Sustainable Cities and Society 89 (2023) 104351

Contents lists available at ScienceDirect

Sustainable Cities and Society

journal homepage: www.elsevier.com/locate/scs

A Review of Reinforcement Learning for Controlling Building Energy Systems From a Computer Science Perspective☆

David Weinberg [a,*], Qian Wang [a,d], Thomas Ohlson Timoudas [c], Carlo Fischione [b]

[a] Department of Civil and Architectural Engineering, KTH Royal Institute of Technology, Teknikringen 78, 114 28, Stockholm, Sweden
[b] Department of Network and Systems Engineering, KTH Royal Institute of Technology, Teknikringen 33, 114 28, Stockholm, Sweden
[c] RISE Research Institutes of Sweden, Division Digital Systems, Computer Science, Isafjordsgatan 28 A, 164 40, Kista, Sweden
[d] Uponor AB, Hackstavägen 1, 721 32, Västerås, Sweden

ARTICLE INFO

Keywords:
Building Energy System
HVAC
Heating
Cooling
Reinforcement learning
Machine learning
RL
ML

ABSTRACT

Energy efficient control of energy systems in buildings is a widely recognized challenge due to the use of low temperature heating, renewable electricity sources, and the incorporation of thermal storage. Reinforcement Learning (RL) has been shown to be effective at minimizing the energy usage in buildings with maintained thermal comfort despite the high system complexity. However, RL has certain disadvantages that make it challenging to apply in engineering practices. In this review, we take a computer science approach to identifying three main categories of challenges of using RL for control of Building Energy Systems (BES). The three categories are the following: RL in single buildings, RL in building clusters, and multi-agent aspects. For each topic, we analyse the main challenges, and the state-of-the-art approaches to alleviate them. We also identify several future research directions on subjects such as sample efficiency, transfer learning, and the theoretical properties of RL in building energy systems. In conclusion, our review shows that the work on RL for BES control is still in its initial stages. Although significant progress has been made, more research is needed to realize the goal of RL-based control of BES at scale.

1. Introduction

According to a study by the International Energy Agency (IEA) in 2022, buildings account for 30% of global final energy usage and 27% of carbon emissions (International Energy Agency, 2022). As buildings contribute significantly to energy usage and emissions, they must be targeted for change in order to create a sustainable future society. The IEA states that the next decade will be critical for implementing the new technologies necessary to make 20% of the building stock zero-carbon-ready by 2030. In this regard, the European Commission's 2021 Digital Compass notes that the goal of accelerating the sustainable transformation is closely tied to digitalization (European Commission, 2021).

Digitalization trends in buildings, such as increased device connectivity and large-scale data collection, necessitate a reconsideration of how building energy systems (BES) are operated. The development of data-driven BES control systems, which have been shown to reduce operating energy usage by 10%–20%, is a crucial enabler for achieving sustainability goals (Vázquez-Canteli & Nagy, 2019). These control systems are particularly important in the transition to an energy-neutral, carbon-free future. On the one hand, the increasing use of low-temperature heating and the integration of renewables into BES require the development of novel control methods to optimize system performance. In addition, coordinating building clusters in local energy communities creates an unprecedented level of system complexity, requiring scalable BES control methods that can be applied to multiple-building systems. With the increasing trend of large-scale data collection in buildings, there is an opportunity to use Machine Learning (ML) approaches to control BES and reduce final energy usage and carbon emissions.

1.1. Background

Today, most building heating and cooling systems are operated using rule-based control logic (Salsbury, 2005). However, these traditional control systems have proved insufficient for operating the

☆ Financial Disclosure: This work is financially supported by Swedish Energy Agency with project grant agreement No. 51544-1. Special thanks to RISE Research Institutes of Sweden for providing in-kind supports.

* Corresponding author.
E-mail addresses: davidwei@kth.se (D. Weinberg), qianwang@kth.se (Q. Wang), thomas.ohlson.timoudas@ri.se (T.O. Timoudas), carlofi@kth.se (C. Fischione).

https://doi.org/10.1016/j.scs.2022.104351
Received 25 May 2022; Received in revised form 11 December 2022; Accepted 11 December 2022
Available online 14 December 2022

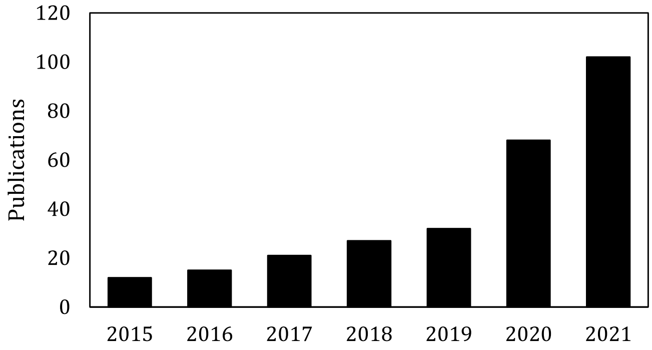

Fig. 1. Number of publications per year in the field of RL for HVAC control. The results were obtained from Web of Science using the search query `topic=(reinforcement learning AND (hvac OR heating OR cooling))`.

next generation of thermal systems, which will use temperature levels much closer to room temperatures and integrate thermal storage solutions. The main challenge is the increased thermal inertia of such next-generation systems–the reduced energy transfer capacity requires longer control horizons to maintain a sufficient thermal comfort. Model Predictive Control (MPC), a method for dynamic long-term planning of control systems, has been proposed as a solution to these problems (Afram, Janabi-Sharifi, Fung, & Raahemifar, 2017; Jin, Baker, Christensen, & Isley, 2017; Manjarres, Mera, Perea, Lejarazu, & Gil-Lopez, 2017). However, such methods require accurate building plant models, which are inherently hard to obtain due to the complex and time-varying dynamics.

Reinforcement Learning (RL) is getting increased attention in the field of building Heating, Ventilation and Air Conditioning (HVAC) and energy network control research. It has proven to be an attractive alternative to MPC due to its model-free approaches. Commonly, RL algorithms do not strictly require a model of the environment, but learn control policies through interaction. While this alleviates the problems of building modelling, RL suffers from other drawbacks, such as the need for massive quantities of data and long training times.

Several advancements have been made in the field of RL in the last decade, and many effective learning algorithms have emerged. Some of these algorithms have been applied to a wide range of applications in BES (Chen, Norford, Samuelson, & Malkawi, 2018; Du, Zandi, et al., 2021; Gao, Li, & Wen, 2020; Wei, Wang, & Zhu, 2017). The number of publications per year on RL for HVAC control can be seen in Fig. 1. Due to the growth of the field, there is a need for an up-to-date literature review summarizing the main research gaps and previous research.

1.2. Previous reviews

Previous reviews focus on RL methodology, but do not consider the full extent of the computer science related challenges of BES control. Vázquez-Canteli and Nagy review the application of RL for demand response applications (Vázquez-Canteli & Nagy, 2019). RL theory focusing on Q-learning and the exploration/exploitation trade-off is treated. Commonly used algorithms and action selections are also presented. Future research directions are identified, including the incorporation of expert knowledge, and reduction of the state–action space to improve the sample efficiency of the RL algorithms. The need for multi-agent systems to facilitate simultaneous control of building clusters is also emphasized. Wang and Hong review the use of RL for applications in building control such as window opening, lighting, and HVAC (Wang & Hong, 2020). Many aspects of RL, such as the popularity of various algorithms, the exploration/exploitation trade-off, and the choice of states and actions are investigated. Reduction of the state–action space and multi-agent systems are suggested as future research directions. Han, et al. reviews the use of RL for control with various comfort objectives such as indoor air quality, noise, and thermal comfort (Han, et al., 2019). Commonly used algorithms as well as exploration strategies are analysed. A discussion on value-based versus policy-based methods is included. Multi-agent RL is identified as a future research direction of significant importance. Although not strictly a review, Nweye, Liu, Stone, and Nagy identify nine practical challenges of RL in grid-interactive buildings (Nweye et al., 2022). Sample efficiency, partial observability, and explainability are, among other topics, identified as important future research directions. An example of off-line learning of an RL controller in the proposed test-environment CityLearn is also provided.

Some previous reviews treat RL in combination with other topics related to building HVAC and energy systems control. Royapoor, Antony, and Roskilly reviews holistic control of a building with HVAC control included as a section (Royapoor et al., 2018). Popular control methods, including RL, are investigated. The use of thermal comfort and occupancy modelling for control to reduce operational energy usage is discussed. A survey on industry familiarity with various control strategies is also included. Hong, Wang, Luo, and Zhang reviews the use of ML in the entire building life cycle, with a section on HVAC control (Hong et al., 2020). Supervised Learning (SL) for personal comfort and occupancy-modelling is treated. MPC, RL and their respective advantages for HVAC control, are discussed. Potential benefits of planning in RL methods are mentioned. Pinto, Wang, Roy, Hong, and Capozzoli review the use of transfer learning in smart buildings (Pinto, et al., 2022). Transfer learning for both RL and various kinds of predictive modelling are treated both in the context of single buildings and building clusters.

