Data-Enabled Optimization of Building Operations

by

Tianyu Zhang

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Computing Science

University of Alberta

Abstract

Retrofitting buildings and optimizing their operation have been at the forefront of global efforts to reduce carbon emissions for the past few decades. Intelligent control of building systems, such as Heating, Ventilation, and Air Conditioning (HVAC), presents two clear benefits: it improves human comfort, and reduces energy consumption and carbon emissions. However, the complex interplay between various building systems, coupled with the high costs associated with data collection, poses a significant challenge for developing accurate models and control schemes that rely on data or building models. To address these challenges, this thesis explores the use of readily available data to (a) train control agents capable of striking an acceptable balance between energy consumption, and thermal and visual comfort of the occupants; (b) evaluate a diverse population of control agents to find the most suitable one for transfer to a new building; (c) establish accurate personal thermal comfort models; (d) learn complex building dynamics; (e) assign space to occupants such that a better trade-off between energy consumption and thermal comfort can be achieved.

To facilitate training and evaluation of learning-based controllers, we implement an open-source simulation platform in Python. This platform, called COBS, enables modeling occupant behavior and learning a control policy in an online fashion by interacting with building systems in simulation. The first contribution of this thesis is learning a near-optimal policy for controlling a subset of actuators and setpoints that are part of multiple building

systems, namely HVAC, shading, and lighting systems, leveraging a model-free Reinforcement Learning (RL) algorithm. We show that this controller achieves a better trade-off between energy consumption and human comfort than controllers that are widely used in commercial buildings today. A notable extension is the introduction of a Multi-Agent Reinforcement Learning (MARL)-based HVAC control policy training and offline evaluation framework. With an emphasis on policy and environment diversity, this framework harnesses the power of transfer learning to ensure robust performance across various buildings, even without retraining.

The next key contribution of this thesis is the introduction of personal comfort models and proposing a data-efficient approach for training these models. Specifically, the models are trained using weak labels derived from occupants' interactions with a Personal Comfort System (PCS), reducing both the need for direct occupant engagement and potential subjective biases. To reduce the training cost for individuals who lack prior data, several group comfort models are trained and combined using an ensemble method.

Lastly, the thesis presents novel and efficient algorithms that can be adopted in a workspace reservation system that assigns desks to long-term and short-term occupants in shared workspaces. These algorithms assign occupants with similar thermal preferences to a zone that fulfills their thermal comfort requirements in the most energy-efficient manner. This leads to a more energy-efficient HVAC operation, while ensuring the satisfaction of thermal comfort constraints.

Collectively, the above contributions could greatly enhance building operations and reduce the carbon footprint of the building sector without requiring expensive retrofits.

Preface

This thesis is an original work by Tianyu Zhang. Some of the chapters are based on conference and journal publications co-authored by the author of this thesis. We list these publications below:

- A part of Chapter 3, in particular Section 3.3, is based on the following poster presentation:

  T. Zhang, O. Ardakanian, "Poster Abstract: COBS: COmprehensive Building Simulator", In *Proceedings of the 7th ACM Conference on Systems for Built Environments*, BuildSys, ACM, November 2020.

  T.Z. designed and implemented the toolkit. T.Z. and O.A. edited the manuscript.

- Chapter 4 is based on this conference paper:

  T. Zhang*, G. Baasch*, O. Ardakanian, R. Evins, "On the Joint Control of Multiple Building Systems with Reinforcement Learning", In *Proceedings of the 12th ACM International Conference on Future Energy Systems*, eEnergy, pp.60–72, ACM, June 2021.

  T.Z. contributed to the conceptualization of the work, implemented the control agents, conducted the experiments, and produced the results. G.B. undertook the literature review, implemented the control agents, and conducted the experiments. O.A. offered technical advice and ideas and advised on T.Z.'s research. R.E. advised on G.B.'s research. All authors participated in drafting and revising the manuscript.

- Chapter 5 is mostly based on the following two papers:

*Equal contribution.

T. Zhang, AK GS, M. Afshari, P. Musilek, M. E. Taylor, O. Ardakanian, "Diversity for Transfer in Learning-based Control of Buildings", In *Proceedings of the 13th ACM International Conference on Future Energy Systems*, eEnergy, pp.556–564, ACM, June 2022.

T.Z. formulated the methodology, executed the experiments, and generated the results. AK.GS. conducted the literature review. O.A. and M.T. provided technical advice and verified proofs. All authors contributed to the preparation of the manuscript.

AK GS*, T. Zhang*, O. Ardakanian, M. E. Taylor, "Mitigating an Adoption Barrier of Reinforcement Learning-Based Control Strategies in Buildings", *Energy and Buildings*, vol.285, 112878, Elsevier, April 2023.

AK.GS. designed and implemented the policy ranking methods and conducted the literature review. T.Z. proposed the framework, assisted with the literature review, designed the experiments, and produced the results. O.A. contributed to the experiment, design and provided proof verification and intuitions. M.T. offered technical advice. All authors edited the manuscript.

- Chapter 6 is based on this journal presentation:

  T. Zhang, J. Gu, O. Ardakanian, J. Kim, "Addressing Data Inadequacy Challenges in Personal Comfort Models by Combining Pretrained Comfort Models", *Energy and Buildings*, vol.264, 112068, Elsevier, June 2022.

  T.Z. conducted the literature review, developed the methodology, conducted the experiments, and produced the results. J.G. assisted in visualizing the results. O.A. proposed the research idea and advised T.Z.'s research. J.K. provided the data set and offered intuitions and technical advice. All authors participated in editing the manuscript.

- Chapter 7 is mostly taken from this unpublished work:

  T. Zhang, O. Ardakanian, "Revisiting Space Planning in Coworking

Spaces", Submitted to *Proceedings of the 15th ACM International Conference on Future Energy Systems*, eEnergy, 2024, under review.

T.Z. undertook the literature review, designed and implemented the methodology, conducted the experiments, generated the results, and edited the manuscript. O.A. contributed to the conceptualization of the work, advised on the project, and edited the manuscript.

Additional publications: Below is a brief overview of papers that were published or submitted during the PhD but have not been included in this thesis:

- T. Zhang, O. Ardakanian, "Investigating the Impact of Space Allocation Strategy on Energy-Comfort Trade-off in Office Buildings", In *Companion Proceedings of the 14th ACM International Conference on Future Energy Systems*, eEnergy, pp.145–149, ACM, June 2023.

  This paper investigates the impact of the space assignment strategy on the energy-comfort trade-off in office buildings and whether it depends on specific building characteristics. Our simulation shows that varying the space assignment strategy in a medium office building can lead to over 3.5%/15.1% change in annual/monthly energy consumption, and over 15% change in average thermal comfort when using the personal comfort model. This finding motivates the joint optimization of HVAC operation and space planning, possibly at different timescales.

- A. Zhumabekov, D. May, T. Zhang, AK GS, O. Ardakanian, M. E. Taylor, "Ensembling Diverse Policies Improves Generalizability of Reinforcement Learning Algorithms in Continuous Control Tasks", In *Proceedings of the Adaptive and Learning Agents Workshop*, ALA, 9 pages, ACM, May 2023.

  This paper introduces a simple ensembling technique for DRL policies with a continuous action space. It aggregates actions by performing weighted averaging based on the uncertainty levels of the policies. We investigate its zero-shot generalization properties in a complex continuous

control domain: the optimal control of home batteries in the CityLearn environment — the subject of a 2022 international AI competition. Our results indicate that the proposed ensemble has better generalization capacity than a single policy. Further, we show that promoting diversity among policies during training can reliably improve the zero-shot performance of the ensemble in the test phase. Finally, we examine the merits of the uncertainty-based weighted averaging in an ensemble by comparing it to two alternative approaches: unweighted averaging and selecting the action of the least uncertain policy.

- T. Zhang, A. Banitalebi-Dehkordi, Y. Zhang, "Deep Reinforcement Learning for Exact Combinatorial Optimization: Learning to Branch", In *Proceedings of the 26th International Conference on Pattern Recognition*, ICPR, pp.3105–3111, IEEE, August 2022.

  This paper proposes a new approach for solving the data labeling and improving solving time in combinatorial optimization based on the use of the RL paradigm. We use imitation learning to bootstrap an RL agent and then PPO to further explore global optimal actions. Then, a value network is used to run Monte-Carlo tree search (MCTS) to enhance the policy network. We evaluate the performance of our method on four different categories of combinatorial optimization problems and show that our approach performs strongly compared to the state-of-the-art machine learning and heuristics-based methods.

- MM. Hossain*, T. Zhang*, O. Ardakanian, "Identifying grey-box thermal models with Bayesian neural networks", *Energy and Buildings*, vol.238, pp.1–11, Elsevier, 2021.

  This paper explores various techniques for establishing a suitable thermal model using time-series data generated by smart thermostats. We show that Bayesian neural networks can be used to estimate parameters of a grey-box thermal model if sufficient training data is available, and this model outperforms several black-box models in terms of the temperature

prediction accuracy. Leveraging real data from 8,884 homes equipped with smart thermostats, we discuss how the prior knowledge about the model parameters can be utilized to quickly build an accurate thermal model for another home with similar floor area and age in the same climate zone. Moreover, we investigate how to adapt the model originally built for the same home in another season using a small amount of data collected in this season. Our results confirm that maintaining only a small number of pre-trained thermal models will suffice to quickly build accurate thermal models for many other homes, and that 1 day of smart thermostat data could significantly improve the accuracy of the models transferred to another season.

Amidst life's many turns,
a guiding light did appear,
both a teacher and friend,
holding me dear.

The journey of a thousand miles begins with one step.

– Lao Tzu, Tao-te Ching.

Acknowledgements

I am greatly indebted to my supervisor, Professor Omid Ardakanian, for his unwavering support throughout my PhD journey and carefully reviewing my thesis. His guidance has been invaluable. He walked me through all the stages of writing this thesis and the papers we published before. Without his patient instruction and expert guidance, the completion of this thesis would be impossible. His profound insights and expertise have significantly shaped my research, and his encouragement and mentorship have been instrumental in my personal and professional growth. Beyond the confines of academia, I am immensely grateful for his understanding and support in various facets of life. His belief in my potential has been a driving force, and I am truly fortunate to have had him as my mentor. Thank you, Dr. Ardakanian, for everything.

I would also like to extend my appreciation to all my co-authors with whom I had the privilege to collaborate on various publications during my PhD journey. Their diverse perspectives, expertise, and constructive feedback have enriched our collective work, pushing the boundaries of our research. Additionally, I wish to acknowledge the countless others who have offered their assistance, insights, and encouragement along the way. Their contributions, both big and small, have left an indelible mark on my academic journey, and for that, I am eternally grateful.

Lastly, I must express my gratitude to my family and my girlfriend. Their sacrifices, encouragement, and endless support have been the foundation upon which all my achievements stand. To them, I owe more than words can express.

Contents

List of Tables

List of Figures

Glossary

ABW Activity-Based Workplace

AHU Air Handling Unit

ANN Artificial Neural Network

ANOVA Analysis of Variance

ARMAX AutoRegressive Moving Average model with exogenous variables

ARX AutoRegressive model with exogenous variables

ASHRAE American Society of Heating, Refrigerating, and Air-Conditioning Engineers

B&B Branch-and-Bound

BCVTB Building Controls Virtual Test Bed

BDQN Branching Dueling Q-Network

BMS Building Management System

CBECS Commercial Buildings Energy Consumption Survey

CDF Cumulative Distribution Function

CFD Computational Fluid Dynamics

CNN Convolutional Neural Network

COBS COmprehensive Building Simulator

DDPG Deep Deterministic Policy Gradient

DNN Deep Neural Network

DQN Deep Q-Network

EN European Standard

FCFS First-Come First-Served

FDR False Discovery Rate

FQE Fitted-Q Evaluation

GK Gaussian Kernel

HVAC Heating, Ventilation, and Air Conditioning

ICNN Input Convex Neural Network

IDF Intermediate Data Format

IPW Inverse Probability Weighting

IQR Interquartile Range

ISO International Organization for Standardization

JSON JavaScript Object Notation

LM Levenberg-Marquardt

LR Linear Regression

LSTM Long Short Term Memory

MAPE Mean Absolute Percentage Error

MARL Multi-Agent Reinforcement Learning

MDP Markov Decision Process

MINLP Mixed-Integer Nonlinear Programming

MIP Mixed Integer Programming

MMDP Multi-agent Markov Decision Process

MPC Model Predictive Control

MSE Mean Squared Error

NAS Neural Architecture Search

OPE Off-policy Policy Evaluation

PCS Personal Comfort System

PID Proportional–Integral–Derivative

PIR Passive Infrared Sensor

PMV Predicted Mean Vote

PPD Predicted Percentage of Dissatisfied

PPO Proximal Policy Optimization

RBC Rule-based Controller

RBF Radial Basis Function

RC Resistor-Capacitor model

RF Random Forest

RL Reinforcement Learning

SAC Soft Actor-Critic

SARL Single-agent Reinforcement Learning

SAT Supply Air Temperature

SCIP Solving Constraint Integer Programs

SMOTE Synthetic Minority Oversampling Technique

SNIP Single-shot Network Pruning

SNIPW Self-Normalized Inverse Probability Weighting

SVM Support Vector Machine

VAV Variable Air Volume

ZCP Zero-cost Proxy

Chapter 1

Introduction

In North America, people spend approximately 87% of their time indoors [87]. This emphasizes the importance of providing a safe and comfortable indoor environment by investing in advanced sensing and control technologies and installing them in buildings. Building systems, including Heating, Ventilation, and Air Conditioning (HVAC), lighting, and shading systems, play a crucial role in satisfying individual comfort requirements, but they also consume a significant amount of energy. Remarkably, building operations contribute to one-third of the world's final energy usage and 26% of global energy-related emissions – with 8% directly from buildings and 18% indirectly from electricity production and heat used in buildings [70]. The ongoing trend of urbanization further increases the need for buildings, growing energy demand and associated carbon emissions.

Enhancing control mechanisms has been recognized for its significant potential for improving building energy efficiency and cutting carbon emissions related to building operations. Reports from industry indicate that the implementation of advanced control systems can lead to energy or cost savings of up to 30% [15]. Furthermore, previous studies suggest that inadequately controlled building systems might be responsible for energy waste amounting to 30 to 50% in commercial buildings [79], [105]. In traditional building control systems, building managers define high-level setpoint schedules and control rules for feedback controllers, such as Proportional-controllers or Proportional–Integral–Derivative (PID) controllers, in various building systems [167]. For this reason, they are commonly called Rule-based Controllers

(RBCs). These schedules are often defined based on the subjective judgment of the building managers and might be adjusted according to feedback from building occupants. However, such controllers generally struggle to maintain thermal comfort of the occupants or lead to higher energy consumption [44]. Therefore, there is an urgent need for energy-efficient, occupant-centric building control strategies that save energy, reduce carbon emissions, and maintain thermal and visual comfort of the occupants.

This thesis is dedicated to *exploring the potential benefits of developing and integrating data-driven space management and building control systems*. The primary objective is to devise methods that are both scalable and applicable to the many different kinds of buildings in the building stock, located in different climates, with the aim of autonomously changing indoor conditions according to the characteristics of every thermal zone, HVAC system, occupancy patterns, and comfort needs of the building occupants.

1.1 Potential for improving building operations

The HVAC system accounts for the highest energy consumption in commercial buildings [127]. By employing advanced control algorithms, the HVAC operation can be optimized to better align with the occupancy schedule and operational requirements of the building, thus minimizing energy wastage. Modern buildings can also be integrated with the power grid, enabling them to modulate their energy consumption in response to real-time electricity prices or demand response signals. This not only leads to financial savings but also promotes energy efficiency. Furthermore, proactive controls, enabled by predicting occupant activities and temperature dynamics, ensure that building systems operate at peak efficiency, further reducing energy costs. Lighting systems that can autonomously dim lights based on occupancy and ambient light conditions can also lead to substantial energy savings.

On the topic of thermal comfort, the future of building control lies in personalization. Systems that adapt to the time-varying occupants' preferences can provide an environment that meets each person's comfort needs and allow more efficient energy use. Incorporating personal comfort models into the

control loop not only ensures higher thermal comfort, but also reduces HVAC operating costs by eliminating the need for the system to overcompensate for generalized settings. In terms of visual comfort, automated shading and adaptive lighting systems allow buildings to maintain optimal illumination levels, balancing natural and artificial light sources to reduce glare and improve visual comfort of the occupants.

Energy efficiency and human comfort can be improved simultaneously with control systems that incorporate occupancy information at multiple spatial and temporal resolutions, ensuring building spaces are conditioned according to their function and use. Moreover, in co-working spaces and Activity-Based Workplaces (ABWs), it might be possible to relocate occupants to the building spaces that have lower marginal energy consumption and closer temperature setpoint to the preferred temperature of the occupants. By harnessing these opportunities, it is possible to pave the way for the adoption of building controls that are not only energy efficient, but also prioritize and enhance the comfort and well-being of the occupants.

1.2 Challenges

Difficulty of model identification. Optimizing energy consumption and occupant comfort is a difficult task. This is partly due to the fact that buildings are complex systems with numerous interconnected systems that operate at different timescales. Developing accurate models and incorporating them into Model Predictive Control (MPC) for optimal building operation is a challenge for several reasons. Specifically, understanding the temperature dynamics within a building requires knowledge of specific physical attributes, such as its size, design, and construction materials. Many of these parameters are elusive for most buildings today. Although energy audits can provide these data, they are labor-intensive and costly. This underscores the importance of leveraging data-driven and learning-based approaches, such as Reinforcement Learning (RL), for building controls. RL offers a way to determine optimal control strategies by directly interacting with the physical or simulated building without needing a detailed understanding of its complex dynamics. Mod-

ern buildings are equipped with sensing infrastructure that monitors various physical quantities. Utilizing this data, control strategies can be developed through learning-based methods, even in the absence of a complete building model.

High data needs and safety issues. The critical challenge in the deployment of RL controllers is the lack or insufficiency of training data. Typically, training these controllers requires extensive training data collected over a long duration. An RL controller is trained through a trial-and-error approach, essentially taking control actions (according to a policy), assessing their performance, and refining the control policy based on these evaluations. This trial-and-error approach requires the building systems to execute the actions proposed by the controller. Such a learning paradigm is called "on-policy learning", where the controller's proposed action is directly executed in the environment. Yet, the application of on-policy learning in real buildings presents practical challenges. Building managers are often hesitant to permit an RL controller to try out actions on the actual building environment post commissioning, given the damage that these arbitrary (and sub-optimal) actions might inflict on electrical and mechanical components of a building system, and discomfort and health issues they might cause for occupants.

A potential solution to address this challenge is to allow RL controllers to learn without direct interaction with the actual building environment until their behavior is considered safe. Training data can be sourced from historical operational logs (known as offline learning) or from data produced by another controller that is currently used in the building system (known as off-policy learning). Value-based RL algorithms excel at learning a value function from log data, which estimates the effectiveness of actions based on previous observations and received rewards. Although offline and off-policy learning algorithms are more data efficient because training data can be gathered under different policies and generally safer because they do not require implementing the action suggested by the policy being learned in the real system, they often do not have the stable performance of on-policy learning algorithms [155].

This is primarily because the action space cannot be fully explored, and the dynamic nature of buildings means that the reward distribution evolves over time. Empirical evidence from previous research suggests that even with careful design of the RL algorithm and optimized features and hyperparameters, on-policy RL controllers would need approximately three years of training data to surpass the performance of RBCs [162].

Sim-to-real transfer. Using a simulation environment alleviates the data requirement for training RL controllers. Nevertheless, discrepancies arise between the training (simulation) and test (real) environments in two primary aspects. First, the simulated building often does not mirror the actual building. If a precise digital twin were available, MPC could be employed for optimal control. Instead, researchers frequently use a simplified digital twin model or utilize a prototypical model with characteristics akin to the actual building. Second, real-world data tends to be noisier than the data used in simulation. Various unforeseen events, such as equipment malfunction or sensor faults and drifts, can degrade the quality of the data. Third, distribution shifts may happen in the real environment due to factors such as changes in occupancy schedule, seasonal variations, etc. These challenges would reduce the performance of RL controllers when deployed on actual buildings.

Competing objectives. In addition to optimizing HVAC control strategies alone, it is important to manage shading and lighting systems in addition to the HVAC system. The primary aim is to reduce the overall energy use while optimizing both thermal and visual comfort for occupants. As the action space expands with the inclusion of more building systems, the complexity of finding a near-optimal control policy increases. Another notable challenge arises from conflicting or competing objectives that must be optimized at the same time. Specifically, achieving the highest comfort often results in increased energy use, which runs counter to the goal of energy conservation. Additionally, the complex interplay between various actions and control systems and their respective objectives presents unique challenges. For example, opening blinds during the summer can diminish lighting energy demands but can also increase

solar heat gain, requiring more cooling from the HVAC system.

Lack of sufficient labeled data from each occupant. Modeling personal thermal comfort is also challenged by data inadequacy, which can be attributed to three primary factors [83]. First, individual differences in thermal comfort and satisfaction necessitate the customization of the comfort models for each occupant. Second, while these customized models should be trained using supervised learning, the collected labels often exhibit biases. The concept of 'comfort' is naturally ambiguous, and individuals typically struggle to discern minor fluctuations in temperature or humidity. The phenomenon of 'creeping normalcy' – gradual changes are difficult to separate from noise [39] – further complicates matters, as it hinders individuals from making time-independent judgments. Lastly, data acquisition is often costly, either through surveys or the installation of intrusive sensors on the body of occupants. The first challenge increases the training cost of personal thermal comfort models, while the subsequent issues reduce the volume of data that can be used for training.

Overhead of space planning. Finally, in co-working spaces, ensuring optimal space assignment to long-term occupants and performing real-time space assignment to short-term occupants is quite challenging. Given individuals' preference for staying in the same building space and not being frequently relocated to other spaces, it is essential to differentiate between long-term occupants and short-term visitors. Since short-term occupants make a space reservation on short notice and use that space for a short period of time, the space assignment algorithms must be capable of finding near-optimal solutions in less than a few seconds, attaining the desired quality of service.

1.3 Objectives and contributions

To address the challenges mentioned above and realize the identified potentials, we pursue the following objectives:

1. Building a platform that provides an interface akin to OpenAI Gym, facilitating the real-time interaction between the EnergyPlus building energy simulator and control algorithms written in Python. This plat-

form should enable seamless data exchange between several models and allow benchmarking of various control algorithms on multiple buildings.

2. Empirically evaluating the performance of RL algorithms in comparison to RBCs during both heating and cooling seasons, considering varying resolutions of the occupancy data. This evaluation focuses on joint control of the HVAC system's supply air temperature, blind angle setpoints, and lighting systems for each zone in a test building. The control policy is learned and evaluated through interaction with the digital twin of the test building using COBS.

3. Investigating the performance of control agents trained in a specific building – where policy execution and evaluation is inexpensive and can be done safely – and deployed onto a novel building differing from the original training environment. Control agents are selected for deployment based on the performance estimated using only two weeks of the HVAC log data collected from the target building.

4. Constructing personalized thermal comfort models to establish base models for a given population. These base models, forming the ensemble, are trained using real-world data gathered from 38 participants over the span of 6 months. The ensemble model can then be trained for specific individuals with 6 hours of training data collected from the PCS.

5. Developing a space allocation system that strategically puts long-term occupants in specific zones to optimize HVAC energy consumption while satisfying thermal comfort requirements. The system learns how to estimate energy consumption of each zone using a small amount of the HVAC log data collected from the building, and uses this estimation model to solve an optimization problem.

6. Formulating heuristic algorithms that can quickly assign available desks to short-term occupants, delivering performance comparable to the Mixed Integer Programming (MIP) solution in terms of HVAC energy consumption, space utilization rate, and thermal comfort.

By working towards these objectives, we make several contributions to the areas of sensor networks and building automation. First, we develop COBS to support real-time interactions between EnergyPlus and Python-based controllers. Second, we examine the complicated balance between three competing objectives: building energy consumption, thermal comfort, and visual comfort scores. This exploration is carried out in the context of the joint control of HVAC, lighting, and shading systems using different RL algorithms. Third, we introduce a training framework that allows us to learn a library of optimal and near-optimal HVAC control policies in a single training environment using a diversity-induced RL algorithm. We further explore how to select the most fitting policies from this library for deployment in other buildings that could differ from the training building. Fourth, we conceptualize and implement a personal comfort model that leverages weak labels for training, ensuring a non-intrusive and scalable label collection process. We elucidate the process for utilizing these weak labels to train a meta-model and subsequently refine the meta-model to align with individual preferences. Fifth, we cast workspace allocation to long-term occupants as an MIP problem, and show that, in practice, it can be quickly solved to optimality using the Branch-and-Bound (B&B) method. Additionally, we propose two heuristic algorithms that quickly assign the available space to short-term occupants, achieving performance comparable to the space assignments derived from solving the MIP problem. A pictorial description of these contributions and their connection to each other is provided in Figure 1.1.

1.4 Outline of the thesis

Chapter 2 offers a comprehensive review of the literature, highlighting existing gaps that will be addressed in the subsequent chapters. Chapter 3 describes the standard design of the HVAC systems, introduces fundamental RL concepts, and presents the proposed EnergyPlus wrapper. Chapter 4 investigates the joint control of multiple building systems using RL. Chapter 5 elucidates the methodology for training a library of building control policies through interaction with a single training environment (called the source building), evaluating

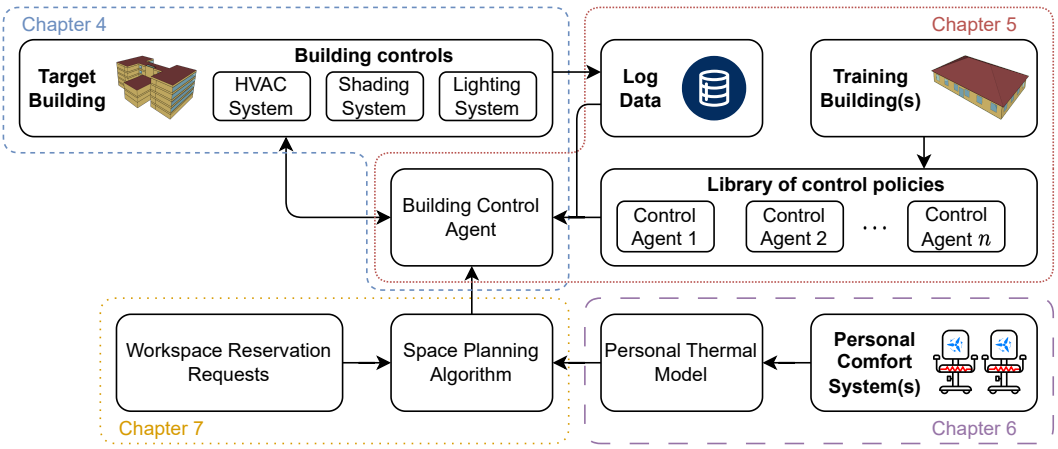


Figure 1.1: A high-level overview of the ideas presented in the thesis

each policy in that library on a small amount of log data obtained from a novel building, and transfer the most suitable policy among these policies to that building. A data-efficient methodology for training personal thermal comfort models is detailed in Chapter 6. Chapter 7 introduces space assignment strategies for both long-term and short-term occupants. The thesis is concluded in Chapter 8, where the limitations of the presented work are outlined, and potential avenues for future research are discussed.

Chapter 2

Literature Review

This chapter reviews the recent studies on heat transfer in the building, establishing personal comfort models, estimating building occupancy at different spatial and temporal resolutions, and space planning in institutional buildings. It also summarizes recent efforts on developing energy-efficient building control strategies. The rest of this chapter is outlined as follows: Section 2.1 provides a detailed overview of three different approaches that have been used for modeling building thermodynamics. Section 2.2 reviews the literature on how thermal comfort has been modeled historically as well as recent work on personal comfort models, followed by the review of building occupancy estimation and space planning practices in Section 2.3. Section 2.4 surveys the related work on local and supervisory level building control methods.

2.1 Modeling heat transfer

The optimal control of the building HVAC system is of great importance as it accounts for a large fraction of the building energy use, and is responsible for maintaining the temperature inside the building within a comfortable range. To optimally control this system, most related work adopts receding horizon control which relies on a model that explains how the room temperature changes as a result of implementing a certain control policy [107] (e.g., increasing the setpoint temperature by 1 degree Celsius or closing the damper for an hour). This has given rise to a large number of studies aiming to solve a *system identification* problem to establish this thermal model through a white-box, black-box, or grey-box approach [37].

The black-box approach requires extensive data to learn the relationship between the current state (*e.g.*, room temperature) and HVAC control input, and the next state [24], [102]. A variety of data-driven techniques have been used in the literature to establish the thermal model. This includes Artificial Neural Network (ANN) [16], [113], [129], ANN with Levenberg-Marquardt (LM) [68], [84], [108], [111], [112], Radial Basis Function (RBF) [48], [136], AutoRegressive model with exogenous variables (ARX) [101], [116], and AutoRegressive Moving Average model with exogenous variables (ARMAX) [115], [125]. These black-box models map a set of features (e.g., previous readings of room temperature and humidity, and its occupancy state) to the room temperature, and their accuracy highly depends on the selected features [43].

Alternatively, the indoor temperature can be estimated by directly applying the laws of thermodynamics. This requires a detailed description of the building, construction materials, and the HVAC system [58]. Such a thermal model can be useful in identifying insulation problems and estimating the whole building energy use. Computational Fluid Dynamics (CFD) is a pure physics-based modeling approach which has been quite successful in predicting environmental quantities such as temperature and humidity [147]. This model has many parameters to tune, such as the wall construction material, thicknesses, and the number of layers. Therefore, the model customized for one building cannot be used in another building as these parameters vary significantly across buildings.

Another type of thermal models is the Resistor-Capacitor model (RC) model which is commonly used for heat transfer analysis in buildings. This grey-box model turns building spaces and multi-layered walls into a number of latent thermal resistances and capacitances. Compared to detailed physics-based models (e.g., models used in EnergyPlus [31]), a low dimensional RC model is less complex, making it easier to identify its parameters from sensor data. Despite its simplicity, it achieves a high accuracy in predicting the indoor temperature. Zhou et al. [176] compare a low dimensional RC model with a physics-based model, and conclude that the RC model can substitute the physics-based model with a negligible loss of accuracy. In this thesis, given

our emphasis on developing model-free RL controllers, we do not need to train thermal models for every zone of the building.

In Chapter 7, we study the relationship between the heating and cooling demand of a zone on a given day and the number of occupants assigned to that zone, the temperature setpoint of that zone, and outdoor air temperature. The goal is to capture this relationship using only limited information about the building, in particular its layout and the capacity of its zones. Although a physics-based or grey-box model can be used to estimate the amount of heat injected or extracted from a zone in one time step, and consequently the total heating/cooling demand of that zone in one day, these models require access to additional metadata and suffer from the accumulation of error over time. For this reason, we take a black-box modeling approach in that chapter. Uniquely, instead of training black-box models that are commonly adopted in the literature [102], we learn a model that is convex in its inputs so it can be incorporated into an optimization problem.

2.2 Modeling human comfort

Thermal comfort reflects an individual's satisfaction with their local thermal environment. The human body needs to emit heat to keep itself functional. This heat transfer is mainly caused by the temperature difference between the human body and ambient environment. If the ambient environment is too hot or too cold, the body cannot release the desired amount of heat, causing discomfort. It is assessed by subjective evaluation using metrics such as thermal sensation, acceptability, and preference. To quantify human perception of thermal comfort, several models have been proposed by researchers. One of the most recognized models is Fanger's Predicted Mean Vote (PMV) [46], which is adopted as the basis of the international standards ISO 7730 [71] and American Society of Heating, Refrigerating, and Air-Conditioning Engineers (ASHRAE) 55 [4]. The PMV model uses two personal (i.e., clothing insulation, metabolic rate) and four environmental (i.e., air and mean radiant temperatures, air velocity, and relative humidity) variables, to calculate a thermal sensation score in the range of -3 to +3. In this scale, +3 indicates

that the environment is 'hot', -3 indicates that the environment is 'cold', and 0 means the environment is 'neutral' which is regarded as thermally comfortable. The ASHRAE defines a comfort range based on the PMV model, which is from -0.5 to 0.5. This range is used to specify conditions for acceptable thermal environments in building design and operation. Another well-known thermal comfort model is adaptive comfort models that are adopted into standards ASHRAE 55 and EN 15251. Compared to the PMV model that ignores the effects of outdoor climate on thermal perception, the adaptive model takes prevailing outdoor air temperature into account to express acceptable indoor temperature [36]. Therefore, the adaptive comfort model is used for naturally-ventilated buildings with operable windows, whereas PMV is typically applied to mechanically-conditioned buildings. In Chapter 6, we use PMV and adaptive model for comparative analysis.

Unfortunately, conventional comfort models, including the PMV and adaptive comfort models, do not accurately represent the individual thermal comfort due to individual differences in thermal perception and requirements [83]. Moreover, they use a fixed set of input variables and do not include additional variables that show relevance to the thermal comfort such as sex and body mass index. They also incorporate many input features that are hard to obtain. Therefore, there is a need for a personal thermal comfort model that can accurately reflect individuals' thermal preferences and can be easily trained even when training data is scant.

2.2.1 Personal comfort models

Developing personal comfort models has become increasingly popular in recent years [103]. According to a literature review [160], a sharp increase can be witnessed in the number of papers in this area in the past five years. Prior work has focused on correlating thermal comfort to various sensor data using numerous approaches, such as Support Vector Machine (SVM) [134], Linear Regression (LR) [59], logistic regression [34], Random Forest (RF) [9], [74], ANN [52], [99], Gaussian process [29], fuzzy rules [75], and adaptive stochastic modeling [61]. They reported significant improvement in thermal comfort

prediction accuracy (84% median accuracy [160]) compared to the conventional comfort models (34% overall accuracy [28]) that are not personalized, such as the PMV and adaptive models. Nevertheless, they require a large amount of survey data to train the model, which is costly to acquire in a reliable fashion in the real world. Without this explicit feedback, personal thermal preferences cannot be learned. Some studies investigate the development of personal comfort models following rudimentary laws of metabolism [90]. This approach requires the knowledge of the building environment and occupants, making it difficult to scale to a large building with a diverse set of occupants. In Chapter 6, we adopt a data-driven modeling approach.

2.2.2 Important features for thermal comfort modeling

Two types of sensor data, which can be related to environmental and personal factors, are commonly used as input features to predict an individual's thermal comfort. The environmental factors include indoor temperature, relative humidity, air velocity, etc. They are typically gathered by sensors deployed in the building to monitor the indoor environment. The personal data can be collected by wearable sensors measuring quantities such as the skin temperature and heart rate [21], [74], or thermal arrays and cameras installed in the room, capturing the body temperature and clothing level [8]. While the use of personal data for thermal comfort modeling offers clear advantages in estimation accuracy [160], the data collection process is expensive, intrusive, and can lead to privacy concerns. For these reasons, we do not utilize personal data in this thesis.

To develop data-driven thermal comfort models, the sensor data must be accompanied with reliable information about human thermal comfort, which will be treated as a label in the model training process. In general, the more trainable parameters a model has, the more training data would be needed to develop a sufficiently accurate model. In fact, all of the 105 papers reviewed in [160] collect subjective feedback from occupants via mobile applications, web applications, or paper questionnaires. However, a huge amount of survey data must be collected to build a complex model (e.g., a neural network)

Figure 2.1: This image from [83] shows heating and cooling elements of the personal comfort system (left side) and the controller with wireless connectivity (right side).

that can relate various input features to each individual's thermal comfort. A possible solution is to group individuals with similar thermal preferences to increase the amount of labeled data for each model, thereby improving the thermal comfort prediction accuracy [74], [95]. Moreover, surveys must be completed at frequent intervals, otherwise the gap in label data makes it impossible to capture temporal dependencies between input features and thermal comfort. In practice, it is cumbersome to acquire frequent and reliable feedback from building occupants. This has hampered the development of time-series models and recurrent neural networks for thermal comfort modeling as it can be deduced from the literature review in [160].

To tackle this issue, in Chapter 4, we create artificial labels from how an individual uses the heating and cooling functions of their Personal Comfort System (PCS), depicted in Figure 2.1. This personal comfort system contains two sets of thermal control devices, one is placed under the chair seat and the other one is installed behind the back of the chair. Each set contains a fan and a heat strip to perform heating and cooling operations. These two sets of devices can be controlled independently, and their operations along with chair occupancy, ambient temperature and humidity, and zonal temperature and humidity are logged at 5-minute intervals. Reference [83] shows that an individual's behavior with this PCS is a strong predictor of their thermal comfort needs and can be used as a proxy for thermal preference. Thus, we

replace the direct feedback obtained via surveys with individuals' heating and cooling behavior measured at the same rate as other sensor data. This helps to preserve the temporal dependencies in the input data, and also reduces the label data collection cost by eliminating the need for survey data.

Note that PCS comes in many different forms, including desktop fans, heated and cooled chairs, and foot and leg warmers. With recent inclusion of PCS in ASHRAE 55, PCS is expected to play an increasing role in building design and operation. Hence, the findings of this thesis can be extended to other types of PCS devices to better understand individuals' thermal preferences in the built environment.

2.2.3 Transfer learning and ensemble methods

A common assumption for thermal comfort modeling is that sufficient training data is available [32], [83]. However, this assumption is not valid for individuals who are new to the building or we do not have enough labeled thermal comfort data for them for various reasons. Transfer learning is an effective approach to address the inadequacy of labeled thermal comfort data. There are two common kinds of transfer learning: inductive learning and transductive learning. The inductive transfer learning approach utilizes training data from different but related tasks in the current learning step. For example, the authors in [119] cast thermal comfort prediction as a regression problem and combine active learning with transfer learning to transfer a population thermal comfort distribution and personalize it using a small number of queries asked from the target individual. In follow-on work [89], a general thermal comfort input space is defined and used to generate enough training data. This model is then personalized as more label data becomes available. This approach still requires a lot of data to personalize the model.

In transductive transfer learning, the thermal model trained in one environment (or for one individual) is transferred to another environment (respectively to another individual) usually after some adaptation, i.e., retraining using a small amount of data. In [140], the authors develop the Convolutional Neural Network (CNN)-Long Short Term Memory (LSTM) model to extract high-

level features from input features and learn feature relations. The authors have shown that this model can be transferred to a different data set after retraining with even higher classification accuracy than the model built from scratch in the test data set. This work has a few shortcomings: the model relies on features that are difficult to measure or obtain, such as the metabolic rate and clothing factor, and that the LSTM model is not used for understanding the temporal relations between input data and thermal comfort.

To leverage the knowledge gained from the source domain to the greatest extent, a common solution is to break down the model into multiple sub-models and only retrain the sub-models that need further training. This idea is adopted in [67], which is the closest line of work to the work presented in Chapter 6 of this thesis. The authors built a base classification model for each data set, which is then used as the feature extractor in a deep neural network to predict the thermal comfort. Hence, the base models are directly transferred to the target domain without retraining and only the fusion network is trained in the target domain. They achieved on average 64.1% estimation accuracy in a data set in which the PMV model's accuracy was 30.4%. However, both [140] and [67] train one general comfort model for all individuals in the data set instead of personalizing comfort models. Moreover, they do not study the sensitivity of the result to the amount of data used to retrain the model. In Chapter 6, we borrow the idea of combining base models to predict the thermal comfort of new individuals. However, we construct several group comfort models and use them instead of the personal comfort models as our base models. This improves the thermal comfort prediction accuracy because it reduces the amount of trainable weights and increases training data for each base model as suggested in [74], [95]. Additionally, we evaluate the impact of having different amounts of training data available for incrementally training the models.

2.3 Estimating the building's occupancy state

Building occupancy is one of the main factors that determine its energy consumption. According to the 2012 Commercial Buildings Energy Consumption

Survey (CBECS) [151], a building that was occupied for all 168 hours in a week consumed 46% more energy per square foot than a building that was occupied for only 80 hours in a week. Incorporating more occupancy information can help to optimize control policies. According to a study, HVAC energy consumption can be reduced by around 25% compared to when it is controlled using a fixed setpoint during the occupancy period without violating thermal comfort requirements [47]. Turley et al. [150] evaluate the energy efficiency and human comfort with different occupancy patterns using MPC, and shows that incorporating the number of occupants in every room is essential for higher energy savings. Unfortunately, such occupancy data is hard to collect from every zone (as it requires special sensors), so most previous studies incorporate binary occupancy information (occupied/vacant) at building or floor-level, assuming that it can be obtained in a reliable fashion.

Many previous studies proposed monitoring and estimating the occupancy state using wired and wireless sensor networks. The vision-based method is one of the most popular approaches to estimate number of occupants thanks to the high accuracy [149]. However, the vision-based approach typically has high computational complexity, illumination and occlusion problems, high installation costs, and privacy issues. Some studies utilized existing surveillance cameras for occupancy estimation [26], [56], but because of the improper angle and low resolution, the accuracy does not meet expectations. Passive Infrared Sensor (PIR) sensors have been utilized in less intrusive solutions [135], [139]. However, they still require the installation of extra sensors. Some studies used WiFi signals to estimate the number of occupants [110], because the WiFi access points are typically installed in buildings. The drawbacks of using WiFi signals include the cases where people carry multiple WiFi-enabled devices, no device, or prefer not to use WiFi because of the rapid development of the fourth and fifth-generation mobile networks. It is also hard to accurately determine the occupants' location given the coverage of a WiFi access point. The carbon dioxide concentration sensor has also been used in previous studies [177], but it suffers from the long delay due to the slow gas mixture rate. Besides, it is necessary to install this sensor in the best location, otherwise

there is no strong correlation between carbon dioxide concentration in a space and its occupancy state.

While many studies use approximate or ground-truth occupancy information in the control loop [172], or predict the building occupancy state [126] for optimal control, changing the spatial distribution of occupants for improved energy efficiency is an underexplored problem. The COVID-19 pandemic caused a sudden shift in workspace utilization strategies. Many workplaces have moved away from the fixed seating arrangement, giving their employees the flexibility to choose their seating either through online reservation or on a first-come, first-served basis. Research indicates that when implemented with careful planning, such adaptable work environments can enhance space utilization, curtail operational expenses as a result of reduced energy usage, and improve teamwork and productivity [13], [20], [141]. The co-working spaces, or the so-called ABW, offer a unique opportunity to decide on the occupancy state of each zone in the building rather than passively estimating the number of occupants in each zone. In Chapter 7, we delve into strategies for space assignment that minimize the whole-building energy use while ensuring a comfortable environment for the occupants of each zone.

The closet work considers objectives that differ from ours. For example, Yip *et al.* [165] investigate how to allocate staff in a hospital to improve service efficiency, and Yang *et al.* [163] focus on occupant assignment to reduce connectivity costs associated with interpersonal communication. In recent work, Deng *et al.* [38] proposed clustering occupants based on their thermal preferences to reduce building energy use. However, they jointly optimize the average thermal comfort of occupants and HVAC energy consumption using an iterative method, assuming a fixed number of occupants in the experiment. We address a more general problem, combining energy consumption with individual thermal comfort in a building that houses a certain number of long-term occupants and a variable number of short-term occupants that may arrive every day, which better represents the occupancy pattern of co-working spaces.

2.4 Control strategies for building systems

Numerous attempts have been made to date to optimally control HVAC, lighting, shading, and other building systems. But, designing an optimal controller is complicated due to the complex building design and structure [66], high variability of energy demand [155], and inaccurate estimation of human comfort [34], [173]. The control strategies can be broadly divided into three categories: rule-based, model-based, and model-free. Table 2.1 shows example control strategies from each category. Regardless of which control strategy is adopted, occupancy information can be incorporated in the control loop to achieve an acceptable trade-off between energy savings and occupant comfort.

In the rule-based approach, control rules and schedules are defined by the facilities manager based on their intuition about how the building occupancy varies over time. It is shown in [7] that using static per-zone schedules can considerably reduce the energy consumption of HVAC. In another study [138], it is shown that a rule-based lighting controller can lower the building energy use by up to 12% without negatively affecting the visual comfort of occupants. Rule-based controllers are easy to implement and do not require training complex models, but their performance is highly dependent on the quality of the rules. In practice, the control performance degrades over time with changes in the zone occupancy schedule and outside air temperature.

In the model-based approach, models for heat transfer, occupancy, and different components of building systems are utilized in the control loop to minimize the energy use over a time horizon subject to a set of constraints. While a high-order heat transfer model can accurately determine the temperature of every zone in the building, proper identification of this model is difficult. Estimating these parameters requires collecting a large amount of data under different operating conditions or running expensive energy audits. Alternatively, low-order thermal models can be built using a data-driven approach if enough training data is available [63], [176]. These models have proven to be useful for MPC, lowering the energy consumption of the HVAC system while maintaining thermal comfort [131], [158]. Model-based reinforcement learning

Table 2.1: A representative subset of related work for different occupancy levels, control points and methods. For similar control scenarios, more recent studies were chosen.

	Thermal Zones	**Occupancy**	**Control Variables**	**Control Method**
[68]	4 Zones	Building-level (Binary)	Temp. setpoint	Rule-based
[138]	20 Zones	Room-level (Count)	Lights (on/off) Blinds (angle)	Rule-based
[131]	1 Zone	N/A	Temp. setpoint	MPC
[153]	3 Zones	Building-level (Estimated Count)	Temp. setpoint	Model-based
[49]	4 Zones	Building-level (Binary)	Temp. setpoint	MPC
[25]	1 Zone	Building-level (Binary)	HVAC (on/off) Window (open/closed)	Q-learning
[27]	1 Zone	Building-level (Binary)	Lights (on/off) Blinds (step changes in angle)	Q-learning
[118]	1 Zone	Not mentioned	HVAC (heating/cooling power) Window (open/closed) Door (open/closed)	Deep Q-learning
[40]	1 Zone	Building-level (Count)	HVAC setpoint Light (dimming level) Blinds (angle) Windows (open pct.)	BDQN
[133]	1 Zone	N/A	HVAC on/off	DDPG
[57]	1 Zone	N/A	Temp. setpoint Humidity setpoint	DDPG
[23]	5 Zones	Room-level (Binary)	SAT setpoint	PPO

techniques have been used to optimize HVAC operation too [41], [168]. While model-based HVAC control strategies have great performance, explaining interactions between multiple building systems requires more complex models which cannot be easily trained, especially in a building with heterogeneous spaces.

