower ratios indicate lower exposure and energy consumption for the DRL agent than DynOpt, indicating comparable or better performance of DRL. Fig. 2 shows the effect of variations in the house characteristics (A: volume, B: thermal permeability, C: penetration factor, and D: deposition rate) on the performance of the DRL agent.

The exposure and energy ratios (Fig. 2A–D) demonstrate that the differences between the two control strategies are minimal as the ratios range between 1.00 and 1.09, except for a few outliers. Both exposure and energy ratios remain in the range of 1.05 ± 0.03, indicating that the total exposure and energy, on average, are just 5% more for the DRL agent-based control than the DynOpt control. The largest difference of 14% in energy and exposure is observed for the house with the largest volume. This could be due to minor deviations in the indoor–outdoor AER, leading to a considerably increased cooling demand. However, the overall trend indicates that personal exposure and energy consumption are comparable across the two control strategies, regardless of changes in house characteristics. These findings suggest that the performance of the DRL agent is relatively

Table 2 Values of different hyperparameters tested for their tuning. The values shown in bold font type are selected for all further simulations. (Table adapted from Mishra *et al.*[23])

Hyperparameters	Values
Hidden layers and nodes	{**128, 256, 128**}, {128, 256, 128, 64}, {128, 256, 256, 128, 64}
Learning rate	**0.001**, 0.01
Discount factor (γ)[a]	**0.99**
Batch size	**512**, 1024
$\varepsilon_{\text{sart}}$, $\varepsilon_{\min}$, $\varepsilon_{\text{decay}}$[a]	**1.0, 0.01, 0.99**
Target network update frequency (m)	100, **200** timesteps
Replay memory size[a]	**20 000**

[a] Fixed values were taken for these parameters.

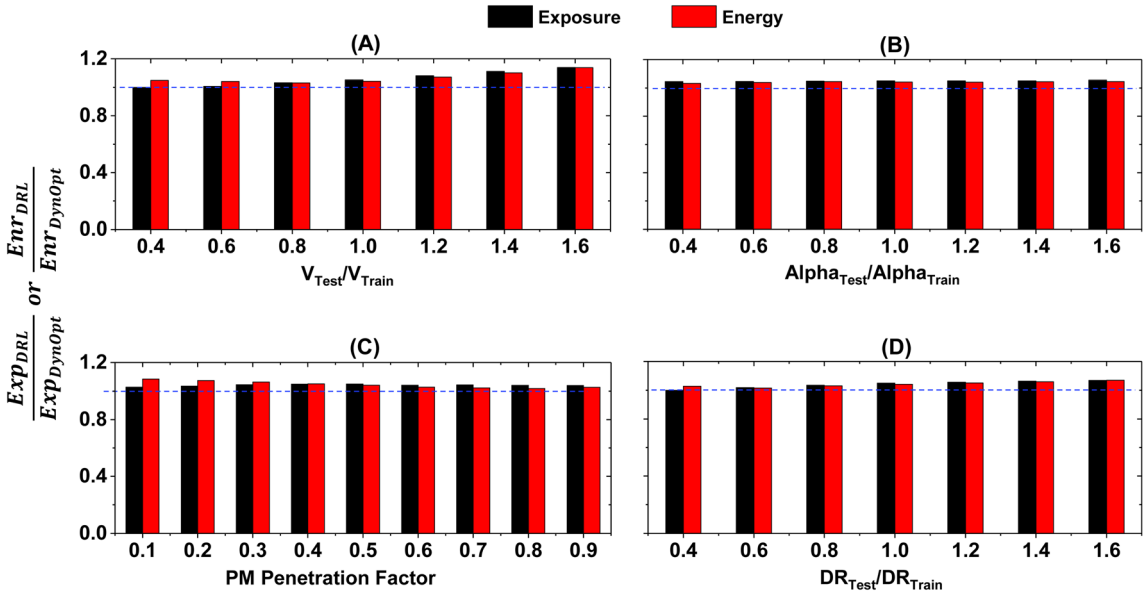

Fig. 2 Variations in ratios of cumulative exposure and energy consumption between DRL agent and DynOpt for changes in (A) volume, (B) thermal permeability, (C) PM penetration factor, and (D) deposition rate of the house. *V*: volume of a house, Alpha: thermal permeability of a house, DR: PM deposition rate in a house, test: testing house, train: training where the agent is trained.

independent of house characteristics under similar climatic conditions. Therefore, a DRL agent trained for a particular house can be deployed to other houses with little to no decline in performance.

The above discussion pertains to cases in which house characteristics were modified one at a time. The results indicate that, within realistic single-parameter perturbations, the trained DRL agent remains stable and near-optimal, and no distinct "turning point" for any individual parameter can be identified across C1–C5. Since isolated parameter variation did not produce significant degradation, it was hypothesized that performance limitation, if present, would emerge from conjugate (interaction) effects among parameters rather than from individual thresholds. Moreover, in the real world, multiple house characteristics are likely to change simultaneously. Therefore, C6 was designed to evaluate combinations at the minimum and maximum bounds of all four house characteristics, yielding a total of 16 combinations (C6_1 to C6_16; Table 1). Table 3 shows the ratios of exposure and energy for the DRL agent and DynOpt in these 16 houses.