1.3. Contributions of this review

The contributions of this review to the existing literature consist of four key aspects:

- Many challenges of applying RL to BES control have their roots in computer science, not in building energy research. We provide a guide for building energy researchers to become familiar with the computer science related challenges, and to identify key research gaps where meaningful contributions can be made.
- Previous treatment of the computer science related challenges of RL for BES control is distributed across a vast body of literature, which is difficult to overlook. We address this by identifying and structuring the main challenges into three categories, outlined in Sections 5–7.
- Previous reviews address only limited parts of the computer science related challenges of RL for BES control. We address this by providing a comprehensive treatment of all the main challenges.
- The study identifies promising future research directions to address the computer science related challenges of applying RL to BES control.

The three categories of challenges that are identified in this review are referred to as *reinforcement learning in single buildings*, *reinforcement learning in building clusters*, and *multi-agent aspects*. To further justify our contributions, the coverage by previous reviews of the three categories can be seen in Table 1. Note that coverage of the main challenges in previous reviews can mean two things: (1) The reviewers investigate previous work concerning the challenges, or (2) discuss them as future research directions. It is evident that no previous review treats all categories of challenges.

1.4. Organization of the review

In Section 2, we introduce the BES and discuss the challenges of control in such systems. This section is followed by a primer on the theory behind RL in Section 3. Given an overview of key theoretical

Table 1
Coverage of the three main challenges in previous reviews. Review work, and identification as a future research direction are denoted by RW and FR, respectively.

Author	Reference	RL in single buildings	RL in building clusters	Multi-agent aspects
Vázquez-Cante and Nagy	Vázquez-Canteli and Nagy (2019)	FR/RW	–	FR
Wang and Hong	Wang and Hong (2020)	RW	–	FR
Han et al.	Han, et al. (2019)	RW	–	FR
Nweye et al.	Nweye et al. (2022)	FR	–	FR
Royapoor et al.	Royapoor et al. (2018)	FR	–	–
Hong et al.	Hong et al. (2020)	FR/RW	FR	–
Pinto et al.	Pinto, et al. (2022)	RW	FR/RW	–

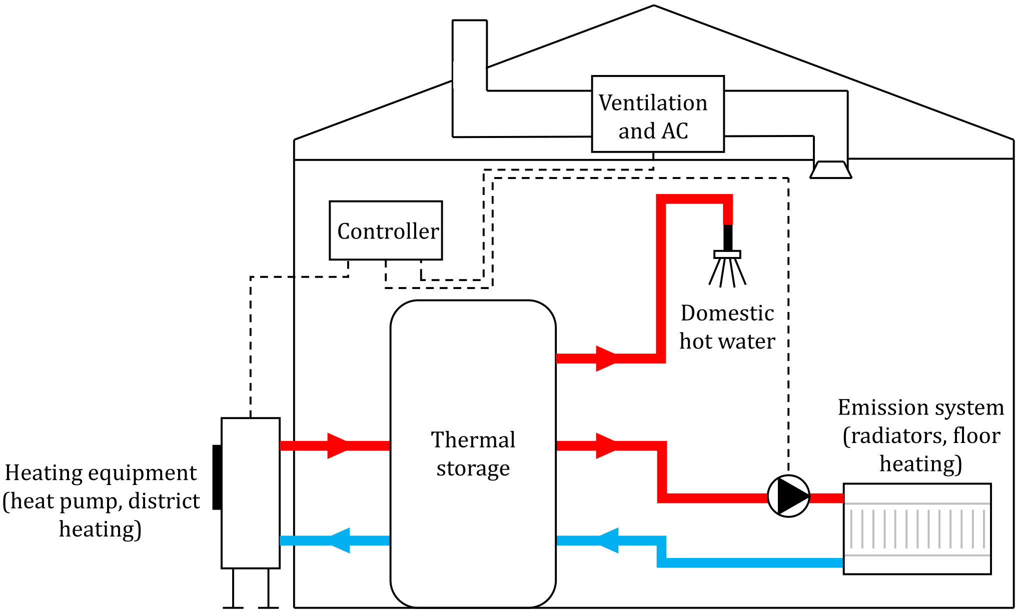
Ventilation and AC
Controller
Domestic hot water
Thermal storage
Emission system (radiators, floor heating)
Heating equipment (heat pump, district heating)

Fig. 2. A generic illustration of a BES, and the associated measurement and control system. The dashed lines represent exchange of sensor measurements and control signals.

aspects, we discuss the computer science perspective, and review state-of-the-art approaches to RL-based BES control in Sections 4–7. In Section 8, we make a future outlook and identify promising research directions. Lastly, we conclude the paper and summarize our findings in Section 9.

2. Problem formulation

The Building Energy System (BES) can be thought of as all systems in a building that use energy. This includes both electrical appliances and lighting, and the thermal energy system. In this review, we limit the definition of BES to encompass heating equipment, thermal storage, emission systems, and ventilation and air conditioning, as well as their auxiliary systems, as shown in Fig. 2. Even though the constituent parts of the BES have their own intricate principles of operation, we will not cover them in detail. The purpose of this section is merely to establish the role of RL in BES control. For case-specific treatment of BES components, we refer the reader to Frederiksen and Werner (2013), Tymkow, Tassou, Kolokotroni, and Jouhara (2020).

All the BES components in Fig. 2 are part of an interconnected system: heat is generated using heating equipment, injected into a thermal storage, and distributed in the building using an emission system. On top of this, ventilation, and air conditioning are added to ensure an acceptable indoor environment. Naturally, all components must be controlled to ensure a comfortable indoor climate, preferably at a low monetary cost while using as little energy as possible. An important example of this is optimal control of thermal storage systems to shift the time of heat production away from times of high load. For optimal operation, the charging of the storage tank must be scheduled such that the load can always be supplied, while keeping the cost and energy usage to a minimum. But the components of a BES are highly interactive, and achieving these goals using traditional control methods is prohibitively difficult. Using Fig. 2 as an example, the controller must coordinate the operation of all components in the BES while taking into account complications such as the system's dynamic response, and time delays in sensor measurements and actuators. In large buildings, where the number of measured and controlled variables increases, this coordination problem becomes highly complex, so the need for new control methods is apparent.

2.1. Data collection in buildings

To control and monitor a BES, operational data must be collected and stored. This is typically done in a Building Management System (BMS), which integrates sensors, actuators, and networking devices to provide information about all subsystems in a building. For control purposes, the BMS operates on multiple levels which represent distinct levels of abstraction of the system. For example, the level closest to sensors and actuators is known as the field level (Levermore, 2000). Devices at the field level receive instructions from higher level controllers to write specified output signals to the actuators. They also read sensor values and passes them back to the higher-level devices. The higher-level controllers receive the sensor data, which is stored and processed to determine suitable control signals to send to the field level devices. This is demonstrated using dashed lines in Fig. 2 where at each BES-component, there are sensors that measure the variables which we want to control. These measurements are communicated to a central controller, which computes the control signals to be applied to the actuators at the BES-components. Even in buildings where BMS systems are not installed, the use of networked sensors and actuators allows for scaled down control solutions. Measurement technology in BES is a vast area on its own, and for brevity we will not discuss it in more detail here. We instead refer the interested reader to Levermore (2000), Sofos, et al. (2020).

These days, BMS systems are common and serve as promising platforms for integrating data-driven control into BES. Stored operational data can be used inside ML workflows to create control systems that autonomously optimize the energy performance of a building. In particular, BMS systems facilitate large scale data collection in clusters of buildings, paving the way for coordination and control of entire energy communities. In the next section, we explain the role of RL in this scenario, and how it can be integrated to solve control problem in complex BES.

2.2. The need for reinforcement learning

The possibility of large scale data collection in buildings allows for different forms of data-driven control. An important example is to use data to calibrate the parameters of a physics-based BES model, in a so-called grey-box modelling arrangement (Li, O'Neill, et al., 2021). The calibrated system model could then be used for optimal control. This approach, taken by many researchers in the MPC community, has been demonstrated to suffer from reliability issues tied to the model's agreement with the real system (Drgoňa, et al., 2020). According to Lee and Zhang, grey-box techniques may be used to automate the parameter estimate process in physics-based modelling (Lee & Zhang, 2021). It is possible that the use of a fairly generic physics-based grey-box model allows for it to be transferred between multiple BES, removing the need for re-modelling of each building. RL should be viewed as a complement to such approaches, with the added benefit of being capable of controlling buildings with complex dynamics, and perhaps removing altogether the requirement to comply with a pre-specified system model.

RL offers a data-driven approach to learning the optimal control signals to send to each component in the BES by interacting with the physical system. No model of the BES is strictly required by many RL algorithms, thanks to its purely data-driven procedures. Moreover, RL is in principle agnostic to the underlying technologies used in the BES—it does not matter whether it uses a boiler, heat pump, district heating, or something else. Even the integration of renewables and thermal storage can be done conveniently using the RL-formulation by including relevant measurement information from the corresponding components in the BES. However, as shall be clarified in the next section, the problem remains identical from an RL point of view.

Table 2
Pros and cons of physics-based, and RL-based control.