In recent years, model-free RL algorithms have been applied to address the optimal control of the HVAC system [155], lighting system [122], and other systems. Instead of relying on a built-in thermal model, they provide the opportunity for trial-and-error learning through direct interactions with building

systems or an external simulated environment. There are three main types of model-free RL algorithms, namely Q-learning (value-based), actor-critic, and policy gradient methods. Value-based RL algorithms are used in many papers such as Q-learning [25], [27] and Deep Q-learning [40], [118], which The Q-learning algorithm updates action values (i.e., Q-values) for each state based on the observation. It is generally more sample-efficient than other model-free RL algorithms. Actor-critic methods are also adopted to control the building system. They learn the control policy as well as the Q-values to update the control policy. Policy gradient algorithms are considered the least sample-efficient model-free RL algorithms, yet there are usually more stable than the other RL algorithms. Of the 77 papers that applied RL to building controls and were reviewed in [155], three-quarters (59) used value-based methods, some (12) used actor-critic methods, and a few (3) used policy gradient approaches. (The remaining 3 were model-based approaches.) Despite the large number of reinforcement learning algorithms that are used in the building control domain, they are seldom compared in terms of their performance, stability, and convergence speed.

HVAC control strategies: In the context of HVAC control, these strategies minimize the building energy use while maintaining a comfortable indoor environment for occupants. Control strategies are typically implemented at two levels. Local control strategies directly control the operation of specific HVAC components, such as the supply air temperature [172] and water temperature [24], bypassing conventional feedback controllers. Supervisory control strategies, on the other hand, tweak specific setpoints, *e.g.*, room temperature setpoints, and leave the control task to conventional feedback controllers [40], [100].

From a different perspective, control strategies can be classified based on the control approaches they used, which can be broadly classified into rule-based, model-based, and model-free control algorithms. Rule-based HVAC controllers are relatively easy to implement and can considerably reduce the building energy consumption [7], [138]. But their performance heavily relies

on the quality of the control rules and setpoints.

In model-based HVAC control, models are used to predict the heat load and energy demand of the building. These models are built using physics-based or data-driven approaches. In MPC, these models are used to minimize energy use and occupant discomfort over a finite time horizon. Such controllers can significantly reduce the energy consumption [131], [150], [158]. Reference [150] evaluates the energy efficiency and human comfort under MPC and shows that occupancy-based MPC can achieve up to 50% energy savings over a constant temperature setpoint control method. Indeed, the performance of MPC depends on the quality of the predictive model(s).

Developing physics-based thermal models is challenging in a large multi-zone building, requiring manual effort or a substantial amount of training data [10]. Even if accurate models are developed and incorporated in the control loop of one building, these models would not work on another building's HVAC system without having to re-define the physics-based models or re-learn the models from new data.

Learning-based control algorithms, such as model-based and model-free reinforcement learning, has immense potential for energy savings and improved indoor environment quality. In recent years, various types of model-free RL have been applied to the HVAC control problem with the goal of finding a near-optimal control policy that minimizes energy consumption while maintaining thermal comfort without modeling the complex building dynamics [23], [76], [155], [172]. An RL agent learns the mapping between the state of the building and an action via trial-and-error. It is shown in [172] that a single RL agent that observes the state of the whole building and controls all setpoints can reduce the total HVAC energy consumption by around 22% compared to a rule-based controller in a small multi-zone building.

Unfortunately, training these RL agents requires a substantial amount of data to sufficiently explore a large or continuous state-action space. As more features are added to the state, the complexity and the number of parameters used to represent the agent grows exponentially. Moreover, a single agent that controls multiple actuators cannot be easily transferred to another build-

ing that has a different state and/or action space. Chen *et al.* [23] reduce the training cost of an RL agent that controls the HVAC system through a differentiable MPC policy that encodes system dynamics and offline imitation learning, using the operational data collected under a default controller. However, learning an accurate model can be challenging in a given building and more historical data would be needed to fully capture the system dynamics. Similarly, offline RL techniques generally require a substantial amount of historical data before they can learn a high-quality policy.

In a recent survey paper, Pinto *et al.* [128] have reviewed the applications of transfer learning to buildings, including papers that use transfer learning to address the data inadequacy challenge in developing learning-based controllers for building systems. This literature review reveals that there is no paper that takes advantage of diversity for transfer learning in this domain. Xu *et al.* [161] address the problem of transferring previously learned HVAC control policies to an unseen building. Their methodology involves decomposing the policy neural network into a transferable front-end network and a trainable back-end network. The front-end network captures building-agnostic behavior, whereas the back-end network needs to be trained on the target building. The back-end network can be trained in an offline fashion through supervised learning by using the log data collected by a default controller on the new building. It can also be trained in an online fashion by deploying a randomly initialized back-end network. Although this approach reduces the training cost of RL to some extent, control performance can still be poor while the back-end network is being trained in the target building. In another line of work, Fazel *et al.* [80] propose augmenting the training data collected from the target building. The authors use generative adversarial networks to learn the building performance profile from the actual data, and generate synthetic data that reflect climate and operation variations, while keeping the building profile the same. However, 1 year data is required to train the generative model, which may not be readily available in all buildings. The temporal dependency between the synthetic data is also not considered, and the synthetic data is generated only according to the learned distribution, so the performance would still be low on out-of-

distribution samples, which downgrade the performance of control policies on certain buildings.

Joint control of building systems: The whole building energy consumption can be further reduced when building systems are jointly controlled compared to when only HVAC is controlled using a learning-based algorithm [40]. Still, the joint control of multiple building systems is challenging because it increases dimensions of state space and action space, and makes it harder to learn a near-optimal policy due to complex interactions between systems. Previous work focuses on zone-level energy optimization through the joint control of lights and blinds [27], HVAC and windows [25], [33], and all these four systems [40], [88]. The state and action space becomes increasingly large as more zones are included; none of these studies address the joint control of building systems in a multi-zone building, and no previous work quantifies additional energy savings compared to the case where these systems are controlled separately.

In Chapter 4, we quantify additional energy savings by adopting RL-based control algorithms to manage the operation of HVAC, shading, and lighting systems, while thermal and visual comfort of occupants in the building are maintained.

Multi-agent building control: MARL-based controllers are proven to be useful in energy-efficient control of building systems [55], [152], and are amenable to transfer learning [117], addressing scalability issues in large buildings. Decentralized control strategies have been used in previous work for specific building components, such as using room/zone agents to control the thermostat setpoint based on the presence of occupants for energy use optimization [35]. Some studies divide the HVAC system into multiple subsystems, then develop a control strategy for each subsystem. For example, Zhao et al. [174] manage the electrical power flow using an electricity agent, and the heating and cooling components are controlled by a heating agent and a cooling agent, respectively. Similarly, Klein et al. [86] use multiple agents to control the HVAC, lighting, appliance, and human separately. Other studies address the scalability issue

by investigating the distributed decision making where the HVAC system is oversimplified [19]. Decomposing the building control task into multiple subtasks will help to reduce the training cost. However, there is no discussion about whether the agent trained in one building can be used to control the systems in other buildings.

Unlike the previous work that control multiple building systems using MARL, in Chapter 5, we decompose the optimal control of the HVAC system into the problem of controlling the environment of individual thermal zones in the building, which can be solved using MARL. In this setting, each agent is responsible for controlling the HVAC components (e.g., control points in the variable air volume system) in one zone of the building. This reduces the size of an agent's state-action space and enables the transfer of policies to other buildings, regardless of the number of zones they have or their floor plan. This is because, at the zone level, most buildings have the same set of sensors and actuators, so the corresponding agents will have an identical state-action space. Control of other building systems can also be broken down into zone-level controls, which we will address in future work.

Diversity in reinforcement learning: In most cases, multi-agent RL algorithms focus on optimizing a single solution for a given training environment. Although this single solution works well in the environment it has been trained in, RL algorithms typically aim to find a single (near-)optimal policy. Although this control policy works well in the environment it has been trained for, it might perform poorly in a new environment or even when the original environment changes. This problem is exacerbated in MARL as agents tend to overfit to their co-players [91]. A interesting approach to address this problem is to incorporate *diversity* when learning a control policy [106]. Diversity approaches can be broadly classified into two categories: environmental diversity [73], [106] and policy diversity [45], [97], [104]. In environmental diversity, variants of the given environment are used to train the agents. The goal is to obtain policies that capture main features of the environment, hence generalize to a broader class of environments [73], [106]. Policy diversity focuses

on finding distinct sub-optimal policies in the training environment. Masood and Doshi-Velez [104] use a maximum mean discrepancy approach to obtain a set of diverse policies. Information theoretic approaches are used in [45], [97] to address the same problem. Furthermore, there is a vast literature on quality diversity for both single and multi-agent approaches (see [51], [132], [164] and the references therein). The goal of quality diversity, is to find a diverse collection of behaviors in which each member is as well performing as possible. Novelty search, with local competition [96] and MAP-Elites [114] are among the quality diversity algorithms that discover different behaviors, while simultaneously improving behaviors that have been discovered already.

Transfer learning: Although policy gradient and actor-critic RL algorithms can solve the continuous control problem of building systems efficiently by interacting with the building, many episodes are generally needed to find a reasonable, near-optimal policy. To reduce the training time in the target environment, one common approach is to transfer and reuse the RL knowledge learned in similar environments, thereby achieving better performance in a smaller number of episodes [146]. Transfer learning has been used in both MPC-based building control [150] and RL-based building control [161], focusing on the knowledge transfer across different seasons in the same building [169], or across various thermal zones of the same building [117]. However, transferring control policies to other buildings has not been adequately explored in previous work (see this survey [155]). Unlike these papers, we train multiple control agents in parallel, each controlling a particular zone of the building. These agents are trained in a controlled built environment, referred to as the training building, before they are transferred to the target building. The obtained policies are then assigned to different zones in the target building. If this assignment is done properly and these policies are *diverse*, they will perform better than policies that are trained from scratch in the target building.

In Chapter 5, we employ diversity training in HVAC control which is cast as a multi-agent reinforcement learning problem. By accounting for diversity,

we argue that the policies that are trained in a controlled built environment can perform better when transferred to the target building.

Chapter 3

Background

This chapter provides an overview of building systems and functional relationships between them, fundamental concepts in reinforcement learning, and the simulation environment developed and used in this thesis. In Section 3.1, we briefly introduce the major components of the building systems, offering insights into their design and functionality. In Section 3.2, the focus shifts to the fundamentals of reinforcement learning, including core principles, and algorithms. Finally, in Section 3.3, we present a simulation platform designed to facilitate the application of RL-based controls to building systems. This platform enables seamless integration of the reinforcement learning agents with the state-of-the-art building energy simulation program.

3.1 Building systems

Commercial buildings are controlled by mechanical and electrical systems, such as HVAC, lighting, and shading. The HVAC system consumes a considerable amount of energy to provide heated, cooled, and conditioned air to occupants, thereby maintaining comfortable and healthy indoor conditions. In Canada, 63% of the energy in commercial and residential buildings is used for space heating and cooling [120]. This number increases to 75% if we include the energy used for lighting. This along with the fact that buildings can store heat due to their thermal mass and exhibit different spatio-temporal occupancy patterns makes the HVAC control problem important and nontrivial. Lights, on the other hand, are often controlled using a reactive strategy because illuminance will change immediately after a control policy is implemented.

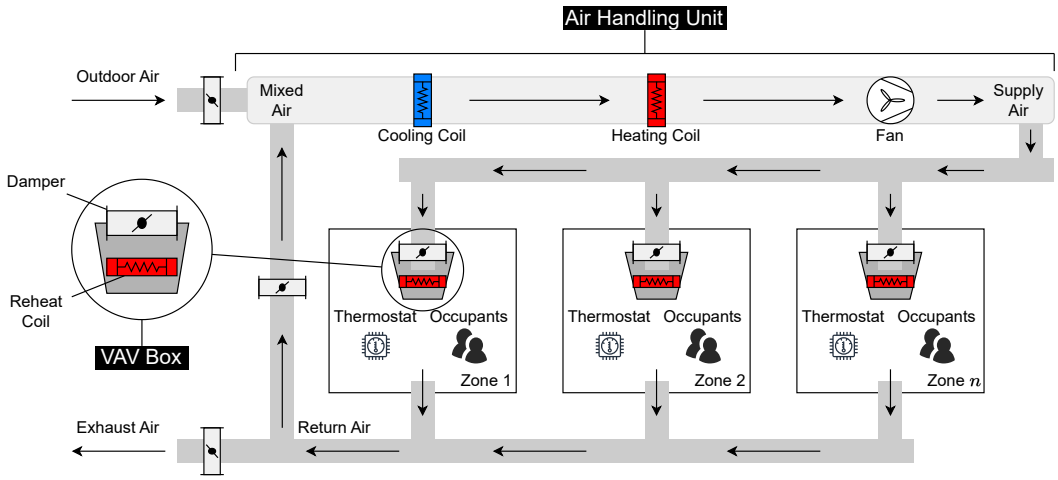


Figure 3.1: A diagram of an AHU feeding a number of VAV systems at terminal zones. The components depicted here match the components of the HVAC system that we consider in this thesis.

Figure 3.1 illustrates a typical HVAC system for a medium size office building. It consists of a centralized Air Handling Unit (AHU) that moves conditioned air through the building via a duct system. In the AHU, the outside air and the return air from zones are mixed together. The mixed air is then heated or cooled to a specified temperature before it is pushed through the duct system by a fan. In larger office buildings, multiple AHUs may be required. To account for losses in the duct system, this temperature is set to be below the desired temperature range of each zone. Variable Air Volume (VAV) systems are often used in office buildings because they allow for zone-specific control with a single AHU. A terminal VAV box exists in each zone, which might be a single room or span multiple rooms. It is responsible for controlling the amount of supply air by opening and closing a damper. A reheat coil may be present in the VAV box to heat the supply air in the zone to the desired temperature. This allows each zone to have its own thermostat with a unique temperature preference, which is often expressed by a thermal comfort range in which no corrective action needs to be taken by the VAV controller. The HVAC can be controlled at the building level, e.g., using the Supply Air Temperature (SAT) setpoint [23], or at the zone level using the thermostat temperature setpoints [40], mass flow rate setpoint, or the actuators in the

respective VAV system. The HVAC energy consumption is the total energy consumed by VAV systems, AHU heating and cooling coils, and the fan.

Auxiliary building systems, such as lighting and blinds, also have a large effect on occupant comfort and whole-building energy use [27]. The lighting system affects the building energy use, the visual comfort of occupants, and to a lesser extend, the thermal comfort of occupants, as lights produce heat when they are on. It may consist of dimmable or non-dimmable lights located in different building spaces. Dimmable lights are normally controlled using a reactive strategy. They can be dimmed linearly between the maximum and minimum light outputs according to the available daylight measured at some point in the zone. In simulation, the daylight illuminance is calculated based on cloud coverage. Blinds are usually mounted on the inside of windows and consist of a series of equally-spaced slats that are oriented horizontally. The blind controller can change the slat angle from 0 to 180 degrees. By controlling the blind angle, it is possible to change the ratio of direct and diffuse solar radiation passing through the blind. Opening or closing blinds thus changes the amount of heat gain and the illuminance level, thereby affecting both visual and thermal comfort conditions.

Lights, blinds, and HVAC systems have complex interactions. As explained, opening blinds during the day will influence the interior daylight illuminance, providing natural lighting and heating up the zone due to the solar radiation. If the zone temperature goes above the desired zone temperature, the HVAC system will supply more cool air to the zone, affecting the total energy consumption of the HVAC system. On the other hand, switching on the lights in a zone will increase illuminance and energy use at the same time. Thus, there are many ways to navigate the three-way trade-off between the energy use, thermal comfort, and visual comfort. While thermal and visual comfort requirements can be satisfied when these systems are controlled independently, this comes at the price of increased energy consumption. The joint control of building systems enables finding a better trade-off between the energy use, and thermal and visual comfort.

3.2 Reinforcement learning

RL focuses on how an agent takes actions in an environment to maximize its cumulative reward. By taking an action and receiving feedback in the form of reward or penalty, the agent gradually learns a decision-making policy. This adaptive learning process makes RL particularly powerful for complex tasks where the best course of action is not immediately evident.

Markov Decision Processs (MDPs) offer a classical framework for sequential decision-making, wherein actions influence not only the immediate reward but also future outcomes. MDPs represent a mathematically refined version of the RL problem, that is essential for theoretical assertions. We use a tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$ to describe an MDP, where $\mathcal{S}$ defines the states the agent could be in, $\mathcal{A}$ are the actions an agent could execute, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \mapsto \mathcal{S}$ is the (stochastic) transition function, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto \mathbb{R}$ is the reward function, and $\gamma \in [0, 1]$ is the discount factor which is set to 0 at the terminal state. The problem can be described as follows: the observation of the environment at time t is called $s_t \in \mathcal{S}$, which is sent to the agent so that it can take an action $a_t \in \mathcal{A}$. Subsequently, the environment responds with a numerical reward signal $r_{t+1} = \mathcal{R}(s_t, a_t)$, the next state's observation s_{t+1} (also denoted as s'), and a termination signal flagging the end state.

An agent interacts with an environment to learn a (stochastic) policy, $\pi : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$, describing how to select actions to maximize the expected long-term discounted reward $G_t = \sum_{k=t+1}^{T} \gamma^{k-t-1} r_k$, where T is the final time step and can be ∞ for non-episodic problems. In general, the agent learns the state value function $v_\pi(s)$ and/or action value function $q_\pi(s, a)$ to estimate the expected reward for the given state and state-action pair through interactions with the environment. The value functions are defined as follows:

$$v_\pi(s) = \mathbb{E}_\pi \left[G_t | s_t = s \right] = \sum_a \pi \left(a|s \right) \sum_{s', r} \mathcal{P} \left(s', r|s, a \right) \left[r + \gamma v_\pi \left(s' \right) \right],$$

$$q_\pi(s, a) = \sum_{s', r} \mathcal{P} \left(s', r|s, a \right) \left[r + \gamma \sum_{a' \in \mathcal{A}} \pi \left(a'|s' \right) q_\pi \left(s', a' \right) \right].$$

The optimal policy π^* is defined as the policy that has better than or equal to all other policies' expected return on all $s \in \mathcal{S}$. Such a value functions is called optimal value function, denoted respectively $v_*(s)$ and $q_*(s,a)$, which are given by:

$$\begin{aligned}
\pi^* &= \operatorname*{argmax}_{\pi} v_\pi(s) \\
&= \operatorname*{argmax}_{\pi} \sum_a \pi(a|s) q_\pi(s,a), \\
v_*(s) &= \max_a \sum_{s',r} \mathcal{P}\left(s', r|s, a\right)\left[r + \gamma v_*\left(s'\right)\right] \\
&= \max_a q_*(s,a), \\
q_*(s,a) &= \sum_{s',r} \mathcal{P}\left(s', r|s, a\right)\left[r + \gamma \max_{a'} q_*\left(s', a'\right)\right] \\
&= \sum_{s',r} \mathcal{P}\left(s', r|s, a\right)\left[r + \gamma v_*\left(s'\right)\right].
\end{aligned}$$

If the MDP is fully known, planning or dynamic programming methods may be used. A model-based RL method learns to estimate $\mathcal{P}$ and $\mathcal{R}$, also it can do planning. A model-free RL method simply learns how to act through a sequence of interactions with the environment in order to maximize reward without building such a model. When the state space is not discrete or is very large, function approximation can be used (e.g., a neural network that approximates the value function or the policy, and is parameterized by θ). The MDP can be augmented to account for multiple agents in the environment, whether they are communicating or not. The three main types of RL are value-based RL, policy-based RL, and actor-critic RL. We explain the differences between them in Section 3.2.3.

3.2.1 Function approximation

Value functions are commonly stored as tables. However, an alternative approach involves approximating this mapping table using a function defined by certain parameters. The function could range from linear functions based on state features to complex multi-layer deep neural networks or even decision trees. This approximation eliminates the need for the agent to traverse every state and action to learn values. Furthermore, when one state experiences

an update, this modification can influence the values of multiple other states. While this method of generalization can enhance the learning process, it may also introduce complexities [142].

Deep Q-Network (DQN): DQN represents the combination of traditional action value function learning (Q learning) and deep neural networks, offering a solution to challenges faced in high-dimensional state and action spaces [109]. DQN incorporates techniques like experience replay that will be introduced next to stabilize and enhance the learning process.

3.2.2 Experience replay

Experience replay captures the agent's interactions over time, storing them in a replay buffer. This buffer collects experiences from numerous episodes within the same environment. For policy updates, the agent periodically draws a subset of experiences, or a mini-batch, from this buffer. While experiences can be sampled uniformly at random, they can also be prioritized, with higher-priority experiences being sampled more frequently. The primary advantage of this replay buffer is improving sample efficiency as it allows repeated use of past experiences. In contrast, without such a buffer, each experience would only be used at most once.

3.2.3 Actor-critic methods

Value-based methods learn the values of actions and/or states to guide action selection. Instead, policy gradient methods focus on learning a parameterized policy that allows autonomously choosing an action given the state, eliminating the need for a value function during this decision-making process. While a value function might still play a role in refining the policy parameter, it is not involved in the action selection process. The policy undergoes updates based on the gradient of some scalar performance measure with respect to the policy parameter. An actor-critic method simultaneously approximates both the policy (referred to as the 'actor') and the value function (the 'critic'), with the latter typically representing a state value function which is used to assess actions.

Soft Actor-Critic (SAC): SAC is an actor-critic off-policy maximum entropy RL algorithm with a stochastic actor [65]. It maximizes both the expected reward and the entropy, allowing the agent to explore more widely and simultaneously consider multiple near-optimal policies. It is shown to have stable performance, and be robust to noise and the choice of hyperparameters. The state value function, soft Q-function, and policy are trained by optimizing:

$$\mathcal{J}_V(\psi) = \mathbb{E}_{s_t \sim \mathcal{D}} \left[\frac{1}{2} \left(V_\psi(s_t) - \mathbb{E}_{a_t \sim \pi_\phi} \left[Q_\theta(s_t, a_t) - \log \pi_\phi(a_t|s_t) \right] \right)^2 \right]$$

$$\mathcal{J}_Q(\theta) = \mathbb{E}_{(s_t, a_t) \sim \mathcal{D}} \left[\frac{1}{2} \left(Q_\theta(s_t, a_t) - R_t + \gamma \mathbb{E}_{s_{t+1} \sim p} \left[V(s_{t+1}) \right] \right)^2 \right]$$

$$\mathcal{J}_\pi(\phi) = \mathbb{E}_{s_t \sim \mathcal{D}, \epsilon_t \sim \mathcal{N}} \left[\log \pi_\phi \left(f_\phi(\epsilon_t; s_t) | s_t \right) - Q_\theta \left(s_t, f_\phi(\epsilon_t; s_t) \right) \right]$$

where π is the policy, ψ, θ, and ϕ are the parameters for state value function, soft Q-function, and policy, R_t is the reward for the (s_t, a_t) pair, γ is the discount factor, p is the state transition probability, $\mathcal{D}$ is the replay buffer, V is the state value, Q is the state-action value, ϵ_t is an input noise vector, and f_ϕ is the unbiased gradient estimator.

Proximal Policy Optimization (PPO): PPO is another state-of-the-art policy-gradient algorithm based on the actor-critic framework [137]. It differs from standard policy gradient algorithms in that it performs multiple epochs of minibatch updates per data sample. PPO has shown strong performance in nearly all reinforcement learning tasks [166] thanks to a clipping method that constrains the update of the behavior policy within a trust region, meaning that the behavior policy remains close to the target policy. This accelerates learning, but the agent might be trapped into a sub-optimal policy. PPO optimizes the clipped surrogate objective given by:

$$L(\phi) = \hat{\mathbb{E}}_t \left[\min \left(\rho_t(\phi, \phi_{\text{old}}) \hat{A}_t, \quad \text{clip}\big(\rho_t(\phi, \phi_{\text{old}}), 1 - \epsilon_t, 1 + \epsilon_t\big) \hat{A}_t \right) \right]$$

with $\rho_t(\phi, \phi_{\text{old}}) = \frac{\pi_\phi(a_t|s_t)}{\pi_{\phi_{\text{old}}}(a_t|s_t)}$, where $\hat{A}_t \doteq G'_t - V_{\pi_{\phi_{old}}}(s_t)$ is an estimator of the advantage function at t. Here G'_t is the discounted reward of the minibatch, $\hat{\mathbb{E}}$ denotes the empirical average over a batch of samples, $V_{\pi_{\phi_{old}}}(s_t)$ is the predicted value of the state s_t under policy $\pi_{\phi_{old}}$, ϵ is a hyperparameter

that discourages making updates that are far from the current policy, and $\text{clip}\big(\rho_t(\phi, \phi_{\text{old}}), 1-\epsilon_t, 1+\epsilon_t\big)$ clips the probability ratio between old and new policies within $[1-\epsilon_t, 1+\epsilon_t]$.

3.2.4 Off-policy Policy Evaluation (OPE)

Off-policy policy evaluation concerns estimating the performance of a given decision-making policy, known as the *evaluation policy*, using historical data that may have been generated by a different *behavior policy*. We denote the historical data as $\mathcal{D} = \{(s_t, a_t, r_t)_{t=1}^n\}$. The most popular OPE methods in the literature are based on importance sampling, examples of which are Inverse Probability Weighting (IPW) [130] and Self-Normalized Inverse Probability Weighting (SNIPW) [143]. In general, SNIPW is shown to be more stable in certain tasks as its value is bounded by the support of the rewards, and its variance is smaller than IPW [77]. Given the evaluation policy π_e and the behavior policy π_b that was used to generate the historical data, the value of π_e (i.e., the expected cumulative reward available from each state–action pair) under IPW and SNIPW is defined as follows:

$$\hat{V}_{\text{IPW}}(\pi_e; \mathcal{D}) \doteq \frac{1}{n} \sum_{t=1}^{n} \rho(e, b) r_t,$$
$$\hat{V}_{\text{SNIPW}}(\pi_e; \mathcal{D}) \doteq \frac{\sum_{t=1}^{n} \rho(e, b) r_t}{\sum_{t=1}^{n} \rho(e, b)},$$

where $\rho(e, b)$ is the importance sample ratio, $\mathcal{D}$ denotes the offline dataset from which the trajectory was sampled, and s_t, a_t and r_t respectively represent the state, action taken, and reward received at time step t. The above-mentioned OPE methods assume that actions are discrete, and use rejection sampling to filter the dataset. However, this approach cannot be extended to work with continuous actions as rejection sampling does not work in the continuous setting [78]. To overcome this limitation, Kallus *et al.* [78] employ kernel density estimation to calculate the value of a policy, which is given by:

$$\hat{V}_{\text{Kernel}}(\pi_e; \mathcal{D}) \doteq \mathbb{E}\left[\frac{1}{h} K\left(\frac{\text{argmax}_{a'_t} \pi_e(a'_t | s_t) - a_t}{h}\right) \frac{r_t}{\pi_b(a_t | s_t)}\right].$$

Here K is the kernel function, such as the Gaussian kernel, and h is the bandwidth, which is a hyperparameter. When a Gaussian kernel is adopted,

we refer to this method as GK.

Fitted-Q Evaluation (FQE) [92] is another policy evaluation method that is shown to perform relatively better than other OPE methods [54]. However, a drawback of FQE is that a neural network needs to be trained for each policy to estimate its value, increasing the computational cost significantly.

3.2.5 Policy evaluation via a proxy

The performance of a neural network can be estimated using a low-cost or Zero-cost Proxy (ZCP), a concept stemming from the field of Neural Architecture Search (NAS) [1], [94]. The underlying principle involves utilizing a mini-batch of data to determine the gradient of loss for each layer. Subsequently, these gradients are consolidated and the result is used as a heuristic measure to evaluate the performance of the neural networks.

Lee *et al.* [94] introduce a saliency metric, called SNIP, that approximates the change in loss when a connection is removed. This helps identify connections in the network that are important to the given task before training the network, using a mini-batch of data. While SNIP was originally proposed for network pruning, it can be used as a proxy for NAS, based on the observation that a neural network that attains a higher SNIP will perform better in the given task [1]. SNIP is defined as

$$\mathcal{S}_{SNIP} \doteq \left| \frac{\partial \mathcal{L}}{\partial \theta} \odot \theta \right|,$$

where $\mathcal{L}$ is the loss function of the neural network with parameters θ, and $\odot$ denotes the Hadamard product operation. Abdelfattah *et al.* [1] empirically evaluate various ZCP metrics to compare their efficiency in ranking neural networks. They also propose a new metric, called *gradnorm* (GN), which can be used for NAS, and is defined as the sum of the Euclidean norm of the gradients after back-propagating the loss computed from a mini-batch of data.

However, in RL, the parameterized policy needs to run on the target environment to calculate the loss value for the use of ZCPs, which is prohibitive. GS *et al.* [64] modify the ZCPs by borrowing importance sampling from OPE, making possible the use of ZCP methods to rank RL policies.

3.3 Simulation environment for RL-based control of buildings

Executing control policies in a simulated environment can help overcome the barriers to real-world deployment of a new control policy by enabling comprehensive evaluation of this policy in buildings with various sizes and occupancy schedules, possibly located in different climates. Furthermore, policy evaluation in a simulated environment is useful for the design of RL controllers that improve a policy in an iterative fashion and benefit from offline training. Existing building energy simulators, such as EnergyPlus [31], can provide an accurate building energy analysis over multiple days within seconds, but they suffer from two fundamental problems. First, they require the full control policy before running a simulation. This prevents users from writing code that interfaces with other simulators and incorporating external models in their control algorithm. Second, they only focus on the number of people present in each zone, neglecting their movements and effects of their actions on the environment. To simulate occupants' actions, it is necessary to track each occupant inside the building and simulate actions (e.g., turning lights on, opening blinds, adjusting temperature setpoints) conditioned on their location. This is crucial as in most cases occupants must be in the proximity of an actuator or a control panel in order to operate it. Due to these shortcomings, recent studies [23] build standalone simulators using BCVTB [157] to generate data for offline learning.

In this thesis, we introduce COBS[1], an open-source and modular simulation platform in Python which is designed to support fine-grained control over the states of multiple building systems (which can be modeled separately) in each step of simulation. It provides the ability to include and exchange data between multiple models, e.g., for state prediction or estimation, allows benchmarking control algorithms across many buildings, and facilitates online learning. Additionally, COBS utilizes an occupancy simulator that generates

[1]COBS can be downloaded from `https://github.com/sustainable-computing/COBS/`. The documentation is available at `https://cobs-platform.github.io`.

realistic individual trajectories and samples interactions between occupants and building systems from conditional probability distributions. These probabilities can be learned from datasets that capture how occupants interact with building interfaces.

COBS is a simulation platform for occupant-centric control of buildings. The objectives of COBS are to improve reproducibility of building control policies, make possible real-time interactions between the control agent and building environment, and enable benchmarking multiple control algorithms under different occupancy conditions. To this end, COBS includes an occupancy schedule generator that utilizes a queueing network to generate realistic occupancy movements and actions in a given building.

3.3.1 Architecture

COBS has been developed with three design goals. First, it must return the observed building state at each time step, and execute user-specified actions to update this state for the next time slot. This is imperative for implementing learning-based control algorithms. Second, it must provide a simple interface for the inclusion of state estimators and predictive models (for room temperature, occupancy, solar radiation). Adding predictions to the state is essential for proactive control of building systems. Third, it needs to interface with data synthesizers and brokers to obtain traces, e.g., for occupant movements and actions. This is in contrast to existing building simulation packages, such as EnergyPlus, which take as input pre-defined schedules for occupants, windows, lights, etc.

Figure 3.2 shows the overall architecture of COBS. It takes the building IDF file created by modeling software, such as SketchUp, OpenStudio, and EnergyPlus, and combines it with the output of an occupancy simulator which determines the location of each occupant and any actions they may perform in that location. The resulting model is used to implement actions in the event queue and simulate the building state using EnergyPlus. The platform takes advantage of a priority scheduling algorithm to schedule various time-stamped events. The simulated building state is then modified and augmented

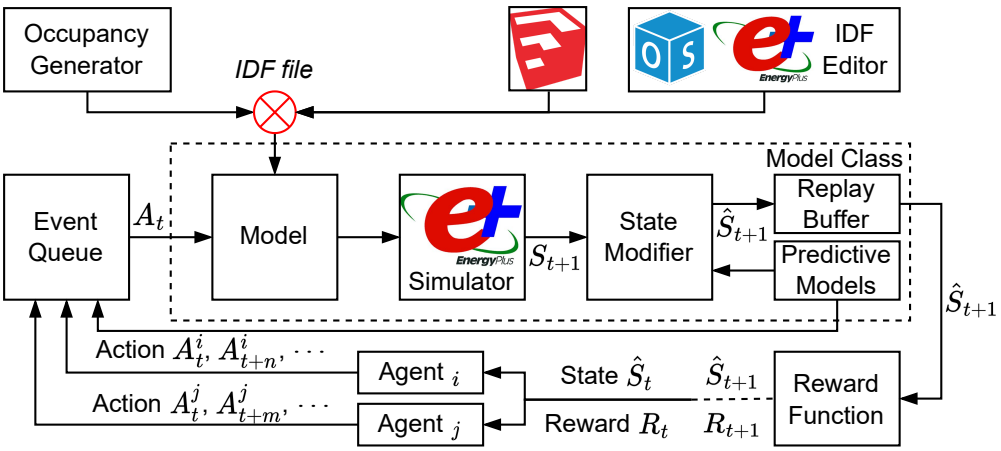


Figure 3.2: COBS's architecture

by several estimation and prediction models. This updated state is then used for reward calculation (given a user-defined function) and sent to the control agent. All actions, rewards, and modified state variables for each time slot are stored in a replay buffer for ease of access in the future.

COBS enables the agent to learn an optimal control policy through direct interaction with the simulated building environment. This is particularly useful for implementing RL algorithms to optimally control HVAC and lighting systems or window blinds, an area that has received increasing attention in recent years [33]. Unlike the environment used in RL, COBS provides the agent with not only historical and real-time data but also future predictions. This improvement makes the design of RL algorithms easier and opens the door to the integration of several models with the building control agent. The platform consists of three main components which are described below.

- **Model Class** consists of a replay buffer, a building model in IDF format, and several models for estimating, predicting, and modifying the state returned by EnergyPlus. The platform provides methods like `reset` and `step` following the same structure as in OpenAI Gym [14]. Thus, RL agents written based on Gym can be ported to our platform with minimal changes. The platform uses multiprocessing and locking mechanisms to support real-time interaction between agents and EnergyPlus, thereby ensuring the action is implemented before simulating the next state.

- **EventQueue Class** uses a priority queue to store and schedule all actions at specified times. The queue determines the order of execution of different actions in each time slot according to their priority and insertion time in the queue. Agents and predictive models can access the queue to retrieve future events and make decisions based on them if needed.

- **OccupancyGenerator Class** exploits a queueing network simulator to produce realistic occupancy schedules. The queueing network is constructed according to the floorplan of the building, probabilities of visiting different spaces upon leaving a space, and the average time spent in each space (terminal zone). We elaborate on this process in Section 3.3.2.

3.3.2 Simulating occupants' movements and actions

Several studies suggest that occupants' presence and actions can greatly affect the energy use and number of thermal comfort violations [6]. For instance, occupants can open/close windows and doors, resulting in a considerable change in the carbon-dioxide concentration, air flow, and room temperature. Control systems respond to this change in different ways to maintain the temperature around the setpoint and meet comfort requirements. This implies that using a fixed occupancy schedule and neglecting actions to evaluate control policies may lead to different conclusions.

To address the challenge of collecting occupancy data, including occupant presence and actions, we propose an occupancy trace generator that draws on queueing theory to generate movement trajectories for occupants. Furthermore, we use the synthesized trajectories to simulate the occupant actions based on the control knobs that exist in the room where the occupant is and the conditional probability provided in the form of a JSON file.

We treat each occupant as a job in the queueing network and each zone as a First-Come First-Served (FCFS) queueing system with infinite servers and exponentially distributed service times. Therefore, each zone is an $M/M/\infty$ queue, and the whole building can be modeled as an open queueing network comprised of N queues which are connected according to the floor plan of the building. Occupants arrive to the building following a Poisson process with a

rate that depends on time of the day. Concretely, the arrival rate is relatively higher in the morning when people are expected to go to work than it is in the afternoon.

The stay time in each zone depends on its function for each occupant. Occupants stay longer in their designated office space and shorter in other spaces in the building. Movements inside the building are governed by a probability which is higher for returning to their office and lower for visiting other spaces upon leaving a space. The time spent moving between spaces is also considered. We assume the service in this queueing network can be interrupted by a number of events, including the lunchtime and start of a meeting.

After simulating the rooms visited by each occupant for a given simulation period, we calculate the total number of occupants in each zone and store it in the building model. The occupant location is then used to simulate their actions; they must be sampled separately for each time slot because their occurrence may depend on the current state of the building. Hence, in each time step, we filter out infeasible actions in all occupied zones and decide whether to simulate an occupant's action by sampling from a conditional probability distribution.

Chapter 4

Controlling multiple building systems via reinforcement learning

Commercial buildings are comprised of mechanical and electrical systems that work in tandem to provide a healthy, safe, and comfortable environment for occupants. These systems have complex interactions with each other, and consume a large amount of energy. In this chapter, we apply three model-free deep reinforcement learning algorithms to jointly control HVAC and blind systems in a multi-zone test building, in scenarios with and without automatic dimming of the lights in response to daylight levels. The control agents are trained through interactions with a building simulator that generates traces for the movement of occupants. We investigate the three-way trade-off between energy use, thermal comfort, and visual comfort, and discuss how the joint control of the building systems could provide a better trade-off compared to when they are controlled separately. We compare the performance of the proposed control algorithms assuming the availability of occupancy data with two spatial resolutions, and confirm through experiments that a better trade-off can be achieved should zone-level occupancy information become available. By incorporating zone-level occupancy information, we show that 11.0% and 31.8% more energy can be saved respectively in heating and cooling seasons over existing rule-based baselines that control the same building systems.

4.1 Introduction

Commercial buildings consume a significant amount of energy worldwide. Driven by the global threat of climate change, extensive research has been

done in the past few decades to explore how to save on the energy used by building systems, while maintaining thermal and visual comfort of occupants. In particular, various rule-based, model-based, and model-free control techniques have been employed to obtain energy-efficient operation policies for HVAC, lighting, and blind systems. Rule-based techniques are based on a set of control rules defined by the facilities manager. Model-based techniques take advantage of a physics-based or data-driven dynamic model that explains the state evolution (e.g., heat transfer, air flow, occupant movement), whereas model-free techniques aim to learn a control policy through interactions with building systems or a simulated environment. Model-free techniques are more promising when the goal is to control multiple building systems with complex interactions that cannot be precisely modeled [40].

Despite the tremendous progress toward energy-efficient control of building systems, there are several important questions that are yet to be addressed. We outline these research questions below:

1. **How does the joint control of building systems affect the whole-building energy use?** Due to the complex interactions between building systems, the control decisions made in one system could affect the performance of the other ones. For example, closing blinds in an overheated zone may reduce the energy use of the HVAC system during the day, but this comes at the price of increasing the energy use of the lighting system because lights must be turned on to satisfy the visual comfort requirement. Dimming lights, on the other hand, reduces the amount of energy used for lighting but it may also change the HVAC energy consumption as it influences the internal heat gain. Modeling interactions between building systems in addition to the uncertainty of the environment is indeed a difficult task. To contain complexity, related work either controls a single building system [23], [117], [175], neglecting the interplay between this system and the other systems, or considers the interactions between two or more systems in a single zone [25], [27], [40]. To our knowledge, there is no work that quantifies the amount of

energy that can be saved in a multi-zone building when building systems are controlled jointly.

2. **What are the best trade-offs between energy use, thermal comfort, and visual comfort?** The trade-off between energy use and thermal comfort has been widely studied in the context of optimal HVAC control. However, there is little work that navigates the three-way trade-off between energy saving, thermal comfort, and visual comfort. This is a barrier to the deployment of the control techniques in real buildings as the facilities manager cannot easily trade energy savings for extra comfort (and vice versa). Ideally, they should be able to tweak some weight parameters to make trade-offs within Pareto-efficient choices.

3. **Will incorporating zone-level occupancy information noticeably change the performance of a control policy?** To make possible higher energy and cost savings without compromising comfort, most control techniques incorporate occupant presence or count information at the building level. This makes sense because estimating the number of occupants in each zone is difficult without having a number of sensors installed there. Should this information become available, the thermal and visual discomfort can be calculated for each individual occupant that is present in a given zone. However, it is unclear whether incorporating high spatial resolution occupancy data could help achieve a better trade-off between energy consumption, thermal comfort, and visual comfort.

4. **How does the performance of a given control policy vary across seasons?** The outside air temperature can affect the performance of optimal HVAC control algorithms, so studies often train agents for more than one season. For the joint control of building systems, the difference between seasons can become even more prominent. For example, to improve thermal comfort and reduce energy use, blinds should be open during the day in winter to heat up the building, and vice versa in

the summer. Understanding how the control performance varies across seasons and whether there is a specific model-free control algorithm that outperforms others in all cases requires a comprehensive evaluation of building controls in both heating and cooling seasons. This has not been explored in previous work and will provide insight into control strategy selection.

To address these questions, this chapter studies the joint control of HVAC, lighting, and blind systems in a five-zone test building modeled in EnergyPlus [31]. We use state-of-the-art deep RL algorithms to determine the optimal control policies for HVAC and blinds, while a daylight auto-dimming strategy is used for lighting control. These algorithms are suitable for this problem because they can handle large state and action spaces, and learn the complex interactions between multiple building systems[1]. We make three specific contributions:

- We utilize three model-free RL algorithms to train agents that can jointly control the supply air temperature and blind angle setpoints for every zone in our test building. These include two actor-critic algorithms, namely PPO and SAC, and a Q-learning algorithm, called Branching Dueling Q-Network (BDQN). We evaluate these algorithms in different scenarios in terms of the achieved reward, convergence speed, and stability, following the guidelines provided in [159].

- We investigate the three-way trade-off between energy consumption, thermal comfort, and visual comfort. We discuss the best weight factors for the terms in the reward function; these weights will allow for maximizing energy savings while keeping thermal and visual discomfort below specified thresholds.

- We compare the performance of these algorithms with existing baselines in heating and cooling seasons with building-level and zone-level

[1]Code is available at `https://github.com/sustainable-computing/COBS-joint-control`

occupancy information. We show that the energy use would be further reduced if we knew the occupancy state of all zones in a building. This highlights the importance of monitoring or estimating the occupancy state of every zone through multimodal sensor fusion. Incorporating zone-level occupancy information, we show that 11.0% and 31.8% more energy can be saved, in heating and cooling seasons respectively, over existing rule-based baselines that control the same building systems.

4.2 Methodology

We explore rule-based and RL-based joint control of the building systems for a 5-zone office building in Pittsburgh, Pennsylvania in January (heating season) and July (cooling season). The floor area of this building is 5,000 square feet and it has been used in previous work [23], [138]. The control setpoints that we adjust using different algorithms are supply air temperature setpoint and blind angle setpoint. The building is simulated in the EnergyPlus [31] environment and is controlled via the COBS described in Section 3.3 which interacts with EnergyPlus. COBS is used to programmatically execute rule-based control scenarios and to train the RL agents. The office building we control is depicted in Figure 4.1, and the relevant design details are described in the following sections. We now present our control scenarios, problem formulation, and model-free RL algorithms.

4.2.1 Simulation environment

HVAC design: As explained in Section 3.1, the HVAC system is a packaged VAV system with one heating coil and one cooling coil in addition to VAV reheat coils. Similar to [23], we use the SAT setpoint as the HVAC control point. Other VAV setpoints are controlled using a feedback control strategy. The VAV reheat coils are turned off in the cooling season.

Occupancy: Two occupancy conditions are considered, namely building-level and zone-level occupancy schedules. The building-level occupancy schedule assumes that all zones are occupied from 8:00 and 18:00. The zone-level occupancy schedule determines the number of occupants that are present in

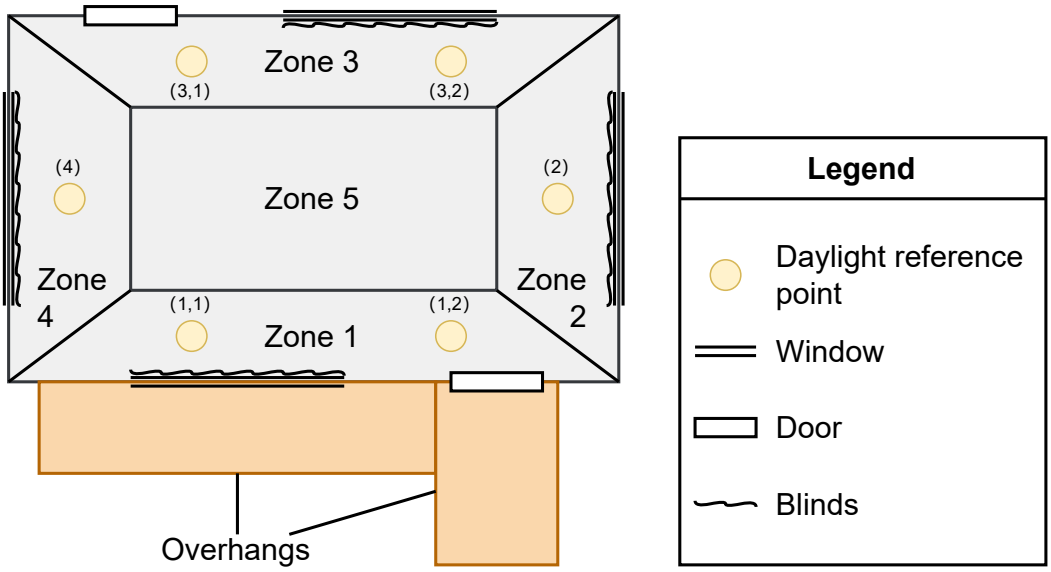


Figure 4.1: The layout of the medium office building studied in this chapter, including the daylighting reference points.