In the cases where the house characteristics vary between two extremes, the exposure ratios remain within 0.92 and 1.07 for all houses, except for two cases (C6_14 and C6_16). Overall, the average ratios for exposure are slightly greater than one (1.03 ± 0.07), while a comparatively higher value (1.09 ± 0.06) is observed for the total energy ratios. Based on the observed standard deviation (0.07), the exposure ratios (1.19) for the two houses, C6_14 and C6_16, lie outside the central cluster of values and clearly separate them from the remaining distribution. Therefore, these cases are treated as outliers and share three extreme characteristics: maximum house volume, maximum deposition rate, and maximum penetration factor. These represent the corner of the multidimensional house parameter space where infiltration is maximized due to a high penetration factor, dilution effect due to high volume, and altered removal dynamics at high deposition. The simultaneous presence of these three maxima creates a compounded condition that is not encountered in single-parameter variations (C1–C5). Therefore, their deviation reflects a conjugate interaction effect rather than isolated parameter sensitivity. The energy consumption ratios also demonstrated similar variations (1.00 to 1.20), again demonstrating a compounded effect of multiple house characteristics.

To further analyse this compound effect, a multivariate regression analysis is performed, linking variations in exposure ratios to changes in house characteristics. Fig. 3 shows the exposure-energy ratio behaviour obtained from the multivariate analysis. The exposure ratio demonstrates a strong linear association ($R^2 = 0.90$) with changes in house characteristics. The individual coefficients for changes in volume (0.074), penetration factor (0.080), and deposition rate (0.070) are comparable, indicating that no single driver dominates and that the agent's performance depends on multiple house characteristics. Simultaneously, changes in the thermal permeability have the least impact on the exposure, with an individual coefficient of 0.001. These observations suggest that the exposure ratio exhibits a stable, well-defined linear dependence on the selected independent variables, with a distributed multi-factor influence. On the other hand, for energy performance, multivariate regression only explains a part of the variability. This indicates that the energy dynamics involve more complex non-linear interaction and control trade-offs.

In brief, the indoor dynamics of PM and the thermal environment are governed by the synergistic effects of various house parameters, and it is challenging to attribute the observations to a specific characteristic. These results indicate that the DRL agent, when transferred to different houses under similar ambient conditions to those of the training house, performs

Table 3 Ratios of both total exposure and energy consumption corresponding to DRL and DynOpt for different combinations of extreme values (maximum and minimum) of house characteristics. Minimum and maximum values of house characteristics are shaded in green and red, respectively[a]

Case ID	Volume factor ($V_{testing}/V_{training}$) ($V_{training}$ = 250 m^3)	PM deposition factor ($\lambda_{testing}/\lambda_{training}$) ($\lambda_{training}$ = 1.6 h^{-1})	Thermal permeability factor ($\alpha_{testing}/\alpha_{training}$) ($\alpha_{training}$ = 0.068 kJ s^{-1} C^{-1})	PM penetration factor (p_{DASS})	Exposure ratio Exp_{DRL}/Exp_{DynOpt}	Energy ratio Enr_{DRL}/Enr_{DynOpt}
C6_1	0.4	0.4	0.4	0.1	0.92	1.08
C6_2	0.4	0.4	0.4	0.9	0.99	1.15
C6_3	0.4	0.4	1.6	0.1	0.93	1.06
C6_4	0.4	0.4	1.6	0.9	0.99	1.11
C6_5	0.4	1.6	0.4	0.1	1.00	1.05
C6_6	0.4	1.6	0.4	0.9	1.02	1.07
C6_7	0.4	1.6	1.6	0.1	1.00	1.05
C6_8	0.4	1.6	1.6	0.9	1.03	1.05
C6_9	1.6	0.4	0.4	0.1	1.00	1.14
C6_10	1.6	0.4	0.4	0.9	1.04	1.00
C6_11	1.6	0.4	1.6	0.1	0.99	1.12
C6_12	1.6	0.4	1.6	0.9	1.04	1.01
C6_13	1.6	1.6	0.4	0.1	1.07	1.20
C6_14	1.6	1.6	0.4	0.9	1.19	1.13
C6_15	1.6	1.6	1.6	0.1	1.07	1.17
C6_16	1.6	1.6	1.6	0.9	1.19	1.11

[a] V: Volume of a house, λ: PM deposition rate, α: thermal permeability of a house, p_{DASS}: PM penetration factor of dedicated air supply system (DASS). 'Testing' and 'training' refer to the houses where the agent is being tested and trained.

comparably to a fully physics-driven strategy (DynOpt), with some outliers in extreme cases. Therefore, the proposed DRL agent can be transferred to different houses under the same climatic conditions after sufficient training. The next challenge in transferring the DRL agent for indoor environment control is ascertaining its performance under varying ambient conditions that differ from those during training, and the subsequent section discusses the same.