	Pros	Cons
Physics-based	• Incorporation of expert knowledge • Interpretable system dynamics	• Difficult in complex systems • Governing physical laws might not exist
RL-based	• No system-model required • Can control very complex systems	• Requires large quantities of data • Few stability- or robustness guarantees

3. A primer on reinforcement learning

Clearly, from a control system perspective, certain variables of interest in the BES must be measured. For instance, it is natural to measure the indoor air temperature, as it is directly related to the thermal comfort. But many other variables might also be useful in the control problem, such as the relative humidity and the thermal energy usage of the building. In RL jargon, the collection of such relevant measured information is referred to as the system *state*, and is denoted by the real valued vector $\mathbf{s}_t \in \mathcal{S}$. The subscript t is a discrete time index, and the set $\mathcal{S}$ of possible state values is called the *state space*. Similarly, there are actuators in the BES such as valves, pumps and power regulators which allows for manipulating the system. In RL, the control signals sent to the actuators are referred to as *actions*, and are denoted by the vector $\mathbf{a}_t \in \mathcal{A}$. Here, the set $\mathcal{A}$ is the set of possible actions, called the *action space*. RL algorithms have the goal of finding the optimal action $\mathbf{a}_t$ to take given a measured state $\mathbf{s}_t$. More formally, they attempt to find a probability density function $\pi(\mathbf{a}_t|\mathbf{s}_t)$ known as a *policy*, which describes the optimal actions in a given state. The optimal policy may be found by solving

$$\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^{t} r(\mathbf{s}_t, \mathbf{a}_t)\right], \tag{1}$$

where $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is a so-called *reward function* (Sutton & Barto, 2018), and $\gamma \in [0, 1)$ is a discount factor. This reward function measures the immediate benefit of taking the action $\mathbf{a}_t$ in state $\mathbf{s}_t$, so the objective is to find a control policy π that maximizes the expected long-term reward. Note that in the equation above, $\mathbf{a}_t$ is a random variable with density $\pi(\mathbf{a}_t|\mathbf{s}_t)$. The density of $\mathbf{s}_t$ is given by the state-transition dynamics of the system, which are typically assumed to be Markovian, resulting in a so-called Markov Decision Process (MDP). We shall denote the transition dynamics by $T(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$.

In practice, the reward function may be difficult to specify for a given problem, and the resulting optimal policy π changes accordingly. The form of the reward function is typically left to the control systems engineer as a design choice, which is non-trivial in general. In the case of BES control, there is a consensus that a suitable reward function should penalize high energy usage in the BES, as well as violations of any thermal comfort requirements. For a more detailed overview of the choice of reward function, we refer the reader to Han, et al. (2019), Vázquez-Canteli and Nagy (2019), Wang and Hong (2020).

3.1. Value-based and policy-based algorithms

RL provides algorithms to solve the problem in Eq. (1) without explicit knowledge of the dynamics T. Instead, RL algorithms rely on interactions with the physical system to update the control policy π towards the optimum, essentially by trial and error. In fact, there are several benefits of using RL over physics-based modelling, but as shall be clarified later in the review, there are also drawbacks. The pros and cons of RL compared to physics-based modelling are summarized in Table 2.

There is a plethora of algorithms for finding the optimal policy for a specific system using RL. They are typically classified as either value-based, or policy-based algorithms. Value-based algorithms typically try to find a policy which maximizes the so-called action value function, which is defined as

$$Q^{\pi}(\mathbf{s}, \mathbf{a}) := \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^{k} r_{t+k+1} \middle| \mathbf{s}_t = \mathbf{s}, \mathbf{a}_t = \mathbf{a}\right]. \tag{2}$$

A prominent approach is to use a parametrized representation $Q_\theta : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$, such as a neural network, for the action-value function. The parameters would be estimated by minimizing the mean squared error between the true action-value function and the approximation.

$$\min_{\theta} \mathbb{E}\left[\frac{1}{2}\left(Q_\theta(\mathbf{s}, \mathbf{a}) - Q(\mathbf{s}, \mathbf{a})\right)^2\right]. \tag{3}$$

The target Q is typically not known exactly, but an estimate can be provided by running the controller in the system. The parameters of Q_θ would then be updated using stochastic gradient descent or variations thereof Sutton and Barto (2018). For instance, this is the general idea behind the famous Deep Q-Networks (DQN) algorithm (Mnih, et al.,

Table 3
Classification of popular RL algorithms.

	On-policy	Off-policy
Value-based	SARSA (Sutton & Barto, 2018)	DQN (Mnih, et al., 2013)
Policy-based	TRPO (Schulman, Levine, et al., 2017), PPO (Schulman, Wolski, et al., 2017)	DDPG (Lillicrap, et al., 2019), SAC (Haarnoja et al., 2018)

2013). Once the Q-function has been estimated, the optimal policy may be extracted by choosing the action with the highest Q-value in a given state.

Policy-based algorithms, on the other hand, directly parametrize the policy as a function $\pi_\theta : \mathcal{S} \rightarrow \mathcal{A}$, or a probability distribution over actions $\pi_\theta(\mathbf{a}|\mathbf{s})$. The optimal parameters are found by directly solving the problem

$$\max_\theta \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(\mathbf{s}, \mathbf{a})\right] , \tag{4}$$

where the expectation is over the stationary state distribution $\mathbf{s} \sim \rho$ and the policy $\mathbf{a} \sim \pi_\theta$ (Sutton & Barto, 2018). There are many popular policy-based algorithms in RL, such as Trust Region Policy Optimization (TRPO), Proximal Policy Optimization (PPO), Deep Deterministic Policy Gradient (DDPG) and the Soft Actor Critic (SAC) (Haarnoja, Zhou, Abbeel, & Levine, 2018; Lillicrap, et al., 2019; Schulman, Levine, Moritz, Jordan, & Abbeel, 2017; Schulman, Wolski, Dhariwal, Radford, & Klimov, 2017).

3.2. On-policy and off-policy algorithms

Another useful classification of RL algorithms is whether the data used to improve the current policy must be collected in the system using that same policy. On-policy algorithms have this requirement, whereas off-policy algorithms allow for the use of data collected in the system using another policy.

In BMS systems, sensors are typically read with a sampling time interval of several minutes—sometimes even hourly. This slow collection of data is, as will be shown in the following sections, problematic as RL algorithms need large quantities of data to find the optimal control policy. Hence, off-policy algorithms appear attractive for BES control due to their ability to reuse all data collected throughout the learning process. With on-policy algorithms, the duration of data collection increases, as it is required to collect larger quantities of data after every policy update. In Table 3, a summary of the workings of some popular RL algorithms is shown.

4. The computer science perspective

Many of the commonly encountered challenges in applications of RL for BES control do not have their roots solely in energy systems research, but also in computer science. Therefore, in this review, we aim to investigate these challenges by an interdisciplinary approach. In the context of the problem at hand, this means studying the limitations of RL in BES and how these issues can be resolved. We do this by dividing the review into 3 main sections:

- RL in single buildings
- RL in building clusters
- Multi-agent aspects

Firstly, the section on RL in single buildings addresses the approaches taken by building energy researchers to alleviate the issue of learning inefficiencies in RL. RL algorithms typically require massive quantities of data and extensive interaction with the building before convergence is achieved (Yu, 2018). This implies that the training process could be prohibitively long for RL to be practically feasible in real-life BES. In this topic, we restrict our review to solutions treating single buildings in isolation—we do not allow the reuse of data from other buildings. The main remedies turn out to be related to the choice of learning algorithm, the use of learnt plant models, as well as incorporation of expert knowledge.

Secondly, in the section on RL in building clusters we review approaches taken by building energy researchers which focus on exploiting data from other buildings when deploying RL in building retrofitting practices, or in newly built. Here, unlike in the first topic, we focus on approaches which transcend the isolated treatment of single buildings. Instead, we review methods that rely on transferring information for efficient learning in clusters of buildings. The most prominent approaches are related to various forms of transfer learning—either by directly transferring control policies, or by transferring plant models between buildings.

Lastly, the section on multi-agent aspects treats simultaneous learning of RL-controllers in building clusters using multi-agent RL. Under imposed privacy restrictions or severe communication constraints, agents might not be able to mutually share information, so we outline the distinction between independent and joint learning. Moreover, we discuss the safety and scalability aspects of multi-agent control of BES in building clusters.

In summary, these three sections provide an overview of the most important challenges of applying RL to BES control. They also provide detailed descriptions of the tools and approaches adopted to address these challenges. However, it is critical to remember that although the challenges are treated in isolation here, in engineering, they may arise simultaneously. For example, in the case of using transfer learning for RL in building clusters, it may be advisable to still use some approaches outlined in the section on RL in single buildings. Similarly, even if one is faced with a multi-agent RL problem, it is essential to still consider the potential benefits of, e.g., transfer learning to speed up the training process. Also, even though the challenges of applying RL to BES control are considered here on the scale of single buildings and building clusters, they could in principle arise on smaller scales. For example, the issue of controlling the BES in a multi-zone building can be solved by having separate controllers for each thermal zone. In such an arrangement, each thermal zone could possibly utilize information from the other zones in a transfer learning-like arrangement. The task now strongly resembles RL in building clusters, but on the scale of a single building. In the next three sections, the three main challenges mentioned earlier will be reviewed in detail.