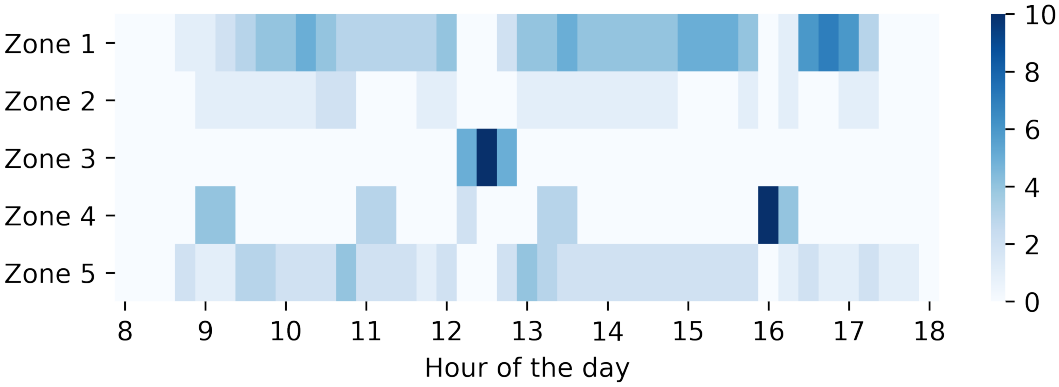


Figure 4.2: The number of occupants in each zone during working hours.

each zone at any given point in time. This helps model the amount of heat emitted by occupants and better assess thermal and visual comfort of the occupants in each zone. We use the COBS platform to generate several zone-level occupancy schedules. Figure 4.2 illustrates the number of occupants in each zone throughout a day.

Window blinds: White painted metal blinds are present on the windows in all four perimeter zones. Each slat is 2.5 cm wide and the separation between slats is 1.875 cm. We assume that windows are not motor-operated, hence the blind angle and position can be adjusted without any constraints.

Table 4.1: Control scenarios and corresponding baselines.

HVAC	Blinds	Lighting	Baseline
SAT setpoint	Always open	Not controlled	(1)
SAT setpoint	Always open	Auto dimming	(3)
SAT setpoint	Using the same setpoint	Not controlled	(2)
SAT setpoint	Using the same setpoint	Auto dimming	(4)
SAT setpoint	Using different setpoints	Not controlled	(2)
SAT setpoint	Using different setpoints	Auto dimming	(4)

Daylighting: Zonal illuminance values are used for rule-based lighting control and to evaluate the visual comfort of the occupants. They are measured at desk height (76.2cm) using daylighting reference points in EnergyPlus, the positions of which are specified in Figure 4.1. Zone 5 does not have any daylighting reference points because auto dimming does not occur in zones without windows. We turn on the lights when Zone 5 is occupied and turn them off when occupancy is zero.

4.2.2 RL problem formulation

In this section we describe the MDP framework, including state and action spaces, and the reward function. At each time step, the building and its surrounding environment are in some state s_t. The agent exerts a control action a_t to control building systems. This action causes a random state transition to s_{t+1}. The RL agents are trained in six specific scenarios for controlling HVAC, blind and lighting, as outlined in Table 4.1. Through interactions with the simulated environment, each agent learns an optimal *policy* π, that is a sequence of control actions starting from state s. When blinds are controlled, the agent either learns a policy that adjusts all the blind setpoints in the same way, or a policy that adjusts them independently. Note that lighting is not controlled by the RL agent. Hence, there is either no lighting control or the auto dimming strategy is adopted.

State: The state at time t, denoted by s_t, consists of the following observations: the temperature in each zone including the plenum ($^{\circ}C$), the number of occupants in each zone (for building-level occupancy schedule, all zones share

the same value of 0 or 1, indicating whether the building is occupied), the hour of the day (0-24), the slat angle of the blinds (degrees) in each of the four zones that have windows, the ambient temperature ($°C$), and the site solar radiation (W). In addition to these observations, it contains the ambient temperature and site solar radiation for the next twelve 15-minute time steps. These forecast are assumed to be perfect. Thus, each state consists of 18 observations and 24 predicted values.

Action: The action at time t, denoted by a_t, determines the control decision made in each building system. The action space differs depending on the control scenario and the agent type. The control scenario determines the number of control points while the agent type affects the range of possible actions pertaining to a control point. We always control the SAT setpoint for each control scenario in a range of $[-20, 20°C] + T_{MA}$, where the T_{MA} is the mixed air temperature. The blinds can be controlled with different setpoints, jointly according to the same setpoint, or not controlled at all; in the latter case it is assumed that blinds are not available in the building. The action for each blind is between 0 and 180.

The SAC agent (described in Section 3.2.3) considers a continuous action space for each control point, while other RL agents consider discrete actions[2]. In particular, we discretize the action for SAT setpoint to 20 and blinds to 18 evenly spaced values. Therefore, in the most complex control scenarios, where we control the blinds using different setpoints, the action space is 5-dimensional for the SAC agent and there are 2,099,520 ($18^4 \times 20$) possible actions for other agents. This large action space makes it difficult to find the optimal policy.

To effectively find the optimal policy with this large action space, we deploy a feature sharing neural network for each agent. That is, instead of having a large number of cells for all possible actions, after a few hidden layers we create multiple branches in the neural network. The number of branches is the same as the number of control points we have in each scenario. For example,

[2]We got better results when we discretized the action space for the other two agents.

we have 5 distinct branches when we have different setpoints for blinds (4 branches for blind setpoints and one for the SAT setpoint). The same idea was used in [40] to reduce the size of neural networks.

Reward: The reward function balances three competing objectives: the total facility energy consumption including both the HVAC system and lights (denoted by E), the occupant thermal comfort (denoted by T_c), and the occupant visual comfort (denoted by V_c). It can be written as follows:

$$R = -\rho_E Norm(E) - \rho_T Norm(T_c) - \rho_V Norm(V_c) \tag{4.1}$$

where ρ_E, ρ_T and ρ_V are weight factors (reward parameters) that represent the relative importance of different terms in the reward function. These parameters can take values from $\{0.1, 0.4, 0.7, 1.0\}$. We consider all reward functions that are obtained by assigning these values to the parameters in a combinatorial fashion. The $Norm()$ function is defined as:

$$Norm(x) = (x - x_{min})/(x_{max} - x_{min}). \tag{4.2}$$

It is used to scale each term in the reward function. The process used to calculate E, T_c and V_c is described next.

Energy consumption: Since both HVAC and lighting systems run on electricity only in our test building, we use the total electricity consumed by HVAC and lighting systems as a measure of the total facility energy use:

$$E = E_{HVAC} + E_L \tag{4.3}$$

where E_{HVAC} is the electricity consumed by the HVAC system and E_L is the electricity consumed by the lights located in zones that have windows[3] (both are expressed in Wh). Note that we do not take into account the energy consumed to operate the blinds since it is negligible compared to the other components.

Thermal comfort: The occupant thermal comfort is calculated according to the PMV specified by Fanger's model [46], which has been used in building

[3]We ignore the electricity consumption of lights in Zone 5, which does not have a window, since we cannot affect this energy consumption.

control since the 1960s. The PMV predicts the average vote of a group of people on a 7-point index ranging from $+3$ = hot to -3 = cold. Both the ISO 7730 standard [71] and ASHRAE [4] recommend maintaining $|PMV|$ below 0.5. Thus, we calculate T_c at a given time step as follows:

$$T_c = \frac{\sum_i O_i \cdot T_{ci}}{\sum_i O_i} \tag{4.4}$$

where T_{ci} represents the thermal comfort in zone i given by:

$$T_{ci} = \begin{cases} 0, & |PMV_i| \leq 0.5 \\ |PMV_i| - 0.5, & otherwise. \end{cases} \tag{4.5}$$

PMV_i and O_i indicate respectively the PMV value and occupancy state of zone i. O_i is 1 when zone i is occupied and 0 otherwise.

Visual comfort: In this study, visual comfort is determined using the illuminance rates at the daylighting reference points (see Figure 4.1). A penalty is applied when the illuminance rates either do not meet or exceed engineering standards for visual comfort. According to the Illuminating Engineering Society of North America, the comfort range for office lighting is between 300 lux and 750 lux [69]. Thus, the visual comfort reward for zone i is given by

$$V_{ci} = \begin{cases} 0 & 300 \leq \mathbb{E}\left[I_i\right] \leq 750 \\ 300 - \mathbb{E}\left[I_i\right], & \mathbb{E}\left[I_i\right] < 300 \\ \mathbb{E}\left[I_i\right] - 750, & \mathbb{E}\left[I_i\right] > 750, \end{cases} \tag{4.6}$$

where $\mathbb{E}\left[I_i\right]$ is the expected illuminance rate in zone i. The illuminance values, I_i, are obtained from the daylighting reference points labeled in Figure 4.1. We take the average of the illuminance values of the reference points in each zone and denote it by $\mathbb{E}\left[I_i\right]$. Notice that the illuminance value will never fall below 300 lux during the occupancy time as the indoor artificial light can always provide enough illuminance when they are on. Then, the total visual reward V_c is calculated as follows:

$$V_c = \frac{\sum_i O_i \cdot V_{ci}}{\sum_i O_i}. \tag{4.7}$$

4.2.3 Deep reinforcement learning algorithms

We use three model-free RL algorithms and one model-based RL algorithm to control building systems. These algorithms are SAC, PPO, BDQN, and model-based BDQN.

SAC (described in Section 3.2.3): In this study, we use Adam optimizer with a learning rate of 0.0003. We set the discount factor to 0.99 and consider a batch size of 256. We use a squashed Gaussian policy with two hidden layers and 256 cells in each layer for the actor network. For the critic network, we use a network with two 256-cell hidden layers with the leaky rectified linear unit (ReLU) as the activation function. We use automatic entropy tuning which allows the agent to automatically balance exploitation and exploration.

PPO (described in Section 3.2.3): We use two hidden layers with 100 units each layer, utilizing the leaky ReLU activation function for both actor and critic networks. After two hidden layers, the actor network has multiple branches, one for each actuator type. We set the learning rate to 0.0005 and the discount factor to 0.99.

Branching Dueling Q-Network (BDQN): BDQN is a branching variant of the dueling double deep Q-network [156]. It is an off-policy algorithm which is shown to outperform various algorithms such as Deep Deterministic Policy Gradient (DDPG) in high dimensional action spaces tasks [145]. For comparison with previous work, we use exactly the same settings that are used in [40]. The Q-value for each branch d and the maximum accumulated reward can be written as:

$$Q_d\left(s, a_d\right) = V\left(s\right) + \left(A_d\left(s, a_d\right) - \frac{1}{n}\sum\nolimits_{a'_d \in \mathcal{A}_d} A_d\left(s, a'_d\right)\right)$$

$$R_d = R + \gamma \frac{1}{N}\sum\nolimits_d Q_d\left(s', \arg\max_{a'_d \subseteq \mathcal{A}_d} Q_d\left(s', a'_d\right)\right)$$

where $\mathcal{A}_d$ is the set of actions that can be taken on branch d, and A_d represents the advantage function.

Model-based BDQN: We have also considered a model-based version of BDQN for comparative analysis. This design mirrors the structure of the model-free BDQN, but it uses an additional DNN that consists of four hidden layers, each with 300 cells. This network, which uses the rectified linear unit (ReLU) as its activation function, is designed to model the building dynamics. It takes the current state and action as the input and predicts the next state

and anticipated reward. The model continuously learns the building's dynamics from the historical trajectories collected over time. The BDQN updates its policy parameters based on the interactions with the building model; this is called planning in the RL literature. In this study, the planning step is set to 12. We label this model-based BDQN as 'Planning'.

4.2.4 Training RL agents

We split the task into two seasons: winter and summer. The winter season model only uses January data to train and test, and the summer season model only uses July data to train and test. We assume that each episode is one month long and is comprised of 2,976 15-minute time steps. We use 400 episodes to train the RL agents in each season. EnergyPlus is used to simulate the building environment after each epoch. We use the historical weather data in Pittsburgh to get the outdoor temperature and solar radiation for the current time step and future predictions.

4.3 Evaluation metrics and baselines

We evaluate the RL agents in six different control scenarios and compare their performance with four existing baselines. Four metrics are used for performance evaluation: the total electricity consumption of the month, average thermal comfort over the month, thermal comfort violation rate of the month, and visual comfort violation rate of the month. The thermal comfort violation rate is defined as the percentage of time that the absolute value of PMV averaged over all occupied zones is greater than 0.5 when the building is occupied. We define the visual comfort violation rate similarly.

We consider four rule-based baselines that are implemented in EnergyPlus for each season: (1) HVAC only, (2) HVAC & blinds, (3) HVAC with auto-dimming, and (4) HVAC & blinds with auto-dimming. The performance of the RL agents is compared to the respective rule-based baselines based on the control scenario (see the last column of Table 4.1).

HVAC: For all baselines, the supply air temperature is controlled by EnergyPlus using the SetpointManager:Warmest/Coldest object that at-

tempts to meet the heating load for multiple zones at a time. Details of the control strategy can be found in the EnergyPlus documentation[4]. In short, the setpoint manager calculates the average SAT that is required to meet the zones' heating/cooling loads based on the supply air mass flow rates, and adjusts the SAT setpoint accordingly.

Blinds: When blind control is included, predefined EnergyPlus programs that are designed to reduce heating and cooling load are used. Specifically, the blinds are closed in the heating season if it is nighttime and the outdoor temperature is below a setpoint. In the cooling season, the blinds are kept open at night, and closed during the day only if the solar radiation on the window exceeds a setpoint. The setpoints were chosen by trying a wide range of values to find the ones that performed the best in terms of the whole-building energy use, and thermal and visual comfort.

Daylighting: When lighting control is included, the DAYLIGHTING:CONTROLS object is used so that the overhead lights dim continuously as the daylight illuminance increases[5]. The lights are always turned off during the night with and without auto-dimming.

4.4 Results

In this section we first evaluate the performance of the four baselines and four RL-based control strategies. We present the trade-offs between whole-building energy consumption, thermal comfort, and visual comfort, and discuss which agent yields better trade-offs for each control scenario. We discuss the best trade-off that can be achieved using each RL algorithm and compare them with rule-based baselines. Finally, for a fixed set of the reward parameters, we explain how incorporating zone-level occupancy information would impact the trade-off curves in both heating and cooling seasons, and compare the control agents in terms of the reward they eventually achieve, their convergence speed, and stability across several random runs.

[4]Refer to `https://bigladdersoftware.com/epx/docs/9-3/input-output-reference/group-setpoint-managers.html#setpointmanagerwarmest`

[5]The overhead lights dim linearly when the illuminance increases and stay on with the minimum input power if illuminance surpasses a certain threshold.

4.4.1 A closer look at baseline strategies

We analyze the performance of the four baselines presented earlier; they are rule-based control strategies that incorporate occupancy information. The black stars in Figure 4.4 show the performance of these baselines in respective control scenarios in both heating and cooling seasons with different occupancy schedules. To save space, we only discuss the results obtained when zone-level occupancy information is incorporated. Numerical values are provided in Table A.1 in the appendix. In the cooling season, using rule-based controllers for HVAC and blinds (Baseline 2) or using auto-dimming in addition to rule-based HVAC control (Baseline 3) reduces the total energy consumption by 12% and 28% compared to Baseline 1 which controls HVAC only. Controlling HVAC and blinds with auto-dimming (Baseline 4) yields 32% more savings than controlling HVAC alone (Baseline 1) and around 5% more savings than controlling HVAC with auto-dimming (Baseline 3). This is because blinds can reduce the solar heat gain during the daytime and provide sufficient natural lighting, thereby lowering the energy use.

Controlling blinds and HVAC with a rule-based strategy (Baseline 2) in the heating season also helps reduce the total energy use by 15% over Baseline 1. Yet, unlike the cooling season, adding auto-dimming to Baseline 1 does not reduce the energy use. This is likely because lighting gives off excess energy as heat, hence turning off the lights results in higher heating requirements from the HVAC system. Controlling HVAC and blinds together with auto-dimming (Baseline 4) enables the highest energy savings in the heating season, i.e., 18% reduction in energy use over Baseline 1.

In conclusion, our results show an average energy savings of 26% across both seasons when all three systems are controlled (Baseline 4) compared to when only HVAC is controlled (Baseline 1). This observation motivates the joint control of building systems using more advanced control strategies. In terms of thermal comfort, all baselines were able to meet the ASHRAE PMV requirement. However, their performance is rather poor in terms of visual comfort because the default blind control strategy only closes the blinds at

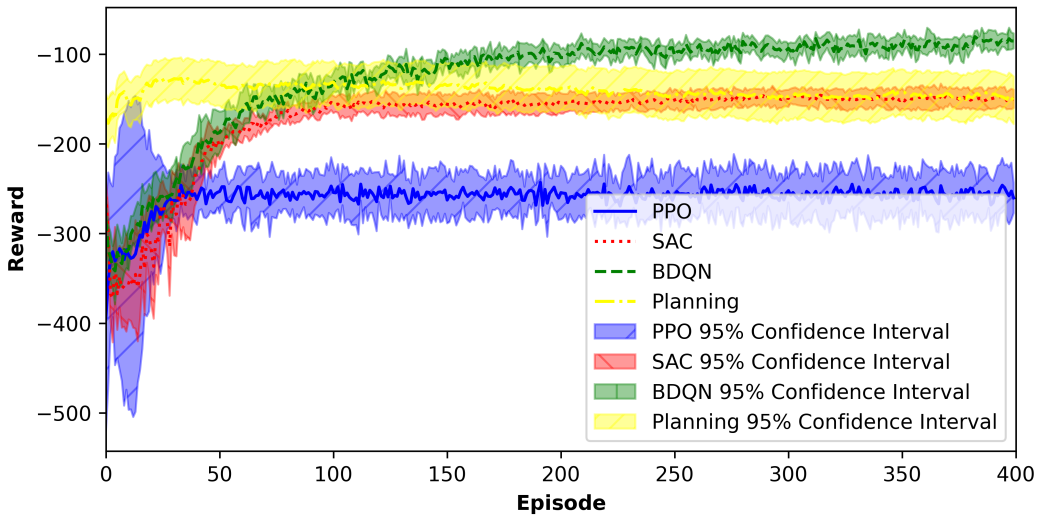


Figure 4.3: Performance comparison of four RL algorithms on the building control domain. The mean and 95% confidence interval of the episode reward are computed based on 10 independent runs in the cooling season.

night. As a result, illumination is always high in the perimeter zones.

4.4.2 Performance, convergence rate and stability

Figure 4.3 illustrates the total reward accumulated in each episode when RL agents control the SAT setpoint and 4 blind setpoints, and lights are auto-dimmed. The episode reward is averaged across 10 runs with different random seeds. The shaded region around the average episode reward depicts the 95% confidence interval. The four RL agents are trained for 400 months and then tested over a period of 200 months in our simulated building. As it can be seen the agents have stable performance in the testing period.

We first compare the performance of the model-free BDQN and the model-based BDQN. As illustrated in Figure 4.3, planning shows faster convergence. This can be attributed to the data efficiency of the model-based BDQN algorithm, which learns from simulated experiences. However, while BDQN takes longer to converge, it achieves a higher reward and has a tighter confidence interval. BDQN learns directly from real experiences, which mitigates the estimation bias often introduced by model-based RL during planning. Since the model-free approach eventually outperforms the model-based algorithm, our subsequent discussions in this thesis will exclusively focus on model-free RL

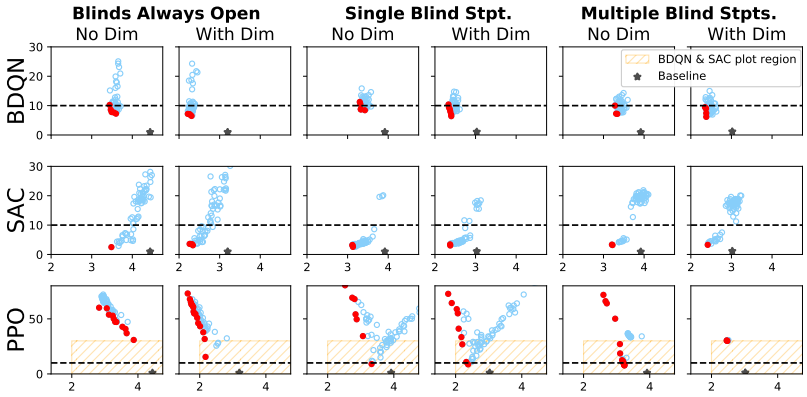
algorithms.

Focusing on model-free RL algorithms, it is evident that BDQN converges to the highest reward, followed by SAC. Looking at the convergence speed, PPO, SAC, and BDQN agents converge at around 30, 100, and 200 episodes, respectively. SAC and BDQN agents show more stable performance (narrower confidence interval) compared to the PPO agent. They have better sample complexity and can effectively use experiences from previous episodes to update the policy. Unfortunately this means that their running time is higher than PPO and they use more memory. In particular, SAC and BDQN agents finish a run in 38 and 31 hours respectively on a server with Intel Xeon E5-2650 v4 (2.2GHz CPU) and NVIDIA Tesla P100 GPUs, and need around 8GB of memory. For PPO, on the other hand, it takes only 7 hours to run on the same server using 4GB of memory.

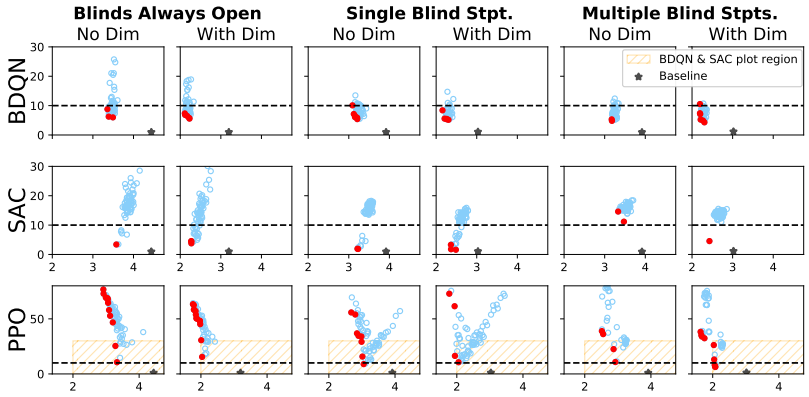
While Figure 4.3 only shows the average reward per episode in the cooling season with zone-level occupancy information and a specific reward parameter setting ($\rho_E = 1$, $\rho_T = 1$ and $\rho_V = 0.4$), we witnessed similar convergence behavior for other reward parameter settings, months, and occupancy schedules.

4.4.3 Identifying three-way trade-offs

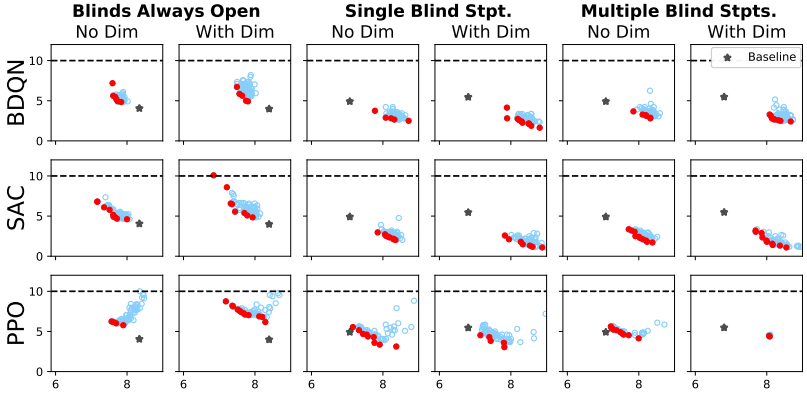
As described in Section 4.2.2, we assess the control performance of RL agents for various combinations of reward parameters $\rho_E, \rho_T, \rho_V \in \{0.1, 0.4, 0.7, 1.0\}$. Figure 4.4 shows the trade-offs between energy use and thermal comfort offered by the three RL agents in six different scenarios with two types of occupancy schedules. The visual comfort is the third dimension which is not shown in this figure. Each reward parameter setting yields a specific trade-off between the competing objectives, which is depicted by a circle in this figure. The Pareto optimal values are painted in red, and the baseline strategy for each scenario is marked with a black star. Notice that in the cooling season, the result for PPO spreads widely. Therefore, the axis limits for PPO are different from SAC and BDQN. Hatch-filled orange rectangles indicate the axis limits of SAC and BDQN plots on PPO plots. Among the three RL algorithms we considered, PPO seems to be the most sensitive to reward parameters.



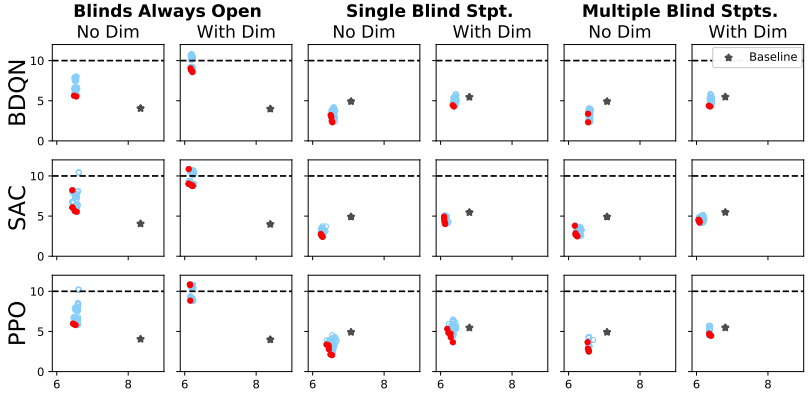
(a) Cooling season with building-level occupancy schedule



(b) Cooling season with zone-level occupancy schedule

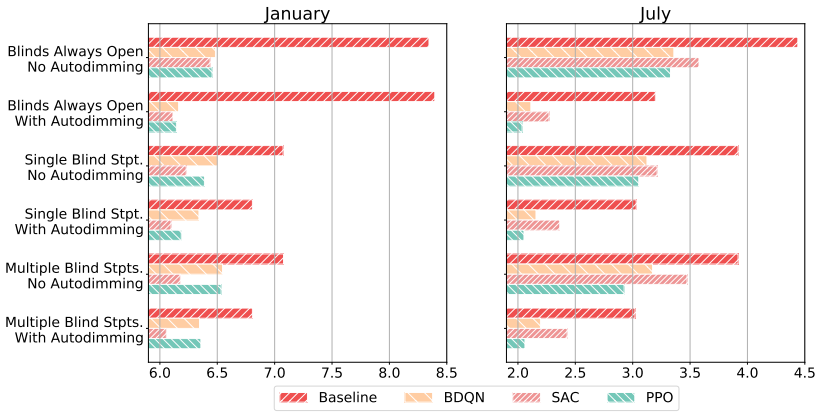


(c) Heating season with building-level occupancy schedule

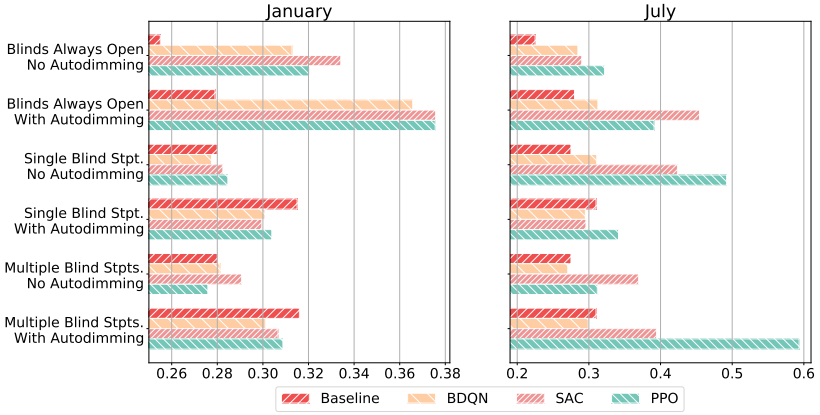


(d) Heating season with zone-level occupancy schedule

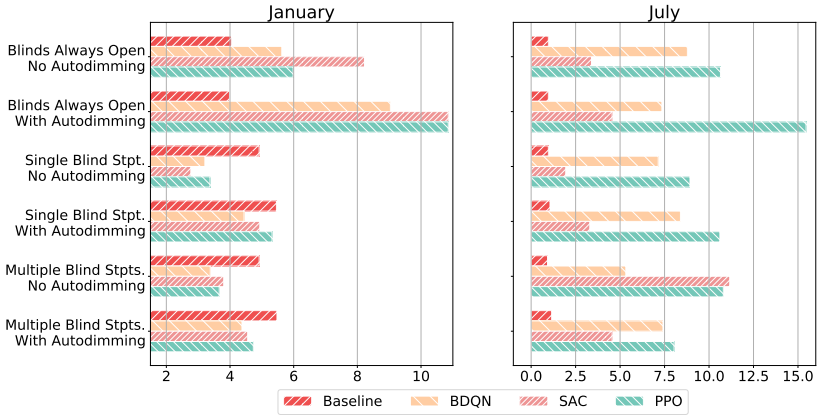
Figure 4.4: The PMV violation rate (y-axis) versus the monthly electricity consumption in MWh (x-axis) for different reward parameters. Points on the Pareto frontier are colored red and baselines are marked with black stars. The horizontal line shows ASHRAE's threshold (10%) for thermal comfort violation [4].



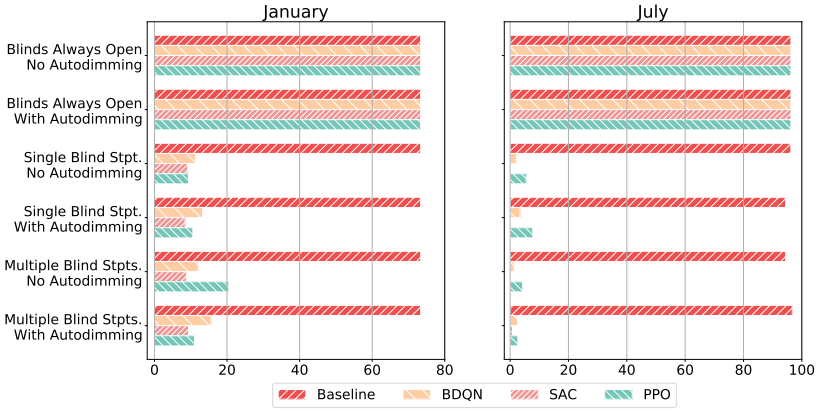
(a) Energy Consumption (MWh)



(b) Thermal Comfort ($|PMV|$)



(c) Thermal Comfort Violation (%)



(d) Visual Comfort Violation (%)

Figure 4.5: Comparison of different RL agents in different control scenarios using a zone-level occupancy schedule. The results are obtained using the best set of reward parameters for each RL agent. The x-axis is exaggerated.

Nevertheless, we observe that for all three agents it is possible to navigate the three-way trade-offs by tweaking the reward parameters.

Best trade-offs: To determine the reward parameter setting that yields the 'best' trade-off, we first filter out the parameter settings that result in a PMV violation rate higher than 10% (ASHRAE's threshold [4]). We then choose the parameter setting that minimizes the whole-building energy use among the remaining choices. If the PMV violation rate exceeds 10% for all parameter settings, we choose the parameter setting that minimizes the product of the whole-building energy use and excess discomfort (i.e., the PMV violation rate minus 10%). The trade-off that corresponds to this parameter setting is called the best trade-off. For simplicity, the illumination violation rate is not considered in the process of finding the best trade-off as it is typically in the acceptable range.

Figure 4.5 provides a comparison between the best trade-offs achieved by each RL agent in different scenarios. Numerical values are provided in Tables A.1 and A.3 in the appendix. Compared to the baselines, the RL agents can save a significant amount of energy while meeting both thermal and visual comfort requirements in most cases. In scenarios where the blind setpoint is controlled, all agents achieve a significant improvement in visual comfort compared to the baselines in both seasons. This implies that the RL agents are able to learn how to use blinds to limit the amount of glare from sunlight. SAC has the lowest visual comfort violation rate in all scenarios. It is worth mentioning that in the scenario where the SAT setpoint and multiple blind setpoints are controlled with auto-dimming, the best RL agent can reduce the whole-building energy use by 11% in heating season and 31.8% in the cooling season over Baseline 4.

4.4.4 Incorporating occupancy information

We evaluate the control performance using both building-level and zone-level occupancy information. Figures 4.4a and 4.4c show the whole-building energy use in cooling and heating seasons along with the thermal comfort violation rate when the RL agents incorporate building-level occupancy information.

Figures 4.4b and 4.4d show the same result this time assuming that the agents incorporate the zone-level occupancy information. It can be readily seen that better trade-offs can be achieved in the heating season when the control agents incorporate zone-level occupancy information. Specifically, BDQN, SAC, and PPO agents can save respectively 3.3%, 18%, and 14% more energy when they take into account zone-level occupancy information rather than building-level occupancy information. The zone-level occupancy information allows the control agents to meet thermal and visual comfort requirements by conditioning only a subset of zones that are occupied. This reduces the energy consumption in HVAC and lighting systems. Interestingly, incorporating zone-level occupancy information does not appear to offer much in terms of energy savings in the cooling season. We attribute this to the fact that in Pittsburgh less energy is consumed to keep the room temperature within the comfort range in the cooling season than in the heating season. Hence, a smaller amount of energy can be saved by not conditioning the unoccupied zones.

Another important observation is that the RL agents cannot always beat the rule-based baselines when they rely on building-level occupancy information. For this reason, we only present the results when a zone-level occupancy schedule is used in the remainder of this section. The performance results for both cases can be found in the appendix (Tables A.1-A.4).

4.5 Discussion

We now return to the four research questions raised in the introduction, followed by a discussion of the key differences between the three model-free control strategies. Many of our findings are novel and provide valuable insight for future research.

How does the joint control of building systems affect the whole-building energy use? The paper that came closest to addressing this question is [40], where the BDQN agent achieved savings compared to rule-based methods, showing the potential of applying model-free RL to the joint control of building systems. However, their baselines included only rule-based HVAC control, even though rule-based blind and lighting strategies have been proven

to offer significant savings. For example, in our building the rule-based control of all systems (Baseline 4) reduced the whole-building energy use by 26% on average across both seasons compared to the control of HVAC only (Baseline 1). Our work shows for the first time that RL-based control saves even more energy than rule-based HVAC and blind control, and that incorporating autodimming increases savings even further. Furthermore, we show that this is true even when generalized to the multi-zone scenario.

We provide numerical evidence in Table A.1 that motivates the installation of dimmable lights and motorized blinds. Dimmable light can always lower the total energy consumption, especially during the summer, and blinds can slightly reduce the energy use as well. Figure 4.4 shows a minor improvement in controlling the blinds with a single setpoint over separate setpoints. This can be used to reduce the action space dimension, simplifying the problem.

What are the best trade-offs between energy use, thermal comfort, and visual comfort? The tradeoffs between energy use and thermal comfort are plotted directly in Figure 4.4. With regards to tuning these two objectives, BDQN and SAC are less sensitive to the reward parameters, whereas PPO is highly sensitive to the reward parameters. Interestingly, Figure 4.5 shows that all of the RL agents (including PPO) easily improved visual comfort over the rule-based baselines. One way to interpret this is that there is a lot of room for improvement in rule-based blind control strategies, with respect to visual comfort. Overall, visual comfort is relatively easy to optimize without much tuning, but the trade-off between energy use and thermal comfort is more complicated to navigate.

Will incorporating zone-level occupancy information noticeably change the performance of a control policy? As highlighted in Section 4.4.4, the inclusion of zone-level occupancy offered noticeable energy savings over building-level occupancy in all cases except for SAC in the cooling season. Figure 4.4 shows that when blinds are included in the heating season, zone-level occupancy is actually required to achieve lower energy use than the rule-based baseline which takes occupancy into account. Based on this result (and the simplicity in aggregating zone-level data up to the building-level) we

argue that incorporating zone-level occupancy information to train RL agents is a viable energy reduction strategy.

Our back-of-the-envelope calculation shows that we can save approximately 1.04 MWh in two months (January and July) by incorporating zone-level rather than building-level occupancy information. With extrapolation, the annual energy and cost savings will be respectively 6.24 MWh and $437, assuming a flat rate of 7¢/kWh. This can offset the cost of buying and installing occupancy sensors in the 5 zones.

How does performance vary across seasons? A trend in RL papers for building control is to present results for two seasons and conclude that the agent can find an optimal control policy for both. Our results show that the reality is more complicated. Not only do the energy savings vary across the seasons, but so does the contribution of the building systems to the savings, the potential benefit from fine-grained occupancy data, and the relative performance of different model-free approaches. This is a conundrum for the practitioner who aims to implement RL in real buildings: if the performance varies drastically between seasons, how can one select a generalizable approach? This question warrants attention in future work.

Which RL algorithm works best? We designed a custom control system for multiple building systems using three popular deep reinforcement learning algorithms that can tackle problems with large state and action spaces. BDQN was adopted from previous work [40], where it was shown to have outstanding performance controlling multiple building systems of a single-zone building. To our knowledge, SAC and PPO were not previously applied to control multiple building systems.

We show here that SAC outperforms BDQN in the heating season (in all scenarios except one) with regard to energy savings. Considering thermal comfort, PPO is not able to satisfy the thermal comfort in the cooling season for most cases with average thermal comfort violation rate of 10.8%; SAC exceeds the threshold once and BDQN can always maintain the thermal violation rate under the threshold. Turning our attention to the effort needed to tune reward parameters, PPO is highly sensitive to these parameters, whereas BDQN

and SAC are less sensitive to the reward parameters. Also, PPO converges remarkably faster than SAC, and SAC is slightly faster than BDQN. As the requirements might differ from case to case, there is no clear winner among these three RL agents. SAC and BDQN seem to offer more promising results if one can afford the one-time computation cost of training the agents.

4.6 Conclusion

This chapter benchmarked multiple model-free reinforcement learning agents and baseline control strategies in a simulated multi-zone building with both zone and building-level occupancy schedules in winter and summer months. We evaluated the effect of controlling different building systems on whole-building energy consumption using different reward parameters, and provided useful insight for practitioners regarding how to make trade-offs within Pareto-efficient choices. Specifically, we showed better trade-offs can be achieved when RL agents rely on zone-level occupancy information rather than building-level occupancy information. We made two important observations when zone-level occupancy information was used by the agents. First, we found that 11.0% and 31.8% more energy can be saved respectively in heating and cooling seasons over existing rule-based baselines that control the same building systems. Second, we found that when lights are dimmed automatically and the RL agent jointly controls HVAC and blinds, the whole-building energy use can be reduced by up to 5.9% and 38.7% respectively in heating and cooling seasons over the case that the RL agent only controls the HVAC system.

Chapter 5

Diversity for transfer in learning-based control of buildings

The application of reinforcement learning to the optimal control of building systems has gained traction in recent years as it can cut the building energy consumption and improve human comfort. Despite using sample-efficient reinforcement learning algorithms, most related work requires several months of sensor data and operational parameters of the building to train an agent that outperforms existing rule-based controllers in a large multi-zone building. Moreover, exploring the large state and action spaces can result in poor indoor environmental quality for occupants. In this chapter, we propose to reduce the training cost of a policy gradient reinforcement learning algorithm by learning a library of control policies on a training building and taking advantage of both environmental and policy diversity. To transfer these policies to a target building, which can be different from the training building, we develop a simple method to assign the best pretrained policy in the library to each zone of the target building. We show that even without retraining the transferred policies on the target building, they can reduce the HVAC energy consumption by 40.4% compared to a fixed-schedule baseline and by 48.97% compared to agents trained on the target building for 5,000 months. The plausibility of our results underscores the importance of using diversity and transfer learning in multi-agent reinforcement learning settings and could pave the way for the adoption of reinforcement-learning based controllers in real buildings.

5.1 Introduction

Extensive research has been done in the past few decades to enable optimal and adaptive control of the HVAC system by applying MPC techniques and more recently RL algorithms [23], [42], [155]. The common goal is to reduce the building energy use and improve the overall occupant comfort and satisfaction. However, both MPC and RL-based control techniques have major drawbacks that have limited their adoption in real buildings. In particular, MPC relies on an accurate model that captures complex dynamics of the building. Identifying this model is nontrivial in large multi-zone buildings due to limited observability and lack of sufficient excitation [10]. Model-free RL techniques, on the other hand, require many interaction episodes in large multi-zone buildings to train an optimal policy. This post-deployment exploration is not affordable in practice as it might cause discomfort or health problems for occupants.

One approach to reduce the training cost is to transfer control policies learned over a sufficient number of episodes in a controlled environment, which can be a real or simulated building, to the target building. Since the training and target buildings might differ with respect to their structure, floor plan, HVAC system, and occupancy pattern, the near-optimal policy found in the training building can perform poorly on the target building. Previous work that studies the relationship between generalization and *diversity* in RL (as shown in Section 2.4) suggests that agents trained with diversity exhibit stronger generalization to novel environments [106]. Inspired by this, in this study we cast HVAC control as a multi-agent reinforcement learning problem where diversity is incorporated in the training process of each agent. Specifically, each agent is responsible for controlling a thermal zone and the agents compete with each other to reduce the total HVAC energy consumption. These agents, which are trained on a controlled building, are then transferred to the target building and assigned to the zones in that building. The assigned policies can be retrained on the target building to adapt them to the environmental condition and occupancy pattern of the respective zones. We show that this post-deployment adaption is not essential when we incorporate diversity.

This chapter also investigates how to evaluate policies in the policy library when transferred to a novel environment, i.e., a thermal zone in a new building. We borrow ideas from NAS and OPE to evaluate policies, using a small batch of log data (e.g., just a few weeks worth of data) from the new building. The log data is collected when the new building is controlled using a default controller, e.g., a rule-based or reactive controller. We show that the proposed policy evaluation approach gives us a reliable estimate of how these policies might perform on the new building, thereby enabling us to assign a subset of them to zones in that building. Our approach entails policy clustering, policy evaluation and ranking based on different scores, sampling, and transferring to respective zones in the target building.

Our contribution is threefold:

- We design a new loss function for policy diversity (defined in Section 5.3), and obtain a collection of sub-optimal policies by incorporating environmental and policy diversity in the training process of RL agents using a model-free policy gradient algorithm.

- We propose a novel two-stage algorithm that combines policy clustering and evaluation, and uses policy ranking methods to efficiently identify high-quality policies among policies in the policy library. Our algorithm requires just two weeks of log data collected from the novel target building.

- We show through simulation that agents pretrained with diversity perform well when they are transferred to a novel environment, even without adaptation. They outperform the agents that are originally trained in the target building in more than 5,000 episodes (months). This suggests that utilizing diversity in transfer learning can substantially reduce the training cost in the target building.

5.2 MARL-based control of HVAC

We consider an HVAC system that consists of one or multiple AHUs and VAV systems as explained in Section 3.1. The HVAC control can be viewed as a

sequential decision-making problem where the control agent interacts with the building and its occupants by performing a sequence of actions, e.g., changing the zone setpoint(s), and receiving a reward to help improve the agent's policy. While a single agent can control the entire building (all actuators in AHUs and VAV systems), it prevents the policy from being transferred to a new building that has a different state-action space, e.g., contains more VAV systems. As a result, we study the HVAC control problem in a MARL setting where each agent is responsible for controlling a single zone; the building is controlled by several independent agents, each making decisions about their respective zone.

We define our implementation of the Multi-agent Markov Decision Process (MMDP) as a tuple $(N, \mathcal{S}, \mathcal{A}_{i,i\in\{1,\dots,N\}}, \mathcal{R}_{i,i\in\{1,\dots,N\}}, \mathcal{P}, \mathcal{H})$ where:

- N is the number of agents.
- **State space $\mathcal{S}$** is a set of all possible states s representing the observation at a given time. We include the readings of six sensors in the state of each zone: the controlled zone mean temperature ($^\circ C$), mean humidity (%), outdoor temperature ($^\circ C$), solar radiation (W), binary occupancy state of the controlled zone, and hour of the day ($0 - 23$).
- **Agent i's action space $\mathcal{A}_i$** contains all possible actions a_i that can be taken by agent i in state s at a given time. We define the action as the minimum damper position of the VAV system in each zone. The minimum damper position is a value in $[0.1, 1]$, where 0 indicates that the damper is closed and 1 indicates that the damper is fully opened. For example, if the agent assigns 0.2 to the minimum damper position, the damper can be opened between 20% and 100%. Each agent only controls the damper position for their respective zone, and the AHU control points and all other VAV control points are adjusted by the controller in EnergyPlus, using the *predictive system energy balance* method [31]. The existence of the reheat coil can help to fulfill the zone temperature requirement. We denote $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_N$.
- **Agent i's reward function $\mathcal{R}_i$** is a mapping from states and actions to

Table 5.1: Description of the states and action of each agent.