3.3 Variability in agent's performance under high ambient conditions

The ambient temperature and RH observed during the field investigation for the three experimental days analyzed in this

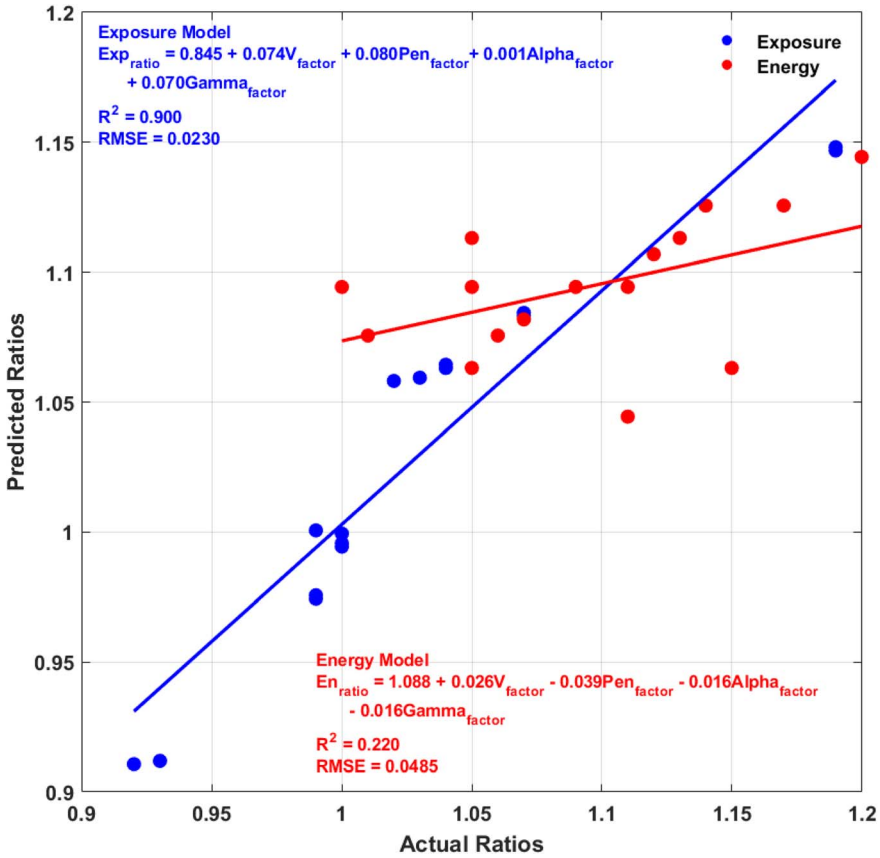

Fig. 3 Multivariate regression of exposure and energy ratios with changes in house characteristics.

study ranged between 26–33 °C and 40–73%, respectively. The ambient PM concentration (μg m^{-3}) variations during the three experimental days were 7.6 ± 3.0, with a maximum of 18.5 μg m^{-3}. However, to evaluate the adaptability of the DRL agent, it is critical to assess the agent's performance across houses under varying ambient conditions. Ambient conditions at various geographical locations are simulated by modifying the ambient temperature, RH, and PM concentration profiles, with the upper and lower bounds of these parameters adjusted to simulate greater variation. These wide ranges of ambient parameters are considered a proxy of weather conditions at different locations. The maximum and minimum values of adjusted temperature, RH, and PM concentration, along with their respective variations, are 44 °C and 28 °C (34 ± 4), 80% and 40% (63 ± 10), and 110 μg m^{-3} and 18 μg m^{-3} (55.5 ± 14.9), respectively. While these variations might not represent all ambient conditions, the framework and findings of this study could help understand the transferability and scalability potential of DRL agents for real-world integration.

In addition to the initially trained DRL agent in the training house (C1 in Table 1), three more DRL agents are trained in the same house under varying ambient conditions. These four agents are referred to as $DRL_{Original}$, the original agent trained under normal ambient conditions; $DRL_{ExtTempRH}$, an agent trained under higher variations in ambient temperature and RH; $DRL_{ExtPMTempRH}$, an agent trained under wider ranges of ambient PM, temperature, and RH; and DRL_{ExtPM}, an agent trained under wider ranges of ambient PM concentration. The ratios of exposure and energy, defined in Section 3.2, have been estimated for these agents under four ambient conditions (original ambient conditions, high ambient PM, high ambient temperature, and RH, and high ambient PM, temperature, and RH) in 17 houses (C1 and C6_1 to C6_16 from Table 1). Fig. 4 demonstrates these ratios for all DRL agents under high ambient conditions.

In Fig. 4, exposure and energy ratios of one signify equal exposure and energy consumption for the DRL agent and the DynOpt. Since both ratios are depicted on the same vertical axis, the composite bar is defined as the sum of exposure and energy ratios for any DRL agent. This bar should remain within the limits marked by dashed lines (shown in Fig. 4). Exceedances of these limits indicate a higher value of exposure or energy for the DRL agent relative to the DynOpt. The upward or downward shifting of the composite bar demonstrates the trade-off between exposure reduction and energy penalty.

Under normal ambient conditions (Fig. 4A), all agents ($DRL_{Original}$, $DRL_{ExtTempRH}$, $DRL_{ExtPMTempRH}$, DRL_{ExtPM}) demonstrate similar levels of exposure and energy compared to DynOpt, barring a few houses with high PM penetration factors (p_{DASS} = 0.9). For $DRL_{Original}$, the exposure and energy ratios vary in the range of 1.03 ± 0.07 and 1.09 ± 0.05, respectively (Fig. 4A). At the same time, DRL_{ExtPM} has the highest exposure compared to DynOpt, with the corresponding ratios in the range of 1.26 ± 0.04, and $DRL_{ExtTempRH}$ has the maximum energy