5. Reinforcement learning in single buildings

We now turn our attention to the problem of data collection, and learning in BES. Data collection in BES can be very time-consuming because of the low measurement sampling frequency. The total time between sensor readings is often in the order of several minutes. This is troublesome as most RL methods require substantial quantities of data to produce useful results. To capture the seasonal variations inherent in the surrounding weather conditions, it is potentially necessary to collect data over several years. This would lead to unacceptably long training times for practical applications. However, the solution is not as simple as sampling more often, as the underlying reason for the low sampling frequency is the slow dynamics of the BES. Oversampling in such a system would not capture any meaningful information, because the important variations are taking place on a time-scale of hours, or even days. Moreover, in a BES control problem, RL methods naturally require direct interaction with the building installations. Given the suboptimal behaviour of a randomly initialized controller, there could be periods of time when the building becomes uninhabitable due to the controller's stochastic exploration of heating and cooling settings. If, however, some

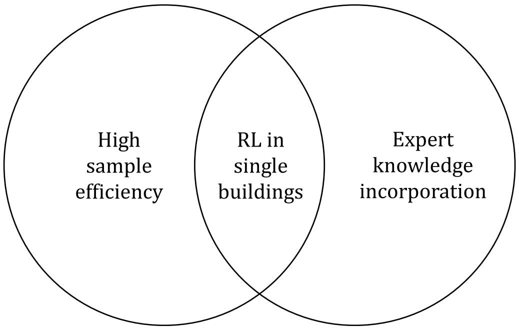

Fig. 3. Illustration of RL in single buildings as the intersection of high sample efficiency and incorporation of expert knowledge.

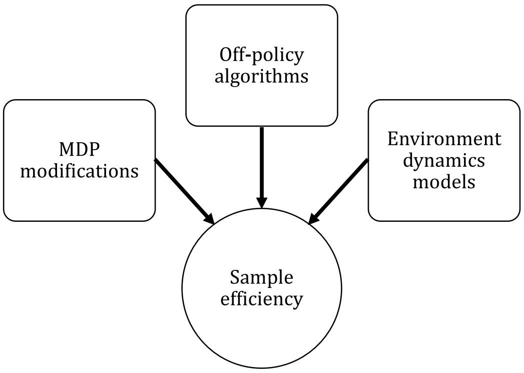

Fig. 4. The main approaches taken in BES control research to increase sample efficiency of RL algorithms.

exploratory actions are known to be unreasonable a priori, considering them less often might speed up the training phase. Hence, incorporating expert knowledge and guidance before deployment can be of time-saving benefit.

The observations above motivate the need for RL algorithms with a short training time and a high final average reward within a single building. More precisely, an algorithm suitable for *RL in single buildings* should possess the following two virtues: it should utilize the collected data as efficiently as possible, and it should incorporate prior knowledge before training as shown in Fig. 3. In this section, previous work treating the two virtues in the context of BES control will be reviewed.

5.1. Sample efficiency

Sample efficiency is the first virtue of an RL algorithm for BES control. Because the data collection process in buildings can be slow, it is desirable for an algorithm to exhaust the collected data as much as possible, even if it leads to a higher computational load. Moreover, it is important to consider what variables are monitored. For a discussion about what measurements should make up the system state in an RL algorithm for BES control, we refer to the review by Wang and Hong (2020). Suitable algorithms should obtain the highest possible final performance from the smallest number of environment interactions. The approaches taken in BES control to increase the sample efficiency are illustrated in Fig. 4 and discussed in detail below.

5.1.1. MDP modifications

High dimensionality of the action space $\mathcal{A}$ has been pointed out as one of the fundamental issues of RL algorithms in BES control in several publications: Gao et al. emphasize that algorithms such as Q-learning, DQN and SARSA are unsuitable for the HVAC control problem due to their inability to handle continuous action spaces (Gao et al., 2020). Discretization of the action space increases its cardinality exponentially, and exploring the state–action space $\mathcal{S} \times \mathcal{A}$ during training requires more interaction samples. Moreover, the maximization over action-values as in the DQN algorithm may be intractable for large or continuous action spaces. Gao et al. also conclude that using DDPG–which is compatible with continuous action spaces–leads to a fast convergence and to a high final average reward. Du, Zandi, et al. confirm these findings by noting that DDPG converges almost one order of magnitude faster than DQN in terms of the number of training episodes (Du, Zandi, et al., 2021). Evidently, using continuous actions is not always an option if discrete setpoints are all that is available. In such cases, the size of the action space $\mathcal{A}$ can still be reduced through other approaches. Jiang, et al. propose a preprocessing of the actions suggested by a DQN agent (Jiang, et al., 2021). During peak load hours, the indoor temperature setpoint is clipped at the comfort boundary. Any higher setpoint suggested by the DQN agent would inevitably lead to a lower reward. Jiang, et al. report that such reduction of the size of the action space helps produce better agents that consistently outperform the non-reduced version.

5.1.2. Off-policy algorithms

On-policy or off-policy operation is another determinant of the sample efficiency of an RL algorithm. Many researchers have used DQN and DDPG, which are both off-policy algorithms, for BES control (Du, Zandi, et al., 2021; Gao et al., 2020; Jiang, et al., 2021; Wei et al., 2017). They have the benefit of being able to utilize a buffer of collected experience that can be used throughout training, despite having been collected using an old policy. Biemann, Scheller, Liu, and Huang compare several on-policy and off-policy methods on a problem of controlling the temperature setpoints and fan mass flow rates in a simulated datacenter with two thermal zones (Biemann, Scheller, et al., 2021). It is concluded that the off-policy SAC agent converges one order of magnitude faster than its on-policy counterparts in terms of the number of environment interactions. Due to its high sample efficiency, they hypothesize that the SAC algorithm will have a key role to play in the next generation of HVAC control systems. On the other hand, successful implementations of on-policy algorithms have also been produced (Chen, Cai, & Bergés, 2020). To increase the data efficiency of on-policy algorithms, it is possible to use recurrent policies. Biemann, Liu, Zeng, and Huang demonstrate that doing so can drastically increase the sample efficiency of the algorithm (Biemann, Liu, et al., 2021). The reason is pointed out as being that the state process is non-Markovian, so that an internal state must be maintained with a recurrent neural network.

5.1.3. Model-based RL

Model-based RL is an alternative measure to improve sample efficiency. The idea is that knowledge of the environment dynamics $T(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$ can be leveraged to reduce the required number of environment interactions. Using a learned model of the environment dynamics for optimizing actions is strongly related to MPC (Afram & Janabi-Sharifi, 2014; Drgoňa, et al., 2020). In fact, MPC can be regarded as a special instance of model-based RL where instead of finding an optimal policy, one directly maximizes over an action sequence, c.f. Eq. (1):

$$\max_{\mathbf{a}_t \dots \mathbf{a}_{t+H}} \quad \sum_{i=t}^{t+H} r(\mathbf{s}_i, \mathbf{a}_i) \tag{5}$$

$$\text{s.t.} \quad \mathbf{s}_{i+1} = f_\phi(\mathbf{s}_i, \mathbf{a}_i)\,. \tag{6}$$

In the optimization problem above, $f_\phi : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ is a parametrized model of the environment dynamics. Kurte, Amasyali, Munk, and Zandi compare MPC and DQN for indoor temperature setpoint control (Kurte et al., 2021). It is found that the relative energy cost savings are approximately 60% larger for the MPC controller than for the DQN agent. This is to be expected since the MPC controller has access to a detailed model of the building physics. Model-based RL also includes other aspects such as learning a policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$ as opposed to directly optimizing over an action sequence. One advantage of doing so is that model free methods can be used for learning. Chen et al. use MPC and an environment dynamics model to perform planning of actions for controlling the supply-water temperature from a district heating connection (Chen et al., 2020). It is implemented in conjunction with a PPO agent to improve a parametrized policy. This leads to large improvements in terms of sample efficiency over the purely model free alternative.

Real time learning of a dynamics model is also possible. For instance, the system developed by Nagy, Kazmi, Cheaib, and Driesen simultaneously learns a neural network model of the environment dynamics combined with MPC to perform planning of the supply power to an air-source heat pump (Nagy et al., 2018). They conclude that the model-based method is consistently better in terms of both sample efficiency and the resulting average reward. Zhang, Kuppannagari, Kannan, and Prasanna confirm the findings by noting that the combination of a neural network dynamics model and MPC converges to its final average reward an order of magnitude faster than PPO for temperature setpoint control in a simulated datacenter (Zhang et al., 2019). Comparable results are obtained by Ding, Du, and Cerpa who use model predictive path integral control in combination with learning a dynamics model using a neural network to control VAV system temperature setpoints in a five-zone building (Ding et al., 2020). Their proposed model-based controller also reduces the training time compared to a PPO agent by an order of magnitude.

5.2. Expert knowledge

Expert knowledge incorporation in the learning process is the second virtue of an algorithm for RL in single buildings. In some cases, in RL, the training time can be reduced by using information that is available a priori. This is particularly useful for developing control system in building stock with renovation practices. The prior knowledge can take on many forms in BES control. For instance, Vázquez-Canteli, Ulyanin, Kämpf, and Nagy pre-train the action value function network of a DQN-controller for the temperature setpoint of an air-to-water heat pump using the fitted Q-iterations algorithm (Vázquez-Canteli et al., 2019). The algorithm amounts to running the DQN algorithm offline with an experience buffer filled with data from an already existing rule-based controller. Approximately 20 days of historical data are used for offline training. The pre-training gives the DQN a head start, and it performs on the same level as the rule-based controller at deployment. After deployment, learning continues online, and the DQN agent shortly outperforms the rule-based controller in terms of electricity cost. Another approach to pre-training is investigated by Chen et al. A parametrized policy is pre-trained using SL on data from an existing district heating supply-water temperature controller (Chen et al., 2020). It is then deployed and refined using PPO.