State	Zone mean temperature	$^{\circ}C$
	Zone mean humidity	%
	Zone occupancy	Binary
	Outdoor temperature	$^{\circ}C$
	Solar radiation	W
	Hour of the day	Integer
Action	VAV damper position	%

real numbers, i.e. $\mathcal{R}_i : \mathcal{S} \times \mathcal{A}_i \to \mathbb{R}_i$. The reward is designed to evaluate how the action chosen in the given state improves or degrades the control system performance. To minimize the total energy consumption of the building, we define the reward of each agent as the energy use of the respective VAV system with a negative sign.

- **Transition function** $\mathcal{P}$ governs transition dynamics from the previous state to the next state given the selected action. This transition function is provided by EnergyPlus [31].

- **Control horizon** $\mathcal{H}$ defines the length of each episode. Although the HVAC system is always controlled by the agent, we consider a fixed time interval to evaluate a policy. Specifically, we use one month in the heating season, with 15-minute increments, to define one episode.

Table 5.1 summarizes all state variables and the action for each agent. Given the MMDP, RL algorithms aim to find a policy π_i that maximizes the agent's expected cumulative reward G_i over some time horizon $\mathcal{H}$:

$$G_i = \mathbb{E}\left[\sum_{t=0}^{\mathcal{H}} \gamma \mathcal{R}_i(s_t, \underset{a_t \in \mathcal{A}_i}{\operatorname{argmax}}\, \pi_i(a_t|s_t))\right]. \tag{5.1}$$

We set the $\gamma = 1$ in this task because the control horizon is fixed and finite, and the goal is to minimize the total monthly energy use rather than an arbitrary energy cost function. We use competitive agents where each agent receives and maximizes their own reward. Although it might be hard to get convergence, the agents are independent which is suitable for transfer learning as only a

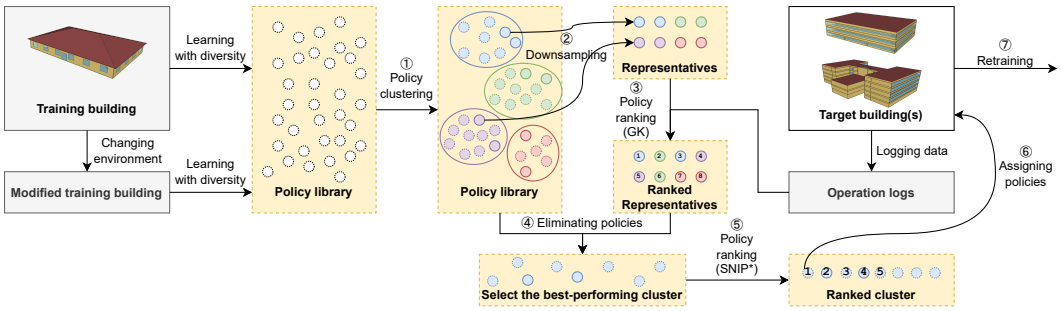


Figure 5.1: Schematic overview of the proposed methodology where circled numbers show different steps of the policy evaluation and assignment method.

subset of agents can be transferred to a new environment. An alternative approach is using cooperative agents that maximize a shared reward, which can be the whole-building energy consumption [18]. However, training these agents is more challenging because all agents simultaneously affect the shared reward, and assigning credit to individual agents is nontrivial.

5.3 Methodology

In this section, we present our methodology for MARL-based HVAC control that involves transfer learning and diversity training.[1] We first describe our implementation of PPO to train each RL agent, followed by our approach for constructing a library of diverse policies. Lastly, we describe an algorithm for transferring policies in the policy library to an unseen target building, and selecting the best policy for each zone in that building. We use a clustering algorithm and various policy evaluation methods to efficiently identify the most suitable policies for controlling the target building using the historical data. Once these policies are assigned to the respective zones in the target building to control VAV systems, we retrain them on the target building in an online fashion. The overall proposed methodology is shown in Figure 5.1.

5.3.1 PPO-based control agent

As described in Section 3.2.3, we use PPO to train control agents, with $\epsilon = 0.2$. For both actor and critic networks, we use two hidden layers with 64 units in each layer and the hyperbolic tangent activation function. We use Gaussian

[1]Code is available at `https://github.com/sustainable-computing/building-MARL`

Algorithm 1 Building a policy library

Require:

$\mathcal{B}$: set of zones in the training multi-zone building;
W: set of diversity weights;
`EPISODES`: number of episodes;
`EPOCHS`: number of time steps per episode;
env: target building environment wrapper

Ensure:

Π: Policy library

$\Pi = \varnothing$ ▷ initialize the policy library
$\mathcal{B}' = \mathcal{B}$.get_variants() ▷ Change $\mathcal{B}$
for w **in** W **do**
  for z **in** $\mathcal{B}^*$ **do**
    Initialize policy $\pi_{z,w}$
    for ep **in** $1, \ldots,$ `EPISODES` **do**
      $S_0 \leftarrow$ env.reset() ▷ Reset environment
      for t **in** $0, \ldots,$ `EPOCHS` **do**
        $A_t \leftarrow \varnothing$
        for z **in** $\mathcal{B}^*$ **do**
          $A_{t,z} \sim \pi_{z,w}(S_{t,z})$ ▷ Sample action
          $A_t \leftarrow A_t \cup \{A_{t,z}\}$
        end for
        $S_{t+1}, R_{t+1} \leftarrow$ env.step(A_t) ▷ Take step
        Store $(S_t, A_t, S_{t+1}, R_{t+1})$ in replay buffer
      end for
      for z **in** $\mathcal{B}^*$ **do**
        Optimize $\pi_{z,w}$ with loss $L_{\text{PPO}} + wL_{\text{diversity}}$
        using the replay buffer
      end for
    end for
    $\Pi \leftarrow \Pi \cup \{\pi_{z,w}\}$
  end for
end for

distribution for the actor network policy parameterization, and set the learning rate to 0.0003 and the batch size to 2,976. Although both real and simulated buildings can be used for training, in this work we train RL agents on a building simulated using EnergyPlus [31], with COBS (introduced in Section 3.3) as the environment interface. In our setup, each agent observes and controls one thermal zone by changing the damper position of the respective VAV system.

Algorithm 2 Naive policy evaluation and assignment

Require:

$\mathcal{B}_{\text{target}}$: set of zones in the target building;
Π_{rule}: set of rule-based policies for all $z \in \mathcal{B}_{\text{target}}$;
Π: policy library generated from Algorithm 1;
env: target building environment wrapper

Ensure:

Π^*: mapping of optimized policies for each zone

for z **in** $\mathcal{B}_{\text{target}}$ **do**
  $\Pi_{\text{evaluate},\cdot} \leftarrow \Pi_{\text{rule}}$
  for π **in** Π **do**
    $\Pi_{\text{evaluate},z} \leftarrow \pi$
    $G_{\pi,z} \leftarrow \text{env.evaluate}(\Pi_{\text{evaluate}})$ ▷ Calculate energy consumption
    $\text{Score}(\pi, z) \leftarrow G_{\pi,z}$
  end for
  $\pi_z^* \leftarrow \text{argmax}_\pi \text{Score}(\pi, z)$ ▷ Best policy for the zone
  $\Pi^* \leftarrow \Pi^* \cup \{\pi_z^*\}$
end for

5.3.2 Policy library

To build our policy library, we use a one-story building that contains 5 thermal zones as our training building. In addition to learning the optimal control policy for each zone, we use two kinds of diversity to construct a policy library. The process is defined below.

Optimal policies: We first learn a (near-)optimal policy for each zone of the training building using the PPO algorithm. This results in 5 policies as the training building contains 5 zones. These 5 policies are trained in parallel and competitively for each zone in the multi-agent framework.

Diverse policies: The policy found by the standard reinforcement learning algorithms is optimized for the given training environment. But they may not perform well in novel environments. The ability to identify a set of near-optimal policies that are different from one another enables us to explore the space of reasonable control policies, increasing the chance of learning a policy that better generalizes to new environments [104].

To find such near-optimal policies, we propose augmenting the loss function of the policy gradient algorithm with an additional term, denoted $L_{diversity}$.

This term forces the current policy to behave differently from the previously learned policies on the given state, while maximizing the cumulative reward. In other words, while the policy that is trained with the modified loss function is different from the optimal policy and other learned policies, its performance is as close as possible to the performance of the optimal policy. This method can be used to produce multiple diverse policies to expand the policy library Π. This loss term, $L_{diversity}$, can be written as:

$$\mathcal{L}_{diversity} = -\frac{\sum_{\pi' \in \Pi_{learned}} \sum_{(s,a) \in \exp} \frac{max\left(\frac{\max(\pi(a|s),\pi'(a|s))}{\min(\pi(a|s),\pi'(a|s))}, \bar{\rho}\right)}{|G_{\exp}(s) - V_{\pi'}(s)|}}{|\Pi_{learned}|}, \tag{5.2}$$

where π is the behavior policy we are updating, $\bar{\rho}$ is the upper bound on the probability ratio, *exp* is the state–action trajectory generated by the behavior policy and stored in the replay buffer, i.e., state-action tuples considered in the current episode, $G_{exp}(s)$ is the cumulative reward of this trajectory starting from the state s, $V_{\pi'}(s)$ is the estimated state value for state s from a learned policy, and $\Pi_{learned}$ is a set of learned policies from which the behavior policy must differ. Note that $max\left(\frac{\pi(a|s)}{\pi'(a|s)}, \frac{\pi'(a|s)}{\pi(a|s)}\right)$ captures the differences between the behavior policy (π) and a previously learned policy (π') by calculating the probability ratio of taking a certain action a given that we are in state s under the two policies. The term $|G_{\exp}(s) - V_{\pi'}(s)|$ measures the estimation bias of a learned policy given the current trajectory. A large value implies that the learned policy disagrees with the experience under the behavior policy. Learning from such experiences naturally distinguishes the behavior policy from the learned policy. Therefore, we lower the diversity loss to encourage learning. In general, $L_{diversity}$ is small when the behavior policy estimates the action probabilities differently from the learned policies, or the learned policies disagree with the trajectory taken under the behavior policy. This loss is averaged over all policies that have been learned so far.

We modify the loss function of the PPO algorithm to include the diversity loss:

$$L' = L_{PPO} + wL_{diversity}, \tag{5.3}$$

where w is a hyperparameter that yields a trade-off between optimality and diversity. It plays the same role as the population diversity factor that was introduced in [123]. It is worth noting that the above equation is a generalization over the standard PPO loss function, as $L' = L_{PPO}$ when $w = 0$. In our implementation, $\Pi_{learned}$ contains only the policy trained for the same zone with $w = 0$, hence the policies trained with a non-zero diversity loss are not forced to be different from each other. When w is large, the algorithm gives more importance to diversity and may sacrifice optimality of the policy. Conversely, a small w causes the algorithm to find a policy that is nearly optimal, even if it is not adequately different from the optimal policy.

Policies for diverse environments: To introduce environmental diversity, small changes are typically made to the training environment such that the general learning task is unchanged, but a more diverse set of environments are considered for training [106]. In this study, we add blinds to cover windows in the zones of the training building to reduce the solar heat gain. We also update the occupancy pattern of each zone to remove the time intervals when a zone becomes unoccupied (e.g., lunchtime) during core business hours As the building environment has changed, the policies learned (with and without the diversity term) in this new environment are expected to be different from the ones learned in the original environment. These policies are also added to the policy library.

Algorithm 1 describes the steps to generate a library of diverse policies when given a building with $\mathcal{B}$ being the original thermal zones, $\mathcal{B}'$ being the zones that are created using environmental diversity, and the set of diversity weights denoted W. As mentioned in Section 5.3.2, we introduce environmental diversity by adding blinds to each zone. In Line 13 and 14, the action of each agent at time t, denoted $A_{t,z}$, is sampled from its policy $\pi_{z,w}$, and gets appended to the overall set of actions, A_t. Assuming that one policy is generated per diversity weight, Algorithm 1 adds a total of $|W| \cdot (|\mathcal{B}| + |\mathcal{B}'|)$ different policies to the policy library.

5.3.3 Policy selection

After building the policy library for the training building, we assign these policies to each zone of the target building and evaluate them. Algorithm 2 describes the steps to choose the best policies from the policy library. Specifically, we evaluate each policy in each zone of the target building to find its cumulative reward, assuming dampers in all other zones are controlled using a fixed schedule. We then select the policy that yields the highest cumulative reward as the best policy for that zone. After identifying the best initial policy for all zones, it is possible to retrain the selected policies in a competitive MARL setting by interacting directly with the target building. This retraining step is useful to adapt the transferred policies from the policy library to the training building.

Evaluating all policies in each zone of the target building is indeed computationally expensive. To address this problem, we propose a policy clustering method that groups similar policies in the library. We can then sample a few policies from each cluster and evaluate their performance using the policy evaluation methods discussed in Section 3.2.4 and 3.2.5. This allows us to understand the performance of other policies that belong to the same cluster. We describe these steps below.

Policy clustering: The goal of this clustering is to identify policies that might have similar performance in the given task. Since we do not know the performance of each policy in the target environment, we cluster policies according to their behavior in the training environment(s).

We represent each policy in the policy library using a feature vector of length m. This vector is constructed by sampling $m-1$ states from the distribution of states visited when the policy was being learned in the training environment, and appending the initial state of the target environment. We then use the actions that would be taken from these m states under this policy to obtain the feature vector of length m. We set m to 10 in this study. Given the policy representation in an m dimensional space, we use K-Means to cluster all policies in the policy library. The elbow method is used to determine the

number of clusters. Specifically, we keep increasing the number of clusters starting from one cluster and calculate the *inertia* of the current clustering result. The inertia is defined as the sum of the squared differences of all samples from the respective cluster center. We stop when the inertia starts decreasing linearly.

After the clusters are formed, we select n representative policies from each cluster. This includes the policy that is closest to the cluster center and $n-1$ randomly selected policies from that cluster. The closet policy to the cluster center is picked as it may represent the average performance of the cluster in the training environment(s), and the other randomly picked policies increase our confidence in the evaluation result. We set n to 5 in this study.

Ranking policies using historical data from the target building Assuming that the historical data $\mathcal{D}$ is collected from the target building under the behavior policy π_b which can be the existing rule-based controller. We adopt the OPE method GK introduced in Section 3.2.4, and the ZCP method SNIP introduced in Section 3.2.5. To distinguish the SNIP under ZCP from the SNIP under OPE, we called the modified ZCP version that is adopted in this study SNIP*.

Selecting policies In Figure 5.1, there are two places where policy evaluation is performed, namely Step 3 and Step 5. In Step 3, we rank the representative policies from each cluster to obtain the ranking of clusters, whereas in Step 5, we only rank the policies from the top cluster. The best-performing policy from the top cluster is then transferred over to the novel target environment. All steps shown in Figure 5.1 are repeated for each zone in the target building to identify the policy that should be transferred and used for that particular zone.

5.3.4 Policy transfer and retraining

After assigning the best policy to each zone in the target building, we retrain all policies using the multi-agent reinforcement learning framework in an online fashion. Updating the policies through interaction with the target building allows the transferred policies to adapt to the target building environment

even further.

5.4 Training and target buildings

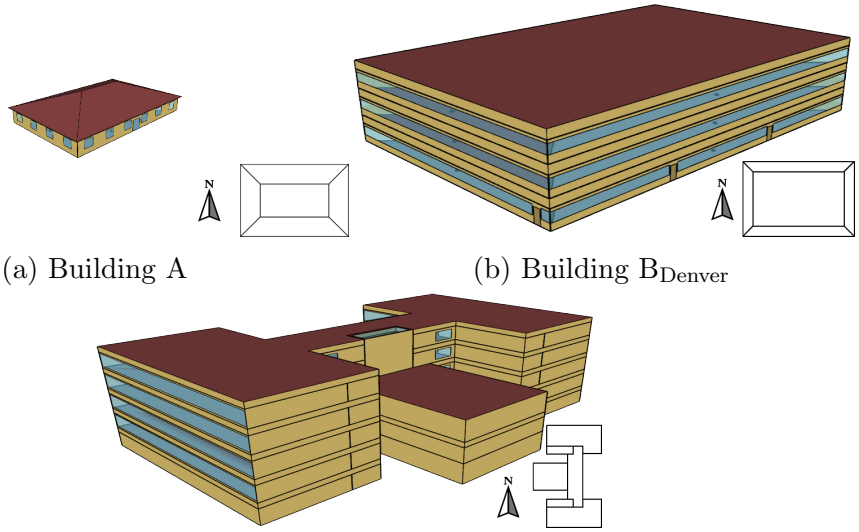
To study the efficacy of the proposed methodology, we evaluate it using the EnergyPlus model of three buildings, including a real campus building. Each building has a unique occupancy schedule which is encoded in the EnergyPlus model. We assume that if a control policy outcompetes other policies with respect to the HVAC energy use reported by EnergyPlus [31] without degrading thermal comfort, it also outcompetes them in the real building, should it be controlled using this policy.[2]

- **Building A** is a small office prototype building as defined by ASHRAE Standard 90.1 [2]. Figure 5.2a shows the floor plan and 3D model of this building. It contains five thermal zones (4 perimeter zones and 1 core zone) and is located in Denver, Colorado. Each zone is conditioned using a dedicated AHU and contains a VAV system. The total floor area of this building is 511.16 m^2.

- **Building $\mathbf{B_{Denver}}$** is a medium office prototype building as defined by ASHRAE Standard 90.1 [2]. It contains 15 thermal zones across three floors and is located in Denver, Colorado. Figure 5.2b depicts the floor plan of this building. There are 4 perimeter zones and 1 core zone on each floor. Each floor is conditioned using an AHU and all zones are equipped with a VAV system. Its total floor area is 4,982.19 m^2.

- **Building $\mathbf{B_{SanFrancisco}}$** is the same building as B_{Denver} with two main differences: 1) it is located in San Francisco, California and 2) its orientation is rotated by 45 degrees (clockwise). We make these changes so as to investigate whether any of the learned policies works well after transfer to a building with a different orientation in a different climate.

[2]We could not possibly deploy the many control policies we considered in this chapter on real buildings to run the microbenchmarks. As a result we evaluated them using EnergyPlus. In practice, the training building might be an EnergyPlus model, but the target buildings are physical buildings.



(a) Building A

(b) Building B_{Denver}

(c) Building C

Figure 5.2: The 3D view and floor plan of the buildings considered in this chapter where north is marked on each floor plan

- **Building C** is a medium campus building representing the model of the building that houses the Department of Energy Engineering at Sharif University of Technology in Tehran, Iran.[3] It contains 26 thermal zones spread across five floors, 11 of which are equipped with a VAV system. The HVAC, lighting, and blind systems are modeled such that they match the design of these systems in the physical building. We assume the building is located in San Francisco, California, because weather data is lacking for its actual location. The total floor area of this building is 5,051 m^2.

Note that the 3D views are scaled in Figure 5.2 to demonstrate the relative size of these buildings. Building A and Building B have similar floor plans, yet their HVAC systems are different and their core zones have different sizes.

5.5 Experiment results

In this section we describe the experiment setup, validate different parts of our methodology using microbenchmarks, and finally make a comparison with

[3]Model is downloaded from `https://github.com/DOEE-BMS/EnergyPlus-Model`

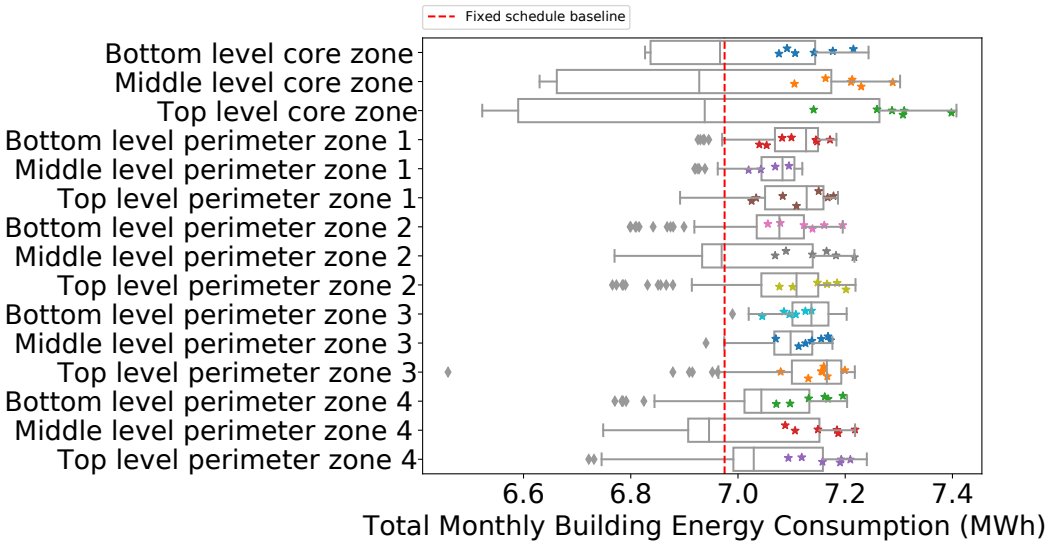


Figure 5.3: Performance of the top 100 policies selected from the whole policy library, Π_{both}, in each zone of the target building $\mathrm{B}_{\mathrm{Denver}}$. The total energy consumption for each policy on each zone in the target building $\mathrm{B}_{\mathrm{Denver}}$ is evaluated for one month, where all other zones are controlled using a fixed schedule baseline. Policies trained with the diversity weight of zero are marked by stars.

baseline control methods in terms of the total HVAC energy use. We start off by evaluating the efficacy of diversity training using naive policy transfer algorithm described in Algorithm 2, then evaluating the proposed transfer learning approach on building $\mathrm{B}_{\mathrm{Denver}}$, and then run experiments on building $\mathrm{B}_{\mathrm{SanFrancisco}}$ and the model of a real building (Building C).

5.5.1 Implementation details

We simulate the building operation using EnergyPlus 9.3 [31] with the actual weather data, and use COBS [170] to interface with the simulation environment. The control policies are trained using PyTorch [124]. The EnergyPlus model uses a 15-minute simulation time step, and each episode is one month. This is equivalent to 2,976-time steps. We fix the training and test periods to January to eliminate the seasonal effect in our simulation.[4] For each experiment, we consider 15 independent runs to calculate the average performance.

Building A is used to build the policy library considering both policy and

[4]Studying seasonal effects is deferred to future work.

environment diversity as outlined in the previous section. All policies are trained using PPO under the MARL framework for 1,000 episodes to ensure convergence. The training cost on the training building is not important because it is a presumably a simulated building or a controlled environment designed for this purpose. Note that we ignore the occupants' thermal comfort in the reward function because the power consumption of the VAV system's reheat coil and the supply air temperature are constantly adjusted by EnergyPlus according to the damper position to satisfy the thermal comfort requirement. The temperature setpoints, however, remain the same for all control scenarios for fair comparison.

We consider three policy diversity weights $w \in \{0.1, 1, 10\}$ to identify near-optimal policies. These policies are forced to be different from the optimal policy π^* ($w = 0$) that is learned for the given zone (hence $\Pi_{learned} = \{\pi^*\}$). This results in 800 policies in the policy library — 10 random seeds for training $\times$ 4 training environments $\times$ 5 zones per environment $\times$ 4 diversity weights. We set the upper bound on the probability ratio $\bar{\rho}$ to 100.

Baselines We consider four baselines: 1) the default controller implemented in the building model, 2) zone-level control policies learned via interaction with the target building (without transfer learning) using the MARL framework, 3) a control policy that decides on the minimum damper position of all zones and is learned through interaction with the target building (without transfer learning) in the Single-agent Reinforcement Learning (SARL) framework, and 4) zone-level control policies learned on the training building and transferred to the target building assuming an oracle produced the optimal assignment of policies to zones in the target building. The last baseline is unrealistic and gives a lower bound on the building energy consumption using the proposed methodology. We could not implement this baseline because identifying the best policy for each zone requires exhaustive search and expensive evaluation. The first baseline is a controller that can be readily used (or is actually being used in case of Building C) — if we beat this baseline, it means that our policies can reduce the HVAC energy use without sacrificing thermal comfort.

5.5.2 Energy saving potentials using diversity training

To better understand the best possible effect of training with different types of diversity, we use 3 strategies to assign pre-trained policies to the zones in the target building B_{Denver} using: (1) selecting from $\Pi_{environment}$ which is the set of policies trained on Building A and variations with the diversity weight of zero (environmental diversity only); (2) selecting from Π_{policy} which is the set of policies trained on original Building A with different diversity weights (policy diversity only); (3) selecting from Π_{both} which is the whole policy library and includes policies trained on Building A and variations with different diversity weights (incorporating both environmental and policy diversity). We compare these policies with the policies that are learned from scratch (i.e., starting with no prior knowledge) for every zone of the target building. The set of these policies is denoted by $\Pi_{scratch}$.

To find the best possible pre-trained policy for each zone, we need to assign each policy in $\Pi_{environment}$, Π_{policy}, or Π_{both}, to all zones of the target building B_{Denver} and calculate the HVAC energy consumption. This results in 800^{15} possible combinations, making it impossible to perform exhaustive search. To address this problem, we use a best-response approach that finds the best pre-trained policy for every zone in the target building B_{Denver} assuming that the other zones are controlled using a fixed schedule, which is defined by ASHRAE 90.1 [62]. Specifically, we calculate the total HVAC energy consumption over 1 month and choose the policy that results in the minimum energy use as the best policy for that zone. This approach is described in Algorithm 2. While it is not guaranteed to identify the best set of policies for the entire building, it provides a good indication of the expected performance.

Figure 5.3 shows the distribution of the monthly energy consumption when we use different policies from the whole policy library to control every zone in the target building B_{Denver}. Due to the large spread in energy consumption when considering all policies, we only present the top 100 best-performing policies. We can make two observations based on Figure 5.3. First, the spread in core zones is larger than that of the perimeter zones, implying that there

is a higher potential for energy savings in the core zones. Moreover, some agents can realize significant energy savings over the baseline in the core zones. We attribute this to the more complex heat dynamics in the core zones as they are adjacent to multiple perimeter zones. The RL agents that learn such complex dynamics can greatly reduce the energy consumption. Second, training with diversity helps to identify policies that are sub-optimal in the training building but perform well in the target building $\mathrm{B}_{\mathrm{Denver}}$. The stars in Figure 5.3 represent the performance of policies that are trained considering environmental diversity only. These stars are dispersed over the right tail of each distribution, indicating a higher energy consumption than the policies that are obtained using non-zero diversity weights. Comparing with the fixed-schedule baseline (the red dashed line), we find that all stars are located on the right side of the baseline, which means that the (near-)optimal policies ($w = 0$) found in the training building always perform worse than the fixed-schedule baseline in the $\mathrm{B}_{\mathrm{Denver}}$. However, policies trained with diversity in the training building can perform better than the baseline as the left whisker is generally on the left side of the vertical line. The second observation supports our argument about incorporating diversity in transfer learning.

Comparing different types of diversity: Figure 5.4 shows the $\mathrm{B}_{\mathrm{Denver}}$'s HVAC energy consumption per episode. The policies are selected from $\Pi_{scratch}$, $\Pi_{environment}$, Π_{policy}, and Π_{both}. They are then trained or retrained in $\mathrm{B}_{\mathrm{Denver}}$ for 5,000 months. The energy consumption is averaged over ten runs with different random seeds and the shaded region around the average shows the 95% confidence interval. The (black) dashed line indicates the baseline performance using a fixed-schedule, defined by ASHRAE 90.1, for all zones.

It can be readily seen that selecting and transferring policies from $\Pi_{environment}$, Π_{policy}, and Π_{both} to building $\mathrm{B}_{\mathrm{Denver}}$ results in a lower monthly energy consumption than learning policies from scratch ($\Pi_{scratch}$). Selecting policies from Π_{both} is better than Π_{policy}, but the curves cannot be easily distinguished in this figure as the difference is small. Moreover, the transferred policies converge faster than the policies learned from scratch on the building $\mathrm{B}_{\mathrm{Denver}}$, i.e.,

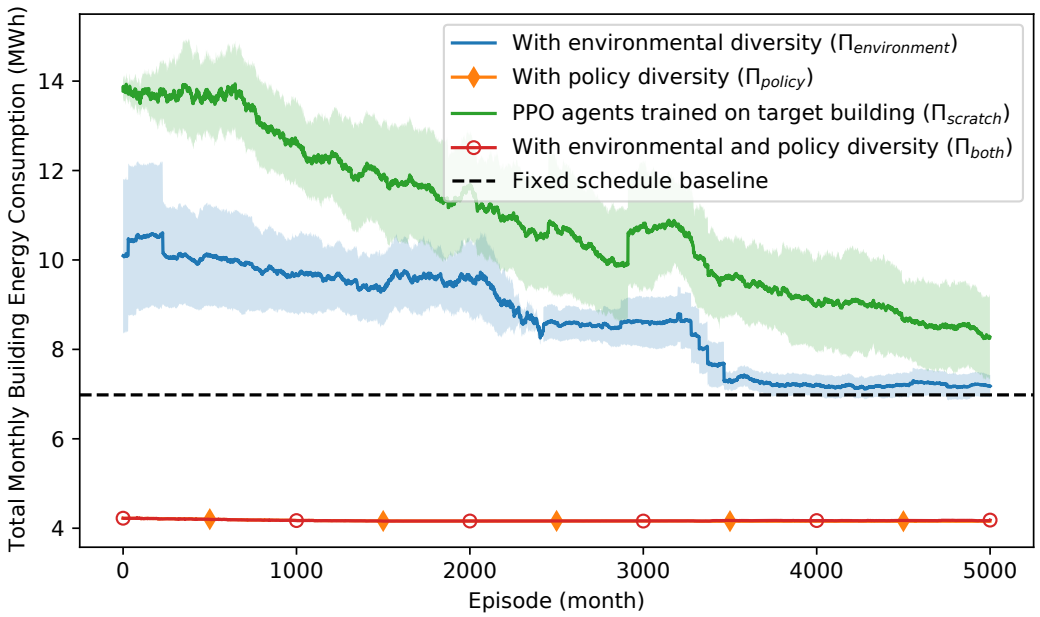


Figure 5.4: Performance comparison of four initial policy selection methods on building B_{Denver}. The mean and 95% confidence interval of the total monthly energy consumption are computed based on 10 independent runs. Note that the y-axis is aggregated.

$\Pi_{scratch}$. Specifically, $\Pi_{environment}$, Π_{policy}, and Π_{both} converge at around 3,500, 1,500, and 1,500 episodes, respectively. $\Pi_{scratch}$ does not seem to converge even after 5,000 episodes. The policies learned from scratch might be stuck in local optima due to the complexity of the multi-agent environment and control task. Incorporating diversity can help the transferred policies get closer to the global optimum by starting from a better point.

Figure 5.4 also demonstrates the need for policy diversity. Without policy diversity, $\Pi_{scratch}$ and $\Pi_{environment}$ performs worse than the fixed-schedule baseline. Π_{policy}, and Π_{both}, on the other hand, saving 40.11% and 40.40% more energy than the baseline, respectively. Interestingly, even without retraining on B_{Denver}, the performance of Π_{policy} (4.22 MWh) and Π_{both} (4.22 MWh) is much better than the performance of $\Pi_{environment}$ (10.09 MWh) and the fixed-schedule baseline (6.98 MWh).

Moreover, compared to the MARL agents trained only on B_{Denver} for 5,000 episodes, 48.97% more energy savings can be achieved by incorporating diversity in the policy library and assigning suitable policies to B_{Denver} without

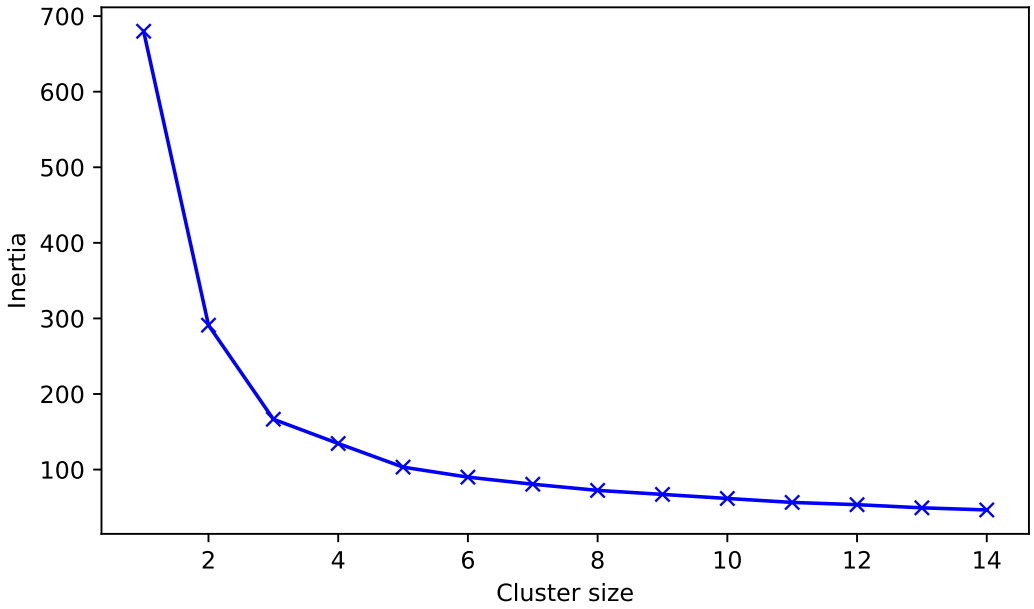


Figure 5.5: Changes in inertia for different cluster sizes.

retraining.

We conclude that the proposed transfer learning framework helps to find better control policies at a lower cost in a novel environment. We observe that introducing environmental diversity is less advantageous than policy diversity, and including both yields almost the same result as incorporating policy diversity only. It is worth noting that the performance of Π_{policy} and Π_{both} does not improve as much as the other two sets of polices over time. This might be because they contain policies that are near optimal for the zones in the target building, leaving little room for improvement when they are retrained. This implies that policies from Π_{policy} and Π_{both} can be used with minimum adaptation, bringing the training cost on the target building to nearly zero.

5.5.3 Policy clustering analysis

Next we examine our policy clustering result to see whether policies that were in the same cluster had similar performance in the target building. Ideally the top cluster should contain the majority of well-performing policies. We again consider B_{Denver} for this microbenchmark. Figure 5.5 shows the result used by the elbow method to choose the optimal number of clusters for a randomly

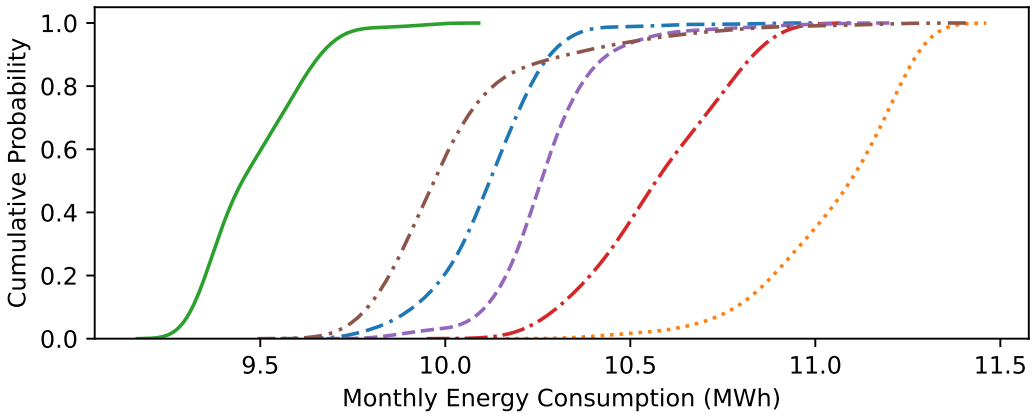


Figure 5.6: The cumulative density plot for the distribution of policies' performance when we form six clusters on a select zone in building B_{Denver}. Each line represents the distribution of one cluster. A similar result can be obtained from other zones in the building as well.

picked zone in B_{Denver}. It can be seen that increasing the cluster size from six to seven would reduce the inertia by the same amount as increasing it from five to six. Thus, the elbow method returns six clusters.

Clustering all policies into six clusters allows to eliminate 83% of the policies after the first round of policy evaluation. We assess the risk of incorrectly removing well-performing policies by plotting the energy performance distribution for all clusters in Figure 5.6. The x-axis represents the total monthly energy consumption if the policy is selected as the behavior policy for the given zone, and all other zones are controlled using the default controller. The left-most curve in Figure 5.6 shows the empirical Cumulative Distribution Function (CDF) of policies that belong to the top cluster. Interestingly, more than 50% of these policies keep the total monthly energy consumption below 9.5MWh. This is while the other clusters barely include a policy that keeps the total monthly energy consumption below 9.5MWh. This implies that the left-most curve represents the best performing cluster and neglecting policies in other clusters should not affect the HVAC energy consumption. Although there is a small overlap between the top three clusters, for every zone, there is at least one policy in the top cluster that is better than all the policies in these two clusters.

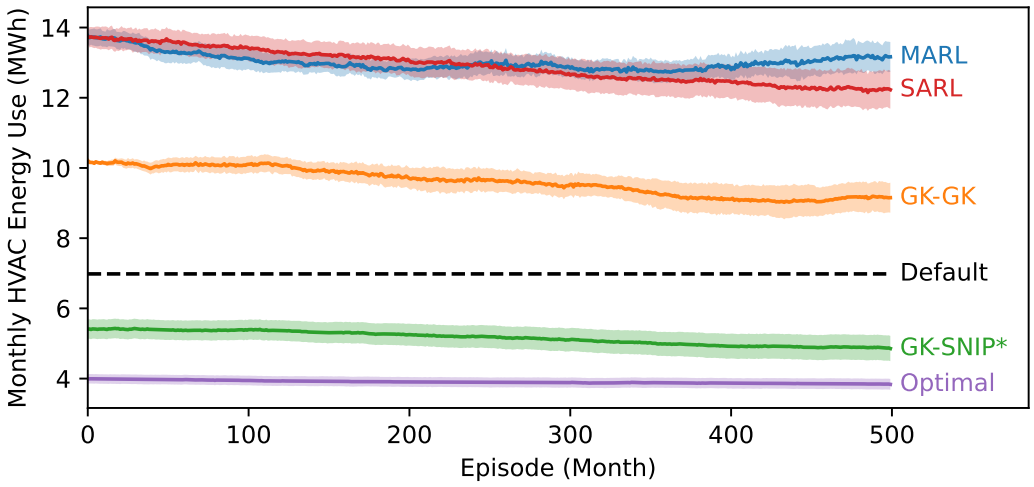


Figure 5.7: Learning curve of different controllers on Building B_{Denver}. Each solid line shows the average performance of 15 runs and the shaded area shows one standard error from the mean. The y-axis is exaggerated.

Recall that we sample $n = 5$ policies from each cluster to estimate the performance of each cluster. From Figure 5.6, we conclude that even if we sample only 1 policy from each cluster, the chance of incorrectly identifying the best performing cluster is slim. Sampling 5 policies would further reduce the probability of misidentifying the top cluster.

5.5.4 Policy transfer to B_{Denver}

GS [64] suggests that combining GK for cluster ranking with SNIP* for policy ranking within the top cluster can provide the best result. We refer to this setting as GK-SNIP* and compare with four baselines introduced in Section 5.5.1 to evaluate the efficacy of the proposed methodology.

Figure 5.7 shows the performance of our proposed method with other baselines on the target building B_{Denver}, which is different from the training building, Building A, in terms of the floor area and HVAC design. However, both buildings are located in the same city and they have relatively similar floor plans. All policies are either selected from the policy library or initialized randomly (SARL and MARL). Regardless, they are (re)trained for 500 episodes (months). Policies that need extensive training are not suitable for deployment on real buildings. For instance, the SARL controller trained on the target

building (without transfer learning) reaches the same level of performance as the optimal policies assigned from the policy library only after 15,000 episodes, i.e., 1,250 years after the deployment!

It can be readily seen that the proposed methodology provides a reasonable assignment for all zones in B_{Denver}. The performance of the proposed GK-SNIP* policy ranking method at episode 0 (5.41 MWh) is 22.5% better than the default controller that is presumably designed by HVAC engineers (6.98 MWh). It is also significantly better than SARL (13.74 MWh) and MARL (13.77 MWh). This implies that GK-SNIP* can be applied to select policies that have reasonable performance on the target building. The optimal assignment has an initial total energy cost of 3.99 MWh. The difference between the proposed policy selection method and the optimal selection is partly due to how we sample policies from the top cluster. Note that the policies assigned to the target building under the optimal assignment do not benefit significantly from retraining. Specifically, the total HVAC energy consumption reduces by 3.8% (from 3.99 MWh to 3.84 MWh) after 500 episodes. We believe this is because there is not much room for improvement as we are already close to the minimum HVAC energy consumption that could be realized by a controller in this building given its occupancy schedule.

The proposed policy selection method, GK-SNIP*, can improve by 10.2%, reaching the total monthly energy consumption of 4.86 MWh after 500 episodes of training on B_{Denver}. This is 30.4% less than the energy consumption of the default controller. Policies trained only on B_{Denver} (not transferred from Building A) fail to reach a level of performance that is comparable with the default controller at the end of the 500 episodes. SARL reaches 12.23 MWh and MARL reaches 13.17 MWh of monthly energy consumption. We also witness an increase in the energy consumption under MARL after around 200 episodes. This might be because agents are not collaborating with each other. As a result, they start to cancel out each other's action (aka fighting zones), increasing the total HVAC energy use.

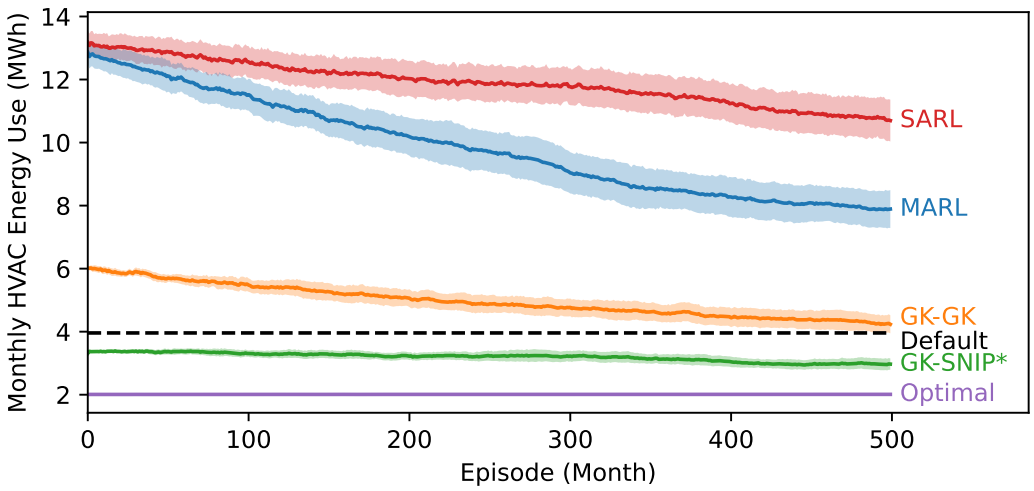


Figure 5.8: Learning curve of different controllers on Building $B_{SanFrancisco}$. Each solid line shows the average performance of 15 runs and the shaded area shows one standard error from the mean.

5.5.5 Policy transfer to other buildings

To further validate our proposed methodology, we consider two target buildings ($B_{SanFrancisco}$ and C) that have some major differences with the training building (Building A). Building $B_{SanFrancisco}$ is located in a warmer climate compared to the training building. Moreover, it differs from the training building in terms of the floor area and HVAC design. Building C is a real building and has several differences with the training building, including its size, occupancy, floor plan, HVAC design, and weather conditions.

Figure 5.8 shows the performance result in Building $B_{SanFrancisco}$. The total energy consumption in all cases is lower than Figure 5.7 because we are looking at a winter month with a higher average outside temperature in San Francisco, reducing the heating demand of the building. Most of the observations made in Section 5.5.4 are true in this case. Before retraining, the proposed policy selection method, i.e., GK-SNIP*, yields 16.4% lower monthly energy consumption (3.31 MWh) than the default controller (3.96 MWh). The optimal assignment yields the lowest monthly energy consumption at episode 0 (2.01 MWh), which is 49.2% lower than the default controller. After 500 episodes of training, the policies assigned by GK-SNIP* reduce the total HVAC en-

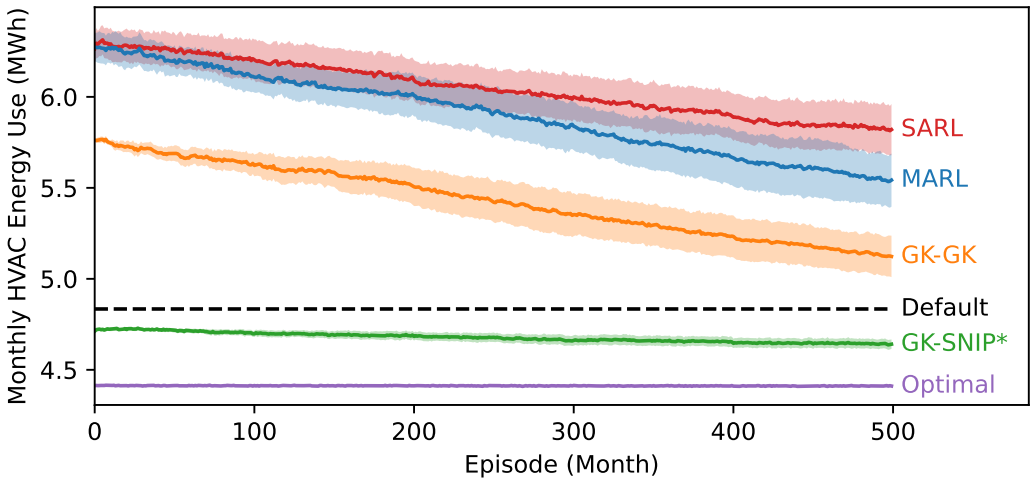


Figure 5.9: Learning curve of different controllers on Building C. Each solid line shows the average performance of 15 runs and the shaded area shows one standard error from the mean. The y-axis is exaggerated.

ergy consumption by 10.3%, reaching 2.97 MWh. This is 25.0% lower than the energy consumption of the default controller and 50.8% higher than the optimal.