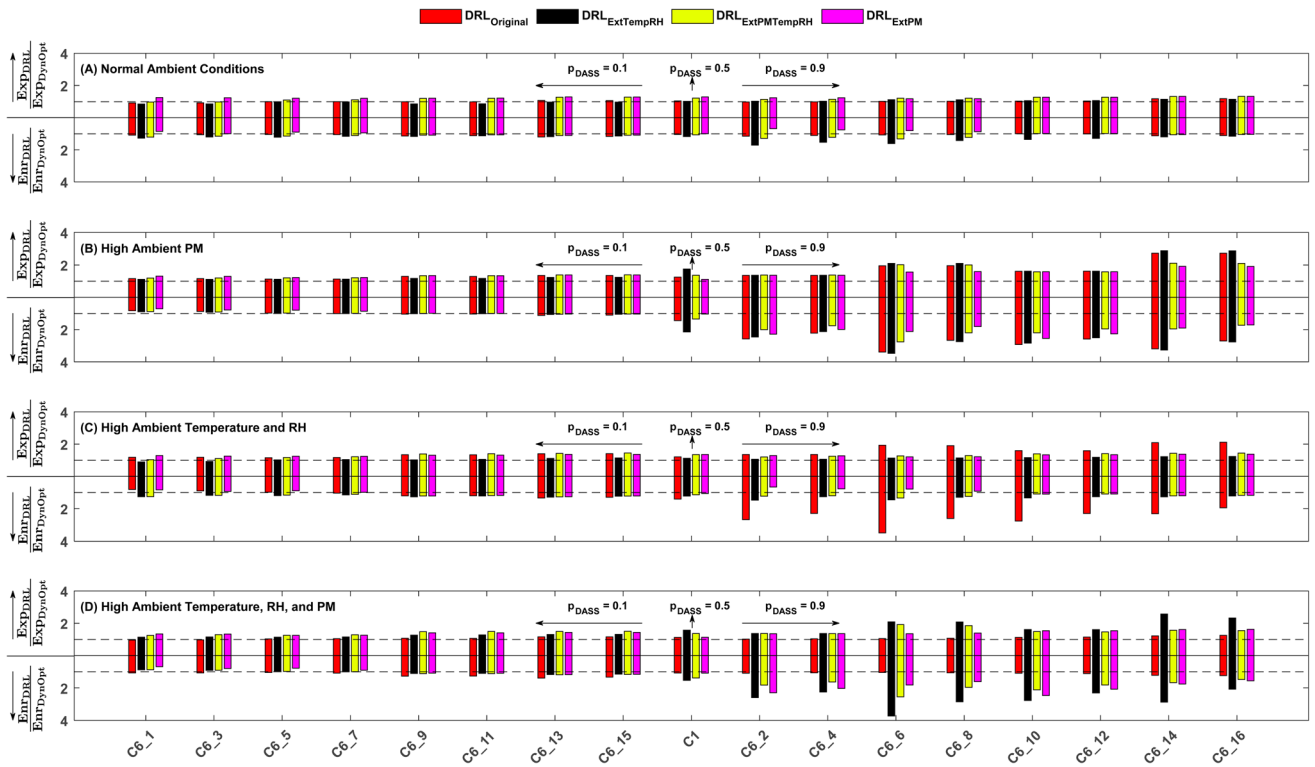

Fig. 4 Ratios of cumulative exposure (above *x*-axis) and energy consumption (below *x*-axis) for DRL agents for 17 combinations of house characteristics (C1 and C6_1 to C6_16) under (A) original ambient conditions, (B) high ambient PM, (C) high ambient temperature and RH, and (D) high ambient PM, temperature, and RH. p_{DASS}: PM penetration factor.

ratios (1.29 ± 0.17) under the normal ambient conditions. Apart from these two notable deviations, the exposure and energy ratios are comparable for other DRL agents in different houses. However, a slight drift of the composite bar in the upward or downward direction exists, representing the inherent trade-off between exposure reduction and energy consumption while ensuring thermal comfort in houses. These variations in energy and exposure ratios for DRL agents with different household characteristics demonstrate that $DRL_{Original}$ can perform reasonably well across a wide range of values of household characteristics. These findings indicate that well-trained DRL agents can be deployed in the field without training for each house.

Under varying ambient conditions (Fig. 4B–D), results can be categorized into two groups based on the PM penetration factor (p_{DASS}). The left and right halves of the plots show results for houses with PM penetration factors of 0.1 (low infiltration) and 0.9 (high infiltration), respectively. The other house characteristics for odd-numbered and their immediate next even-numbered houses are the same. For example, houses C6_1 and C6_2 have the same characteristics (volume, deposition rate, and thermal permeability) except for the p_{DASS} of 0.1 and 0.9, respectively. The original house, C1, is shown at the center and has a p_{DASS} of 0.5.

Under different configurations of high ambient conditions (temperature, RH, and PM), the exposure and energy ratios for all DRL agents are significantly lower in the houses with a lower PM penetration factor ($p_{DASS} = 0.1$). The average exposure ratios for p_{DASS} of 0.1 are 1.23 ± 0.13, while the same ratio varies in the range of 1.57 ± 0.41 for houses with high PM infiltration ($p_{DASS} = 0.9$). Energy ratios also demonstrate similar trends, wherein for p_{DASS} of 0.1, the average energy ratios are 1.06 ± 0.15, and for p_{DASS} of 0.9, the variations are in the range of 1.95 ± 0.71. Looking explicitly at the performance of $DRL_{Original}$, the average ratios of exposure and energy ratios for low-infiltration houses are 1.14 ± 0.11 and 1.10 ± 0.14, respectively, *i.e.*, lower than the average exposure ratio (1.23) and higher than the average energy ratio (1.06) for all agents. Also, a clear differentiation can be made between the houses with low and high PM infiltration in terms of energy and exposure ratios for $DRL_{Original}$. This distinct difference between exposure and energy ratios between low and high PM infiltration houses demonstrates an exacerbated decline in the DRL agents' performance due to the high infiltration of ambient PM. Therefore, transferring $DRL_{Original}$ to houses with high PM infiltration may result in suboptimal control. However, the agent performed well in houses with low PM infiltration rates under varying ambient conditions. Variability in agents' performance in houses with high PM infiltration may arise from fluctuations in the predicted indoor–outdoor AER. For example, increased indoor–outdoor AERs have a lesser impact on exposure and energy consumption in houses with lower PM infiltration.