Providing the agent with more granular information about the environment is another way of incorporating expert knowledge in an RL algorithm. Du, Li, et al. use multitask learning to jointly learn the separate tasks of heating and cooling a building by controlling the indoor temperature setpoint (Du, Li, et al., 2021). A binary task ID is fed to the policy network of a DDPG agent to signify whether heating or cooling is to be performed. The resulting learning process is twice as fast as the single task implementation. Moreover, restricting the action space $\mathcal{A}$ using the ideas discussed earlier is another way of introducing information about the environment. If certain actions are known a priori to be unreasonable, or even forbidden, they can be eliminated.

6. Reinforcement learning in building clusters

Reusing information from previous building installations to ensure efficient training of new RL agents in BES control is highly desirable. Rather than training an agent from scratch, it can be given a "warm start" using transfer learning from previous successful implementations. However, all buildings are unique; both in terms of envelope and design, climate, heat loss status and internal heat gains. What works in one building might not work in another. Moreover, the state–action space $\mathcal{S} \times \mathcal{A}$ may vary between buildings due to the changing availability of measured state information and controllable variables. As a result, the possibility of transferring information between buildings might diminish because this drastically changes the structure of the problem.

Identical state–action spaces in similar building types can also be imagined. Using digital twins, which are defined as digital representations of the building installation and energy systems, for simulating operation is an important example of this: Training the RL agent in the simulated environment and then deploying it in the real building can be seen as a special case of transferring information between buildings. In that case, the main challenge is instead the discrepancy between the simulated and real-life building performances. But even if the simulated dynamics are validated and consistent, the effects of exogenous variables and disturbances can have a profound impact when transferring a controller to a new environment. The RL controller could as a result be forced into parts of the state–action space unseen during training.

RL in building clusters concerns the performance gained by transferring knowledge from one building to another. Controlling buildings are separate, but related tasks which can be exploited in the training process. It should be noted that RL in building clusters and in single buildings are distinctly different. The end goal is the same–to reduce the training time or increase the final average reward of the RL controller–but the means of achieving it are different. RL in single buildings concerns the utilization of data and knowledge available within a single building. RL in building clusters, on the other hand, focuses on transferring information between buildings, as visualized in Fig. 5. Here, the focus will be on transfer learning for RL applications, but for a more general review on transfer learning for smart buildings, we refer the reader to the publication by Pinto, et al. (2022).

6.1. Transferring policies

Transferring information from one building to another is required to facilitate RL-based BES control in building clusters. However, the information term is ambiguous regarding what is in fact being transferred. If there exists a policy that works well for controlling one building, transferring it to a new building with a different state–action space $\mathcal{S}\times\mathcal{A}$ requires a corresponding change to the policy and value function approximations. It is possible that some learned features about the states can be transferred among buildings and the approximation be fine-tuned for a new task with a new action space. However, to the authors' best knowledge, this remains unexplored in the context of BES control.

Assuming a fixed state–action space is common, but changing dynamics and exogenous variables are still novel challenges. Wei et al. note that despite two buildings being identical, surrounding climate characteristics such as weather patterns have an impact on the RL controller's learning process when applied to VAV-system control (Wei et al., 2017). This suggests that even if an RL agent is trained successfully for one building, it cannot be seamlessly transferred to a similar building in another climate without further fine-tuning. This raises the important question of whether performance can be improved by transferring policies between buildings without re-training from scratch. Du, Zandi, et al. transfer a learned policy to several new simulated buildings with varying thermal mass to investigate generalizability and robustness to changing transition dynamics when controlling indoor temperature setpoints in a building (Du, Zandi, et al., 2021). The transferred RL controller outperforms a rule-based controller in terms of temperature violation-time, with a slight increase in electricity cost for all test buildings. On the other hand, compared to a fixed setpoint

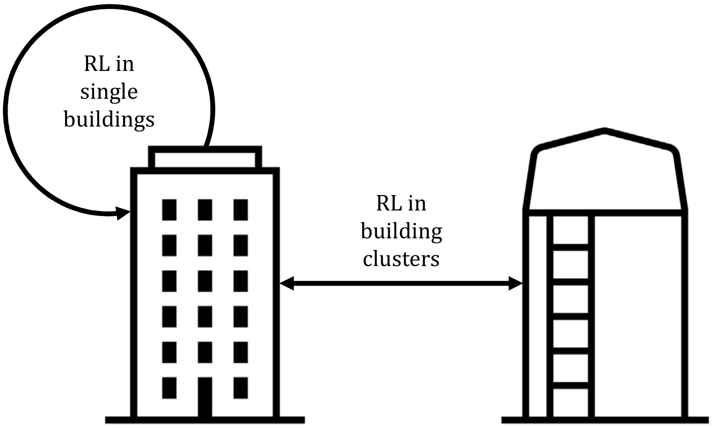

Fig. 5. RL in building clusters concerns the transfer of information between buildings to increase control performance. The arrows represent flow of information.

controller, the RL controller reduces the electricity cost by 15% with a slight increase in temperature violation-time. The issue is further studied by Biemann, Scheller, et al. who investigate the robustness to varying weather conditions when controlling the setpoint of a cooling coil and the fan air flow in a simulated datacenter (Biemann, Scheller, et al., 2021). By randomizing the weather patterns used in each training episode to be one from a few different climates, the resulting agent becomes more robust to such changes. It is also reported that the electricity consumption is consistently reduced by approximately 15% compared to a rule-based controller when evaluated on weather conditions not seen during training. Zhang, et al. perform a limited study of transferring policies between buildings with varying user preferences and appliance parameters (Zhang, et al., 2020). It is found that the training time is reduced by transferring policies when the buildings are similar. The advantages diminish when the buildings are more distinct.

6.2. Transferring dynamics models

In model-based RL, information about the environment can be transferred instead of policies. In a new building, efficient learning of the environment transition dynamics $T(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$ is enabled by transfer learning. On a high level, the idea is that a model of T in an old building can be transferred to new buildings. This would reduce the training time and increase the prediction performance compared to training the model from scratch. Fan, et al. use transfer learning to create predictive models for building energy usage (Fan, et al., 2020). Models are trained on a set of source buildings and are then transferred and fine-tuned on another set of target buildings. Remarkably large performance improvements of up to 60% in terms of relative reduction of the prediction error are observed. The improvements are most apparent when the available datasets at the target buildings are limited in size. Similarly, Qian, Gao, Yang, and Yu use transfer learning to bridge the sim-to-real gap by pre-training an energy prediction model in a simulator and using real building data for fine-tuning (Qian et al., 2020). An increase in prediction performance of 10% can be observed, and the benefit is the highest when only limited real-world measurement data is available.

7. Multi-agent aspects

In the context of a local energy community, it is likely that the energy systems in several buildings need to be synchronized for renewable energy sharing purposes, such as geothermal and photovoltaic production. From an RL perspective, controlling multiple buildings simultaneously translates to a multi-agent control problem. It is a system of many agents interacting simultaneously with a common environment, and with each other. One of the main difficulties concerning Multi-Agent Reinforcement Learning (MARL) is that the environment transition dynamics T depend on the joint action of all agents. Hence, if a single agent does not have information about what the others are doing, the environment will appear non-stationary because, in general, $T(\mathbf{s}'|\mathbf{s}, \mathbf{a}_1, \mathbf{a}_2 \dots \mathbf{a}_N) \neq T(\mathbf{s}'|\mathbf{s}, \mathbf{a}_1, \mathbf{a}'_2 \dots \mathbf{a}'_N)$. Another complication of MARL is the problem of credit assignment. In the case of agents receiving a collective reward, it is not clear how this reward should be divided among the agents. Hence, a single agent can be solely responsible for a high collective reward, despite the other agents not acting optimally. For these reasons, the naive application of the single-agent approaches from the previous sections in a multi-agent setting may be problematic.

There are also practical implications of introducing multi-agent control systems. For instance, rather than installing powerful computational hardware locally in each building, the system designer has the choice of confining the data processing to a single location as depicted in Fig. 6. By transmitting data to a central processing server, the need for investing in hardware for every building is mitigated. However, depending on the quantity of data and the rate of transmission, communication can introduce a major bottleneck, especially for communication intensive methods in RL. As processing power is becoming increasingly affordable and available, using distributed processing on the edge is a viable alternative to central server processing in smart buildings. Furthermore, with companies becoming increasingly restrictive about sharing their collected data, the imposed privacy restrictions can make transmission of raw data impossible.

7.1. Independent- and joint learning

It is imaginable that, given the ever-increasing available computational power, one could formulate the control of many buildings in a network as one big, central RL problem. Accordingly, the joint action of every building in existence would be controlled using a single policy. Under the assumption of not having the communication constraints outlined earlier, this effectively outlines a single-agent problem. However, as emphasized in the section on RL in single buildings, the state–action space would grow exponentially with the number of agents in the system. In the case of representing a policy or value function using a neural network, the training would require vast amounts of data to converge. Wei et al. recognizes this problem in the context of control of the air flow in a VAV-system in a building with multiple thermal zones (Wei et al., 2017). They propose maintaining separate policies for each thermal zone, and thus implicitly formulating it as a multi-agent problem. This transforms the problem into training several neural networks of manageable sizes, which leads to a larger reduction of energy cost and temperature violations compared to the single-agent formulation. The RL control system is evaluated in simulations of three different buildings using weather data from two distinct locations,

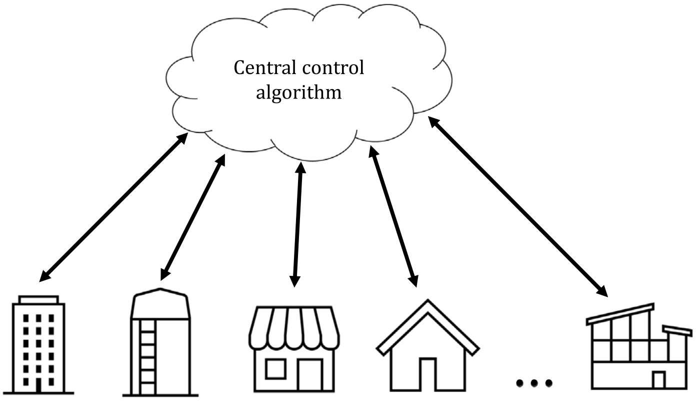

Fig. 6. Illustration of a generic multi-agent system with centralized control. The arrows indicate flow of information.

namely Riverside and Los Angeles. Compared to the single-agent formulation, the multi-agent system consistently achieves a larger relative cost reduction. Compared to a rule-based control of the VAV-system, the multi-agent RL controller gives a reduction of the monetary energy cost of 20%–70%. The wide range of cost reductions are believed to arise due to the difference in weather patterns in the two test locations.