Figure 5.9 compares the performance of the proposed method with the four baselines in Building C. We see the same trend here too. GK-SNIP* performs better than the default controller, SARL, and MARL, and is slightly worse than the optimal assignment. The default controller consumes 4.83 MWh of energy in one month, whereas the proposed GK-SNIP* reduces it to 4.71 MWh before retraining and to 4.64 MWh after 500 episodes. These numbers are 4.413 MWh and 4.411 MWh for the optimal assignment.

Our experiment on all buildings supports the claim that diversity-induced RL offers clear benefits for transferring policies to a novel target building, and that the proposed GK-SNIP* policy selection method can efficiently identify policies, among the policies in the policy library, that perform well in the novel target building using only 2 weeks of historical data. The transferred policies consistently outperform the default controller in terms of the HVAC energy use without sacrificing thermal comfort. This is the case even before these policies are retrained to adapt to the new environment.

5.6 Discussion

We presented a MARL-based HVAC control strategy that incorporates environmental and policy diversity to better explore the space of reasonable control policies in a controlled built environment, which can be a real building used for training or a simulator. We showed that the policies learned by accounting for diversity can generalize better to novel environments, for example buildings that have a different structure, floor plan, HVAC system, and occupancy pattern from the training building. We ran experiments on COBS and EnergyPlus to quantify energy savings that can be realized in a large multi-zone target building using environmental diversity, policy diversity, and both, and to investigate whether the policies should be retrained in each case.

Our result indicates that by transferring and deliberately assigning diverse policies to the zones in the target building, the monthly energy use of the HVAC system can be reduced by up to 40% over the default fixed-schedule controller in EnergyPlus, if the policy can be evaluated using Algorithm 2, and that retraining may not be needed when we incorporate policy diversity. We then proposed an offline policy selection algorithm that effectively identifies high-quality policies to transfer into a novel building, even using as little as two weeks of operational log data, to deal with the impractical issue of Algorithm 2. The proposed policy selection algorithm resulted in 4.0-30.4% energy saving than the default controller.

Note that although we used simulated buildings in our experiment to represent the training and target buildings, the proposed methodology can be applied to real buildings.

Chapter 6

Data-efficient personal comfort modeling

Occupant thermal-comfort complaints are the biggest operational headache of facilities managers. Many of the complaints can be attributed to the diverse nature of individuals' thermal comfort needs which are not accounted for in the de facto standard for thermal comfort. This has motivated research on developing data-driven personal comfort models and incorporating them in control loops. But the progress on this front has been hampered by the lack of sufficient ground-truth thermal comfort data to train accurate thermal comfort models. To address this problem, in this chapter we explore how artificial labels, indicating individuals' true thermal preference, can be generated from their heating and cooling behavior with a personal comfort system. Furthermore, we use clustering to identify individuals with similar comfort requirements in a rich dataset collected from 37 individuals in an office building, and develop a small number of group comfort models, each achieving a high accuracy in predicting the thermal comfort of individuals within the respective cluster. The pretrained group comfort models are then combined using an ensemble method to create a general thermal comfort model that can accurately predict the thermal comfort of any individual without knowing their thermal preferences or group membership a priori. We evaluate the efficacy of two ensemble methods as more training data becomes available and show that they outperform two conventional comfort models (PMV, Adaptive) and the personal comfort model that is developed from scratch for a particular individual. Specifically, the best ensemble comfort model yields on average

71% accuracy in predicting individuals' thermal preference using only 6 hours of training data, excluding no occupancy periods.

6.1 Introduction

Providing thermal comfort is one of the primary goals in the design and operation of a building's HVAC system. There is strong evidence that occupants that are more satisfied with their surrounding thermal environment have improved health and productivity [50], [93].

There has been several attempts in recent years to address these shortcomings by developing models that better reflect the thermal comfort of a specific individual [82]. The so-called *personal comfort models* are often developed using additional features, including measurements of the central HVAC system [89] and the data collected by sensors embedded in wearable devices [67], [98] and PCS, e.g., a heated and cooled chair [83] or a desk fan [59]. These measurements are then related to occupants' feedback acquired via surveys, indicating their satisfaction with or preference for their local thermal environment (e.g., comfortable, want cooler, want warmer). Unfortunately, conducting surveys at regular intervals is costly and can be deemed intrusive too. It is also hard to secure consistent occupant feedback in the long run as the participation rate tends to decay over time. Consequently, only a small amount of labels indicating an individual's true thermal comfort is typically available, posing a significant challenge for the development of personal comfort models with many trainable parameters, e.g., neural networks. A related problem is training an accurate personal comfort model when sufficient training data is unavailable, for example, because the occupant is new to the building (aka the "cold start" problem).

This chapter aims to address (a) the inadequacy of training data by generating artificial labels from individuals' heating and cooling behavior with a PCS, and (b) the cold start problem in the development of personal comfort models by combining a number of pretrained *group comfort models* through an ensemble method. Each group comfort model is an expert neural network that is trained on data from a group of individuals who have similar thermal com-

fort requirements. It outputs a probability distribution over thermal comfort labels for any individual that could be a member of this group. By creating an ensemble of these experts, it is possible to build an accurate thermal comfort model for a given individual after observing their behavior for a short period of time.

To show the feasibility of these ideas and validate the proposed methodology, we use the PCS chair dataset from a field study described in [81], which is the largest PCS study to date. We postulate that the 37 occupants in this dataset are a representative sample of the office building occupants with regard to their thermal comfort requirements[1] and that this dataset can be divided into groups of occupants with similar thermal comfort requirements (although the division into these groups is not known in advance).
Our contribution is fourfold:

- We propose a new method of generating artificial labels from occupants' heating and cooling behavior to increase the size of training data for personal comfort models.

- To facilitate the training of data-driven comfort models, we identify relevant features for explaining an individual's thermal comfort among the data gathered by PCS, HVAC, and HOBO[2] sensors. We develop a recurrent neural network, specifically a LSTM model, to predict thermal preference of each individual using the identified features.

- We explore the possibility of grouping individuals based on the performance of their personal comfort model when applied to predict thermal preference of other individuals. Specifically, we apply a hierarchical clustering algorithm to obtain clusters using the proposed distance measure. Once clusters are formed, we train a new LSTM model on the data that belongs to all individuals within the same cluster and refer to it as

[1]Should there be occupants with unique comfort needs that are not included in that dataset, it is possible to extend our ensemble model by adding a new group thermal comfort model.

[2]HOBO is an independent environmental data logger installed at each occupant's workstation as described in [81].

the group comfort model. We evaluate the efficacy of these models in predicting an individual's thermal preference by comparing them with various baselines.

- We apply two ensemble methods, namely stacked ensemble and mixture of local experts [72], to combine the output of the pretrained group comfort models. These methods discover the relevance of the group comfort models for predicting thermal preference of a person even if they are new to the building, thereby improving accuracy and robustness in a classification task where labeled data is limited. We compare the predictive power of the *ensemble comfort models* with PMV and adaptive comfort models along with personal comfort models that are trained from scratch for this person. We show that the mixture of experts ensemble comfort model outperforms the other models and achieves an accuracy of around 71% after observing an individual's heating and cooling behavior for up to 6 hours, excluding the time that the chair is unoccupied, and without relying on survey data.

6.2 Data set

We use the data set from a field study described in [81], which is the largest PCS study to date. This data set was also used in [83] to develop personal comfort models and compare them with PMV and adaptive comfort models. It consists of three types of data, namely survey data denoted $\mathcal{D}_{survey}$, sensor data denoted $\mathcal{D}_{sensor}$, and metadata denoted $\mathcal{D}_{meta}$. The data was collected between April and October 2016 in an office building located in Northern California in three overlapping phases. Each phase was about 4 months long with respectively 10, 17, and 10 participants. Therefore, the data set consists of a total of 37 people participated in the case study: 17 male and 20 female participants, out of which 30 were in an open workspace and 7 had private offices. We ignore the metadata $\mathcal{D}_{meta}$ as it is not used in this study.

The survey data $\mathcal{D}_{survey}$ is collected three times per day for a period of twelve weeks following a one-week adjustment period. All subjects were asked to select their current thermal comfort preference from one of the following

Table 6.1: Modeling Feature list

Sensor	Feature	Unit
HVAC	① Room air flow	ft^3/min
	② Room damper position	%
	③ Room discharge air temp.	°F
	④ Room heating output	%
	⑤ Room maximum airflow	ft^3/min
	⑥ Room minimum airflow	ft^3/min
	⑦ Room heating setpoint	°F
	⑧ Room cooling setpoint	°F
	⑨ Room temp.	°F
Weather Station (WS)	① Outdoor air temp.	°C
	② Prevailing mean outdoor air temp.	°C
HOBO at each workstation	① Air temp.	°C
	② Operative temp.	°C
	③ Slope in air temp.	°C/h
	④ Relative humidity	%
	⑤ Slope in relative humidity	%/h

three options: 'want warmer'; 'want cooler'; 'no change'. The survey completion time is also recorded in the data set. We omit the description of other features as they are not used in this chapter.

The sensor data $\mathcal{D}_{sensor}$ includes data from four sources: the PCS chair; the central HVAC system; the HOBO sensor located at an occupant's workstation; and a nearby weather station. The PCS data is collected at 20-second intervals. The HOBO and HVAC data are logged at 5-minute intervals. The weather station data is recorded hourly. We resample the data from different sources at 1-minute rate and align them by selecting data points that are nearest to regular intervals of 1 minute.

We use the forward filling method to fill all missing values. If $|x_{t+1}/x_t - 1| > \tau$, then x_{t+1} is considered an outlier. For temperature and humidity readings, τ is set to 10% and 20% respectively. Outliers are treated in the same way as missing values. Table 6.1 lists all features that we used in this chapter.

6.2.1 Generating artificial labels

After data cleaning, there are 4743 survey responses available in the $\mathcal{D}_{survey}$ for the 37 occupants. This gives us, on average, fewer than 130 thermal comfort labels per individual, which is not enough to train an accurate personal comfort model. What exacerbates the problem is that some individuals infrequently participated in the surveys and the number of labels available from them is less than 50. This presents a significant barrier to the development of complex data-driven models, for example neural networks.

To address this problem, we consider individuals' heating and cooling behavior with the PCS chair as a proxy for their thermal comfort preference (i.e., the label). Specifically, it is easy to determine when individuals use the heating or cooling function because the control intensity of chair fans and heaters is continuously recorded in $\mathcal{D}_{sensor}$ (in a scale from 0% to 100%). Figure 6.1 illustrates the amount of labels collected over time from one individual. It can be readily seen that there is a 97 fold increase in the amount of labels available for an individual if we use their heating and cooling behavior as a proxy for their thermal preference instead of using survey responses. The total amount of artificial labels generated per individual is about 10k. Given that each label represents a 1-min interval, 10k labels amount to 7 days of PCS data, excluding the instances where the PCS chairs are unoccupied by the occupants.

In previous work [81], [83] we examined the correlation between occupants' true thermal preference and PCS operation. We have corroborated that individuals' heating and cooling behavior with the PCS chair can be treated as indirect feedback in a vast majority of cases. Thus, it can be used to generate thermal comfort labels with an acceptable loss of accuracy. Notice that we do not use occupant behavior with the PCS as an input feature for thermal comfort prediction (as the authors did in [83]) since it is used to generate labels for training models.

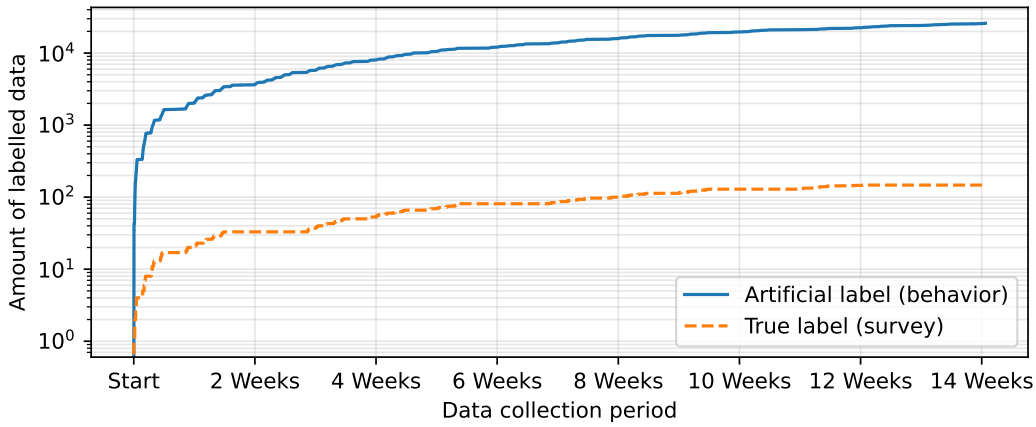


Figure 6.1: The total amount of labeled thermal comfort data that becomes available from a sample individual over time. Note that the y-axis is logarithmic scale.

6.2.2 Dealing with imbalanced data

One crucial step in the preprocessing of thermal comfort data is to ensure that class labels are evenly distributed in the training set. This is because the abundance of samples from one class (e.g., 'no change') can swamp samples from the other classes ('want warmer/cooler') as they have the same weight in the loss function. To give an example, in a multi-class classification problem where the ratio of samples in the majority class to all samples is $\alpha\%$, a naive classifier that always predicts the majority label will attain $\alpha\%$ classification accuracy. The data set we use in this chapter, like other real-world thermal comfort data sets [67], is prone to the imbalanced data problem. For example, even after generating artificial labels, thermal comfort labels are not evenly distributed in our data set and for some individuals the value of α is greater than 85%.

There are a few different ways to address the data imbalance issue. One approach is to tweak the loss function such that mispredicting minority classes is worse than mispredicting the majority class. Since this has to be done for each individual differently, it creates problems when we group some individuals together. The other two common solutions are reducing the number of samples in the majority class (undersampling), or synthesizing samples for minority classes (oversampling). In this chapter, we use a combination of undersampling

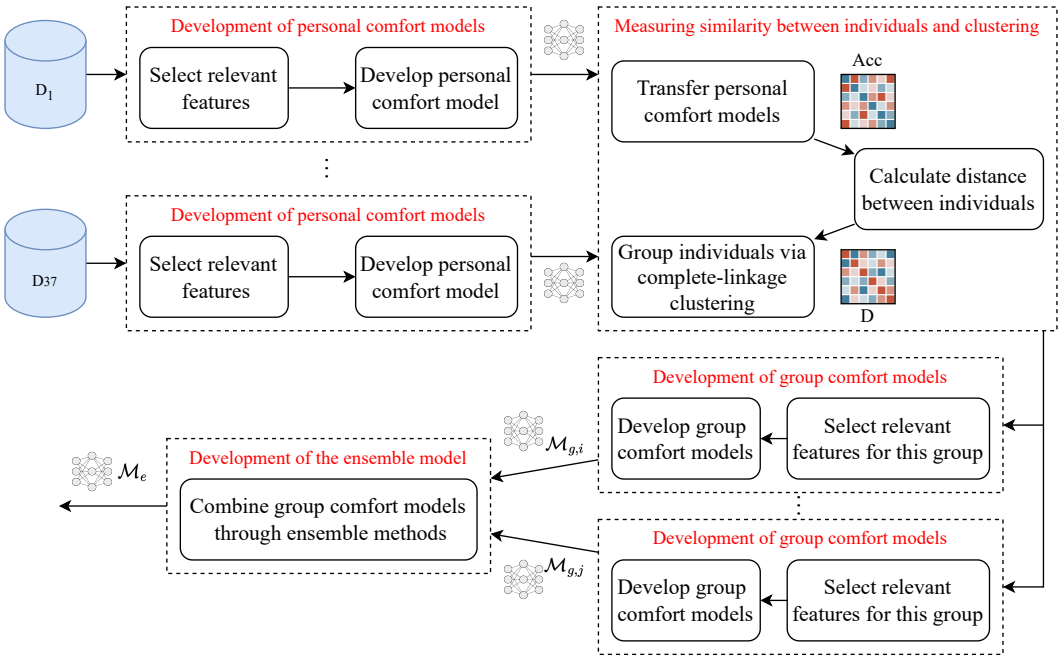


Figure 6.2: The proposed methodology.

and oversampling techniques to balance the comfort data such that the size of the data set remains the same. For undersampling, we ignore a sequence of samples (i.e., a segment of time-series) with the same label in a way that it does not affect temporal correlations in the data. For oversampling, we use the Synthetic Minority Oversampling Technique (SMOTE) [22], which selects a random sample from a minority class, identifies its k-nearest minority-class neighbours, then generates new samples using a convex combination of the selected sample and a randomly picked sample among its neighbours. There are two individuals (ID 13 and 19) who do not have the 'want warmer' label at all. We only balance the other two labels for them.

6.3 Methodology

In this section, we describe our methodology for building base learners (i.e., group comfort models) and combining them to predict an individual's thermal preference using only a small amount of labeled data (i.e., artificial labels created from PCS heating/cooling behavior). We assume artificial labels are added in the preprocessing step and class labels are now equally distributed in the data set. Figure 6.2 shows an overview of our approach which entails

(1) finding relevant features to predict thermal preference of each individual; (2) training the personal comfort model using these features; (3) measuring similarity between individuals by applying the personal comfort model of each individual to predict thermal preference of the other individuals; (4) clustering similar individuals into a group according to this distance measure; (5) training a group comfort model for all individuals within the same cluster; and (6) finally combining these pretrained group models to predict thermal preference of a new person using an ensemble method. We describe each of these tasks below.

6.3.1 Finding relevant features

We consider two different sets of features to train and evaluate the performance of personal comfort models. The first set consists of all features included in $\mathcal{D}_{sensor}$. The second set consists of the relevant features for predicting thermal comfort of each individual. To select these features we use the Benjamini-Hochberg procedure [12]. This procedure selects the relevant features so as to keep the False Discovery Rate (FDR) below a certain threshold, where FDR is defined as the expected proportion of type I errors in hypothesis testing. In other words, it ensures the percentage of irrelevant features that are called relevant out of all hypothesis tests (e.g., univariate statistical tests) is less than the given threshold.

The rationale for selecting only a subset of features for each individual is that reducing the number of features should reduce the amount of trainable weights, lower inference time, and improve robustness of the learning algorithm. The rationale for selecting all features is that the first few layers of the trained neural network could automatically perform feature selection. Thus, manually selecting a subset of feature may not be necessary given that we use a neural network model. For the cases where feature selection is done prior to model development, we select the relevant features for each individual following the Benjamini-Hochberg procedure; this leads to possibly different feature sets to predict an individual's thermal preference. Thermal comfort is influenced by individual differences [154], and therefore individuals would

have a different set of most important features since their comfort perception is influenced by the unique combinations of personal factors and environmental conditions at their local workstation. For example, mean radiant temperature might be a more important feature for an individual who seats next to a window.

To select the relevant features, we first run the Analysis of Variance (ANOVA) test to obtain the f-statistic and p-value associated with each feature. We then rank all features by sorting them in ascending order of p-value. The Benjamini-Hochberg procedure finds the largest rank that controls FDR at level τ_{FDR}. Hence, the relevant features are the features that satisfy the following constraint:

$$\frac{\text{Feature rank} \times \tau_{\text{FDR}}}{\text{Total number of features}} > \text{Feature p-value.}$$

Figure 6.3 illustrates this procedure for a sample individual. The x-axis shows the rank of each feature according to its p-value and the slope of the red dotted line is $\frac{\tau_{\text{FDR}}}{\#\text{ features}}$. The features that fall below this line are the relevant features that are selected. For example, there are six relevant features for this individual. We summarize the identified relevant features for each individuals in Figure 6.4.

6.3.2 Developing personal comfort models

We use the LSTM model to predict an individual's thermal preference. LSTM is a powerful model for univariate time-series forecasting due to its ability to capture temporal dependencies in time-series data. Our comparison with Deep Neural Network (DNN) and RF models in Section 6.4 shows that it is important to take into account these temporal dependencies. We use DNN as a baseline to show the significance of capturing temporal dependencies in thermal comfort data. Similarly, we use RF as a baseline because it is a suitable model for classification when the data set is small, and that it had the best performance in the same data set among the models developed in [83].

Each LSTM model has two LSTM layers with 10 cells in each layer. The input is a subset of features in $\mathcal{D}_{sensor}$ that are selected as discussed in the next section, and the output layer is a fully-connected dense layer that consists

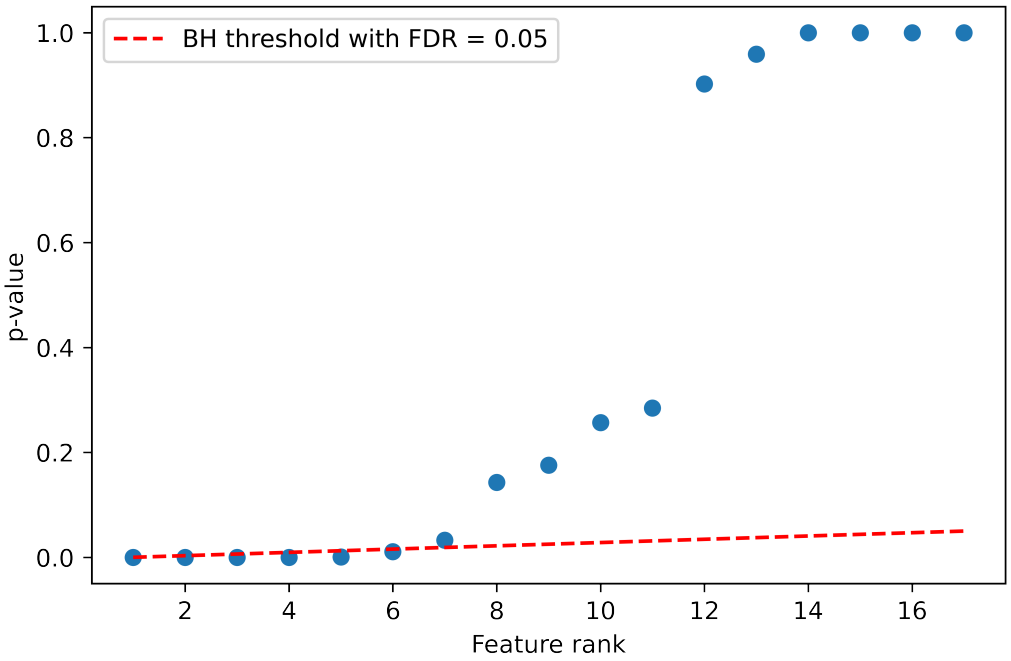


Figure 6.3: The Benjamini-Hochberg procedure for selecting the relevant features for an individual using $\tau_{\text{FDR}} = 0.5$.

of three neurons with softmax activation. Hence, it outputs a probability distribution over the three possible thermal preference labels (i.e., no change, want cooler, want warmer). Note that there is a cell state in LSTM that stores information from previous time steps. We use the Adam optimizer [85] to optimize the weights during the training time.

We denote the LSTM model trained for each individual i by $\mathcal{M}_i$, and the sensor data collected from this individual by $\mathcal{D}_i$. We have:

$$\cup_{i\in\{1\cdots k\}}\mathcal{D}_i = \mathcal{D}_{sensor}, \qquad \cap_{i\in\{1\cdots k\}}\mathcal{D}_i = \emptyset,$$

where k is the total number of individuals in our data set. We select the relevant features in $\mathcal{D}_i$, as discussed earlier, for personal comfort modeling by applying a feature mask to the data set of each individual.

The structure of the DNN model, which is used as a baseline, is similar to the LSTM model. We simply replace all LSTM layers with dense layers containing the same number of cells.

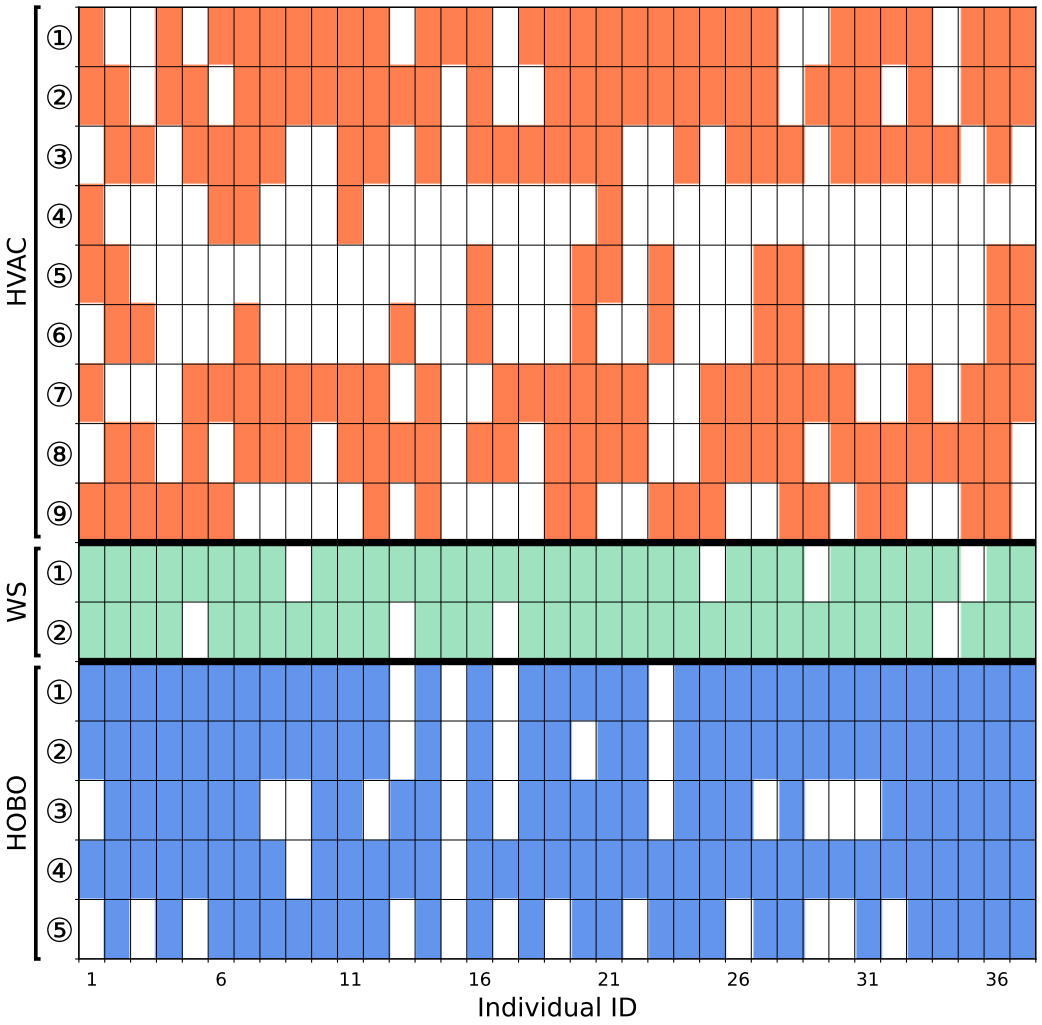


Figure 6.4: Relevant features used as input to each personal comfort model. Colored cells represent the selected features. Circled numbers refer to the order in which features are listed in Table 6.1.

6.3.3 Measuring similarity between individuals

We use the LSTM model described in the previous section as the personal comfort model. To determine the number of training epochs, we monitor the training accuracy and terminate the training process when the accuracy does not improve for 5 consecutive epochs. The batch size is set to 60, which is equivalent to 1 hour of data.

Once the personal comfort model is trained for each individual, we transfer their personal comfort model to other individuals and calculate the prediction accuracy of the transferred model. Specifically, suppose $\mathcal{M}_i$ is the model

trained on $\mathcal{D}_i$; we evaluate its performance when it is applied to predict thermal preference of another individual j ($j \neq i$) using the respective thermal preference labels in $\mathcal{D}_j$. This prediction accuracy is denoted by $\text{Acc}_{i,j}$ which is the element in row i column j of the accuracy matrix Acc. Note that labels in $\mathcal{D}_i$ and $\mathcal{D}_j$ are balanced, as described in Section 6.2.2, to avoid biased estimation.

6.3.4 Clustering individuals

After identifying $\mathcal{M}_i$'s, we group together *similar* individuals. Here the dissimilarity between two individuals is defined in terms of the prediction error when the personal comfort model of one individual is applied to predict thermal preference of the other individual. By grouping individuals, we are essentially expanding the data set used for training each of the group comfort models. The individuals with inadequate training data will benefit most from this grouping because the group comfort model can achieve a much higher accuracy than the respective personal comfort model.

We define the distance between two individuals as the maximum prediction error of the transferred personal comfort models. That is

$$\text{D}_{i,j} = \text{D}_{j,i} = 1 - \min(\text{Acc}_{i,j}, \text{Acc}_{j,i}), \tag{6.1}$$

where $i, j \in \{1 \cdots k\}$ and $i \neq j$. This way we can construct a symmetric distance matrix denoted by D from the matrix Acc which is asymmetric. Note that $\mathcal{M}_i$ might have different input features than $\mathcal{M}_j$ when we use the relevant features for each individual.

To complete the distance matrix D, we set diagonal elements to 0, intuitively because each individual is identical to themselves. Then, the symmetric distance matrix D is used to form the clusters using complete-linkage clustering. Complete-linkage clustering is an agglomerative clustering technique where each individual is initially in its own cluster. Clusters are sequentially combined into larger clusters, starting from the two clusters that have the minimum distance from one another, and the distance matrix is updated every time two clusters are merged into one. Specifically, the distance between two clusters is defined as the distance between those two individuals (one in

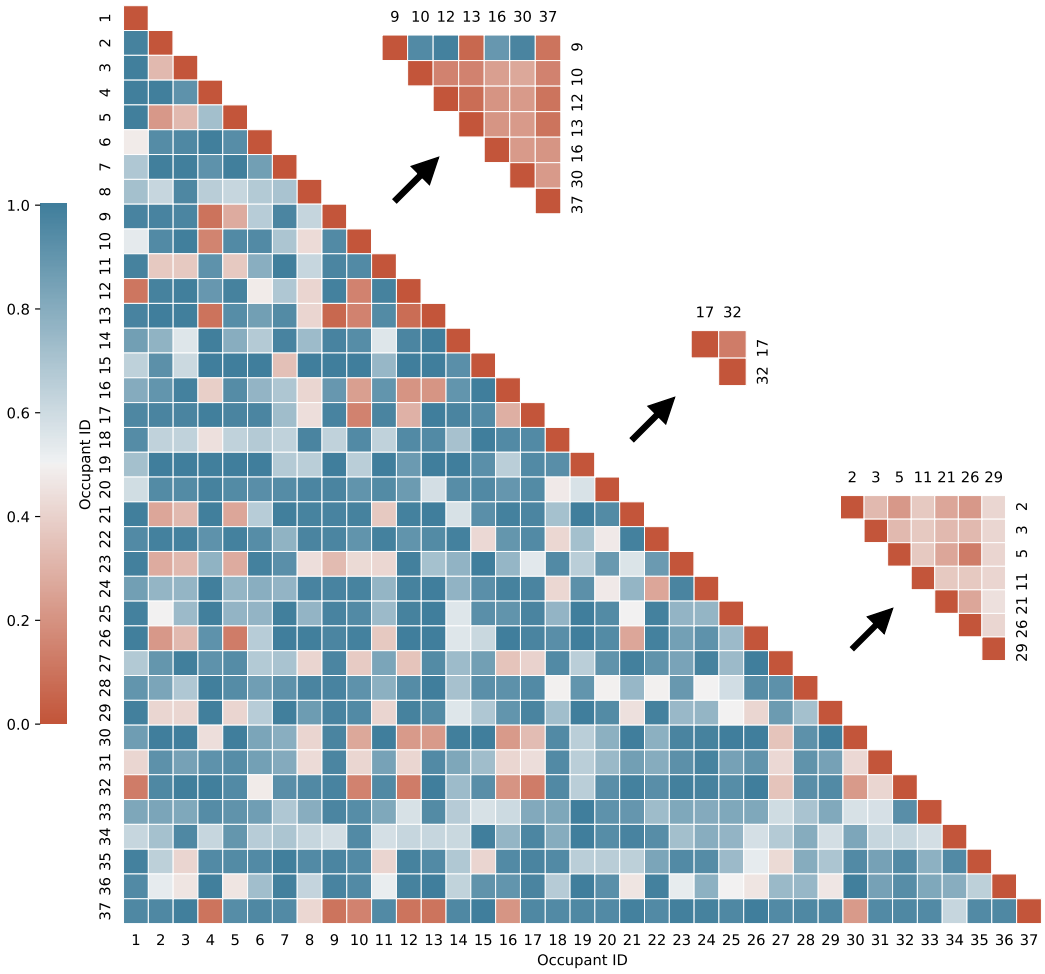


Figure 6.5: Sample agglomerative clustering result based on 37 occupants. The lower triangular portion of D is shown here together with the upper triangular portion of D of three specific clusters (out of the 15 resulting clusters).

each cluster) that are farthest away from each other. We stop merging clusters when all the pairwise distances (between clusters) are greater than a certain threshold. We set this threshold to 0.4 and treat the clusters that are not merged as our final clusters. The threshold is set based on the average estimation accuracy over all personal comfort models. When we use the relevant features only to train personal comfort models, this clustering results in 16 clusters, 8 of which contain only one individual. However, when we use all features to train personal comfort models, we obtain 15 clusters, 6 of which contain only one individual.

Figure 6.5 depicts the lower triangular portion of the symmetric distance matrix constructed for all 37 individuals before clustering them and three example distance matrices corresponding to three clusters of size 7, 2, and 7 individuals, respectively. Observe that individuals that are grouped together will have smaller pairwise distances if the corresponding group comfort model (introduced next) is used to predict their thermal preference. This implies that group comfort models work quite well for individuals within the respective clusters.

6.3.5 Developing group comfort models

To build the group comfort model $\mathcal{M}_{g,c}$ for each cluster c, we adopt the same neural network architecture and training procedure as the LSTM-based personal comfort models (i.e., $\mathcal{M}_i$'s). We first create the data set $\mathcal{D}_{g,c}$ for each cluster c, where $\mathcal{D}_{g,c} = \cup_{i \in c} \mathcal{D}_i$. In the next step, we re-evaluate the importance of input features for each group comfort model. To this end, we run the Benjamini-Hochberg procedure again, this time considering $\mathcal{D}_{g,c}$, to obtain the relevant features for the group model. Figure 6.6 shows the set of features that emerged as relevant features for each cluster. While the set of relevant features typically differs from one cluster to another, most clusters consistently include the following features: room air flow, discharge air temperature, and setpoints (measured by HVAC sensors), mean outdoor air temperature (from the weather station), and all measurements by the HOBO sensor installed at the workstation. This explains that individual's thermal perception is influ-

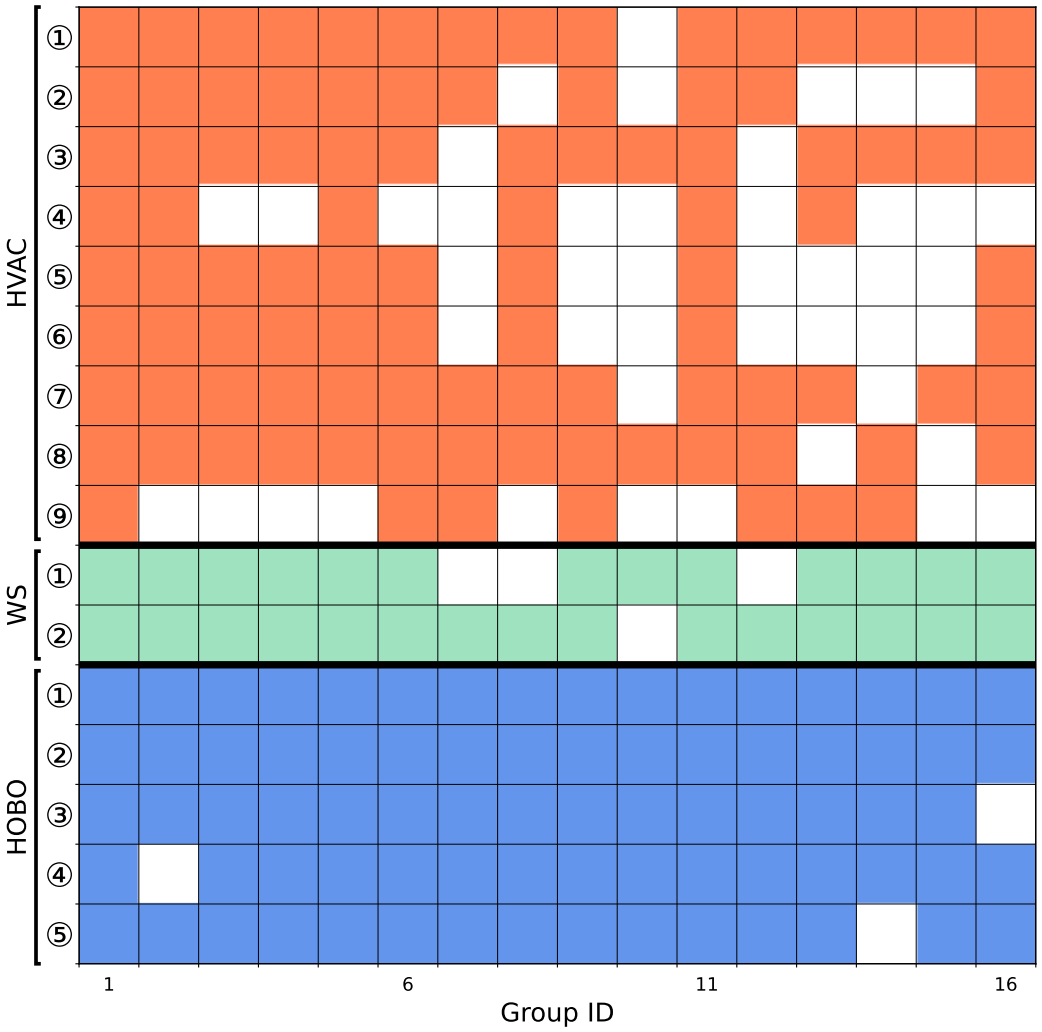


Figure 6.6: Relevant features used as input to each group comfort model. Colored cells represent the selected features. Circled numbers refer to the order in which features are listed in Table 6.1.

enced by a combination of local, room, and outdoor environmental conditions. Lastly, we train the group comfort model $\mathcal{M}_{g,c}$ for every cluster c using all labeled thermal comfort data in $\mathcal{D}_{g,c}$.

6.3.6 Combining group comfort models

Although each group comfort model can predict thermal preference of individuals within that group with high accuracy, they cannot be used for individuals that are new to the building because their group membership is not known a priori. Additionally, since enough labeled data (i.e., artificial labels) is not available from an individual who is new to the building, an accurate personal comfort model cannot be built from scratch. To address this problem, after training a group comfort model for every cluster (i.e., the base learner), we ensemble them to predict thermal preference of a person that is new to the building.

We note that it is possible to ensemble personal comfort models rather than group comfort models. However, this would increase the number of parameters in the ensemble model that must be trained, lowering the accuracy of our method when training data is limited and undermining its ability to address the cold start problem.

Next we introduce two different methods to ensemble the LSTM models that pertain to different groups.

Stacked ensemble Figure 6.7 shows the architecture of the ensemble model that uses a neural network to combine multiple base models, each being a pre-trained LSTM model. The model takes all sensor data available in $\mathcal{D}_{sensor}$ as input, including the HVAC sensor data and the HOBO sensor data. The feature masks are used to filter out the features that are not among the relevant features for each group so that $\mathcal{M}_{g,c}$ receives only the appropriate input features. The output of each group comfort model $\mathcal{M}_{g,c}$ is the probability distribution over the three thermal preference labels. These probabilities are then fused using a dense layer with 10 neurons (referred to as the meta-model). Fusing these probabilities before producing the final output enables the model to learn the latent relationship between the outputs of different group com-

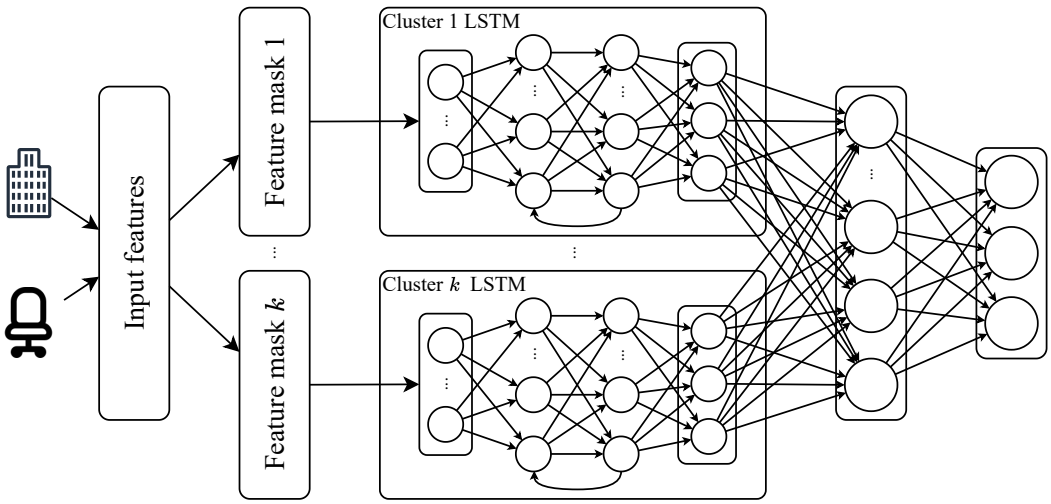


Figure 6.7: Architecture of a stacked ensemble. The input to the meta-model (the last layer before output) is the stacked output of group thermal comfort models.

fort models, which is not considered in other ensemble methods that will be presented next.

In the training stage, all the weights in group comfort models are marked immutable. Therefore, the model only needs to update the weights for the last two layers. We use the same optimizer that is used to train $\mathcal{M}_{g,c}$ to update the meta-model's weights.

Mixture of experts The mixture of experts [72] is an ensemble learning method in which a gating network learns how to assign responsibilities to the experts (i.e., base learners). Concretely, the gating network computes the probability of assigning an input data point to each expert, which is an element of $\mathbf{w}$, based on the relative performance of the experts for that data point. The architecture of the gating network is shown in Figure 6.8. Similar to the stacked ensemble, we only update weights for the gating network during training and do not modify weights of the base models. The loss of this gating network is defined as:

$$\ell = -log \sum_{c} w_c e^{-\frac{1}{2}\|\mathcal{M}_{g,c}(f_c(\mathrm{X}))-\bar{\mathbf{y}}\|^2},$$

where X is a batch of input data, f_c is the feature mask, and $\bar{y}$ is the associated label. The main difference between the mixture of experts and stacked ensem-

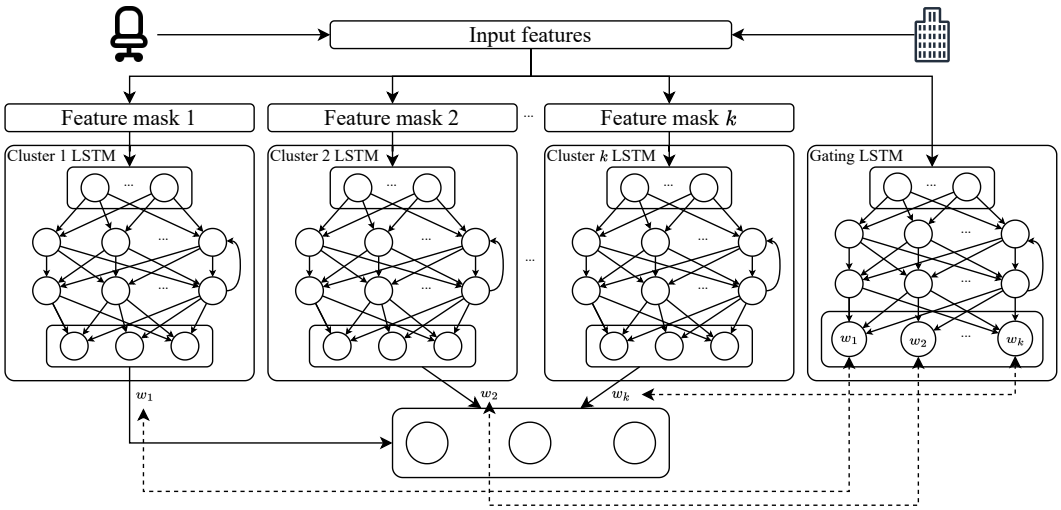


Figure 6.8: Architecture of a mixture of experts consisting of a gating network that assigns responsibilities to local experts.

ble is that the base models compete with each other in the former whereas they collaborate with each other in the latter to produce the final output. This competitiveness is because of the loss function we use to train the gating network.

6.4 Results

In this section we study the efficacy of different comfort models in a classification task, i.e., thermal comfort prediction, using the methodology laid out in Section 6.3. We adopt accuracy, which is defined as the percentage of correct predictions over all predictions, as our evaluation metric. We also present the confusion matrix to provide insight into which classes are confused more often. We split the data collected from each individual i into training and test sets which are denoted by $\mathcal{D}_i^{train}$ and $\mathcal{D}_i^{test}$, respectively. Specifically, the first half of $\mathcal{D}_i$ is used for training and the rest for testing. We balance class labels in $\mathcal{D}_i^{train}$ using the technique described in Section 6.2.2. However, for fair comparison with related work, we do not balance class labels in $\mathcal{D}_i^{test}$ and only use artificial and true labels for the evaluation of each comfort model.