From the preceding discussion in Sections 3.2 and 3.3, the following critical observations can be made on the at-scale deployability potential of a DRL agent trained with a limited range of household characteristics:

(1) The performance of $DRL_{Original}$, when transferred to houses with different characteristics under normal ambient conditions, is reasonably comparable to DynOpt, indicating that a sufficiently trained DRL agent can be transferred to other houses for optimal indoor environment control under normal ambient conditions.

(2) Under varying ambient conditions (shown in Fig. 4B–D), a proxy for different geographical locations, $DRL_{Original}$ performs better in houses with lower PM infiltration. The performance differences in houses with low and high infiltration demonstrate high variability under varying ambient conditions, highlighting the challenges associated with wide-scale deployability.

(3) The PM infiltration in a house depends on the ventilation mechanism. When the house is positively pressurized, the inflow of ambient air occurs *via* the DASS, and infiltration of ambient PM can be reduced by installing an air filter in the DASS. In this scenario, $DRL_{Original}$ can still be deployed, as demonstrated by its performance in houses with low infiltration. However, when the house is negatively pressurized, ambient air enters through cracks and openings, making it difficult to restrict the infiltration of pollutants. Therefore, in leaky houses (with more cracks and openings), it may be challenging to integrate the trained DRL agent.

4. Conclusion

In this work, a deep reinforcement learning agent trained on a single house is transferred to different houses with varying characteristics under both normal and high ambient conditions to gauge the scalability and transferability of such agents for optimizing pollutant exposure and energy consumption. The agent controls indoor environment dynamics by changing the indoor–outdoor AER through DASS. The transferability of any agent is primarily governed by two factors: house characteristics and ambient conditions. Therefore, this study evaluates the performance of the trained DRL agent under multiple combinations of house characteristics and ambient conditions. Firstly, the agent's performance variability across house characteristics is assessed in digital twins of multiple houses with varying volumes, thermal permeabilities, deposition rates, and PM penetration factors. Thereafter, the agent's performance is evaluated during high ambient conditions (temperature, RH, and PM concentration).

The study results demonstrate that the trained agent ($DRL_{Original}$) can be transferred to houses with varying characteristics under normal ambient conditions. The ratios of exposure (1.05 ± 0.03) and energy (1.05 ± 0.03) between the DRL agent and the dynamic optimization remain close to one, indicating acceptable performance. Under different ambient conditions, the original agent's performance is sub-optimal compared to the dynamic optimization control for houses with high PM infiltration. In contrast, for low PM infiltration, the agent performs comparably to the dynamic optimization strategy, with exposure and energy ratios of 1.14 ± 0.11 and 1.10 ± 0.14, respectively. These trends suggest that PM penetration affects the agent's performance at locations with different ambient conditions from the original locations. Therefore, manual intervention is needed in houses that allow high PM penetration, such as installing an air filter in the indoor–outdoor ventilation unit, enabling the transfer of the original DRL agent to other houses under high ambient conditions.

While this study shows the potential of DRL and similar agents to be transferred to different houses with minor or no additional intervention and training, the real-world integration and performance evaluation of these agents is imperative to assess the challenges associated with field deployment. Moreover, exposure in houses varies spatially, so the assumption of a well-mixed indoor air does not remain valid in all conditions.[54,55] Therefore, multi-agent control systems are needed to reduce personal exposure levels, considering the multi-zonal representation of a house. Furthermore, the performance of the reinforcement learning agents needs to be assessed under ambient conditions throughout the year to develop a solution suitable for all weather conditions. Simultaneously, uncertainty quantification and extensive benchmarking of their performance are critical for enhancing confidence in the agents' predictions. Moving forward, a collaborative effort from multi-disciplinary stakeholders is necessary to integrate these agents to improve health and safety in buildings. Advances in control algorithms, wide-scale field deployment, and performance benchmarking of these agents could promote their adoption at the community scale.

Conflicts of interest

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

Abbreviations

ACH	Air changes per hour
AER	Air exchange rate
DASS	Dedicated air supply system
DQN	Deep Q network
DRL	Deep reinforcement learning
DynOpt	Dynamic optimization
HVAC	Heating, ventilation, and air conditioning
PM	Particulate matter
RH	Relative humidity

Data availability

The input data used for training and the code for creating a skeleton of the DQN agent are available at https://github.com/nishmishra15/dopeec.

Supplementary information (SI) is available. See DOI: https://doi.org/10.1039/d5va00438a.

Acknowledgements

This work was supported by the Ministry of Education, India (Grant# AICoE/2024/AIS/2100000017) and Scheme for

Promotion of Academic and Research Collaboration (Grant# SPARC/2024-2025/ENSU/P3292). The authors thank Prof. Marina E. Vance, University of Colorado, Boulder, USA, and Prof Atila Novoselac, University of Texas at Austin, USA, for experimental data from the UTest House. Partial funding from the Indian Institute of Technology, Gandhinagar, is also acknowledged.