It is important to note the distinct difference between agents learning independently and jointly. It is possible to formulate the multi-agent control problem as each agent solving a task independent of one another, but with a reward function that is dependent on the joint performance of all agents. Vázquez-Canteli et al. define the reward function of each agent to be the sum of the monetary cost of all the agents' individual energy usages (Vázquez-Canteli et al., 2019). In contrast to the purely independent learning case, there is now an incentive for agents to avoid using a lot of energy simultaneously, such that the agents must learn to coordinate. The algorithm is evaluated on a task of controlling the temperature setpoints of air-to-water heat pumps in two large residential building equipped with PV panels. It is demonstrated that the multi-agent formulation leads to lower energy cost compared to rule-based controllers. When photovoltaic arrays are included in only one of the buildings, the multi-agent controller outperforms independent control in terms of energy cost. This is believed to be due to coordination of the photovoltaic production so that the building without photovoltaic panels may consume more electricity from the grid when the other is self-sufficient. Li, Zhang, et al. also study MARL with joint learning for HVAC control, but uses a more sophisticated measure of thermal comfort, namely the Predicted Mean Vote (PMV) metric (Li, Zhang, et al., 2021). When applied to controlling the indoor temperature- and humidity setpoints of a multi-zone laboratory building, it is concluded that MARL outperforms rule-based control by up to 15% in terms of operational energy usage without sacrificing thermal comfort. Moreover, the MARL formulation outperforms the approach of using a single agent to control all actions in terms of the PMV. This again suggests that coordination between agents can be beneficial in BES control.

Finally, the choice of independent or joint learning depends on whether the actions of one agent affect the others. If it is believed that $T(\mathbf{s}'|\mathbf{s},\mathbf{a}_1,\mathbf{a}_2\ldots\mathbf{a}_N) = T(\mathbf{s}'|\mathbf{s},\mathbf{a}_1)$, then independent learning can be performed because there is no need for coordination. However, this would be oversimplifying, as parts of the heating/cooling system might be common to many buildings. Thermal storage and PV production are examples of such common utilities that create the need for coordination. Moreover, if a power or energy limit is persistent in the system, coordination is needed to avoid exceeding this constraint.

7.2. Safety and scalability

Apart from learning efficiency, an especially important aspect of RL algorithms for BES control is their scalability. As the number of buildings in an energy community increases, the control problem should not become intractable. MARL methods have the desirable property of preserving computational tractability, even in systems with many agents. Yu, et al. conclude that MARL is indeed scalable by evaluating their proposed algorithm in a system of 30 agents (Yu, et al., 2021). They draw a conclusion which can be summarized as follows. In a system of 30 agents where each agent has access to 10 discrete actions, using a single-agent formulation–amounting to central control of all actions in the entire system–would yield 10^{30} different actions that can be applied in the system. Such an enormous number of actions would make the problem impossible to solve in practice. When instead formulated as a MARL problem, each agent learns a separate policy, so each agent only needs to consider 10 different actions. This makes the system scalable even as more agents are added.

Another important aspect of safe and reliable heating and cooling systems in local energy communities is the power capacity limit. It is undesirable for agents to require power simultaneously. Even if RL-control of a building cluster leads to a reduced collective energy signature, there may be other undesirable effects related to the power grid supplying electricity to the BES. An example of such a situation is when the heat in the building cluster is supplied by heat pumps. Vazquez-Canteli, Henze, and Nagy study several grid impact factors when applying multi-agent SAC to controlling BES with heat pumps, photovoltaic arrays, and thermal storage (Vazquez-Canteli et al., 2020). It is found that their MARL algorithm shows dramatically reduced peak electrical load compared to rule-based control. The ramping of the power is also reduced, suggesting less sharp peaks. Effects on the power grid are particularly important to consider because power capacity shortage is an increasing problem in general.

8. Future outlook

The research on RL for BES control is undoubtedly still in its early stages. In this review, we have identified the three main computer science related challenges in the area. Each of the three challenges still has unexplored research directions that require attention. In this section, we will highlight such directions that the authors believe are promising and meaningful.

8.1. Reinforcement learning in single buldings

An algorithm with high sample efficiency is crucial in BES control. A promising approach which possesses such a virtue is model-based RL. When using MPC as a basis for model-based RL, the biggest bottleneck is the requirement for an accurate environment model. However, if the learning of the model were to happen online, this problem would be alleviated. Moreover, online learning of a model, integrated with a model-free method for learning a policy in a so-called DYNA scheme, has been successful in many other applications of RL (Feinberg, et al., 2018; Janner, Fu, Zhang, & Levine, 2019; Sutton, 1991). However, it is not clear how detailed the environment dynamics model must be in the case of BES control. BES are simple systems in the sense that one might not need a complicated model to describe their dynamics, but they are susceptible to change over time, nonetheless. This would require some degree of adaptivity in the models used for planning in an RL algorithm.

Pre-training of policies by imitating already existing suboptimal controllers is another approach. There are however issues of such strategies that remain untreated in the context of BES. The most prominent drawback of pre-training is that it might affect exploration when learning the final policy. Therefore, application of the current state of the art in imitation learning along with systematic empirical studies of pre-training would also provide valuable insights.

8.2. Reinforcement learning in building clusters

Transferring information between buildings is highly desirable in practice and deserves more attention in the future. However, there are major obstacles like differing state–action spaces between buildings, making such a process difficult. Even under the assumption of a fixed state–action space, there is need for algorithms that produce generally capable agents that can be deployed with reasonable results “out of the box”. Fast fine-tuning of the policy would then be required to obtain optimal behaviour in as little time as possible. A promising approach to obtain agents that are robust to changing environment dynamics is domain randomization (Tobin, et al., 2017). Such a training framework would be useful when making the sim-to-real transfer–i.e., training an agent in a simulation and then deploying it in the real word–due to the mismatch between simulation and reality. Another approach is to use meta-RL to train agents with the goal of quickly adapting to new tasks, rather than being good at one task (Finn, Abbeel, & Levine, 2017; Nichol, Achiam, & Schulman, 2018).

In connection to the topic of model-based RL, it would be desirable to use transfer learning when learning a model of the environment dynamics. Doing so could potentially alleviate some scalability issues of, e.g., MPC. Comprehensive investigations into the effects of transferring dynamics models as well as entire policies between buildings would be beneficial. The ideas of meta-learning can potentially be useful here as well.

8.3. Multi-agent aspects

Multi-agent RL is a promising approach to control communities with multiple buildings and allows for coordination to reduce the collective energy usage on the community level. The popular approach of defining a common reward function to incentivize coordination between agents does however have its limits. For instance, defining the reward as the sum of individual rewards can create unfairness among the agents. Further research into reward constellations that take fairness into account are needed to ensure proper functionality in a multi-agent system. Another topic that requires more attention is that coordination between agents is not always possible. It can be due to the lack of a central server for data processing or privacy constraints which makes the transmission of raw data impossible. Hence, there should be investigation into the effects on the collective performance of having completely decentralized control. Such a scenario would be especially important to consider in the context of power grid effects such as peak loads.

8.4. Theoretical guarantees

The past research on RL for BES control is to a considerable extent empirical. But to gain further intuition about the challenges of sample efficiency and performance in terms of the final reward, those problems should be studied theoretically. One of the important theoretical questions that needs answering is whether a bound can be placed on the sample complexity of an RL algorithm under reasonable assumptions. Such a bound would offer insight into the number of environment interactions needed to guarantee a certain expected reward. Furthermore, it is also desirable to upper bound the performance degradation in terms of the expected reward of transferring an agent to from one building to another when the buildings have different transition dynamics. If such a bound were obtained, it would be helpful when studying the transfer of policies between BES.

8.5. Explainability

There exists an increasing body of literature on attempting to explain or visualize why RL agents take the actions that they do (Gunning, et al., 2019; Puiutta & Veith, 2020). This interest is not purely academic, but a strong demand from industry is also persistent due to the ethical–and sometimes legal–obligations to motivate the actions taken in autonomous systems. BES are no exception, and studies on which features are influential when controlling them would provide a better understanding of the problem. Moreover, explainability could aid in feature selection when designing the state space in RL for BES control, which could in turn improve the sample efficiency of the algorithms.