Consider an individual i that belongs to cluster c. We evaluate the personal, group, and ensemble comfort models on $\mathcal{D}_i^{test}$. Note that while their personal comfort model, $\mathcal{M}_i$, is trained on $\mathcal{D}_i^{train}$, their group comfort model, $\mathcal{M}_{g,c}$, is

trained on $\mathcal{D}_{g,c} \setminus \mathcal{D}_i^{test}$. This guarantees that the training set does not overlap with the test set. Finally, the ensemble comfort model is built by combining all the group comfort models. We assume that each group comfort model $\mathcal{M}_{g,c'}$ is trained on $\mathcal{D}_{g,c'}$, except for the group comfort model $\mathcal{M}_{g,c}$ that is trained on $\mathcal{D}_{g,c} \setminus \mathcal{D}_i$. This is necessary as we intend to evaluate the ensemble model for an individual that is new to the building, so neither their training data nor their test data can be used to train the group models that will serve as base models in the ensemble model. Finally, if individual i is the sole member of cluster c, we neglect $\mathcal{M}_{g,c}$ when we build the ensemble model because $\mathcal{D}_{g,c} \setminus \mathcal{D}_i$ would be an empty set in this case.

For the sake of comparison, we calculate and report the prediction accuracy of the PMV and adaptive models for each individual too. We use the `pythermalcomfort` Python package [144] to calculate the PMV value and acceptable operative temperature under ISO 7730 [71] and ASHRAE 55 [4] standards. To compare the results of conventional and personal comfort models on the same scale, we convert PMV into thermal preference classes based on the following assumptions: $|PMV| \leq \tau$ is 'no change'; $PMV > \tau$ is 'want cooler'; and $PMV < -\tau$ is 'want warmer'. Because both the ISO 7730 standard [71] and ASHRAE 55 [4] recommend maintaining $|PMV|$ below 0.5, we set $\tau = 0.5$ in this study. To convert the output of the adaptive model into thermal preference classes, we assume acceptable operative temperature within 80% acceptability limits to be 'no change'; and greater/less than the upper/lower 80% acceptability limits to be 'want cooler/warmer', respectively.

6.4.1 Evaluating personal comfort models

Figure 6.9 depicts the accuracy distribution of three personal comfort models, namely LSTM, DNN, and RF, and two conventional comfort models, namely PMV and Adaptive, when they are applied to predict the thermal comfort of every individual in our data set. It also depicts the accuracy distribution of a group comfort model which we discuss in the next section. Each personal comfort model is trained twice on $\mathcal{D}_i^{train}$, first with the relevant features of individual i then using all features. The width of the box shows the Interquartile

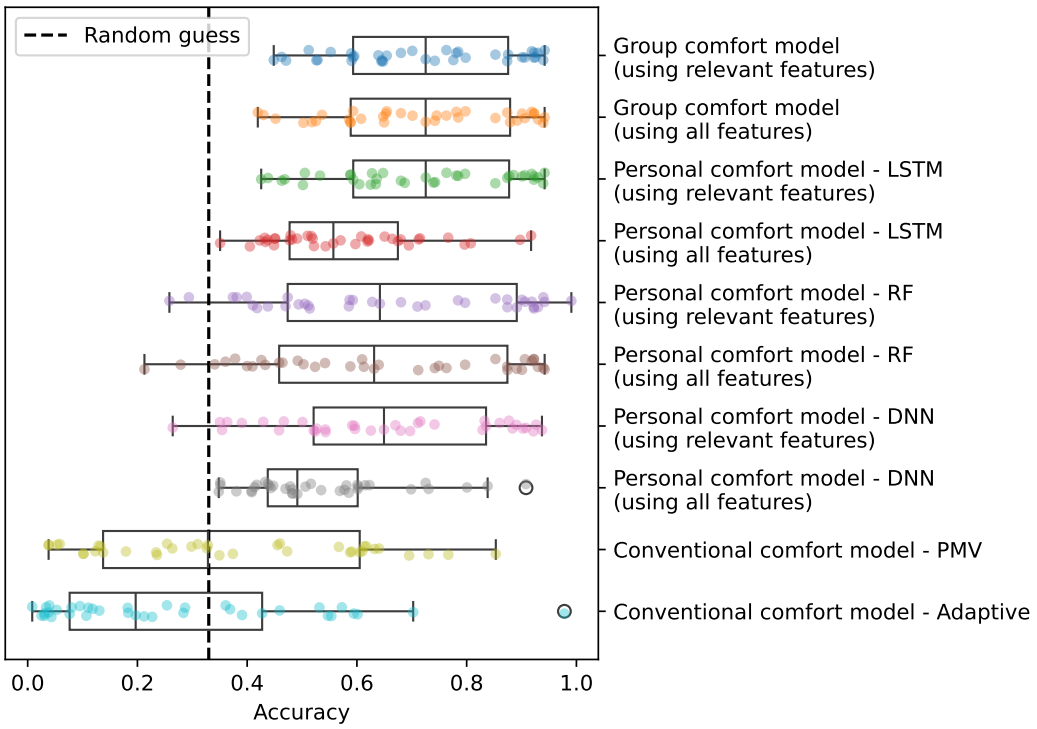


Figure 6.9: Thermal comfort prediction accuracy of different individuals using group, personal, and conventional comfort models.

Range (IQR), the vertical line in the middle is the median of the distribution, and the whiskers extend to show the rest of the distribution excluding outliers. Each dot represents the prediction accuracy of a thermal comfort model for one individual.

Compared to conventional comfort models, personal comfort models give more accurate predictions for most individuals as it is evident from this figure. In fact, the median accuracy of both conventional comfort models is worse than random guess (i.e., 33.3%), which confirms the finding from [28]. Turning our attention to the personal comfort models that are trained using the relevant features for each individual and all features, we witness that performing feature selection is helpful as the median shifts to the right in all cases. We conclude that removing irrelevant features allows for training a more accurate model, especially when training data is limited and the neural network models cannot easily weed out such features.

It can also be seen that the LSTM model, when trained using the relevant

features, yields a higher median accuracy than DNN and RF. This implies that capturing temporal dependency enhances the performance of personal comfort models. It also has higher precision and recall than DNN and RF on average. Specifically, using the relevant features only, the average precision of LSTM, DNN, and RF is respectively 70.41%, 63.53%, and 63.70%, and the average recall of these models is respectively 69.43%, 63.43%, and 63.82%. Note that the two RF models have superior performance for a number of individuals, but in general they do not outperform the LSTM model that takes the relevant features as input for each individual; this is also evident from a comparison of their median values. Moreover, the thermal comfort prediction accuracy has a wider spread (IQR) under the RF models. Although the RF models predict the thermal comfort of some individuals with over 90% accuracy, their accuracy is even lower than random guess for some other individuals. This justifies the use of the LSTM model to build group comfort models which we discuss next.

Figure 6.13b shows the confusion matrix for the LSTM model. The value written inside each cell indicates the percentage of data with a given true label that is predicted to have a label that might be the same or different from the true label. Notice that the share of 'no change', 'want warmer', and 'want cooler' labels in the test set is respectively 28.24%, 29.82%, and 41.94% (written below each column), suggesting that the test set is not heavily imbalanced.

6.4.2 Evaluating group comfort models

We evaluate the performance of group comfort models when they are used to predict the thermal comfort of an individual. Assigning a group comfort model to an individual is carried out based on the assumption that the group membership is known a priori. The first two box plots in Figure 6.9 show the distribution of accuracy when the group comfort model is used to predict an individual's thermal preference. It can be readily seen that it outperforms all personal comfort models in terms of the median of the accuracy distribution using either one of the feature sets. The only exception is the LSTM model

that takes relevant features as input, even in that case the performance of this model is on par with the group comfort model. The superior performance of the group comfort model can be due to a few reasons. A group comfort model is trained on labeled thermal comfort data from all individuals within the same group, increasing the total amount of training data compared to what is available for training each personal comfort model. The increased size of training data can also help to alleviate the class imbalance problem which is common in thermal comfort data sets. The only drawback is generalization, i.e., the group comfort model is not custom-made for each individual. This implies that the personal comfort model can potentially outperform the group comfort model when training data is sufficient, which is not the case here as we are tackling the cold start problem.

Despite the higher accuracy of group comfort models, they cannot be used to predict the thermal preference of an individual that is new to the building since we do not know their group membership yet. We address this problem by combining pretrained group comfort models using different ensemble methods. In essence, the learning algorithm utilizes $\mathcal{D}_i^{train}$ to incrementally teach the ensemble model which base model(s) should be used to predict the thermal preference of this particular individual. We create ensemble models by combining group comfort models rather than personal comfort models for two main reasons. First, the LSTM-based group comfort model proves to have similar performance compared to the LSTM-based personal comfort model (see Figure 6.9). Second, by combining multiple individuals into one group, we can effectively reduce the number of models to consider and consequently the amount of parameters in the meta-model (or gating network) to train using $\mathcal{D}_i^{train}$. This can enable the ensemble model to converge to a high accuracy level using less training data as we discuss in the next section.

6.4.3 Addressing the cold start problem

We now show the benefit of using ensemble comfort models when it comes to predicting thermal preference of an individual who has provided no feedback about their local thermal environment.

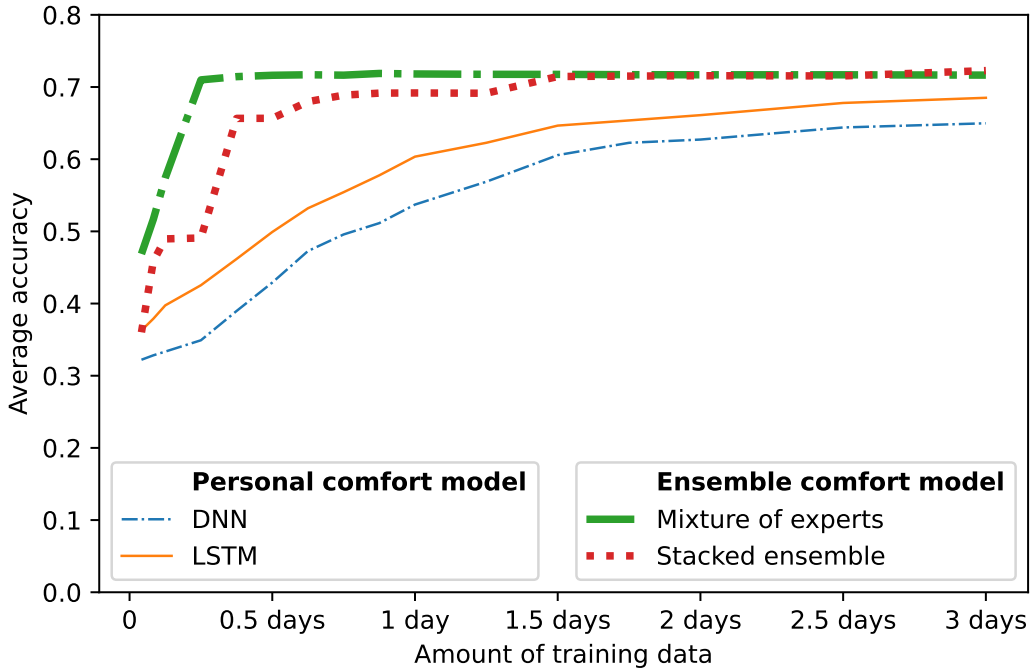


Figure 6.10: The learning curves of different comfort models trained using the relevant features for each individual. Each curve represents the thermal comfort prediction accuracy of a specific model averaged over individuals who had enough data.

Figure 6.10 shows the prediction accuracy of different kinds of comfort models (averaged over all individuals who had enough data for this experiment) as we increase the amount of training data that is available for each individual from 1 hour to 3 days (in chair occupancy time). It can be readily seen that the accuracy of the two ensemble models converges to a relatively high value when the amount of training data surpasses 1.5 days. Another observation is that the learning curve of the mixture of experts has a steeper slope in the first few hours than other models, indicating that it can achieve a sufficiently high accuracy using only 6 hours of training data. This suggests that the mixture of experts ensemble model is more capable of addressing the cold start problem than the stacked ensemble. We also find that all ensemble comfort models converge to above 71% average accuracy, whereas LSTM and DNN-based personal comfort models that are trained from scratch reach respectively the average accuracy of 68.5% and 64.9% with 3 days of training data. Note that the accuracy of personal comfort models appears to increase slowly after they are provided with 1.5 days of training data. The LSTM-based personal comfort model initially has comparable performance with the stacked ensemble, but the stacked ensemble improves quickly as the amount of training data surpasses 6 hours. The stacked ensemble eventually reaches a prediction accuracy that is even slightly higher than the other ensemble model, precisely 72.3% with 3 days of training data.

The ensemble models are expected to eventually reach the same level of accuracy as the group comfort model when the group membership is known by learning how to assign responsibilities to the base models or fuse their outputs. Our experiment shows that with 3 days of training data the mixture of experts and stacked ensemble models converge to the average accuracy of 71.7% and 72.3%, respectively. This is just about 1.5% and 0.8% lower than the average accuracy of the group comfort model (73.1%) assuming that the group membership is known a priori. We anticipate that the gap will shrink further as more training data becomes available beyond the 3 days used in this experiment.

Figure 6.11 depicts changes in the accuracy of the LSTM-based personal

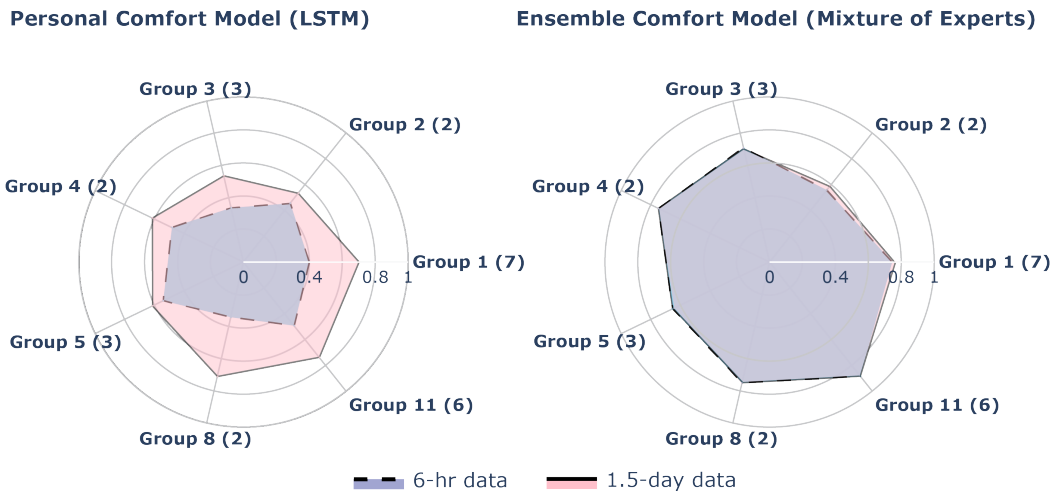


Figure 6.11: The performance comparison of different kinds of comfort models as more training data becomes available. The purple and pink areas show respectively the thermal comfort prediction accuracy when only 6 hours and 1.5 days of training data is available. The accuracy of the LSTM-based personal comfort model improves 20% on average when 1.5 days of training data becomes available. This accuracy improvement is only 0.65% for the mixture of experts ensemble comfort model. The results are grouped by the group each individual belongs to.

comfort model averaged over individuals in the same group, and the mixture of experts ensemble model when we increase the amount of training data from 6 hours (when the best ensemble model converges) to 1.5 days (when all models start to converge). We only show the results for the groups that have more than one member (7 groups in this category). For most groups, the mixture of experts ensemble[3] has already converged using 6 hours of training data and providing more training data does not lead to a significant performance boost. The only exceptions are individuals in Group 1, 2 and 5 who experience a modest performance boost when 1.5 days of training data becomes available. On the contrary, the LSTM-based personal comfort models that are trained from scratch do not reach their highest potential when only 6 hours of training data is available. However, there is a performance boost when 1.5 days is used to train the personal comfort model. In Group 8, we even observed an increase of more than 30%, bringing the average accuracy closer to the accuracy

[3]For brevity, we only plot the results of one ensemble comfort model, i.e., the mixture of experts, in this figure.

attained by the ensemble comfort model. This warrants further investigation into why training a personal comfort model from scratch for a certain group of individuals could be a better idea than the other groups. We defer this to future work.

Figure 6.12 compares the performance of different comfort models as more training data becomes available. We consider 37 individuals that had more than 1.5 days of data in our dataset. There are three bars for each individual depicting the accuracy of the following thermal comfort models: the LSTM-based personal comfort model, the mixture of experts ensemble model, and the group comfort model (assuming we knew the group membership a priori). The lower segment of each bar shows the accuracy when the respective model is trained using 6 hours of data. The upper segment (stacked on top of the other segment) shows the accuracy improvement when the model is trained using 1.5 days worth of data. This figure shows how much the accuracy of the LSTM-based personal comfort model and the mixture of experts ensemble model will increase for each individual as more training data becomes available. We find that the accuracy of the mixture of experts ensemble model does not noticeably increase for most individuals when more that 6 hours of training data becomes available. Overall, the mixture of experts and stacked ensemble attain respectively the average accuracy of 71.0% and 49.1%, when 6 hours of training data is available, while the LSTM-based personal comfort model yields 42.5% accuracy on average for the same amount of training data.

To better understand the performance of personal and ensemble comfort models, it is useful to look at the confusion matrix together with the classification accuracy that we presented in the chapter. Figure 6.13 shows the confusion matrix for the LSTM-based personal comfort model and the mixture of experts model. Each confusion matrix shown here is created by averaging over individuals who had enough data. We note that the percentage written next to each row/column label shows the share of the respective true/predicted label. Figure 6.13a and 6.13c show the confusion matrices for the LSTM-based personal comfort model and the mixture of experts ensemble model when 6 hours of training data is available. It can be readily seen that the ensemble

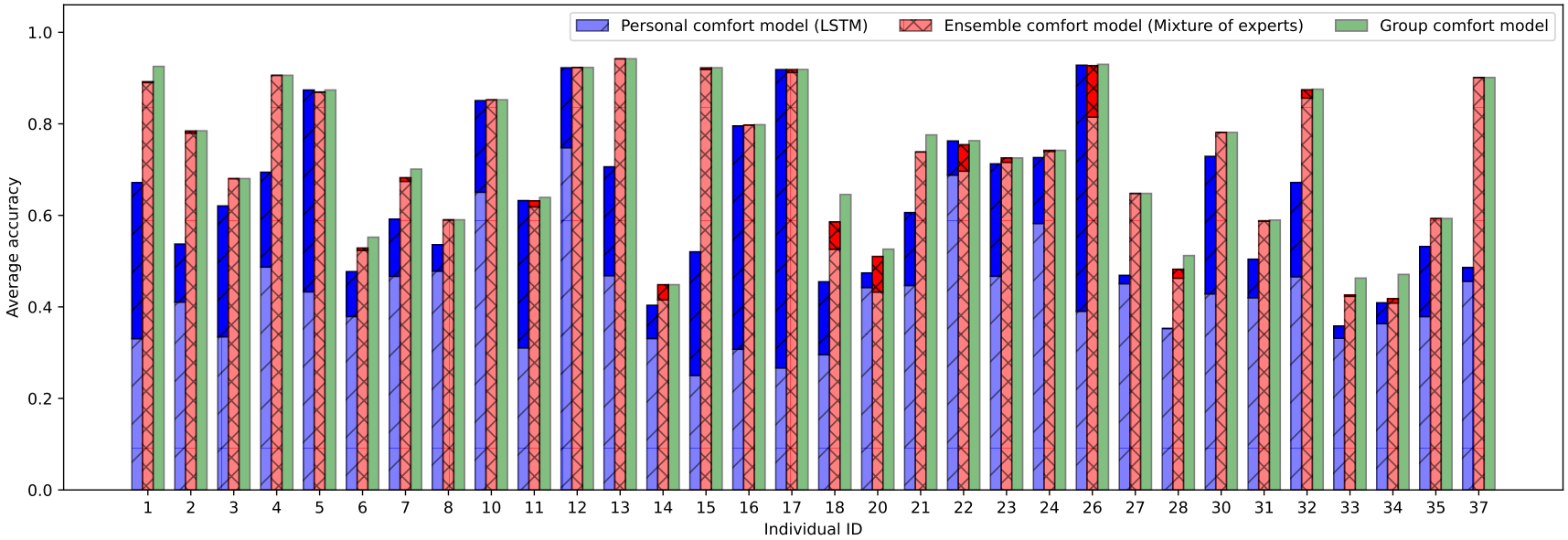
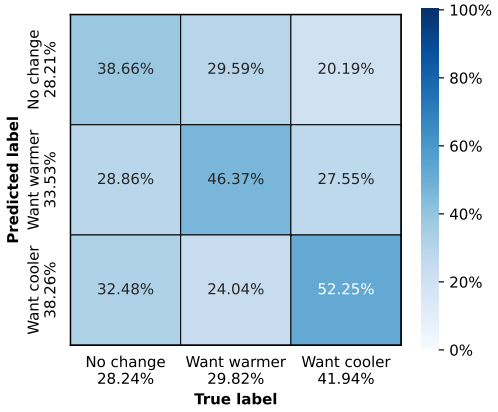
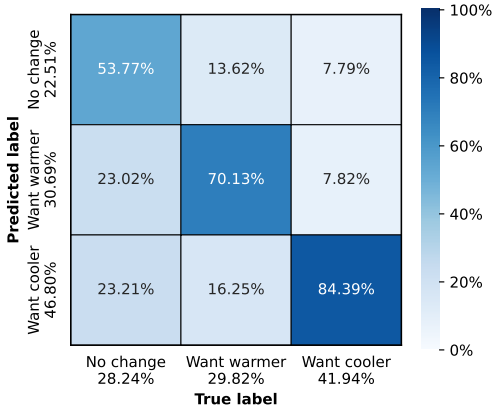


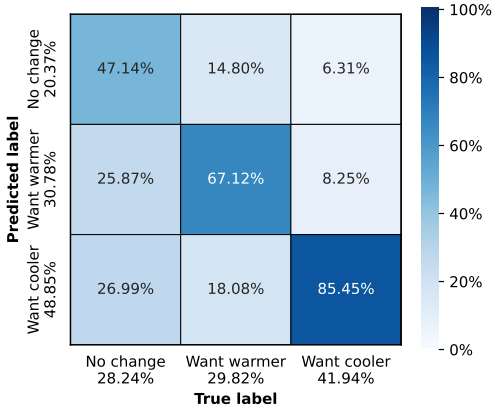
Figure 6.12: The performance comparison of different kinds of comfort models as more training data becomes available. The lower segment of each bar shows the thermal comfort prediction accuracy when only 6 hours of training data is available. The upper segment shows improvement in the average accuracy when 1.5 days of training data becomes available.



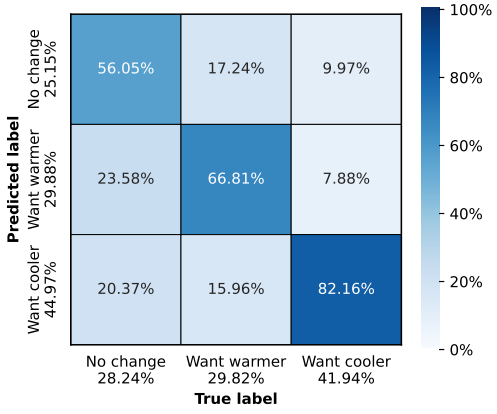
(a) LSTM-based personal comfort model with 6 hours of training data



(b) LSTM-based personal comfort model with all training data



(c) Mixture of experts ensemble model with 6 hours of training data



(d) Mixture of experts ensemble model with all training data

Figure 6.13: The confusion matrix obtained on the test set for (a, b) the LSTM-based personal comfort model and (c, d) the mixture of experts ensemble model trained using the relevant features for each individual with different amounts of training data.

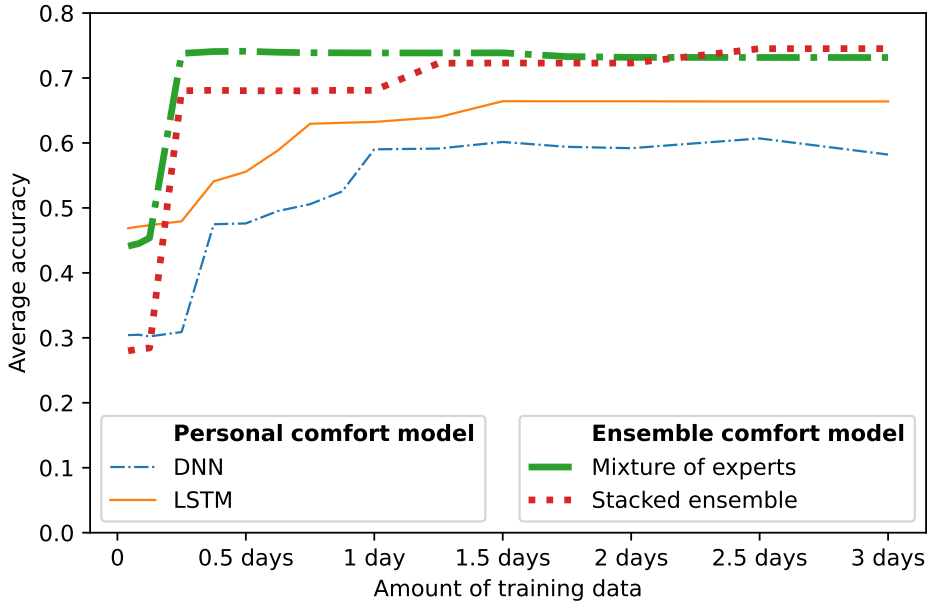
model has better performance than the personal comfort model for all three class labels. Moreover, it significantly reduces the possibility of confusing 'want warmer' and 'want cooler' labels (the entries in Row 3 Column 2 and Row 2 Column 3), which has more serious consequences as it can affect HVAC operation. This result alludes to the potential of ensemble comfort models to address cold start as they can predict thermal preferences with over 71% accuracy using 6 hours of training data, excluding no occupancy periods.

6.5 Discussion

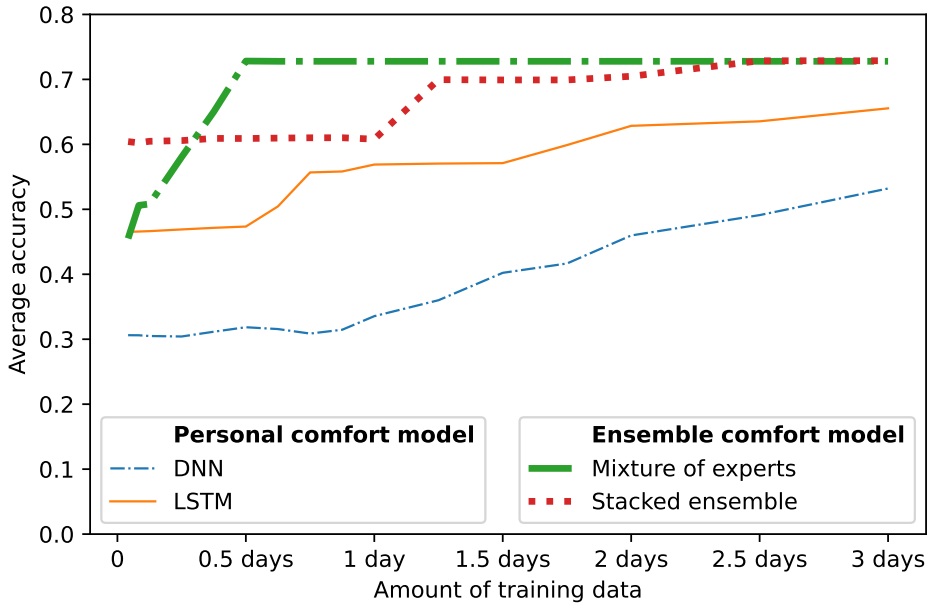
6.5.1 Revisiting relevant features for predicting thermal preferences

Understanding which features are relevant to a prediction model is important because it can help to (1) improve the computational overhead and accuracy of the predictive model especially when labeled data is scarce, and (2) optimize data collection efforts. At individual levels (Figure 6.4), features from the HOBO sensors and weather station are more frequently identified as relevant compared to those from the HVAC sensors. A similar observation is made at group levels (Figure 6.6) where relevant features are frequently selected from the HOBO sensors and weather station. However, certain features from the HVAC sensors are considered important to many individuals/groups, such as room air flow, room damper position, room heating setpoint, and room cooling setpoint. This may be due to the narrow deadbands in this building that drive the frequency of air exchange and the volume of air flow. Practically speaking, selecting important features for model development is not always based on their contribution to accuracy but rather on the data collection cost. Although HOBO sensors can better capture what individuals experience in their workstation, they require separate installations since they are not part of the building's existing sensor network.

For those whose workstation is closely located to their zonal thermostat, thermostat readings may reasonably describe individuals' thermal conditions. To verify this hypothesis, we select 10 individuals whose workstation is close to the thermostat and evaluate the performance of different comfort models



(a) HOBO sensor data was available



(b) HOBO sensor data was not available

Figure 6.14: The learning curves of different comfort models trained using the relevant features for each individual when the HOBO sensor data is ignored and when they are utilized. Only 10 individuals whose workstations were close to the HVAC sensors are selected. The curves show the thermal preference prediction accuracy of a specific model averaged over these individuals.

for these individuals only (see Figure 6.14). The figure shows that not having features from HOBO sensors will increase the amount of training data that is needed for all models, especially the personal comfort models, to converge. Nevertheless, with 12 hours of training data, the mixture of experts ensemble reaches an accuracy of 73%. We observe that each model eventually converges to nearly the same accuracy level regardless of the HOBO sensor data availability. More precisely, upon convergence, the difference between the accuracy of the model trained with the HOBO sensor data and the accuracy of the model trained without this data is minimal (less than 1%, except for the DNN-based personal comfort model which has an accuracy loss of 5%). Based on this observation, we conclude that the absence of HOBO sensors will not significantly affect the performance of comfort models we introduced in this study for occupants that sit close to a thermostat. This can be attributed to the fact that for these individuals the data collected by HVAC sensors is a reasonable proxy for the data collected by HOBO sensors.

6.5.2 Using comfort proxies to generate artificial labels

Previous work [81], [83] has shown that the choice of heating vs. cooling via PCS is a strong predictor of one's thermal preferences. This is not surprising since many "smart" thermostats (e.g., Nest) use this concept to train their temperature preference model. Based on this finding, we generate artificial labels (i.e., want warmer/want cooler/no change) by using PCS behavior data and develop a set of thermal comfort models (personal, group, and ensemble). We show the advantage of this approach by: (1) increasing the amount of labeled data per person compared to survey feedback (e.g., by 97 folds as shown in Figure 6.1), (2) avoiding the need for occupant surveys that can be intrusive, and (3) allowing the use of complex algorithms that require large training data, such as neural networks. We also show that capturing temporal dependency in thermal preferences via continuous PCS data can enhance prediction accuracy of the model (i.e., LSTM-based comfort models).

The same approach is applicable to other heating and cooling devices that allow personal control over thermal environment, such as desk fans, radiant

heaters, and smart thermostats with individualized accounts. With the growing adoption of IoT devices in homes and buildings, it is possible to collect real-time data that can be traced back to individuals for continuous preference learning and dynamic setpoint controls. Physiological conditions (e.g., heart rate, skin temperature) can also act as proxy variables to infer individuals' thermal comfort [30], [60], [98]. Hence, future studies should explore the feasibility of generating artificial labels based on wearable sensors for the training of personal comfort models.

6.5.3 Importance of ensembling pretrained comfort models

Lastly, we try to shed light on why it is possible to quickly build an accurate classification model by ensembling a set of pretrained group comfort models. We believe that there are two main reasons. First, the group comfort models are trained using sufficient data, so by combining these models and keeping their weights fixed we can build a powerful and complex neural network with only a small number of trainable weights. This model can be easily trained even when training data is limited. Second, classification models that are trained from scratch inevitably have high variance due to the small size of the training set. Ensemble methods can reduce this variance, leading subsequently to higher performance in the classification task when training data is limited.

Insufficient data is an inherent problem in thermal comfort modeling, especially when we have to develop a model for a new person. Ensembling pretrained group comfort models can help to address the cold start problem by quickly building a comfort model with a reasonable accuracy using only 6 hours of training data. We believe that it will be a useful tool to learn individuals' thermal preference in a new building or an existing building with transient populations.

6.6 Conclusion

The emergence of occupant-centric controls in the building domain has fuelled research on the development of thermal comfort models that are accurate, adaptive, and customized for building occupants. In this chapter, we address

two major challenges of developing such models, namely the lack of sufficient labeled thermal comfort data and low prediction accuracy of thermal comfort models when occupants are new to the building. The proposed approach entails generation of synthetic labels from occupants heating and cooling behavior and development of personal comfort models for each occupant using the augmented data set. The personal comfort models were applied to predict thermal preference of other occupants as a basis for measuring the pairwise distance between them and accordingly clustering them. Using a rich data set from a field study with 37 individuals, we were able to identify a small number of clusters, each containing one or multiple individuals with similar thermal preferences, and trained a group comfort model using data from individuals in each cluster. We then combined these pretrained group comfort models through various ensemble learning methods. Our result suggests that the best ensemble comfort model, namely the mixture of experts ensemble, can reach its peak performance when 6 hours worth of data (excluding no occupancy periods) becomes available from a new occupant, while much more data would be needed to train a personal comfort model for this occupant.

Chapter 7

Space planning in flexible workspaces

Energy consumption in office buildings, especially in coworking spaces, can be substantially reduced through joint optimization of space use and heating and cooling demands. This chapter addresses this underexplored research problem in a coworking space that offers long-term and daily plans. We train an input convex neural network to estimate the energy consumed by the HVAC system in a single day to condition a given zone of the building. Due to the convexity of this model in its inputs, we can formulate a convex mixed-integer program to optimize HVAC energy consumption by deciding how to assign desks to occupants and adjust zone temperature setpoints. Considering a medium-sized office building as the coworking space, we show that this optimization problem can be solved to near-optimality relatively quickly, hence it can be used to make decisions regarding long-term bookings. For daily bookings, we design heuristic algorithms that take the solution of the optimization problem and assign the remaining space, while ensuring the satisfaction of thermal comfort constraints. By incorporating these algorithms in the workspace reservation system, we quantify the potential savings that can be achieved.

7.1 Introduction

Most HVAC control algorithms that have been developed to date treat the occupancy state of the building or individual zones within the building as an exogenous variable and use a general thermal comfort model, such as the Fanger comfort model [46], which ignores individual differences in thermal comfort and satisfaction. To optimize HVAC energy consumption without

sacrificing thermal comfort, previous work has focused on dynamically adjusting the room temperature setpoint using a supervisory control system, *e.g.* a reinforcement learning agent [40], that works in conjunction with conventional feedback controllers, or directly controlling a subset of actuators that exist in the AHU and VAV boxes, from supply air temperature and flow rate to damper and reheat valve positions [23], [64]. This results in a trade-off between energy use and thermal comfort. As shown in [171], a better trade-off is achievable if one can change the spatial distribution of the occupants in the building and simultaneously adjust the temperature setpoint of every zone according to the thermal comfort needs of the actual occupants of that zone. This calls for joint optimization of space use, and heating and cooling energy consumption.

Despite having great promise, occupants cannot be freely relocated in all commercial buildings to increase savings and comfort for several reasons: (a) some building spaces have a unique function or contain special equipment; (b) organizational dependencies often dictate what spaces might be used by a group of occupants; (c) relocating occupants could negatively affect their performance. However, many office buildings, including coworking spaces, consist of shared workspaces that have the same function and can be accessed by individuals that book their desks independently. The coworking spaces, in particular, offer a range of plans and pricing models to cater to the diverse needs of their customers. The most common plan offered to members, *e.g.* customers who have a monthly or longer-term subscription, is the dedicated desk or office plan where members will have their own desk in a shared workspace or their own private office. Non-members can buy a day pass on short notice (known as on-demand booking), which provides temporary access to the workspace on a first-come, first-served basis. It is generally acceptable to assign different desks to an individual across multiple bookings, if they buy day passes.

Optimizing HVAC energy consumption in coworking spaces is a non-trivial problem. This is primarily due to the difficulty of modeling the latent relationship between the energy consumed by the HVAC system to condition a given zone, and the number of occupants and temperature setpoints of that

zone and adjacent zones. Even when an accurate model can be identified via a data-driven approach, if the model is nonlinear and nonconvex, computational issues will arise in solving an optimization problem that has this model as the objective function. Apart from that, due to integer decision variables, the optimization problem will still be NP-Hard even if we embed a convex model in the objective function. We make the following contributions in this chapter:

- We train an input convex neural network to estimate the total amount of energy that must be consumed by the HVAC system in a single day to condition a zone with a specific temperature setpoint and number of occupants. We show that this surrogate model achieves higher accuracy than other alternatives. Using this surrogate model and probabilistic thermal comfort profiles of the occupants, we cast optimal HVAC operation and space allocation as a convex Mixed-Integer Nonlinear Programming (MINLP) [11], where the objective function is convex and the feasible set is convex when integrality is relaxed. This problem is solved relatively quickly to near-optimality by a branch-and-bound algorithm.

- We propose two efficient heuristic algorithms for assigning desks to short-term occupants and show that their solutions are comparable with that of the online version of MINLP, yet they run faster and can scale to larger buildings.

- We evaluate the performance of the proposed algorithms by analyzing the total HVAC energy consumption, average thermal comfort of building occupants, and the ratio of rejected reservation requests to received requests. Our result indicates that these algorithms can increase the profit of a coworking office building located in Toronto, Canada, by accepting around 100 more on-demand bookings without increasing energy consumption and associated costs!

7.2 Problem statement

The assignment of occupants to thermal zones is anticipated to have significant impact on the best trade-off that can be found between HVAC energy use and

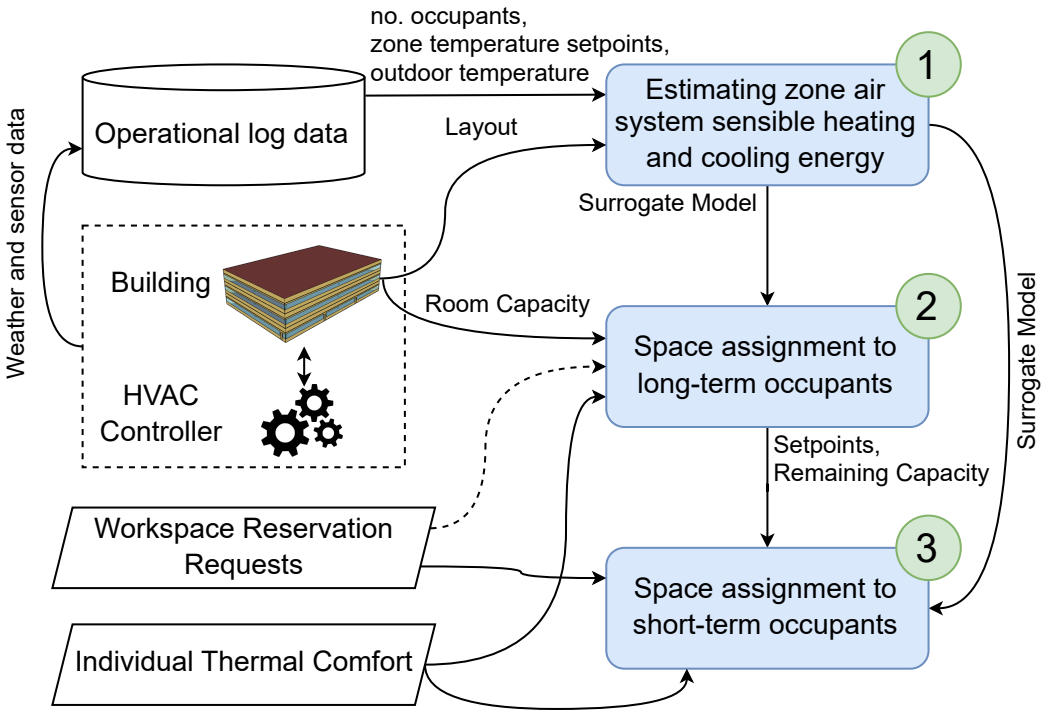


Figure 7.1: A schematic representation of our methodology.

occupant thermal comfort [171]. We study this impact in an office building that offers a coworking space. The shared workspaces in this building are rented out to long-term and short-term occupants who bought monthly subscriptions and day passes, respectively. We seek to minimize HVAC energy consumption while satisfying the minimum thermal comfort and space capacity constraints by changing the assignment of spaces/desks to occupants and adjusting the zone temperature setpoints. This approach capitalizes on three things: the diversity of individual thermal preferences, the unique characteristics of each thermal zone, and differences in marginal energy consumption of the HVAC system in each zone caused by placing more occupants there.

7.2.1 Assumptions

We make the following assumptions which serve as the foundational basis for subsequent analyses in the study.

1. The building is divided into multiple zones, each having its own thermostat and local control system. This enables zone-level control of the HVAC system by adjusting the temperature setpoint and other control

knobs, *e.g.* the damper position.

2. The HVAC system has two modes of operation. In the *active* mode, from 6am to 10pm, the fan is on and its speed is determined by a controller in the AHU. In the *inactive* mode, from 10pm to 6am of the next day during which the building is closed to occupants, the fan operates at the minimum speed to save energy. Similarly, the heating and cooling setpoints can be defined by our algorithm in the active mode, but they are set to 15.6°C and 26.7°C in the inactive mode to save energy.

3. The layout of the building showing the location of each room and its adjacent rooms, the capacity of each room, and the mapping between rooms and thermal zones are known.

4. The coworking space offers only two types of plans. We classify the building occupants into two distinct groups according to the plan they purchase. To minimize discomfort caused by relocation, *long-term occupants* are offered dedicated desks in a shared workspace (zone) for the term of their subscription. Thus, they will continue to use the same space that is assigned to them on the first day. *Short-term occupants*, however, can reserve desks in a shared workspace (zone) for a single day by purchasing a day pass the previous day. So if an individual buys a day pass two days in a row, they may be assigned two different desks. Group booking, which is crucial for team collaboration, is allowed for both types of occupants. Desks assigned to a group of occupants are guaranteed be in the same workspace (zone).

5. Occupants will use the space that is assigned to them and will not occupy other spaces for a considerable amount of time.

6. Occupants disclose their true thermal preference, *i.e.*, indoor temperature that they are most comfortable with, when they purchase the plan via the workspace reservation system.

7. The total amount of heat emitted by occupants or introduced by computers and appliances that they use in the zone is almost the same for every occupant regardless of their demographic information.

8. A small amount of log data, *e.g.* 14 days of data, is available from the building. The log data collected from physical and virtual sensors installed in the building or the surrounding environment are aggregated and stored in the Building Management System (BMS). Weather data is also assumed to be available from the same geographical location in that time period.

7.2.2 Test building

To evaluate our space assignment and supervisory control algorithms, we conduct a simulation study on a 3-story, 15-zone medium office building – a commercial reference building model developed by the US Department of Energy [2]. This building has a gross area of 4,982.19m^2 and a total capacity of 267 people. Each floor, consisting of four perimeter zones (two large and two small) and a central core zone, has a dedicated AHU. The capacity of core zones, large perimeter zones, and small perimeter zones is 53, 11, and 7 people, respectively. All zones are equipped with a VAV system to facilitate zone-level control. Except the room temperature setpoints, all HVAC control points are controlled by the EnergyPlus default feedback controller, using the *predictive system energy balance* method [31]. Building operations are simulated using EnergyPlus 9.3 [31] and meteorological data from Denver, Colorado. We use COBS (described in Section 3.3) to interface with the simulation environment. We highlight the energy saving potential by performing the optimization in July; however, the underlying concept can be applied to any office building and in any season, as long as the same amount of recent log data is available from that building.

7.3 Methodology

We propose efficient algorithms that can be utilized in the reservation system of a coworking space to make decisions regarding which spaces/desks to assign

to occupants and the temperature setpoint of every individual thermal zone in the building during the time that HVAC is in the active mode of operation. Ideally, these decisions must be made in such a way that a good trade-off between HVAC energy consumption and occupant comfort is attained.

We cast the optimal HVAC operation and space allocation problem as a convex mixed-integer nonlinear program [11] – an optimization problem in which the objective function is convex and the feasible set is convex when integrality is relaxed. Despite the large number of decision variables, in practice, this combinatorial optimization problem, which is a subclass of mixed-integer convex programming, can be solved relatively quickly using the branch-and-bound method or its variants. To this end, the first step of our methodology involves training an accurate surrogate model to estimate the contribution of a given zone to HVAC energy consumption on a particular day, given the outdoor air temperature and operational log data collected in the previous planning period. If this model is a convex function of its inputs, it can be incorporated into the objective function of the combinatorial optimization problem that is solved at two different timescales to assign desks to long-term and short-term occupants in a shared workspace.

Assuming long-term occupants buy a monthly plan, the first optimization problem is solved once a month to assign dedicated desks to these occupants and determine the temperature setpoint of each occupied zone. Once the solution is found, we update the remaining capacity of each zone for this month and start processing next-day reservation requests of prospective short-term occupants. Specifically, dedicated desks in zones that have enough capacity will be assigned to short-term occupants, and temperature setpoints are adjusted to meet the thermal comfort requirements of all occupants. This will be done without relocating long-term occupants.

Figure 7.1 depicts the three main steps of our methodology. Note, operational log data collected in the current planning period can be used to fine-tune or retrain the surrogate model for the next period.