References

1 N. K. Mishra and S. Patel, *Need for a Holistic Approach to Assessing Sustainable, Green, and Healthy Buildings*, Environment & Health 2024, DOI: 10.1021/ENVHEALTH.4C00161.

2 A. C. Pinho-Gomes, E. Roaf, G. Fuller, D. Fowler, A. Lewis, H. ApSimon, *et al.*, Air pollution and climate change, *Lancet Planet Health*, 2023, 7, e727–e728, DOI: 10.1016/S2542-5196(23)00189-4.

3 J. Eom, M. Hyun, J. Lee and H. Lee, Increase in household energy consumption due to ambient air pollution, *Nat. Energy*, 2020, (5), 976–984, DOI: 10.1038/s41560-020-00698-1.

4 F. B. Bennitt, S. Wozniak, K. Causey, S. Spearman, C. Okereke, V. Garcia, *et al.*, Global, regional, and national burden of household air pollution, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021, *Lancet*, 2025, **405**, 1167–1181, DOI: 10.1016/S0140-6736(24)02840-X.

5 N. K. Mishra, P. Biswas and S. Patel, Future of clean energy for cooking in India: A comprehensive analysis of fuel alternatives, *Energy Sustainable Dev.*, 2024, **81**, 101500, DOI: 10.1016/J.ESD.2024.101500.

6 D. K. Farmer and M. E. Vance, Indoor air: sources, chemistry and health effects, *Environ. Sci.:Processes Impacts*, 2019, **21**, 1227–1228, DOI: 10.1039/C9EM90035G.

7 C. Arata, P. K. Misztal, Y. Tian, D. M. Lunderberg, K. Kristensen, A. Novoselac, *et al.*, Volatile organic compound emissions during HOMEChem, *Indoor Air*, 2021, **31**, 2099–2117, DOI: 10.1111/INA.12906.

8 S. Patel, S. Sankhyan, E. K. Boedicker, P. F. Decarlo, D. K. Farmer, A. H. Goldstein, *et al.*, Indoor Particulate Matter during HOMEChem: Concentrations, Size Distributions, and Exposures, *Environ. Sci. Technol.*, 2020, **54**, 7107–7116, DOI: 10.1021/ACS.EST.0C00740/SUPPL_FILE/ES0C00740_LIVESLIDES.MP4.

9 D. K. Farmer, M. E. Vance, D. Poppendieck, J. Abbatt, M. R. Alves, K. C. Dannemiller, *et al.*, The chemical assessment of surfaces and air (CASA) study: using chemical and physical perturbations in a test house to investigate indoor processes, *Environ. Sci.:Processes Impacts*, 2025, **27**(6), 1551–1572, DOI: 10.1039/D4EM00209A.

10 A. K. Thakur and S. Patel, Indoor Air Quality in Urban India: Current Status, Research Gap, and the Way Forward, *Environ. Sci. Technol. Lett.*, 2023, **10**, 1146–1158, DOI: 10.1021/ACS.ESTLETT.3C00636/SUPPL_FILE/EZ3C00636_SI_001.

11 L. A. Wallace, T. Zhao and N. E. Klepeis, Indoor contribution to PM2.5 exposure using all PurpleAir sites in Washington, Oregon, and California, *Indoor Air*, 2022, **32**, e13105, DOI: 10.1111/INA.13105.

12 D. M. Lunderberg, Y. Liang, B. C. Singer, J. S. Apte, W. W. Nazaroff and A. H. Goldstein, Assessing residential PM2.5 concentrations and infiltration factors with high spatiotemporal resolution using crowdsourced sensors, *Proc. Natl. Acad. Sci. U. S. A.*, 2023, **120**, e2308832120, DOI: 10.1073/PNAS.2308832120/SUPPL_FILE/PNAS.2308832120.SAPP.

13 *Buildings – Energy System – IEA n.d.*, https://www.iea.org/energy-system/buildings, accessed January 1, 2025.

14 N. K. Mishra, M. E. Vance, A. Novoselac and S. Patel, Dynamic optimization of personal exposure and energy consumption while ensuring thermal comfort in a test house, *Build. Environ.*, 2024, **252**, 111265, DOI: 10.1016/J.BUILDENV.2024.111265.

15 M. Elhami, S. S. Goodarzi, S. Maleki and B. Sajadi, Three-objective optimization of the HVAC system control strategy in an educational building to reduce energy consumption and enhance indoor environmental quality (IEQ) using machine learning techniques, *J. Build. Eng.*, 2025, **105**, 112444, DOI: 10.1016/J.JOBE.2025.112444.

16 F. Guo, S. woo Ham, D. Kim and H. J. Moon, Deep reinforcement learning control for co-optimizing energy consumption, thermal comfort, and indoor air quality in an office building, *Appl. Energy*, 2025, **377**, 124467, DOI: 10.1016/J.APENERGY.2024.124467.

17 T. Yang, L. Zhao, W. Li, J. Wu and A. Y. Zomaya, Towards healthy and cost-effective indoor environment management in smart homes: A deep reinforcement learning approach, *Appl. Energy*, 2021, **300**, 117335, DOI: 10.1016/J.APENERGY.2021.117335.