8.6. System design

Apart from RL-methodology and theory, there is the issue of designing a scalable computation infrastructure to process measurement data in BES. Today, measurements are typically processed locally–e.g., in a microcontroller–on the building premise. An alternative approach is to instead transmit the measurements to a datacenter for processing and storage in the cloud. Such data collection opens for the use of online processing using RL for controlling the BES without the need for installing computational hardware in the building itself. On the other hand, this would require a channel of communication between the sensors and actuators in the BES, and the cloud provider. Therefore, there is a demand for interoperability between all system components, which is a major challenge in practice.

Moreover, when controlling BES, data naturally arrives sequentially in time as a stream. In the case of controlling a cluster of buildings concurrently, data could arrive asynchronously from any of the buildings and would require fast processing. Moreover, since each building could have unique control logic, the computation cannot be statically allocated but must be dynamic. Considering the design of the data processing pipeline is an important future research direction, as it is crucial for making large-scale control possible.

9. Conclusions

In this review, we have identified the main computer science related challenges in BES control using RL. They grouped into three categories: RL in single buildings, RL in building clusters, and multi-agent aspects. The main conclusions that can be drawn from this review in connection to the three main categories are the following:

- There exist empirical studies of model-based RL methods which show promising results in terms of sample efficiency. However, the requirement for historical data to model the building and the energy systems system remains problematic.
- Transferring policies between buildings with similar state-transition dynamics has empirically shown indications to be highly effective in BES control.
- Multi-agent RL is a promising framework for controlling building clusters when shared utilities such as thermal storage and on-site electricity production are present. However, the challenges of sample efficiency persist.
- There is a shortage of theoretical analysis of the problems encountered in RL for BES control.

Declaration of competing interest

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

Data availability

No data was used for the research described in the article.

References

Afram, A., & Janabi-Sharifi, F. (2014). Theory and applications of HVAC control systems – A review of model predictive control (MPC). *Building and Environment*, *72*, 343–355. http://dx.doi.org/10.1016/j.buildenv.2013.11.016, URL: https://linkinghub.elsevier.com/retrieve/pii/S0360132313003363.

Afram, A., Janabi-Sharifi, F., Fung, A. S., & Raahemifar, K. (2017). Artificial neural network (ANN) based model predictive control (MPC) and optimization of HVAC systems: A state of the art review and case study of a residential HVAC system. *Energy and Buildings*, *141*, 96–113. http://dx.doi.org/10.1016/j.enbuild.2017.02.012, URL: https://linkinghub.elsevier.com/retrieve/pii/S0378778816310799.

Biemann, M., Liu, X., Zeng, Y., & Huang, L. (2021). Addressing partial observability in reinforcement learning for energy management. In *Proceedings of the 8th ACM international conference on systems for energy-efficient buildings, cities, and transportation* (pp. 324–328). Coimbra Portugal: ACM, http://dx.doi.org/10.1145/3486611.3488730, URL: https://dl.acm.org/doi/10.1145/3486611.3488730.

Biemann, M., Scheller, F., Liu, X., & Huang, L. (2021). Experimental evaluation of model-free reinforcement learning algorithms for continuous HVAC control. *Applied Energy*, *298*, Article 117164. http://dx.doi.org/10.1016/j.apenergy.2021.117164, URL: https://linkinghub.elsevier.com/retrieve/pii/S0306261921005961.

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*, Article 562239. http://dx.doi.org/10.3389/fbuil.2020.562239, URL: https://www.frontiersin.org/articles/10.3389/fbuil.2020.562239/full.

Chen, Y., Norford, L. K., Samuelson, H. W., & Malkawi, A. (2018). Optimal control of HVAC and window systems for natural ventilation through reinforcement learning. *Energy and Buildings*, *169*, 195–205. http://dx.doi.org/10.1016/j.enbuild.2018.03.051, URL: https://linkinghub.elsevier.com/retrieve/pii/S0378778818302184.

Ding, X., Du, W., & Cerpa, A. E. (2020). MB2C: Model-based deep reinforcement learning for multi-zone building control. In *Proceedings of the 7th ACM international conference on systems for energy-efficient buildings, cities, and transportation* (pp. 50–59). Virtual Event Japan: ACM, http://dx.doi.org/10.1145/3408308.3427986, URL: https://dl.acm.org/doi/10.1145/3408308.3427986.

Drgoňa, J., Arroyo, J., Cupeiro Figueroa, I., Blum, D., Arendt, K., Kim, D., et al. (2020). All you need to know about model predictive control for buildings. *Annual Reviews in Control*, *50*, 190–232. http://dx.doi.org/10.1016/j.arcontrol.2020.09.001, URL: https://linkinghub.elsevier.com/retrieve/pii/S1367578820300584.

Du, Y., Li, F., Munk, J., Kurte, K., Kotevska, O., Amasyali, K., et al. (2021). Multi-task deep reinforcement learning for intelligent multi-zone residential HVAC control. *Electric Power Systems Research*, *192*, Article 106959. http://dx.doi.org/10.1016/j.epsr.2020.106959, URL: https://linkinghub.elsevier.com/retrieve/pii/S0378779620307574.

Du, Y., Zandi, H., Kotevska, O., Kurte, K., Munk, J., Amasyali, K., et al. (2021). Intelligent multi-zone residential HVAC control strategy based on deep reinforcement learning. *Applied Energy*, *281*, Article 116117. http://dx.doi.org/10.1016/j.apenergy.2020.116117, URL: https://linkinghub.elsevier.com/retrieve/pii/S030626192031535X.

European Commission (2021). 2030 Digital compass: the European way for the digital decade. European Commission. URL: https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:52021DC0118.

Fan, C., Sun, Y., Xiao, F., Ma, J., Lee, D., Wang, J., et al. (2020). Statistical investigations of transfer learning-based methodology for short-term building energy predictions. *Applied Energy*, *262*, Article 114499. http://dx.doi.org/10.1016/j.apenergy.2020.114499, URL: https://linkinghub.elsevier.com/retrieve/pii/S0306261920300118.

Feinberg, V., Wan, A., Stoica, I., Jordan, M. I., Gonzalez, J. E., & Levine, S. (2018). Model-based value estimation for efficient model-free reinforcement learning. [Cs, Stat] arXiv:1803.00101. URL: http://arxiv.org/abs/1803.00101.

Finn, C., Abbeel, P., & Levine, S. (2017). Mode-agnostic meta-learning for fast adaptation of deep networks. [Cs] arXiv:1703.03400. URL: http://arxiv.org/abs/1703.03400.

Frederiksen, S., & Werner, S. (2013). *District heating and cooling*. Lund: Studentlitteratur AB.

Gao, G., Li, J., & Wen, Y. (2020). DeepComfort: energy-efficient thermal comfort control in buildings via reinforcement learning. *IEEE Internet of Things Journal*, *7*(9), 8472–8484. http://dx.doi.org/10.1109/JIOT.2020.2992117, URL: https://ieeexplore.ieee.org/document/9085925/.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics*, *4*(37), eaay7120. http://dx.doi.org/10.1126/scirobotics.aay7120, URL: https://www.science.org/doi/10.1126/scirobotics.aay7120.

Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor. URL: http://arxiv.org/abs/1801.01290. [Cs, Stat] arXiv:1801.01290.

Han, M., May, R., Zhang, X., Wang, X., Pan, S., Yan, D., et al. (2019). A review of reinforcement learning methodologies for controlling occupant comfort in buildings. *Sustainable Cities and Society*, *51*, Article 101748. http://dx.doi.org/10.1016/j.scs.2019.101748, URL: https://linkinghub.elsevier.com/retrieve/pii/S2210670719307589.

Hong, T., Wang, Z., Luo, X., & Zhang, W. (2020). State-of-the-art on research and applications of machine learning in the building life cycle. *Energy and Buildings*, *212*, Article 109831. http://dx.doi.org/10.1016/j.enbuild.2020.109831, URL: https://linkinghub.elsevier.com/retrieve/pii/S0378778819337879.

International Energy Agency (2022). *Buildings*. IEA, URL: https://www.iea.org/reports/buildings.

Janner, M., Fu, J., Zhang, M., & Levine, S. (2019). When to trust your model: model-based policy optimization. In *Proceedings of the 33rd international conference on neural information processing systems* (1122), (pp. 12519–12530). Red Hook, NY, USA: Curran Associates Inc.

Jiang, Z., Risbeck, M. J., Ramamurti, V., Murugesan, S., Amores, J., Zhang, C., et al. (2021). Building HVAC control with reinforcement learning for reduction of energy cost and demand charge. *Energy and Buildings*, *239*, Article 110833. http://dx.doi.org/10.1016/j.enbuild.2021.110833, URL: https://linkinghub.elsevier.com/retrieve/pii/S0378778821001171.

Jin, X., Baker, K., Christensen, D., & Isley, S. (2017). Foresee: A user-centric home energy management system for energy efficiency and demand response. *Applied Energy*, *205*, 1583–1595. http://dx.doi.org/10.1016/j.apenergy.2017.08.166, URL: https://linkinghub.elsevier.com/retrieve/pii/S0306261917311856.

Kurte, K., Amasyali, K., Munk, J., & Zandi, H. (2021). Comparative analysis of model-free and model-based HVAC control for residential demand response. In *Proceedings of the 8th ACM international conference on systems for energy-efficient buildings, cities, and transportation* (pp. 309–313). Coimbra Portugal: ACM, http://dx.doi.org/10.1145/3486611.3488727, URL: https://dl.acm.org/doi/10.1145/3486611.3488727.