7.4 Estimating HVAC energy use

The most crucial step in optimizing space planning practices in office buildings is modeling the relationship between the total heating and cooling energy consumed by the HVAC system to condition each individual zone and the number of occupants assigned to the rooms that comprise this zone, other factors such as the weather condition, and the temperature setpoint of the respective zone and adjacent zones. This model can be incorporated into an optimization problem that is solved numerically. In the absence of a reliable model, the optimal space assignment must be found using a black-box optimization technique, such as Bayesian optimization [53], requiring the painstaking evaluation of a rather large number of assignments in the real building or through simulation on its digital twin. For example, there will be 5^{30} feasible space assignments in a small 5-zone office building housing 30 occupants with diverse thermal preferences. Considering the choices of zone temperature setpoints, black-box optimization would be prohibitively costly, even if the evaluation was done through simulation.

To approximate the energy used by the HVAC system to maintain the temperature of a given zone around its setpoint, we train a surrogate model on the available sensor data, taking into account the number of occupants assigned to this zone and its temperature setpoint, temperature setpoints of adjacent zones, and outside air temperature. By incorporating zone-level surrogate models rather than a surrogate model built for the whole building in the objective function of the optimization problem, we reduce the model complexity, thereby cutting down on the training cost. Since the contribution of each zone to HVAC energy consumption is not measurable, we use the total energy *delivered* by the HVAC system to each zone as a proxy for the total energy *consumed* by the HVAC system to condition that zone. This quantity, which is known as the *zone air system sensible heating and cooling energy*, can be calculated by multiplying the supply air mass flow rate by the difference between supply air temperature and zone air temperature at each time step, and summing up these quantities over one day. Since supply air mass flow rate,

supply air temperature, and zone air temperature are logged by the BMS at regular intervals, the sensible heating and cooling energy can be easily calculated. We note that although the sensible heating and cooling energy depends on the HVAC control strategy, it also varies with the occupant assignment and setpoint schedule under a fixed HVAC control strategy.

Assuming the number of people that occupy each zone, including long-term and short-term occupants, does not change drastically during work hours of each day and the zone temperature setpoint remains stable in that period of time, we train a surrogate model that predicts the total *daily* sensible heating and cooling energy demand. We argue that one-day resolution is the sweet spot because some of the assumptions listed in Section 7.2.1 will be violated if we consider a lower time resolution, and the estimation error will substantially increase if we consider a higher time resolution (*e.g.* a few hours). The surrogate model is trained utilizing two weeks of historical operational log data. For each day, the training data includes the zone-specific temperature setpoint during the active mode of operation of the HVAC system, the number of occupants in the zone, the average outdoor temperatures during the active and inactive modes of operation of the HVAC system, and the per zone sensible heating and cooling energy.

7.4.1 Building a surrogate model for zone sensible heating and cooling energy

As depicted in Figure 7.1, the surrogate model, which estimates the total energy consumed by the HVAC system on a single day to condition a given zone plays an instrumental role in determining the space assignment strategy that minimizes the building energy use while satisfying space capacity and thermal comfort constraints. Moreover, the convexity of this function is essential as it will be embedded in the objective function of the optimization problem [17]. With these in mind, we choose an Input Convex Neural Network (ICNN) [5] as our surrogate model.

To ensure, the ICNN's output y remains a convex function of its inputs x, we impose non-negativity constraints (described below) on the network param-

eters. A *passthrough* layer is also integrated into the neural network, offering a direct link from the input x to the output of every hidden layer. Adding this layer is necessary because the non-negativity constraint restricts the use of hidden units that mirror the identity mapping in neural networks [5]. To optimize efficiency, especially considering the scarcity of data in our experiment, our ICNN architecture encompasses a single hidden layer with 100 neurons:

$$y = f(x;\eta) = \phi\left(W_1^{(\mathrm{h})}\phi\left(W_0^{(\mathrm{h})}x + b_0\right) + W^{(\mathrm{pass})}x + b_1\right), \tag{7.1}$$

where ϕ denotes the nonlinear, non-decreasing, convex activation function (ReLU in this study), and $\eta = \{W_0^{(\mathrm{h})}, W_1^{(\mathrm{h})}, W^{(\mathrm{pass})}, b_0, b_1\}$ collects the model parameters. Note that all elements of $W_{0,1}^{(\mathrm{h})}$ must be non-negative for the convexity guarantee. The convexity of f in x is derived from two facts. First, when convex functions are combined through non-negative summation, the result remains convex. Second, the composition of a convex function with another convex non-decreasing function results in a convex funciton. The $W^{(\mathrm{pass})}$ parameter is incorporated to preserve the model's representational power. Our ICNN is trained with the Adam optimizer.

To evaluate the performance of the ICNN model, we introduce two additional models that can be trained on the same dataset. RF is a non-parametric model that can approximate arbitrary functions and select important features. But its nonconvexity complicates the subsequent optimization problems. In this chapter, we set the number of estimators in RF to 100. Another model is the Piecewise Linear convex regression (P-Linear), which reduces the complexity of solving the optimization process due to its convexity. The P-Linear model that we consider consists of two line segments, the first segment is for the case that the number of occupants is between zero and one, and the other segment is for the case that the number of occupants is greater than or equal to one. When a zone is not occupied, heat transfer mostly happens through walls, windows, and ducts; whereas when it becomes occupied, there is additional heat gain that can be attributed to the occupants, including human body heat dissipation and heat generated by appliances that are used by occupants. This is why we consider two line segments. A potential drawback of

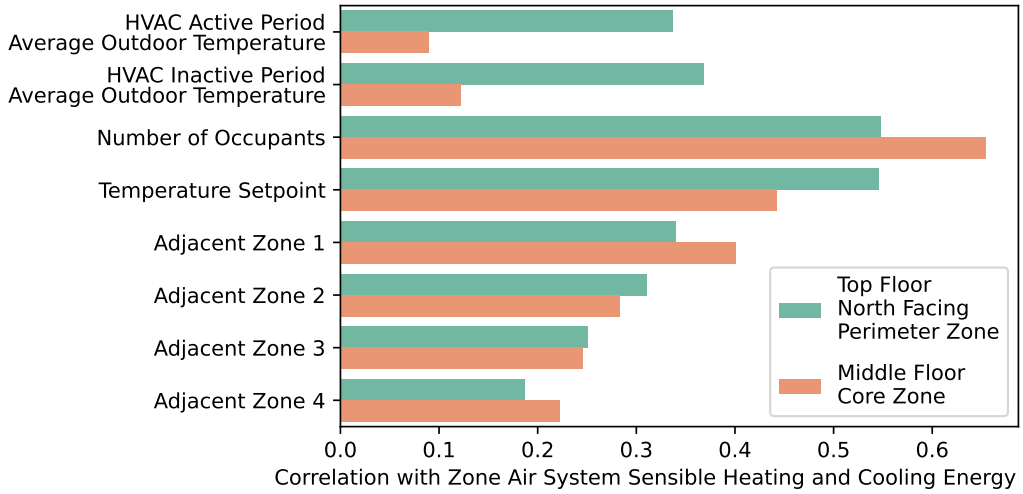


Figure 7.2: Correlation between daily zone air system sensible heating and cooling energy and input features for a perimeter zone and a core zone.

this model is its low accuracy as the latent relationship may not be accurately approximated by a piecewise linear function.

7.4.2 Model training and evaluation

We use 14 days of log data to train our surrogate model and baselines that estimate the total zone air system sensible heating and cooling energy use in one day. The Mean Squared Error (MSE) is used as the loss function. Each zone-level surrogate model is trained using four features, namely the number of occupants assigned to this zone, the zone temperature setpoint, and the average outdoor temperature during the HVAC system's active and inactive modes of operation. Including the average outdoor temperature over two non-overlapping intervals is inspired by the observation that the building exhibits distinct dynamics in these intervals. Specifically, zones are not typically conditioned by the HVAC system from 10pm to 6am of the next day as a result of the deadband widening strategy so the heat exchange is mostly with the outside environment through the building envelope. In addition to these four features, we also include the temperature setpoints of adjacent zones to explore if it improves the model accuracy. We append a '+' sign to the name of the surrogate models that get the additional features to differentiate them from the models that get the main four features. Either way, the model output is

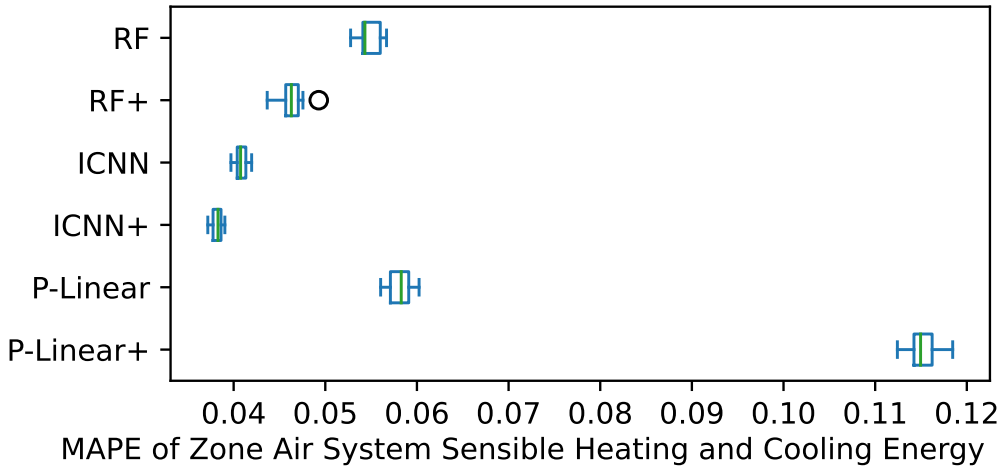


Figure 7.3: The MAPE of different surrogate models with different input features. The results are obtained from ten independent runs.

the total air system sensible heating and cooling energy of the corresponding zone on the given day. Figure 7.2 shows the correlation between the total zone air system sensible heating and cooling load throughout the day and various features for a perimeter zone and a core zone in the building. For clarity, we plot the absolute value of the correlation, since the sign does not matter here. The core zone on the middle floor is not directly connected to the outdoor environment. Thus, it has a weak correlation with the two average outdoor temperatures. The perimeter zone on the top floor, which directly interfaces with the outside environment, has a stronger correlation with the average outdoor temperatures. This plot suggests that all of these features are relevant for the regression task for at least one type of zones. Thus, none of them should be disregarded.

In Figure 7.3, different surrogate models are compared with respect to the Mean Absolute Percentage Error (MAPE). These models were trained using data collected between June 17 and June 30, and subsequently evaluated between July 1 and July 14, assuming that exactly 100 occupants were randomly assigned to building spaces and the assignment changes every day. The temperature setpoints are also randomly defined every day in a range of 20°C to 26°C. To mitigate the impact of randomness, experiments were carried out us-

ing 10 random seeds. Each data point represents the mean MAPE of a model across all zones in one run. We conclude that the ICNN surrogate model, when accounting for the temperature setpoints of adjacent zones, outperforms the other models with a clear margin. Thus, in the remainder of this chapter, we stick with ICNN+ as our surrogate model to estimate HVAC energy consumption per day.

7.5 Optimizing HVAC energy use

We now discuss how to assign desks to different occupants and adjust zone temperature setpoints to optimize the total daily energy consumption of the HVAC system in a coworking space, while satisfying zone capacity and thermal comfort constraints. We decompose the optimization problem into two problems that are solved at different timescales for long-term and short-term occupants. The first optimization problem is solved once a month (*e.g.* before the first day of the month) to assign dedicated desks to long-term occupants, thereby avoiding dissatisfaction caused by multiple relocations. Short-term occupants stay for one day only and submit their reservation requests one day in advance. Consequently, the second optimization problem is solved every time a new request arrives to determine whether short-term occupants can be admitted (given the remaining capacity of each zone and its setpoint temperature) and assign desks to them accordingly.

7.5.1 Modeling personal comfort

Given individual differences in thermal comfort perception, it is imperative to quantify an individual's thermal satisfaction with their environment given its temperature setpoint, and use this as a constraint in HVAC energy optimization. The personal comfort model can be (a) a sophisticated model trained for each individual using explicit feedback they provide about their thermal satisfaction, or implicit feedback they provide through interactions with a personal comfort system [173]; (b) a probabilistic thermal comfort profile for each individual given their preferred temperature and tolerance of thermal discomfort [148]. Since short-term occupants do not spend enough time in the building to provide sufficient feedback for model training or adaptation,

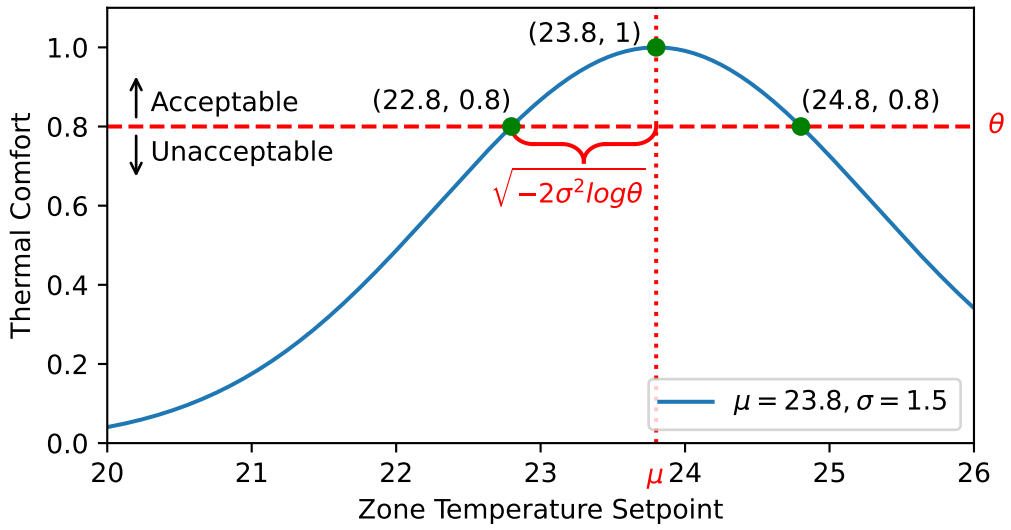


Figure 7.4: Estimated thermal comfort profile for an individual with parameters $\mu = 23.8$ and $\sigma = 1.5$. The dashed line represents the desired thermal comfort threshold of $\theta = 0.8$.

we take the second approach and develop a practical thermal comfort model with just a few parameters. Specifically, for every individual ρ, we postulate that there is an ideal temperature μ_ρ and an associated tolerance range σ_ρ. Following [148], the probability that this individual is comfortable in a zone that has a temperature T is expressed as:

$$p_\rho(T) = e^{-\frac{1}{2}\left(\frac{T-\mu_\rho}{\sigma_\rho}\right)^2}. \tag{7.2}$$

Hence, to keep $p_\rho(T)$ above some threshold θ, the zone temperature setpoint must lie in the following range:

$$\mu_\rho + \sqrt{-2\sigma_\rho^2 \log\theta} \geq T \geq \mu_\rho - \sqrt{-2\sigma_\rho^2 \log\theta}. \tag{7.3}$$

Figure 7.4 illustrates the thermal comfort profile of an individual with $\mu_\rho = 23.8$ and $\sigma_\rho = 1.5$. For $\theta = 0.8$, the acceptable range for the zone temperature setpoint is $[22.8, 24.8]$. If a group of occupants, each with distinct μ_ρ and σ_ρ values, are assigned to the same zone, the temperature setpoint of that zone must lie in the intersection of these ranges to satisfy thermal comfort requirements of all.

We assume that each individual reveals their preferred temperature and tolerance when they book the space. In our experiments, we set all σ_j values to 1.5 based on the observation that 1°C deviation from the preferred temperature is generally deemed acceptable, and sample μ_ρ uniformly from the range between 20°C and 26°C. We set θ to 0.8 to find a reasonable trade-off between HVAC energy consumption and thermal comfort; this aligns with the threshold specified for the Predicted Percentage of Dissatisfied (PPD) index in ASHRAE Standard 55 [3].

7.5.2 Space assignment to long-term occupants

We formulate a convex MINLP by incorporating the ICNN+ model into the objective function and individual thermal comfort profiles in the constraints. The solution to this problem is the assignment of dedicated desks in specific zones to (groups of) long-term occupants and the temperature setpoints of the occupied zones that minimize HVAC energy consumption without sacrificing thermal comfort of the actual zone occupants. This MINLP is defined as follows:

$$
\begin{aligned}
&\underset{\mathcal{X},\mathcal{T}}{\text{minimize}} \sum_{i\in\mathcal{N}} f_i\left(\cdot;\eta\right) \\
\text{s.t.}\quad &\textbf{(C1)} \quad \sum_{j\in\mathcal{M}} \mathcal{X}_{i,j} G_j \le C_i, && \forall i \in \mathcal{N}, \\
&\textbf{(C2)} \quad \sum_{i\in\mathcal{N}} \mathcal{X}_{i,j} = 1, && \forall j \in \mathcal{M}, \\
&\textbf{(C3)} \quad p_j \left(\sum_{i\in\mathcal{N}} \mathcal{X}_{i,j} \mathcal{T}_i \right) \ge \theta, && \forall j \in \mathcal{M}, \\
&\textbf{(C4)} \quad \mathcal{X}_{i,j} \in \{0,1\}, && \forall i \in \mathcal{N}, j \in \mathcal{M}, \\
&\textbf{(C5)} \quad \mathcal{T}_i \in \mathbb{R} \cap \left[t^{lb}, t^{ub}\right], && \forall i \in \mathcal{N}.
\end{aligned}
$$

Here $\mathcal{N} = \{1, \ldots, n\}$ denotes the set of zones in the coworking space, C_i represents the capacity of zone $i \in \mathcal{N}$, $\mathcal{M} = \{1, \ldots, m\}$ denotes the set of long-term occupant groups, where the occupants in each group must be placed in the same zone (*i.e.*, groups are indivisible), and G_j represents the size of group $j \in \mathcal{M}$. If an individual books a single desk, then the corresponding group size will be 1. The objective function is the total energy consumed by

the HVAC system to condition all zones in the building. Thus, it is the sum of the outputs of pretrained surrogate models $f_i(\cdot;\eta)$ for all $i \in \mathcal{N}$.[1] The input features of the surrogate model, which we previously introduced in Section 7.4, are omitted for brevity. Constraint **(C1)** ensures that the total number of long-term occupants assigned to a zone does not exceed its capacity. Constraint **(C2)** ensures that each group is assigned to exactly one zone. The zone temperature setpoint must be chosen such that the thermal comfort of every group member is satisfied with a probability greater than θ if the group is assigned to that zone. So with a slight abuse of notation, Constraint **(C3)** is written for a group of occupants. This constraint is necessary to balance energy savings and thermal comfort. Constraint **(C4)** forces $\mathcal{X}$ to be an $n \times m$ binary matrix indicating the space assigned to each group of long-term occupants. Constraint **(C5)** ensures that the temperature setpoint of each zone, which is a continuous decision variable, remains within reasonable bounds.

We solve the convex MINLP problem using branch-and-bound, which involves conducting a state space search to find the solution. Specifically, it maintains an upper bound, which is the minimum feasible solution, and a lower bound, which is the solution found through relaxation of integrality constraints (C4). The two bounds are refined via a tree search, where each node represents a convex MINLP problem and is branched into two nodes by constraining a chosen integer variable based on its value in the solution of the relaxed problem. The bounds are used to prune the search tree, thereby reducing the number of relaxed problems (NLPs) that must be solved. We note that each relaxed problem can be solved efficiently because of having the sum of convex functions as the objective function. We elaborate on this in Section 7.6.2.

Once the maximum number of iterations is reached or the gap between the two bounds gets smaller than a predefined threshold, the solver returns the smallest feasible solution that has been found. Although it is conceiv-

[1]We do not sum daily energy consumption over one month because due to the long optimization horizon, the same average outdoor temperature forecast obtained from a local weather station would have to be passed to ICNN+ for every day of the month. Thus, the monthly optimization problem would be equivalent to the presented problem.

able that the global optimum may not be found within the maximum search iterations, branch-and-bound typically solves the convex MINLP problem to (near-)optimality when the solver is warm-started [11]. In our experiments, we initialize the upper bound to a feasible solution found by the BestFit-Energy algorithm, which we detail in Section 7.5.3.

7.5.3 Space assignment to short-term occupants

After assigning dedicated desks to long-term occupants and updating the available capacity of every zone, we concentrate on solving the space assignment problem for short-term occupants. Recall that short-term occupants can submit a space reservation request, individually or on behalf of a group, at any time on the day before their intended visit. These requests must be processed in quasi real-time, meaning that admission decisions must be taken immediately[2] and the zone assignment must be done too if they are admitted. As a result, we process these reservation requests sequentially, as soon as they are submitted. Below we present two heuristic algorithms, namely BestFit-Energy and BestFit-Space, that are suitable for real-time decision making along with the Online-MINLP algorithm, which solves a convex MINLP to optimize HVAC energy consumption, this time for a group of short-term occupants.

BestFit-Energy algorithm: This heuristic algorithm assigns desks to short-term occupants trying to minimize the rise in HVAC energy consumption. Specifically, given a group of short-term occupants of size G_k, it assigns desks from a zone i that has enough available capacity, its current temperature setpoint satisfies the comfort requirements of the new occupants, and its f_i would increase by the smallest amount due to the increase in the number of occupants (by G_k). This algorithm prioritizes assigning desks in zones that are already occupied. If a group of occupants must be assigned to a zone i that is currently vacant, because no occupied zone has enough available capacity or the thermal comfort requirement of the group is satisfied in none of the

[2]This is essential for reservation requests that are declined as the user must be notified immediately to make a booking with another coworking space.

occupied zones, the algorithm sets the temperature setpoint of that zone to the value in $[t^{lb}, t^{ub}]$ that minimizes f_i, while satisfying the thermal comfort requirement of the new occupants.

In the case where all assignments that satisfy the capacity constraint violate the thermal comfort constraint, our algorithm considers zones that have enough capacity and attempts to update their temperature setpoints to satisfy the thermal comfort requirements of all occupants that are already assigned to that zone in addition to the new occupants. If there are more than one such zone, the setpoint adjustment will be done eventually for the zone that its f_i will increase by the smallest amount. If no zone temperature setpoint can be adjusted to accommodate this space reservation request, the new group of occupants will not be admitted.

BestFit-Space algorithm: This heuristic algorithm diversifies the temperature setpoints across all zones and assigns desks to short-term occupants such that there is nearly the same number of desks that can be possibly assigned to an arbitrary short-term occupant with any preferred temperature. This helps reduce rejections due to the violation of the thermal comfort constraint, increasing space utilization. Let $\{\mathcal{T}^*\}$ be the set of diverse temperature setpoints that might be assigned to a vacant zone. Refer to Algorithm 4 in the appendix for details about how these diverse setpoints are selected. Initially, short-term occupants are assigned desks in an already occupied zone that has the highest *effective capacity* while satisfying their comfort requirements. The effective capacity, denoted C'_k for zone k, is calculated using the group mean tolerance $\bar{\sigma}$ as follows:

$$C'_k = \sum_{i \in \mathcal{N}} \left(C_i - \sum_{j \in \mathcal{M}} \mathcal{X}^*_{i,j} G_j \right) \max \left(0, 1 - \frac{|\mathcal{T}_k - \mathcal{T}_i|}{2\sqrt{-2\bar{\sigma}^2 \log \theta}} \right)$$

Note that $\mathcal{X}^*$ is a part of the solution to the MINLP solved for long-term occupants. The first term inside the summation is the remaining capacity of the zone and the second term is a scaling factor for each zone's capacity based on the overlap between its temperature setpoint and the temperature setpoint of other zones.

In the case where short-term occupants cannot be placed in any of the zones that are currently occupied, they are assigned desks in the unoccupied zone that will have the smallest increase in f_i with its temperature setpoint being selected from $\{\mathcal{T}^*\}$ to satisfy the comfort requirements of the group. Once a zone's occupancy reaches its capacity, the setpoints in $\{\mathcal{T}^*\}$ will be updated. Short-term occupants will not be admitted if their thermal comfort requirements cannot be satisfied in any zone.

Online-MINLP algorithm: This algorithm assigns desks in one zone to a group of short-term occupants with size G_k and adjusts the zone temperature setpoint by solving the MINLP problem. To this end, the algorithm freezes the elements of $\mathcal{X}$ and simply inserts one column $\mathcal{X}_{:,k}$ at the end ($k = |\mathcal{M}|+1$) that contains the new decision variables. This resuls in the following convex MINLP:

$$\begin{aligned}
\underset{\mathcal{X}_{:,k},\mathcal{T}}{\text{minimize}} \; & \sum_{i\in\mathcal{N}} f_i\left(\mathcal{T}_i, \sum_{j\in\mathcal{M}} \mathcal{X}_{i,j} G_j\right) \\
\text{s.t.} \quad & \textbf{(C1), (C3), (C5)}, \\
& \sum_{i\in\mathcal{N}} \mathcal{X}_{i,k} = 1, \\
& \mathcal{X}_{i,k} \in \{0,1\}, \qquad \forall i \in \mathcal{N}.
\end{aligned}$$

Discussion: Each of the three proposed algorithms has its own strengths and weaknesses. BestFit-Energy and BestFit-Space are sorting-based algorithms, so their running time complexity is lower than Online-MINLP which invokes a MINLP solver. Yet BestFit-Energy struggles to maintain a good thermal comfort level as it prioritizes energy saving over thermal comfort satisfaction. On the other hand, BestFit-Space tends to focus on satisfying the thermal comfort requirements of current and future short-term occupants at the cost of wasting energy. The Online-MINLP algorithm optimally updates the temperature setpoints for each zone, resulting in an energy-efficient solution. We evaluate these algorithms in terms of occupant thermal comfort, energy efficiency, and running time in the next section.

7.6 Experimental results

To thoroughly evaluate the efficacy of our proposed methodology, we carry out a series of experiments on the test building described in Section 7.2.2. The primary goal is to quantify the energy-saving potential of the proposed space assignment strategies compared to the baselines that might be used in practice. Additionally, variations in the daily number of short-term occupants were taken into account to assess its effect on reservation rejection rate, thermal comfort probabilities, and energy consumption under different short-term occupant assignment strategies.

Considering the capacity of the test building (267 people), we assume there is a total of 100 occupants with a monthly subscription which are split into several groups, and the remaining space (desks) can be assigned to short-term occupants, the total number of which may not exceed 250 people per day, again split into several groups. For both types of occupants, the size of each group is uniformly sampled between 1 and 4. We set the thermal comfort threshold θ to 0.8, and t^{lb} and t^{ub} to 20°C and 26°C, respectively. We run our simulation 10 times using different random seeds, each producing a unique pattern of space reservation requests, allowing us to draw a conclusion regardless of the order in which the reservation requests are submitted by prospective short-term occupants.

7.6.1 Space assignment baselines

Three baselines, namely Uniform-Number, Uniform-Ratio, and Random, are considered to better understand the performance of the proposed algorithms. These baselines process space reservation requests sequentially, assigning dedicated desks to either type of occupants. The Uniform-Number strategy tries to evenly distribute occupants across all zones, without exceeding the capacity of each zone. The Uniform-Ratio strategy puts occupants in zones proportional to their capacity, for example all zones will be 20% full. Notably, these two baselines do not factor in thermal comfort, and define the zone temperature setpoint to be the average of preferred temperatures of the occupants in the respective zone. The Random strategy, however, randomly assigns occupants

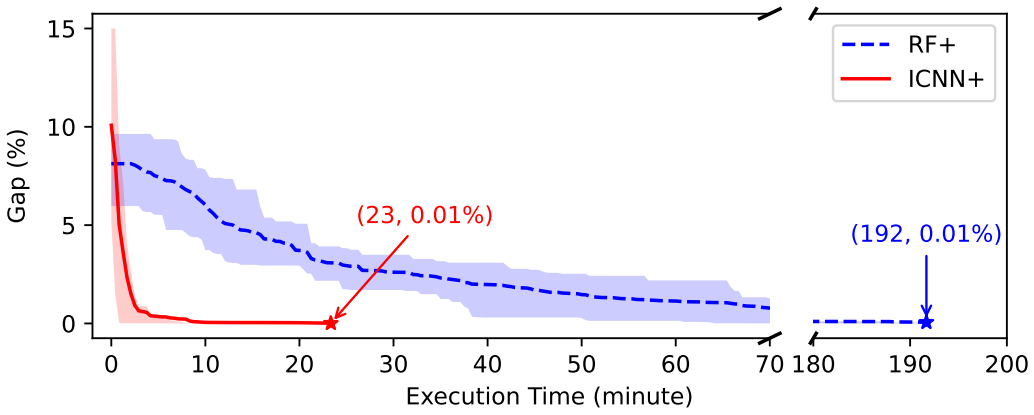
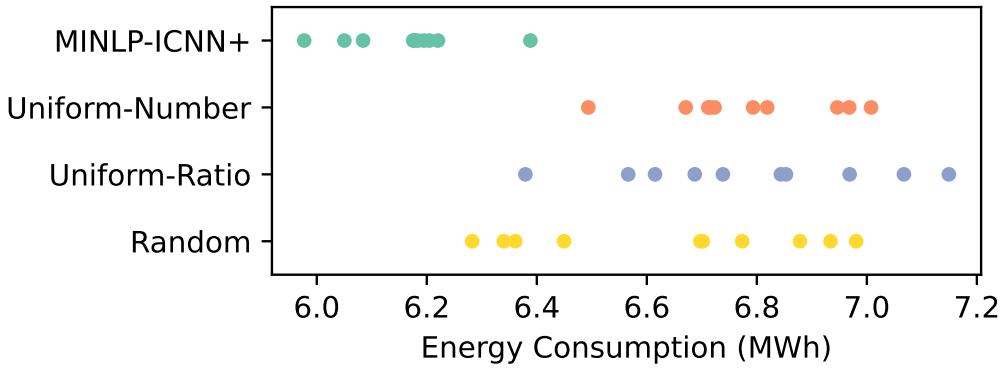


Figure 7.5: Comparing the execution time of the MINLP solver when different surrogate models are used as the objective function. We break the x-axis to show when the gap shrinks for both models. Each curve represents the average execution time, with the shaded region indicating the difference between the 75th and 25th percentiles of the execution time in 10 independent runs.

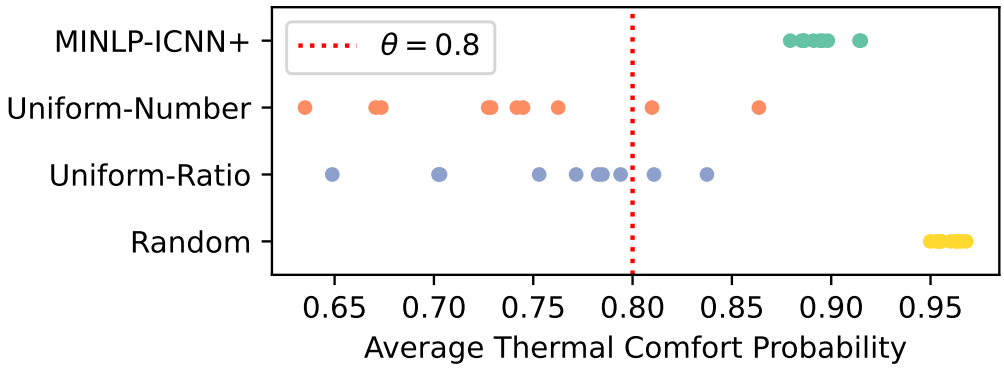
to one of the zones that satisfy their thermal preference without changing the zone temperature setpoint. So it always finds a feasible solution if it admits a group of occupants. When the first occupant (or group of occupants) is assigned to a zone, the zone temperature setpoint is set to their preferred temperature (the average of preferred temperatures, resp.).

7.6.2 Space assignment to long-term occupants

Figure 7.5 compares the time that it takes to solve the MINLP for long-term occupants to near-optimality, *i.e.*, shrinking the relative gap to 0.01%, using two different surrogate models, namely ICNN+ and RF+ described in Section 7.4. To solve MINLP, we use Gurobi installed on a server with an AMD EPYC 7313 16-core processor. The y-axis shows the relative gap which is defined as the difference between the upper and lower bounds divided by the value of the upper bound in each iteration. It can be readily seen that utilizing a convex surrogate model (ICNN+) reduces the execution time by about 88% over the non-convex model (RF+). We believe that the speedup is significant and without using a convex objective function the MINLP cannot be solved in reasonable time for a larger office building with more zones and a higher capacity.



(a) Total HVAC energy consumption



(b) Average thermal comfort

Figure 7.6: Performance of long-term occupant allocation strategies without short-term occupants. Note that the x-axis is exaggerated.

Figure 7.6a shows the cumulative HVAC energy consumption of the building during the simulation period (14 days) under different space allocation and setpoint adjustment strategies, in the absence of short-term occupants. The first row is our proposed algorithm that solves the convex MINLP for groups of long-term occupants, incorporating the ICNN+ surrogate model. The last three rows are our baselines. Each dot represents the result of an independent run.

Upon examining the baselines, it is evident that Random is the most effective one. This suggests that utilizing variable temperature setpoints across the zones can increase energy savings; whereas the setpoints in the other two baselines are usually around 23°C. When comparing MINLP-ICNN+ with the baselines, it becomes evident that the purposed algorithm outperforms the best baseline in terms of energy consumption by more than 6%. Specifically, the average total energy consumption stands at 6.16 MWh (std=0.11) for MINLP-ICNN+ and 6.58 MWh (std=0.34) for Random. These results underscore the efficacy of the proposed algorithm for long-term occupant assignment.

Figure 7.6b compares the average thermal comfort probability of these algorithms.[3] The Uniform-Number and Uniform-Ratio baselines failed to meet the thermal requirement, unlike the other two. The Random baseline achieves the highest thermal comfort probability at the expense of increased energy consumption. This is while MINLP-ICNN+ finds a better trade-off.

7.6.3 Space assignment to short-term occupants

We turn our attention to space assignment to short-term occupants. The experiments are conducted as follows. First, dedicated desks are assigned to 100 long-term occupants using MINLP-ICNN+ and considering 10 random seeds. Next, for each day, space reservation requests from short-term occupants are processed sequentially. Specifically, for every space assignment to long-term occupants, the remaining capacity of each zone is updated and desks are as-

[3]We plot the average thermal comfort probability rather than the distribution of individual thermal comfort because every experiment is repeated multiple times, making it difficult to plot all results. That said, we verified that the thermal comfort probability is indeed above 0.8 for every individual under MINLP-ICNN+ and Random.

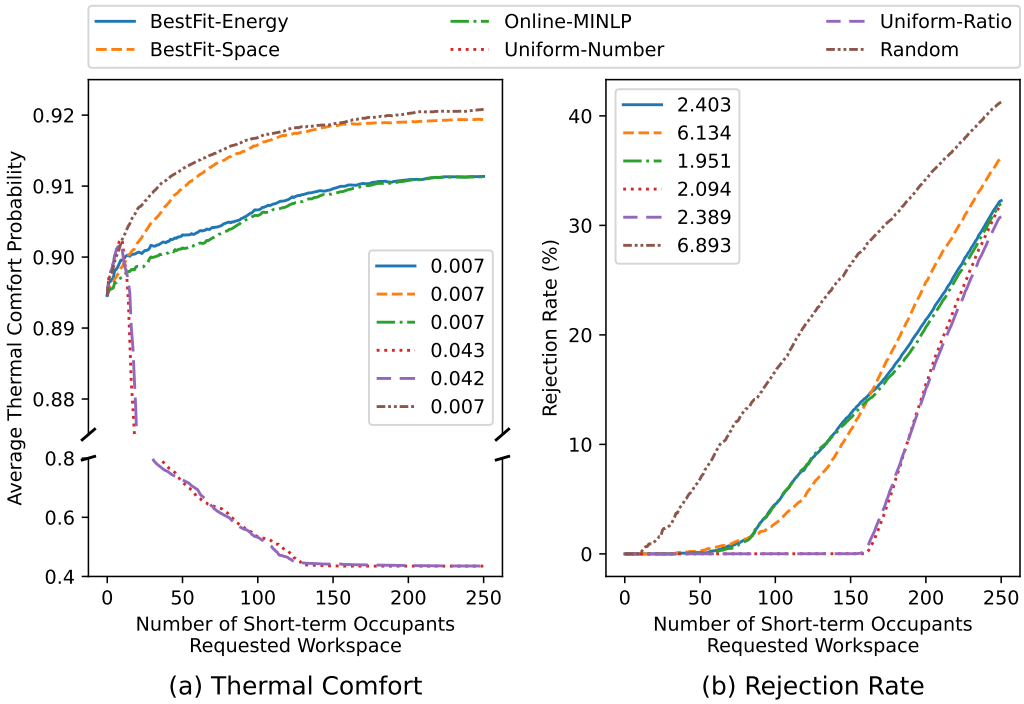


Figure 7.7: Comparison of space assignment algorithms for short-term occupants. For clarity, the standard deviation is not displayed around the mean. Instead it is noted in the legend of each subplot. Note that the y-axis of the left subplot is divided into two segments, each having a different scale.

signed to short-term occupants. We also consider 10 random seeds to generate reservation requests of short-term occupants. Hence, results are reported for 100 runs in total.

Thermal comfort: Figure 7.7a demonstrates how the average thermal comfort probability changes as the number of daily space reservation requests increases. When requests are processed using Uniform-Number and Uniform-Ratio, the average thermal comfort falls rapidly below the threshold (0.8). We attribute this to the fact that these baselines do not account for thermal comfort when assigning desks to (groups of) short-term occupants. However, the other algorithms keep the average thermal comfort probability above the threshold. One notable observation is that BestFit-Space and Random achieve better thermal comfort than BestFit-Energy and Online-MINLP. We attribute this to their intrinsic prioritization of thermal comfort over other objectives.

Rejection rate of reservations: Figure 7.7b shows the rejection rate (pct. of short-term occupant groups that were not admitted) as the number of daily space reservation requests increases. Uniform-Number and Uniform-Ratio reject less requests since they perform space assignment regardless of thermal comfort requirements. In contrast, the Random baseline consistently yields the highest rejection rate, followed by BestFit-Space after processing approximately 150 reservation requests. BestFit-Energy and Online-MINLP exhibit comparable performance, although the rejection rate is slightly lower for Online-MINLP. With fewer than 150 short-term occupants, BestFit-Space rejects fewer requests than BestFit-Energy and Online-MINLP, suggesting that diversifying zone temperature setpoints is a good strategy when space utilization is relatively low. This advantage diminishes as the building occupancy approaches its capacity, *i.e.*, admitting 167 short-term occupants.

Energy consumption: For this particular analysis, we assume exactly 100 reservation requests are received every day over the course of simulation, spanning 14 days. Figure 7.8 is a raincloud plot that shows HVAC energy consumption of different algorithms during the 14 days. It can be seen that BestFit-Energy and Online-MINLP perform better than the other algorithms. BestFit-Space takes the next position. An interesting observation can be made when comparing energy consumption under BestFit-Energy (6.52 MWh, std=0.16) and Online-MINLP (6.52 MWh, std=0.14) for the 200 occupants scenario (100 occupants of each type) demonstrated in Figure 7.8 to the best baseline for space assignment to 100 long-term occupants demonstrated in Figure 7.6a, which is Random (6.58 MWh, std=0.34). Concretely, we find that for the same amount of energy consumption, the coworking space can accommodate up to 100 more occupants using our algorithms.

We also examine a scenario in which the number of reservation requests is much higher than the building capacity so all zones will be full. Expectedly, HVAC energy consumption increases compared to the previous scenario due to the increase in building occupancy. We witness the smallest increase in HVAC energy consumption compared to when there are 100 short-term occupants

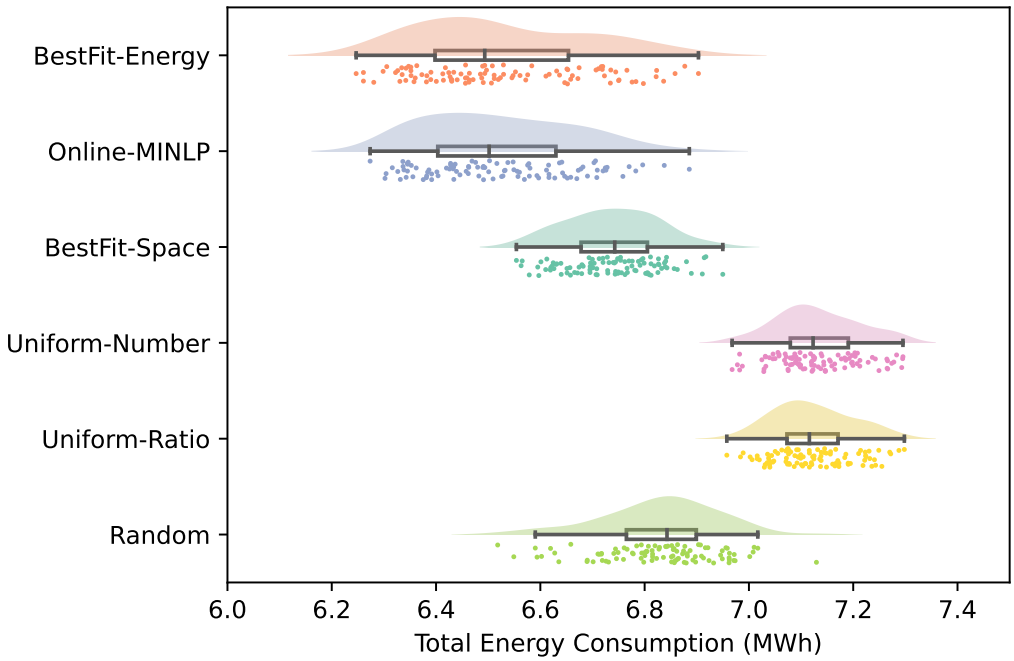


Figure 7.8: Total energy consumption of different algorithms that assign space to short-term occupants over a 14-day planning horizon, assuming 100 workspace reservation requests per day. The boxplot demonstrates the median and interquartile range of the data below the histogram of that data. The x-axis is exaggerated.

under BestFit-Space which is 0.05 MWh (0.76%), while the increase is 0.15 MWh (2.39%) for BestFit-Energy, 0.15 MWh (2.38%) for Online-MINLP, 0.11 MWh (1.64%) for Uniform-Number, 0.12 MWh (1.74%) for Uniform-Ratio, and 0.07 MWh (1.03%) for Random.

Profit analysis: To analyze the profitability of the coworking space, we must take into account the rejection rate and HVAC energy consumption at the same time as both of them will reduce the profit of the coworking space. To understand the confluence of these factors, we take Huddle, a Toronto-based coworking space company, as our case study. Huddle offers a day pass at CA$25.[4] We do our analysis assuming it operates an office building that is identical to our test building, has 100 occupants who purchased a monthly subscription, and receives 100 on-demand reservation requests per day. We

[4]The price is retrieved from `https://www.huddlesharespace.com/` in September 2023.

Table 7.1: Profit analysis for different space assignment algorithms.

Long-term Occupant Assignment	Short-term Occupant Assignment	HVAC Energy Use (kWh/day)	Admitted Short-term Occupants	Average Comfort (%)	On-demand Booking Profit (CA$/day)
(1)	(1)	671.32	100	8.81	2430.85
(2)	(2)	672.45	100	8.81	2430.73
(3)	(3)	619.1	82.44	94.90	1997.23
(4)	(5)	593.37	96.1	91.1	2341.39
	(6)	612.44	97.75	91.85	2380.67
	(7)	592.79	96.12	91.08	2341.95

Algorithms: (1) Uniform-Number (2) Uniform-Ratio (3) Random
(4) MINLP-ICNN+ (5) BestFit-Energy (6) BestFit-Space (7) Online-MINLP

calculate the energy cost using the tiered pricing scheme – one of the three pricing schemes that are in effect in Ontario, Canada [121]. As shown in Table 7.1, we compare 6 different algorithms for space assignment to both types of occupants in terms of the energy consumed by HVAC per day, the average number of admitted occupants that hold day passes, the average thermal comfort probability, and the average profit made per day. All algorithms manage to assign desks to all long-term occupants so we only report the average number of admitted short-term occupants and the profit made from day passes. The algorithms shown in the first two rows fail to meet thermal comfort requirements. Leaving them aside, we find that BestFit-Space consumes more energy than the other algorithms, but admits the highest number of short-term occupants, resulting the highest daily profit. Comparing the last three rows with the third row, where the Random baseline is used to assign desks to short-term occupants, we witness a CA$344.16–CA$383.44 increase in the profit made per day using the proposed algorithms.

Computation overhead: Finally, we analyze the running time of the proposed space assignment algorithms for short-term occupants. BestFit-Energy and BestFit-Space run fast, processing a request within milliseconds on our server. In contrast, the running time of Online-MINLP depends on the solver's termination condition and efficiency. In our setup, using a non-commercial

MINLP solver, namely SCIP, we observe that a (near-)optimal solution is typically found in about 5 minutes for 100 requests (roughly 3 seconds per request). However, when we take advantage of Gurobi with an academic license, this time is reduced to 30 seconds (roughly 0.3 seconds per request). Notice that Online-MINLP processes reservation requests of prospective short-term occupants sequentially, whereas MINLP-ICNN+ assigns desks to 100 long-term occupants at once by solving the optimization problem. This explains why it takes about 23 minutes to find a near-optimal solution in that case.

In summary, BestFit-Energy and Online-MINLP have comparable performance in terms of different metrics. But, unlike Online-MINLP, the heuristic algorithms scale well for large office buildings. If the space utilization is generally low, BestFit-Space is capable of admitting more occupants, thereby increasing the profit of the coworking space company. Otherwise, Online-MINLP and BestFit-Energy manage to admit a greater number of short-term occupants and reduce HVAC energy consumption. Nevertheless, all the proposed algorithms outperform the three baselines in terms of HVAC energy consumption and thermal comfort.