18 L. Yu, Y. Sun, Z. Xu, C. Shen, D. Yue, T. Jiang, *et al.*, Multi-Agent Deep Reinforcement Learning for HVAC Control in Commercial Buildings, *IEEE Trans. Smart Grid*, 2021, **12**, 407–419, DOI: 10.1109/TSG.2020.3011739.

19 D. Bayer and M. Pruckner, Enhancing the Performance of Multi-Agent Reinforcement Learning for Controlling HVAC Systems, *2022 IEEE Conference on Technologies for Sustainability, SusTech*, 2022, pp. 187–194, DOI: 10.1109/SUSTECH53338.2022.9794179.

20 Y. Chen, L. K. Norford, H. W. Samuelson and A. Malkawi, Optimal control of HVAC and window systems for natural ventilation through reinforcement learning, *Energy Build*, 2018, **169**, 195–205, DOI: 10.1016/J.ENBUILD.2018.03.051.

21 K. Al Sayed, A. Boodi, R. Sadeghian Broujeny and K. Beddiar, Reinforcement learning for HVAC control in intelligent buildings: A technical and conceptual review, *J. Build. Eng.*, 2024, **95**, 110085, DOI: 10.1016/J.JOBE.2024.110085.

22 W. Shang, J. Liu, C. Wang, J. Li and X. Dai, Developing smart air purifier control strategies for better IAQ and energy efficiency using reinforcement learning, *Build. Environ.*, 2023, **242**, 110556, DOI: 10.1016/J.BUILDENV.2023.110556.

23 N. K. Mishra, N. Batra and S. Patel, Optimizing Pollutant Exposure, Energy Consumption, and Thermal Comfort in a House *via* Deep Reinforcement Learning Control, *J. Build. Eng.*, 2025, 114074, DOI: 10.1016/J.JOBE.2025.114074.