Lee, Z. E., & Zhang, K. M. (2021). Scalable identification and control of residential heat pumps: A minimal hardware approach. *Applied Energy*, *286*, Article 116544. http://dx.doi.org/10.1016/j.apenergy.2021.116544, URL: https://linkinghub.elsevier.com/retrieve/pii/S0306261921000945.

Levermore, G. J. (2000). *Building energy management systems: applications to low energy HVAC and natural ventilation control* (2nd ed.). London ; New York: E & FN Spon.

Li, Y., O'Neill, Z., Zhang, L., Chen, J., Im, P., & DeGraw, J. (2021). Grey-box modeling and application for building energy simulations - A critical review. *Renewable and Sustainable Energy Reviews*, *146*, Article 111174. http://dx.doi.org/10.1016/j.rser.2021.111174, URL: https://linkinghub.elsevier.com/retrieve/pii/S1364032121004639.

Li, J., Zhang, W., Gao, G., Wen, Y., Jin, G., & Christopoulos, G. (2021). Toward intelligent multizone thermal control with multiagent deep reinforcement learning. *IEEE Internet of Things Journal*, *8*(14), 11150–11162. http://dx.doi.org/10.1109/JIOT.2021.3051400, URL: https://ieeexplore.ieee.org/document/9321466/.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., et al. (2019). Continuous control with deep reinforcement learning. [Cs, Stat] arXiv:1509.02971. URL: http://arxiv.org/abs/1509.02971.

Manjarres, D., Mera, A., Perea, E., Lejarazu, A., & Gil-Lopez, S. (2017). An energy-efficient predictive control for HVAC systems applied to tertiary buildings based on regression techniques. *Energy and Buildings*, *152*, 409–417. http://dx.doi.org/10.1016/j.enbuild.2017.07.056, URL: https://linkinghub.elsevier.com/retrieve/pii/S037877881731321X.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., et al. (2013). Playing atari with deep reinforcement learning. arXiv:1312.5602.

Nagy, A., Kazmi, H., Cheaib, F., & Driesen, J. (2018). Deep reinforcement learning for optimal control of space heating. arXiv:1805.03777.

Nichol, A., Achiam, J., & Schulman, J. (2018). On first-order meta-learning algorithms. [Cs] arXiv:1803.02999. URL: http://arxiv.org/abs/1803.02999.

Nweye, K., Liu, B., Stone, P., & Nagy, Z. (2022). Real-world challenges for multi-agent reinforcement learning in grid-interactive buildings. http://dx.doi.org/10.1016/j.egyai.2022.100202. URL: http://arxiv.org/abs/2112.06127. [cs, eess] arXiv:2112.06127.

Pinto, G., Wang, Z., Roy, A., Hong, T., & Capozzoli, A. (2022). Transfer learning for smart buildings: A critical review of algorithms, applications, and future perspectives. *Advances in Applied Energy*, *5*, Article 100084. http://dx.doi.org/10.1016/j.adapen.2022.100084, URL: https://linkinghub.elsevier.com/retrieve/pii/S2666792422000026.

Puiutta, E., & Veith, E. M. S. P. (2020). Explainable reinforcement learning: A survey. In A. Holzinger, P. Kieseberg, A. M. Tjoa, & E. Weippl (Eds.), *Machine learning and knowledge extraction. Vol. 12279* (pp. 77–95). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-57321-8_5, URL: http://link.springer.com/10.1007/978-3-030-57321-8_5.

Qian, F., Gao, W., Yang, Y., & Yu, D. (2020). Potential analysis of the transfer learning model in short and medium-term forecasting of building HVAC energy consumption. *Energy*, *193*, Article 116724. http://dx.doi.org/10.1016/j.energy.2019.116724, URL: https://linkinghub.elsevier.com/retrieve/pii/S0360544219324193.

Royapoor, M., Antony, A., & Roskilly, T. (2018). A review of building climate and plant controls, and a survey of industry perspectives. *Energy and Buildings*, *158*, 453–465. http://dx.doi.org/10.1016/j.enbuild.2017.10.022, URL: https://linkinghub.elsevier.com/retrieve/pii/S0378778817318522.

Salsbury, T. I. (2005). A survey of control technologies in the building automation industry. *IFAC Proceedings Volumes*, *38*(1), 90–100. http://dx.doi.org/10.3182/20050703-6-CZ-1902.01397, URL: https://linkinghub.elsevier.com/retrieve/pii/S1474667016374092.

Schulman, J., Levine, S., Moritz, P., Jordan, M. I., & Abbeel, P. (2017). Trust region policy optimization. URL: http://arxiv.org/abs/1502.05477. [Cs] arXiv:1502.05477.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. URL: http://arxiv.org/abs/1707.06347. [Cs] arXiv:1707.06347.

Sofos, M., Langevin, J., Deru, M., Gupta, E., Benne, K., Blum, D., et al. (2020). *Innovations in sensors and controls for building energy management: research and development opportunities report for emerging technologies*: *Technical Report NREL/TP–5500-75601, DOE/GO–102019-5234, 1601591*, http://dx.doi.org/10.2172/1601591, URL: https://www.osti.gov/servlets/purl/1601591/.

Sutton, R. S. (1991). Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin*, *2*(4), 160–163. http://dx.doi.org/10.1145/122344.122377, URL: https://dl.acm.org/doi/10.1145/122344.122377.

Sutton, R. S., & Barto, A. G. (2018). *Adaptive computation and machine learning series, Reinforcement learning: an introduction* (2nd ed.). Cambridge, Massachusetts: The MIT Press.

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., & Abbeel, P. (2017). Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems* (pp. 23–30). Vancouver, BC: IEEE, http://dx.doi.org/10.1109/IROS.2017.8202133, URL: http://ieeexplore.ieee.org/document/8202133/.

Tymkow, P., Tassou, S., Kolokotroni, M., & Jouhara, H. (2020). *Building services design for energy efficient buildings* (2nd ed.). New York: Routledge.

Vazquez-Canteli, J. R., Henze, G., & Nagy, Z. (2020). MARLISA: Multi-agent reinforcement learning with iterative sequential action selection for load shaping of grid-interactive connected buildings. In *Proceedings of the 7th ACM international conference on systems for energy-efficient buildings, cities, and transportation* (pp. 170–179). Virtual Event Japan: ACM, http://dx.doi.org/10.1145/3408308.3427604, URL: https://dl.acm.org/doi/10.1145/3408308.3427604.

Vázquez-Canteli, J. R., & Nagy, Z. (2019). Reinforcement learning for demand response: A review of algorithms and modeling techniques. *Applied Energy*, *235*, 1072–1089. http://dx.doi.org/10.1016/j.apenergy.2018.11.002, URL: https://linkinghub.elsevier.com/retrieve/pii/S0306261918317082.

Vázquez-Canteli, J. R., Ulyanin, S., Kämpf, J., & Nagy, Z. (2019). Fusing TensorFlow with building energy simulation for intelligent energy management in smart cities. *Sustainable Cities and Society*, *45*, 243–257. http://dx.doi.org/10.1016/j.scs.2018.11.021, URL: https://linkinghub.elsevier.com/retrieve/pii/S2210670718314380.

Wang, Z., & Hong, T. (2020). Reinforcement learning for building controls: The opportunities and challenges. *Applied Energy*, *269*, Article 115036. http://dx.doi.org/10.1016/j.apenergy.2020.115036, URL: https://linkinghub.elsevier.com/retrieve/pii/S0306261920305481.

Wei, T., Wang, Y., & Zhu, Q. (2017). Deep reinforcement learning for building HVAC control. In *Proceedings of the 54th annual design automation conference 2017* (pp. 1–6). Austin TX USA: ACM, http://dx.doi.org/10.1145/3061639.3062224, URL: https://dl.acm.org/doi/10.1145/3061639.3062224.

Yu, Y. (2018). Towards sample efficient reinforcement learning. In *Proceedings of the twenty-seventh international joint conference on artificial intelligence* (pp. 5739–5743). Stockholm, Sweden: International Joint Conferences on Artificial Intelligence Organization, http://dx.doi.org/10.24963/ijcai.2018/820, URL: https://www.ijcai.org/proceedings/2018/820.

Yu, L., Sun, Y., Xu, Z., Shen, C., Yue, D., Jiang, T., et al. (2021). Multi-agent deep reinforcement learning for HVAC control in commercial buildings. *IEEE Transactions on Smart Grid*, *12*(1), 407–419. http://dx.doi.org/10.1109/TSG.2020.3011739, URL: https://ieeexplore.ieee.org/document/9146920/.

Zhang, X., Jin, X., Tripp, C., Biagioni, D. J., Graf, P., & Jiang, H. (2020). Transferable reinforcement learning for smart homes. In *Proceedings of the 1st international workshop on reinforcement learning for energy management in buildings & cities* (pp. 43–47). Virtual Event Japan: ACM, http://dx.doi.org/10.1145/3427773.3427865, URL: https://dl.acm.org/doi/10.1145/3427773.3427865.

Zhang, C., Kuppannagari, S. R., Kannan, R., & Prasanna, V. K. (2019). Building HVAC scheduling using reinforcement learning via neural network based model approximation. In *Proceedings of the 6th ACM international conference on systems for energy-efficient buildings, cities, and transportation* (pp. 287–296). New York NY USA: ACM, http://dx.doi.org/10.1145/3360322.3360861, URL: https://dl.acm.org/doi/10.1145/3360322.3360861.