24 I. Ajifowowe, H. Chang, C. S. Lee and S. Chang, Prospects and challenges of reinforcement learning based HVAC control, *J. Build. Eng.*, 2024, **98**, 111080, DOI: 10.1016/J.JOBE.2024.111080.
25 A. Chatterjee and D. Khovalyg, Dynamic indoor thermal environment using Reinforcement Learning-based controls: Opportunities and challenges, *Build. Environ.*, 2023, **244**, 110766, DOI: 10.1016/J.BUILDENV.2023.110766.
26 A. Manjavacas, A. Campoy-Nieves, J. Jiménez-Raboso, M. Molina-Solana and J. Gómez-Romero, An experimental evaluation of deep reinforcement learning algorithms for HVAC control, *Artif. Intell. Rev.*, 2024, **57**, 1–39, DOI: 10.1007/S10462-024-10819-X/METRICS.
27 X. Ding, A. Cerpa and W. Du, Exploring Deep Reinforcement Learning for Holistic Smart Building Control, *ACM Trans. Sens. Netw.*, 2024, **20**(3), 1–28, DOI: 10.1145/3656043.
28 M. Zuccotto, A. Castellini, D. La Torre, L. Mola and A. Farinelli, Reinforcement learning applications in environmental sustainability: a review, *Artif. Intell. Rev.*, 2024, **57**, 1–68, DOI: 10.1007/S10462-024-10706-5/METRICS.
29 C. Glanois, P. Weng, M. Zimmer, D. Li, T. Yang, J. Hao, *et al.*, A survey on interpretable reinforcement learning, *Mach. Learn.*, 2024, **113**, 5847–5890, DOI: 10.1007/S10994-024-06543-W/TABLES/9.
30 G. S. A. Krishna, T. Zhang, O. Ardakanian and M. E. Taylor, Mitigating an adoption barrier of reinforcement learning-based control strategies in buildings, *Energy Build*, 2023, **285**, 112878, DOI: 10.1016/J.ENBUILD.2023.112878.
31 X. Fang, G. Gong, G. Li, L. Chun, P. Peng, W. Li, *et al.*, Cross temporal-spatial transferability investigation of deep reinforcement learning control strategy in the building HVAC system level, *Energy*, 2023, **263**, 125679, DOI: 10.1016/J.ENERGY.2022.125679.
32 Z. Wang and T. Hong, Reinforcement learning for building controls: The opportunities and challenges, *Appl. Energy*, 2020, **269**, 115036, DOI: 10.1016/J.APENERGY.2020.115036.
33 A. K. Shakya, G. Pillai and S. Chakrabarty, Reinforcement learning algorithms: A brief survey, *Expert Syst. Appl.*, 2023, **231**, 120495, DOI: 10.1016/J.ESWA.2023.120495.
34 S. Dey, T. Marzullo, X. Zhang and G. Henze, Reinforcement learning building control approach harnessing imitation learning, *Energy AI*, 2023, **14**, 100255, DOI: 10.1016/J.EGYAI.2023.100255.
35 M. Biagiola and P. Tonella, Testing of Deep Reinforcement Learning Agents with Surrogate Models, *ACM Transactions on Software Engineering and Methodology*, Association for Computing Machinery, New York, NY, USA, 2024, vol. 33, no. 3, pp. 1–3, DOI: 10.1145/3631970.
36 N. K. Mishra and S. Patel, Optimizing Trade-off Between Pollutant Exposure and Energy Consumption in Buildings: Uncertainty Informed Reinforcement Learning and Robustness Analysis of Control Policies, *Build. Environ.*, 2026, **291**, 114224, DOI: 10.1016/J.BUILDENV.2026.114224.
37 S. Brandi, M. Fiorentini and A. Capozzoli, Comparison of online and offline deep reinforcement learning with model predictive control for thermal energy management, *Autom. Constr.*, 2022, **135**, 104128, DOI: 10.1016/J.AUTCON.2022.104128.
38 M. Esrafilian-Najafabadi and F. Haghighat, Transfer learning for occupancy-based HVAC control: A data-driven approach using unsupervised learning of occupancy profiles and deep reinforcement learning, *Energy Build*, 2023, **300**, 113637, DOI: 10.1016/J.ENBUILD.2023.113637.
39 M. Genkin and J. J. McArthur, A transfer learning approach to minimize reinforcement learning risks in energy optimization for automated and smart buildings, *Energy Build*, 2024, **303**, 113760, DOI: 10.1016/J.ENBUILD.2023.113760.
40 Z. Deng and Q. Chen, Reinforcement learning of occupant behavior model for cross-building transfer learning to various HVAC control systems, *Energy Build*, 2021, **238**, 110860, DOI: 10.1016/J.ENBUILD.2021.110860.
41 Y. Chen, Z. Tong, Y. Zheng, H. Samuelson and L. Norford, Transfer learning with deep neural networks for model predictive control of HVAC and natural ventilation in smart buildings, *J. Cleaner Prod.*, 2020, **254**, 119866, DOI: 10.1016/J.JCLEPRO.2019.119866.
42 A. Silvestri, D. Coraci, S. Brandi, A. Capozzoli and A. Schlueter, Practical deployment of reinforcement learning for building controls using an imitation learning approach, *Energy Build*, 2025, **335**, 115511, DOI: 10.1016/J.ENBUILD.2025.115511.
43 M. Liu, M. Guo, Y. Fu, Z. O'Neill and Y. Gao, Expert-guided imitation learning for energy management: Evaluating GAIL's performance in building control applications, *Appl. Energy*, 2024, **372**, 123753, DOI: 10.1016/J.APENERGY.2024.123753.
44 L. Yu, Z. Xu, T. Zhang, X. Guan and D. Yue, Energy-efficient personalized thermal comfort control in office buildings based on multi-agent deep reinforcement learning, *Build. Environ.*, 2022, **223**, 109458, DOI: 10.1016/J.BUILDENV.2022.109458.
45 S. Liu, X. Liu, T. Zhang, C. Wang and W. Liu, Joint optimization for temperature and humidity independent control system based on multi-agent reinforcement learning with cooperative mechanisms, *Appl. Energy*, 2024, **375**, 123968, DOI: 10.1016/J.APENERGY.2024.123968.
46 S. Patel, D. Rim, S. Sankhyan, A. Novoselac and M. E. Vance, Aerosol dynamics modeling of sub-500 nm particles during the HOMEChem study, *Environ. Sci.:Processes Impacts*, 2021, **23**, 1706–1717, DOI: 10.1039/D1EM00259G.
47 D. K. Farmer, M. E. Vance, J. P. D. Abbatt, A. Abeleira, M. R. Alves, C. Arata, *et al.*, Overview of HOMEChem: House Observations of Microbial and Environmental Chemistry, *Environ. Sci.:Processes Impacts*, 2019, **21**, 1280–1300, DOI: 10.1039/C9EM00228F.
48 *Ambient Air Quality Data Inventory – Catalog n.d.*, https://catalog.data.gov/dataset/ambient-air-quality-data-inventory, accessed May 6, 2025.
49 Y. An and C. Chen, Energy-efficient control of indoor PM2.5 and thermal comfort in a real room using deep reinforcement learning, *Energy Build*, 2023, **295**, 113340, DOI: 10.1016/J.ENBUILD.2023.113340.

50 M. Han, R. May, X. Zhang, X. Wang, S. Pan, Y. Da, *et al.*, A novel reinforcement learning method for improving occupant comfort *via* window opening and closing, *Sustain. Cities Soc.*, 2020, **61**, 102247, DOI: 10.1016/J.SCS.2020.102247.

51 Y. An, T. Xia, R. You, D. Lai, J. Liu and C. Chen, A reinforcement learning approach for control of window behavior to reduce indoor PM2.5 concentrations in naturally ventilated buildings, *Build. Environ.*, 2021, **200**, 107978, DOI: 10.1016/J.BUILDENV.2021.107978.

52 B. Eysenbach, R. Salakhutdinov and S. Levine, Search on the replay buffer: bridging planning and reinforcement learning, *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Curran Associates Inc., Red Hook, NY, USA, 2019, vol. 32, p. 1366.

53 R. Liu and J. Zou, The Effects of Memory Replay in Reinforcement Learning, *2018 56th Annual Allerton Conference on Communication, Control, and Computing, Allerton*, 2018, pp. 478–485, DOI: 10.1109/ALLERTON.2018.8636075.

54 A. K. Thakur and S. Patel, Characterization of particulate matter in a multizonal residential apartment: transport, exposure, and mitigation, *Environ. Sci.: Atmos.*, 2024, **4**, 1026–1041, DOI: 10.1039/D4EA00080C.

55 A. K. Thakur and S. Patel, Predicting Spatiotemporal Concentrations in a Multizonal Residential Apartment Using Conventional and Physics-Informed Deep Learning Approach, *ACS ES&T Air*, 2025, **2**, 1996–2008, DOI: 10.1021/ACSESTAIR.5C00190.