7.7 Conclusion

We proposed efficient algorithms for joint optimization of space use and zone temperature setpoints in office buildings, especially coworking spaces, to reduce heating and cooling demands, satisfy occupants' thermal comfort, and increase the building owner's profit. Our approach capitalized on three things: the diversity of individual thermal preferences, the unique characteristics of each thermal zone in the building, and differences in marginal energy consumption of the HVAC system in each zone due to placing more occupants there. We developed practical thermal comfort profiles for building occupants using information that they enter in the online system when they book the space, *i.e.*, their preferred indoor temperature and discomfort tolerance. We trained an input convex neural network to predict the daily HVAC energy consumption in a thermal zone and embedded it in the objective function of a convex MINLP. This neural network, which is trained on a dataset that con-

tains two weeks of historical log data, enables solving the optimization problem in a relatively short amount of time to assign desks to long-term occupants. We also proposed two efficient heuristic algorithms for processing on-demand reservation requests in a sequential fashion, and showed that they achieve a reasonable trade-off between energy consumption and thermal comfort without solving an optimization problem.

Chapter 8

Conclusion and future work

Intelligent control of building systems has the potential to enhance occupants' comfort while simultaneously decreasing the overall energy consumption of the building and the associated carbon emissions. In this thesis, model-free RL-based controllers were designed to collaboratively manage the operation of one or multiple building systems. The aim is to curtail total energy consumption while optimizing both thermal and visual comfort. We formulated a personal comfort model that measures individual thermal comfort based on the readings of the HVAC sensors and the sensors embedded in a PCS. Additionally, we introduced a space allocation system designed to assign desks to both long-term and short-term occupants. By grouping individuals with similar thermal preferences together, it enables choosing better temperature setpoints, leading to enhanced energy efficiency and thermal comfort of all occupants.

The contributions of this thesis are as follows:

- We presented the design and implementation of COBS and discussed how it can be used to benchmark control algorithms across a range of buildings, including the prototypical building models released by the US Department of Energy.

- We investigated the three-way trade-offs between energy consumption, visual comfort, and thermal comfort of occupants when using RL agents to jointly control the building HVAC, shading, and lighting system, and evaluated the efficacy of different model-free RL algorithms in comparison to a widely-used RBC and a model-based RL algorithm.

- We adopted the notion of policy and environment diversity to learn a library of control policies from training environments, and employed transfer learning to assign suitable policies from the policy library to each zone of the target building to improve its initial performance, i.e., before retraining on the target environment.

- We used individuals' interaction with a PCS to generate weak labels for training personal comfort models. This data, collected non-intrusively, resulted in a 97-fold increase in the amount of available labels for each individual.

- We proposed a method that groups individuals based on similar thermal preferences and applied two ensemble techniques to combine group thermal comfort models. The ensemble model aids in predicting the thermal preferences of any occupants, especially when limited labeled data is available. We demonstrated that an accurate personal comfort model can be developed using data collected from the PCS over a span of six hours.

- We utilized the zone sensible HVAC heating and cooling load as a proxy to train a model that estimates the HVAC system's energy consumption that is attributable to a given zone based on its temperature setpoint, the number of occupants in the zone, and other relevant features.

- We formulated the space planning problem for long-term occupants as an MINLP problem, and introduced two heuristic algorithms to processing on-demand reservation requests sequentially. We showed that these approaches can reduce building heating and cooling demands, satisfy occupants' thermal comfort, and increase the building owner's profit.

8.1 Future work

With the global interest in decarbonizing all sectors of the economy and the mandate of Canada (and many other countries) to adhere to the Paris climate agreement and achieve net-zero emissions by the year 2050, energy-efficient

operation of buildings is expected to gain more traction in the future. While this thesis makes key contributions that could facilitate the transition toward net-zero buildings, there remain several open problems that must be addressed in future work. Below we list some of the most promising avenues for future work, building on the foundations laid out in this thesis:

1. **Transfer learning for building control:** There exists an opportunity to devise better policy selection methods for identifying the most suitable policy for transfer to a novel building. Seasonal variations can be accommodated through online policy selection and adaptation techniques. Research into training policies for systems with dissimilar action scales is essential. Additionally, the cooperative dynamics between independent decision-making agents presents a promising area for exploration.

2. **Personal comfort modeling:** Determining the number of base comfort models for an ensemble remains an unresolved issue. Constructing the ensemble model demands more data, and the methodology for utilizing only zone-level sensor data to refine individual thermal models warrants further investigation.

3. **Space allocation:** There are three specific problems that warrant further investigation:

   (a) The algorithm for space assignment to short-term occupants warrants additional attention given its online nature and uncertainty about future reservation requests. A thorough analysis of the competitive ratios of the proposed online algorithms, as well as designing new ones, is essential.

   (b) While space assignments are currently done daily, there exists potential for refining the time scale to hourly for greater savings. In addition, examining the duration of occupant stays could also yield significant insights.

   (c) Certain occupants might be open to relocating within the building during their sojourn, should they be provided enough incentive.

The mechanisms for providing such incentives and sharing a portion of the profit made by the building owner, and their subsequent effects on productivity merit a comprehensive study.

In our future work, we also plan to explore the safety and robustness of the controllers developed for different building systems. TThe next step would be to integrate all the solutions we have proposed and deploy them onto actual building environments. By transitioning from simulation results to real-world applications, we seek to empirically validate our findings, ensuring that our results are not only theoretically sound but also practically effective in enhancing building efficiency and occupant comfort. This approach would provide invaluable insights and potentially pave the way for a broader adoption of the proposed strategies across the building stock.

References

[1] M. S. Abdelfattah, A. Mehrotra, Ł. Dudziak, and N. D. Lane, "Zero-cost proxies for lightweight NAS," in *Proceedings of the 9th International Conference on Learning Representations*, ser. ICLR, Virtual Event, Austria, 2021. 37

[2] American Society of Heating, Refrigerating and Air-Conditioning Engineers, *Standard 90.1-2019, Energy Standard for Buildings Except Low-Rise Residential Buildings*. Peachtree Corners, GA, USA: ASHRAE, Inc., 2019. 78, 131

[3] American Society of Heating, Refrigerating and Air-Conditioning Engineers, *Standard 55-2022, Thermal environmental conditions for human occupancy*. Peachtree Corners, GA, USA: ASHRAE, Inc., 2022. 140

[4] American Society of Heating, Refrigerating and Air-Conditioning Engineers and American National Standards Institute, *Thermal environmental conditions for human occupancy*. Peachtree Corners, GA, USA: ASHRAE, Inc., 2004, vol. 55. 12, 52, 59, 61, 111

[5] B. Amos, L. Xu, and J. Z. Kolter, "Input convex neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. ICML, Sydney, NSW, Australia: PMLR, 2017, pp. 146–155. 134, 135

[6] R. V. Andersen, J. Toftum, K. K. Andersen, and B. W. Olesen, "Survey of occupant behaviour and control of indoor environment in danish dwellings," *Energy and Buildings*, vol. 41, no. 1, pp. 11–16, 2009. 41

[7] O. Ardakanian, A. Bhattacharya, and D. Culler, "Non-intrusive occupancy monitoring for energy conservation in commercial buildings," *Energy and Buildings*, vol. 179, pp. 311–323, 2018. 20, 22

[8] A. Aryal and B. Becerik-Gerber, "Skin temperature extraction using facial landmark detection and thermal imaging for comfort assessment," in *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, ser. BuildSys, New York, NY, USA: ACM, 2019, pp. 71–80. 14

[9] A. Aryal, B. Becerik-Gerber, G. M. Lucas, and S. C. Roll, "Intelligent agents to improve thermal satisfaction by controlling personal comfort systems under different levels of automation," *IEEE Internet of Things Journal*, vol. 8, no. 8, pp. 7089–7100, 2020. 13

[10] A. Aswani, H. Gonzalez, S. S. Sastry, and C. Tomlin, "Provably safe and robust learning-based model predictive control," *Automatica*, vol. 49, no. 5, pp. 1216–1226, 2013. 23, 67

[11] P. Belotti, C. Kirches, S. Leyffer, J. Linderoth, J. Luedtke, and A. Mahajan, "Mixed-integer nonlinear optimization," *Acta Numerica*, vol. 22, pp. 1–131, 2013. 128, 132, 142

[12] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995. 100

[13] K. Berelson, F. Simini, T. Tryfonas, and P. Cooper, "Sensor-based smart hot-desking for improvement of office well-being," in *Proceedings of the 1st International Conference on Digital Tools & Uses Congress*, ser. DTUC, Paris, France: ACM, 2018, pp. 1–9. 19

[14] G. Brockman, V. Cheung, L. Pettersson, *et al.*, *Openai gym*, 2016. eprint: `1606.01540`. 40

[15] Building Technologies Division, "Energy efficiency in building automation and control," Siemens Switzerland Ltd, Tech. Rep., 2011. 1

[16] C. Buratti, E. Lascaro, D. Palladino, and M. Vergoni, "Building behavior simulation by means of artificial neural network in summer conditions," *Sustainability*, vol. 6, no. 8, pp. 5339–5353, 2014. 11

[17] S. Burer and A. N. Letchford, "Non-convex mixed-integer nonlinear programming: A survey," *Surveys in Operations Research and Management Science*, vol. 17, no. 2, pp. 97–106, 2012. 134

[18] L. Buşoniu, R. Babuška, and B. D. Schutter, "Multi-agent reinforcement learning: An overview," *Innovations in Multi-Agent Systems and Application*, vol. 1, pp. 183–221, 2010. 71

[19] J. Cai and J. E. Braun, "A generalized control heuristic and simplified model predictive control strategy for direct-expansion air-conditioning systems," *Science and Technology for the Built Environment*, vol. 21, no. 6, pp. 773–788, 2015. 26

[20] C. Candido, L. Thomas, S. Haddad, F. Zhang, M. Mackey, and W. Ye, "Designing activity-based workspaces: Satisfaction, productivity and physical activity," *Building Research & Information*, vol. 47, no. 3, pp. 275–289, 2019. 19

[21] T. Chaudhuri, Y. C. Soh, H. Li, and L. Xie, "Machine learning driven personal comfort prediction by wearable sensing of pulse rate and skin temperature," *Building and Environment*, vol. 170, p. 106 615, 2020. 14

[22] N. Chawla *et al.*, "SMOTE: Synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002. 99

[23] B. Chen, Z. Cai, and M. Bergés, "Gnu-RL: A precocial reinforcement learning solution for building HVAC control using a differentiable MPC policy," in *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, ser. BuildSys, New York, NY, USA: ACM, 2019, pp. 316–325. 21, 23, 24, 30, 38, 44, 47

[24] Y. Chen, Y. Shi, and B. Zhang, "Optimal control via neural networks: A convex approach," in *Proceedings of the 7th International Conference on Learning Representations*, ser. ICLR, New Orleans, LA, USA, 2019. 11, 22

[25] 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 and Buildings*, vol. 169, pp. 195–205, 2018. 21, 22, 25, 44

[26] Z. Chen, Y. Wang, and H. Liu, "Unobtrusive sensor-based occupancy facing direction detection and tracking using advanced machine learning algorithms," *IEEE Sensors Journal*, vol. 18, no. 15, pp. 6360–6368, 2018. 18

[27] Z. Cheng, Q. Zhao, F. Wang, Y. Jiang, L. Xia, and J. Ding, "Satisfaction based q-learning for integrated lighting and blind control," *Energy and Buildings*, vol. 127, pp. 43–55, 2016. 21, 22, 25, 31, 44

[28] T. Cheung, S. Schiavon, T. Parkinson, P. Li, and G. Brager, "Analysis of the accuracy on PMV–PPD model using the ASHRAE global thermal comfort database II," *Building and Environment*, vol. 153, pp. 205–217, 2019. 14, 112

[29] T. C. Cheung, S. Schiavon, E. T. Gall, M. Jin, and W. W. Nazaroff, "Longitudinal assessment of thermal and perceived air quality acceptability in relation to temperature, humidity, and CO2 exposure in singapore," *Building and Environment*, vol. 115, pp. 80–90, 2017. 13

[30] J.-H. Choi and D. Yeom, "Study of data-driven thermal sensation prediction model as a function of local body skin temperatures in a built environment," *Building and Environment*, vol. 121, pp. 130–147, 2017. 124

[31] D. B. Crawley, L. K. Lawrie, F. C. Winkelmann, *et al.*, "EnergyPlus: Creating a new-generation building energy simulation program," *Energy and Buildings*, vol. 33, no. 4, pp. 319–331, 2001. 11, 38, 46, 47, 69, 70, 72

[32] C. Dai, H. Zhang, E. Arens, and Z. Lian, "Machine learning approaches to predict thermal demands using skin temperatures: Steady-state conditions," *Building and Environment*, vol. 114, pp. 1–10, 2017. 16

[33] K. Dalamagkidis, D. Kolokotsa, K. Kalaitzakis, and G. S. Stavrakakis, "Reinforcement learning for energy conservation and comfort in buildings," *Building and Environment*, vol. 42, no. 7, pp. 2686–2698, 2007. 25, 40

[34] D. Daum, F. Haldi, and N. Morel, "A personalized measure of thermal comfort for building controls," *Building and Environment*, vol. 46, no. 1, pp. 3–11, 2011. 13, 20

[35] P. Davidsson and M. Boman, "Distributed monitoring and control of office buildings by embedded agents," *Information Sciences*, vol. 171, no. 4, pp. 293–307, 2005. 25

[36] R. J. de Dear, G. S. Brager, and D. Cooper, "Developing an adaptive model of thermal comfort and preference," *ASHRAE Transactions*, vol. 104, no. 1, pp. 145–167, 1998. 13

[37] B. Delcroix, J. L. Ny, M. Bernier, M. Azam, B. Qu, and J.-S. Venne, "Autoregressive neural networks with exogenous variables for indoor temperature prediction in buildings," *Building Simulation*, vol. 14, pp. 165–178, 2021. 10

[38] M. Deng, B. Fu, C. C. Menassa, and V. R. Kamat, "Learning-based personal models for joint optimization of thermal comfort and energy consumption in flexible workplaces," *Energy and Buildings*, vol. 298, p. 113 438, 2023. 19

[39] J. Diamond, *Collapse: how societies choose to fail or succeed: revised edition*. Penguin, 2011. 6

[40] X. Ding, W. Du, and A. E. Cerpa, "OCTOPUS: Deep reinforcement learning for holistic smart building control," in *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, ser. BuildSys, New York, NY, USA: ACM, 2019, pp. 326–335. 21, 22, 25, 30, 44, 51, 53

[41] X. Ding, W. Du, and A. E. Cerpa, "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*, ser. BuildSys, Virtual Event, Japan: ACM, 2020, pp. 50–59. 21

[42] J. Drgoňa, J. Arroyo, I. C. Figueroa, *et al.*, "All you need to know about model predictive control for buildings," *Annual Reviews in Control*, vol. 50, pp. 190–232, 2020. 67

[43] D. Enescu, "A review of thermal comfort models and indicators for indoor environments," *Renewable and Sustainable Energy Reviews*, vol. 79, pp. 1353–1379, 2017. 11

[44] G. Escrivá-Escrivá, C. Álvarez-Bel, and E. Peñalvo-López, "New indices to assess building energy efficiency at the use stage," *Energy and Buildings*, vol. 43, no. 2-3, pp. 476–484, 2011. 2

[45] B. Eysenbach, A. Gupta, J. Ibarz, and S. Levine, "Diversity is all you need: Learning skills without a reward function," in *Proceedings of the 7th International Conference on Learning Representations*, ser. ICLR, New Orleans, LA, USA, 2019. 26, 27

[46] P. O. Fanger, *Thermal comfort: analysis and applications in environmental engineering*, P. Fanger, Ed. Copenhagen, Denmark: Danish Technical Press, 1970. 12, 51, 126

[47] M. Fasiuddin and I. Budaiwi, "HVAC system strategies for energy conservation in commercial buildings in saudi arabia," *Energy and Buildings*, vol. 43, no. 12, pp. 3457–3466, 2011. 18

[48] P. Ferreira and A. Ruano, "Choice of RBF model structure for predicting greenhouse inside air temperature," *IFAC Proceedings Volumes*, vol. 35, no. 1, pp. 91–96, 2002. 11

[49] P. Ferreira, A. Ruano, S. Silva, and E. Conceicao, "Neural networks based predictive control for thermal comfort and energy savings in public buildings," *Energy and buildings*, vol. 55, pp. 238–251, 2012. 21

[50] W. Fisk and A. Rosenfeld, "Estimates of improved productivity and health from better indoor environments," *Indoor Air*, vol. 7, no. 3, pp. 158–172, 1997. 93

[51] M. Fontaine and S. Nikolaidis, "A quality diversity approach to automatically generating human-robot interaction scenarios in shared autonomy," in *Proceedings of Robotics: Science and Systems XVII*, ser. RSS, Virtual Event, 2021. 27

[52] J. Francis, M. Quintana, N. Von Frankenberg, S. Munir, and M. Bergés, "OccuTherm: Occupant thermal comfort inference using body shape information," in *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, ser. BuildSys, New York, NY, USA: ACM, 2019, pp. 81–90. 13

[53] P. I. Frazier, "A tutorial on bayesian optimization," *arXiv preprint arXiv:1807.02811*, 2018. 133

[54] J. Fu, M. Norouzi, O. Nachum, *et al.*, "Benchmarks for deep off-policy evaluation," in *Proceedings of the 9th International Conference on Learning Representations*, ser. ICLR, Virtual Event, Austria, 2021. 37

[55] Q. Fu, X. Chen, S. Ma, N. Fang, B. Xing, and J. Chen, "Optimal control method of HVAC based on multi-agent deep reinforcement learning," *Energy and Buildings*, vol. 270, p. 112 284, 2022. 25

[56] C. Gao, P. Li, Y. Zhang, J. Liu, and L. Wang, "People counting based on head detection combining adaboost and CNN in crowded surveillance environment," *Neurocomputing*, vol. 208, pp. 108–116, 2016. 18

[57] G. Gao, J. Li, and Y. Wen, "DeepComfort: Energy-efficient thermal comfort control in buildings via reinforcement learning," *IEEE Internet of Things Journal*, vol. 7, no. 9, pp. 8472–8484, 2020. 21

[58] S. Gao, M. Sui, C. Zhang, M. Wang, and Q. Yan, "Thermal model identification of commercial building based on genetic algorithm," in *2019 Chinese Automation Congress*, ser. CAC, Hangzhou, China: IEEE, 2019, pp. 550–554. 11

[59] X. Gao and S. Keshav, "SPOT: A smart personalized office thermal control system," in *Proceedings of the 4th International Conference on Future Energy Systems*, ser. e-Energy, Berkeley, CA, USA: ACM, 2013, pp. 237–246. 13, 93

[60] A. Ghahramani, G. Castro, B. Becerik-Gerber, and X. Yu, "Infrared thermography of human face for monitoring thermoregulation performance and estimating personal thermal comfort," *Building and Environment*, vol. 109, pp. 1–11, 2016. 124

[61] A. Ghahramani, C. Tang, and B. Becerik-Gerber, "An online learning approach for quantifying personalized thermal comfort via adaptive stochastic modeling," *Building and Environment*, vol. 92, pp. 86–96, 2015. 13

[62] S. Goel, M. Rosenberg, R. Athalye, *et al.*, "Enhancements to ASHRAE standard 90.1 prototype building models," Pacific Northwest National Lab. (PNNL), Richland, WA (United States), Tech. Rep., 2014. 82

[63] S. Goyal, C. Liao, and P. Barooah, "Identification of multi-zone building thermal interaction model from data," in *Proceedings of the 50th IEEE Conference on Decision and Control and European Control Conference*, ser. CDC-ECC, Orlando, FL, USA: IEEE, 2011, pp. 181–186. 20

[64] A. K. GS, T. Zhang, O. Ardakanian, and M. E. Taylor, "Mitigating an adoption barrier of reinforcement learning-based control strategies in buildings," *Energy and Buildings*, vol. 285, p. 112 878, 2023. 37, 87, 127

[65] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proceedings of the 35th International Conference on Machine Learning*, ser. ICML, PMLR, vol. 80, Stockholm, Sweden, 2018, pp. 1861–1870. 35

[66] X. Hou, Y. Xiao, J. Cai, J. Hu, and J. E. Braun, "Distributed model predictive control via proximal jacobian ADMM for building control applications," in *Proceedings of the 2017 American Control Conference*, ser. ACC, Seattle, WA, USA: IEEE, 2017, pp. 37–43. 20

[67] W. Hu, Y. Luo, Z. Lu, and Y. Wen, "Heterogeneous transfer learning for thermal comfort modeling," in *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, ser. BuildSys, New York, NY, USA: ACM, 2019, pp. 61–70. 17, 93, 98

[68] H. Huang, L. Chen, and E. Hu, "A neural network-based multi-zone modelling approach for predictive control system design in commercial buildings," *Energy and Buildings*, vol. 97, pp. 86–97, 2015. 11, 21

[69] Illuminating Engineering Society of North America, "Lighting handbook," in Illuminating Engineering Society of North America, New York, USA, 2000. 52

[70] International Energy Agency, *Buildings: A source of enormous untapped efficiency potential*, `https://www.iea.org/topics/buildings`, Accessed: 2022-04-01, 2022. 1

[71] International Organization for Standardization, "Ergonomics of the thermal environment-analytical determination and interpretation of thermal comfort using calculation of the PMV and PPD indices and local thermal comfort criteria," in ISO, 2005. 12, 52, 111

[72] R. Jacobs *et al.*, "Adaptive mixtures of local experts," *Neural Computation*, vol. 3, no. 1, pp. 79–87, 1991. 95, 109

[73] M. Jaderberg, W. M. Czarnecki, I. Dunning, *et al.*, "Human-level performance in 3D multiplayer games with population-based reinforcement learning," *Science*, vol. 364, no. 6443, pp. 859–865, 2019. 26

[74] P. Jayathissa, M. Quintana, M. Abdelrahman, and C. Miller, "Humans-as-a-sensor for buildings–intensive longitudinal indoor comfort models," *Buildings*, vol. 10, no. 10, p. 174, 2020. 13–15, 17

[75] F. Jazizadeh, A. Ghahramani, B. Becerik-Gerber, T. Kichkaylo, and M. Orosz, "Human-building interaction framework for personalized thermal comfort-driven systems in office buildings," *Journal of Computing in Civil Engineering*, vol. 28, no. 1, pp. 2–16, 2014. 13

[76] Z. Jiang, M. J. Risbeck, V. Ramamurti, *et al.*, "Building HVAC control with reinforcement learning for reduction of energy cost and demand charge," *Energy and Buildings*, vol. 239, p. 110 833, 2021. 23

[77] N. Kallus and M. Uehara, "Intrinsically efficient, stable, and bounded off-policy evaluation for reinforcement learning," in *Advances in Neural Information Processing Systems*, ser. NeurIPS, vol. 32, Vancouver, BC, Canada: Curran Associates, Inc., 2019, pp. 3320–3329. 36

[78] N. Kallus and A. Zhou, "Policy evaluation and optimization with continuous treatments," in *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, ser. AISTATS, vol. 84, Lanzarote, Spain: PMLR, 2018, pp. 1243–1251. 36

[79] S. Katipamula and M. R. Brambley, "Methods for fault detection, diagnostics, and prognostics for building systems: A review, part II," *HVAC&R Research*, vol. 11, no. 2, pp. 169–187, 2005. 1

[80] F. Khayatian, Z. Nagy, and A. Bollinger, "Using generative adversarial networks to evaluate robustness of reinforcement learning agents against uncertainties," *Energy and Buildings*, vol. 251, p. 111 334, 2021. 24

[81] J. Kim, F. Bauman, P. Raftery, *et al.*, "Occupant comfort and behavior: High-resolution data from a 6-month field study of personal comfort systems with 37 real office workers," *Building and Environment*, vol. 148, pp. 348–360, 2019. 94, 95, 97, 123

[82] J. Kim, S. Schiavon, and G. Brager, "Personal comfort models: A new paradigm in thermal comfort for occupant-centric environmental control," *Building and Environment*, vol. 132, pp. 114–124, 2018. 93

[83] J. Kim, Y. Zhou, S. Schiavon, P. Raftery, and G. Brager, "Personal comfort models: Predicting individuals' thermal preference using occupant heating and cooling behavior and machine learning," *Building and Environment*, vol. 129, pp. 96–106, 2018. 6, 13, 15, 16, 93, 95, 97,

[84] S. Kim, J.-H. Lee, and J. W. Moon, "Performance evaluation of artificial neural network-based variable control logic for double skin enveloped buildings during the heating season," *Building and environment*, vol. 82, pp. 328–338, 2014. 11

[85] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *In proceedings of the 3rd International Conference on Learning Representations*, ser. ICLR, San Diego, CA, USA, 2015. 102

[86] L. Klein, J.-y. Kwak, G. Kavulya, *et al.*, "Coordinating occupant behavior for building energy and comfort management using multi-agent systems," *Automation in Construction*, vol. 22, pp. 525–536, 2012. 25

[87] N. E. Klepeis, W. C. Nelson, W. R. Ott, *et al.*, "The national human activity pattern survey (NHAPS): A resource for assessing exposure to environmental pollutants," *Journal of Exposure Science & Environmental Epidemiology*, vol. 11, no. 3, pp. 231–252, 2001. 1

[88] D. Kolokotsa, G. Stavrakakis, K. Kalaitzakis, and D. Agoris, "Genetic algorithms optimized fuzzy controller for the indoor environmental management in buildings implemented using PLC and local operating networks," *Engineering Applications of Artificial Intelligence*, vol. 15, no. 5, pp. 417–428, 2002. 25

[89] E. Laftchiev, D. Romeres, and D. Nikovski, "Personalizing individual comfort in the group setting," in *In proceedings of the 35th AAAI Conference on Artificial Intelligence*, ser. AAAI, vol. 35, Virtual Event: AAAI Press, 2021, pp. 15 339–15 346. 16, 93

[90] A. H.-y. Lam, Y. Yuan, and D. Wang, "An occupant-participatory approach for thermal comfort enhancement and energy conservation in buildings," in *Proceedings of the 5th international conference on Future energy systems*, ser. e-Energy, Cambridge, UK: ACM, 2014, pp. 133–143. 14

[91] M. Lanctot, V. Zambaldi, A. Gruslys, *et al.*, "A unified game-theoretic approach to multiagent reinforcement learning," in *Advances in Neural Information Processing Systems*, ser. NeurIPS, vol. 30, Long Beach, CA, USA: Curran Associates Inc., 2017, pp. 4190–4203. 26

[92] H. Le, C. Voloshin, and Y. Yue, "Batch policy learning under constraints," in *Proceedings of the 36th International Conference on Machine Learning*, ser. ICML, vol. 97, Long Beach, CA, USA: PMLR, 2019, pp. 3703–3712. 37

[93] A. Leaman and B. Bordass, "Productivity in buildings: The 'killer' variables," *Building Research & Information*, vol. 27, no. 1, pp. 4–19, 1999. 93

[94] N. Lee, T. Ajanthan, and P. H. S. Torr, "SNIP: Single-shot network pruning based on connection sensitivity," in *Proceedings of the 7th International Conference on Learning Representations*, ser. ICLR, New Orleans, LA, USA, 2019. 37

[95] S. Lee, I. Bilionis, P. Karava, and A. Tzempelikos, "A bayesian approach for probabilistic classification and inference of occupant thermal preferences in office buildings," *Building and Environment*, vol. 118, pp. 323–343, 2017. 15, 17

[96] J. Lehman and K. O. Stanley, "Evolving a diversity of virtual creatures through novelty search and local competition," in *Proceedings of the 13th Annual Genetic and Evolutionary Computation Conference*, ser. GECCO, Dublin, Ireland: ACM, 2011, pp. 211–218. 27

[97] C. Li, T. Wang, C. Wu, Q. Zhao, J. Yang, and C. Zhang, "Celebrating diversity in shared multi-agent reinforcement learning," in *Advances in Neural Information Processing Systems*, ser. NeurIPS, vol. 34, Virtual Event: Curran Associates Inc., 2021, pp. 3991–4002. 26, 27

[98] S. Liu, S. Schiavon, H. P. Das, M. Jin, and C. J. Spanos, "Personal thermal comfort models with wearable sensors," *Building and Environment*, vol. 162, p. 106 281, 2019. 93, 124

[99] W. Liu, Z. Lian, and B. Zhao, "A neural network evaluation model for individual thermal comfort," *Energy and Buildings*, vol. 39, no. 10, pp. 1115–1122, 2007. 13

[100] S. Lu, W. Wang, C. Lin, and E. C. Hameen, "Data-driven simulation of a thermal comfort-based temperature set-point control with ASHRAE RP884," *Building and Environment*, vol. 156, pp. 137–146, 2019. 22

[101] T. Lu and M. Viljanen, "Prediction of indoor temperature and relative humidity using neural network models: Model comparison," *Neural Computing and Applications*, vol. 18, no. 4, p. 345, 2009. 11

[102] M. Luo, J. Xie, Y. Yan, *et al.*, "Comparing machine learning algorithms in predicting thermal sensation using ASHRAE comfort database II," *Energy and Buildings*, vol. 210, p. 109 776, 2020. 11, 12

[103] L. A. Martins, V. Soebarto, and T. Williamson, "A systematic review of personal thermal comfort models," *Building and Environment*, vol. 207, p. 108 502, 2022. 13

[104] M. A. Masood and F. Doshi-Velez, "Diversity-inducing policy gradient: Using maximum mean discrepancy to find a set of diverse policies," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, ser. IJCAI, Macao, China, 2019, pp. 5923–5929. 26, 27, 73

[105] E. Mathews, C. Botha, D. Arndt, and A. Malan, "HVAC control strategies to enhance comfort and minimise energy usage," *Energy and buildings*, vol. 33, no. 8, pp. 853–863, 2001. 1

[106] K. R. McKee, J. Z. Leibo, C. Beattie, and R. Everett, "Quantifying the effects of environment and population diversity in multi-agent reinforcement learning," *Autonomous Agents and Multi-Agent Systems*, vol. 36, no. 1, pp. 1–16, 2022. 26, 67, 75

[107] A. Mirakhorli and B. Dong, "Occupancy behavior based model predictive control for building indoor climate: A critical review," *Energy and Buildings*, vol. 129, pp. 499–513, 2016. 10

[108] L. MMba, P. Meukam, and A. Kemajou, "Application of artificial neural network for predicting hourly indoor air temperature and relative humidity in modern building in humid region," *Energy and Buildings*, vol. 121, pp. 32–42, 2016. 11

[109] V. Mnih, K. Kavukcuoglu, D. Silver, *et al.*, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013. 34

[110] I. P. Mohottige and T. Moors, "Estimating room occupancy in a smart campus using WiFi soft sensors," in *Proceedings of the 43rd IEEE Conference on Local Computer Networks*, ser. LCN, IEEE, Chicago, IL, USA, 2018, pp. 191–199. 18

[111] J. W. Moon *et al.*, "Development of an artificial neural network model based thermal control logic for double skin envelopes in winter," *Building and Environment*, vol. 61, pp. 149–159, 2013. 11

[112] J. W. Moon and S. K. Jung, "Algorithm for optimal application of the setback moment in the heating season using an artificial neural network model," *Energy and Buildings*, vol. 127, pp. 859–869, 2016. 11

[113] J. W. Moon, S. K. Jung, and J.-J. Kim, "Application of ANN (artificial-neural-network) in residential thermal control," in *Proceedings of the 11th International International Building Performance Simulation Association Conference*, ser. IBPSA, Glasgow, Scotland: AIVC, 2009, pp. 27–30. 11

[114] J.-B. Mouret and J. Clune, "Illuminating search spaces by mapping elites," *arXiv preprint arXiv:1504.04909*, 2015. 27

[115] G. Mustafaraj, J. Chen, and G. Lowry, "Thermal behaviour prediction utilizing artificial neural networks for an open office," *Applied Mathematical Modelling*, vol. 34, no. 11, pp. 3216–3230, 2010. 11

[116] G. Mustafaraj, G. Lowry, and J. Chen, "Prediction of room temperature and relative humidity by autoregressive linear and nonlinear neural network models for an open office," *Energy and Buildings*, vol. 43, no. 6, pp. 1452–1460, 2011. 11

[117] S. Nagarathinam, V. Menon, A. Vasan, and A. Sivasubramaniam, "MARCO - multi-agent reinforcement learning based control of building HVAC systems," in *Proceedings of the 11th ACM International Conference on Future Energy Systems*, ser. e-Energy, Virtual Event, Australia: ACM, 2020, pp. 57–67. 25, 27, 44

[118] I. Namatēvs, "Deep reinforcement learning on HVAC control," *Information Technology and Management Science*, vol. 21, pp. 29–36, 2018. 21, 22

[119] A. Natarajan and E. Laftchiev, "A transfer active learning framework to predict thermal comfort," *International Journal of Prognostics and Health Management*, vol. 10, no. 3, 2019. 16

[120] Natural Resources Canada, "Energy use data handbook, 1990 to 2017," in Ottawa, ON, Canada: Natural Resources Canada, 2019. 29

[121] Ontario Energy Board, *Electricity rates*, `https://www.oeb.ca/consumer-information-and-protection/electricity-rates`, 2022. 152

[122] J. Y. Park, T. Dougherty, H. Fritz, and Z. Nagy, "LightLearn: An adaptive and occupant centered controller for lighting based on reinforcement learning," *Building and Environment*, vol. 147, pp. 397–414, 2019. 21

[123] J. Parker-Holder, A. Pacchiano, K. M. Choromanski, and S. J. Roberts, "Effective diversity in population based reinforcement learning," in *Advances in Neural Information Processing Systems*, ser. NeurIPS, vol. 33, Virtual Event: Curran Associates Inc., 2020, pp. 18 050–18 062. 75

[124] A. Paszke, S. Gross, F. Massa, *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, ser. NeurIPS, Vancouver, BC, Canada: Curran Associates, Inc., 2019, pp. 8024–8035. 80

[125] S. Patil, H. Tantau, and V. Salokhe, "Modelling of tropical greenhouse temperature by auto regressive and neural network models," *Biosystems Engineering*, vol. 99, no. 3, pp. 423–431, 2008. 11

[126] Y. Peng, A. Rysanek, Z. Nagy, and A. Schlüter, "Occupancy learning-based demand-driven cooling control for office spaces," *Building and Environment*, vol. 122, pp. 145–160, 2017. 19

[127] L. Pérez-Lombard, J. Ortiz, and C. Pout, "A review on buildings energy consumption information," *Energy and Buildings*, vol. 40, no. 3, pp. 394–398, 2008. 2

[128] G. Pinto, Z. Wang, A. Roy, T. Hong, and A. Capozzoli, "Transfer learning for smart buildings: A critical review of algorithms, applications, and future perspectives," *Advances in Applied Energy*, vol. 5, p. 100 084, 2022. 24

[129] A. Pollard and A. Stoecklein, "Occupant and building related determinants on the temperature patterns in new zealand residential buildings," in *IPENZ Conference 98: The sustainable city*, ser. IPENZ, vol. 2, Auckland, New Zealand: IPENZ, 1998, p. 62. 11

[130] D. Precup, R. S. Sutton, and S. P. Singh, "Eligibility traces for off-policy policy evaluation," in *Proceedings of the 17th International Conference on Machine Learning*, ser. ICML, Stanford, CA, USA, 2000, pp. 759–766. 36

[131] S. Privara, J. Širokỳ, L. Ferkl, and J. Cigler, "Model predictive control of a building heating system: The first experience," *Energy and Buildings*, vol. 43, no. 2, pp. 564–572, 2011. 20, 21, 23

[132] J. K. Pugh, L. B. Soros, and K. O. Stanley, "Quality diversity: A new frontier for evolutionary computation," *Frontiers in Robotics and AI*, vol. 3, p. 40, 2016. 27

[133] Z. Rahimpour, G. Verbič, and A. C. Chapman, "Actor-critic learning for optimal building energy management with phase change materials," *Electric Power Systems Research*, vol. 188, p. 106 543, 2020. 21

[134] R. Rana, B. Kusy, R. Jurdak, J. Wall, and W. Hu, "Feasibility analysis of using humidex as an indoor thermal comfort predictor," *Energy and Buildings*, vol. 64, pp. 17–25, 2013. 13

[135] Y. P. Raykov, E. Ozer, G. Dasika, A. Boukouvalas, and M. A. Little, "Predicting room occupancy with a single passive infrared (PIR) sensor through behavior extraction," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp, Heidelberg, Germany: ACM, 2016, pp. 1016–1027. 18

[136] A. E. Ruano, E. M. Crispim, E. Z. Conceiçao, and M. M. J. Lúcio, "Prediction of building's temperature using neural networks models," *Energy and Buildings*, vol. 38, no. 6, pp. 682–694, 2006. 11

[137] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017. 35

[138] E. Shen, J. Hu, and M. Patel, "Energy and visual comfort analysis of lighting and daylight control strategies," *Building and Environment*, vol. 78, pp. 155–170, 2014. 20–22, 47

[139] S. S. Shetty, H. D. Chinh, M. Gupta, and S. K. Panda, "User presence estimation in multi-occupancy rooms using plug-load meters and pir sensors," in *Proceedings of the 2017 IEEE Global Communications Conference*, ser. GLOBECOM, Singapore: IEEE, 2017, pp. 1–6. 18

[140] N. Somu, A. Sriram, A. Kowli, and K. Ramamritham, "A hybrid deep transfer learning strategy for thermal comfort prediction in buildings," *Building and Environment*, vol. 204, p. 108 133, 2021. 16, 17

[141] T. Sood, P. Janssen, and C. Miller, "Spacematch: Using environmental preferences to match occupants to suitable activity-based workspaces," *Frontiers in Built Environment*, vol. 6, p. 113, 2020. 19

[142] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. Cambridge, MA, USA: MIT press, 2018. 34

[143] A. Swaminathan and T. Joachims, "The self-normalized estimator for counterfactual learning," in *Advances in Neural Information Processing Systems*, ser. NeurIPS, Montreal, QC, Canada: Curran Associates, Inc., 2015, pp. 3231–3239. 36

[144] F. Tartarini and S. Schiavon, "pythermalcomfort: A python package for thermal comfort research," *SoftwareX*, vol. 12, p. 100 578, 2020. 111

[145] A. Tavakoli, F. Pardo, and P. Kormushev, "Action branching architectures for deep reinforcement learning," in *In proceedings of the 32th AAAI Conference on Artificial Intelligence*, ser. AAAI, vol. 32, New Orleans, Louisiana, USA: AAAI Press, 2018, pp. 4131–4138. 53

[146] M. E. Taylor and P. Stone, "Transfer learning for reinforcement learning domains: A survey," *Journal of Machine Learning Research*, vol. 10, no. 1, pp. 1633–1685, 2009. 27

[147] C. Teodosiu *et al.*, "Numerical prediction of indoor air humidity and its effect on indoor environment," *Building and Environment*, vol. 38, no. 5, pp. 655–664, 2003. 11

[148] F. Topak, G. S. Pavlak, M. K. Pekeriçli, J. Wang, and F. Jazizadeh, "Collective comfort optimization in multi-occupancy environments by leveraging personal comfort models and thermal distribution patterns," *Building and Environment*, vol. 239, p. 110 401, 2023. 138, 139

[149] V. Tsakanikas and T. Dagiuklas, "Video surveillance systems-current status and future trends," *Computers & Electrical Engineering*, vol. 70, pp. 736–753, 2018. 18

[150] C. Turley, M. Jacoby, G. Pavlak, and G. Henze, "Development and evaluation of occupancy-aware HVAC control for residential building energy efficiency and occupant comfort," *Energies*, vol. 13, no. 20, p. 5396, 2020. 18, 23, 27

[151] U.S. Energy Information Administration, "Commercial buildings energy consumption survey," in U.S. Government Printing Office, 2012, ch. Consumption & Expenditures. 18

[152] J. R. Vazquez-Canteli, G. Henze, and Z. Nagy, "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*, ser. BuildSys, Virtual Event, Japan: ACM, 2020, pp. 170–179. 25

[153] A. Vishwanath, Y.-H. Hong, and C. Blake, "Experimental evaluation of a data driven cooling optimization framework for HVAC control in commercial buildings," in *Proceedings of the 10th ACM International Conference on Future Energy Systems*, ser. e-Energy, Phoenix, AZ, USA: ACM, 2019, pp. 78–88. 21

[154] Z. Wang, R. de Dear, M. Luo, *et al.*, "Individual difference in thermal comfort: A literature review," *Building and Environment*, vol. 138, pp. 181–193, 2018. 100

[155] Z. Wang and T. Hong, "Reinforcement learning for building controls: The opportunities and challenges," *Applied Energy*, vol. 269, p. 115 036, 2020. 4, 20–23, 27, 67

[156] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas, "Dueling network architectures for deep reinforcement learning," in *Proceedings of the 33rd International Conference on Machine Learning*, ser. ICML, vol. 48, New York, NY, USA: PMLR, 2016, pp. 1995–2003. 53

[157] M. Wetter, "Co-simulation of building energy and control systems with the building controls virtual test bed," *Journal of Building Performance Simulation*, vol. 4, no. 3, pp. 185–203, 2011. 38

[158] D. A. Winkler, A. Yadav, C. Chitu, and A. E. Cerpa, "OFFICE: Optimization framework for improved comfort & efficiency," in *Proceedings of the 19th ACM/IEEE International Conference on Information Processing in Sensor Networks*, ser. IPSN, Sydney, NSW, Australia: ACM/IEEE, 2020, pp. 265–276. 20, 23

[159] D. Wölfle, A. Vishwanath, and H. Schmeck, "A guide for the design of benchmark environments for building energy optimization," in *Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, ser. BuildSys, Virtual Event, Japan: ACM, 2020, pp. 220–229. 46

[160] J. Xie, H. Li, C. Li, J. Zhang, and M. Luo, "Review on occupant-centric thermal comfort sensing, predicting, and controlling," *Energy and Buildings*, vol. 226, p. 110 392, 2020. 13–15

[161] S. Xu, Y. Wang, Y. Wang, Z. O'Neill, and Q. Zhu, "One for many: Transfer learning for building HVAC control," in *Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, ser. BuildSys, Virtual Event, Japan: ACM, 2020, pp. 230–239. 24, 27

[162] L. Yang, Z. Nagy, P. Goffin, and A. Schlueter, "Reinforcement learning for optimal control of low exergy buildings," *Applied Energy*, vol. 156, pp. 577–586, 2015. 5

[163] L. Yang, X. Sun, A. Zhu, and T. Chi, "A multiple ant colony optimization algorithm for indoor room optimal spatial allocation," *ISPRS International Journal of Geo-Information*, vol. 6, no. 6, p. 161, 2017. 19

[164] Y. Yang, J. Luo, Y. Wen, *et al.*, "Diverse auto-curriculum is critical for successful real-world multiagent learning systems," in *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems*, ser. AAMAS, Virtual Event, UK: ACM, May 2021. 27

[165] K. C. Yip, K. W. Huang, E. W. Ho, W. Chan, and I. L. Lee, "Optimized staff allocation for inpatient phlebotomy and electrocardiography services via mathematical modelling in an acute regional and teaching hospital," *Health Systems*, vol. 6, no. 2, pp. 102–111, 2017. 19

[166] C. Yu, A. Velu, E. Vinitsky, *et al.*, "The surprising effectiveness of PPO in cooperative, multi-agent games," in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, ser. NeurIPS, vol. 2, New Orleans, LA, USA: Curran Associates, Inc., 2022. 35

[167] E. Žáčeková and L. Ferkl, "Building modeling and control using multi-step ahead error minimization," in *Proceedings of the 20th Mediterranean Conference on Control & Automation*, ser. MED, Barcelona, Spain: IEEE, 2012, pp. 421–426. 1

[168] C. Zhang, S. R. Kuppannagari, R. Kannan, and V. K. Prasanna, "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*, ser. BuildSys, New York, NY, USA: ACM, 2019, pp. 287–296. 21

[169] T. Zhang and O. Ardakanian, "A domain adaptation technique for fine-grained occupancy estimation in commercial buildings," in *Proceedings of the International Conference on Internet of Things Design and Implementation*, ser. IoTDI, Montreal, QC, Canada: ACM, 2019, pp. 148–159. 27

[170] T. Zhang and O. Ardakanian, "COBS: Comprehensive building simulator," in *Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, ser. BuildSys, Virtual Event, Japan: ACM, 2020, pp. 314–315. 80

[171] T. Zhang and O. Ardakanian, "Investigating the impact of space allocation strategy on energy-comfort trade-off in office buildings," in *Companion Proceedings of the 14th ACM International Conference on Future Energy Systems*, ser. e-Energy, Orlando, FL, USA: ACM, 2023, pp. 145–149. 127, 129

[172] T. Zhang, G. Baasch, O. Ardakanian, and R. Evins, "On the joint control of multiple building systems with reinforcement learning," in *Proceedings of the 12th ACM International Conference on Future Energy Systems*, ser. e-Energy, Virtual Event, Italy: ACM, 2021, pp. 60–72. 19, 22, 23

[173] T. Zhang, J. Gu, O. Ardakanian, and J. Kim, "Addressing data inadequacy challenges in personal comfort models by combining pretrained comfort models," *Energy and Buildings*, vol. 264, p. 112 068, 2022. 20, 138

[174] P. Zhao, S. Suryanarayanan, and M. G. Simoes, "An energy management system for building structures using a multi-agent decision-making control methodology," *IEEE Transactions on Industry Applications*, vol. 49, no. 1, pp. 322–330, 2012. 25

[175] Z. Zheng, Q. Chen, C. Fan, *et al.*, "Data driven chiller sequencing for reducing HVAC electricity consumption in commercial buildings," in *Proceedings of the 9th International Conference on Future Energy Systems*, ser. e-Energy, Karlsruhe, Germany: ACM, 2018, pp. 236–248. 44

[176] D. P. Zhou, Q. Hu, and C. J. Tomlin, "Quantitative comparison of data-driven and physics-based models for commercial building HVAC systems," in *Proceedings of the 2017 American Control Conference*, ser. ACC, Seattle, WA, USA: IEEE, 2017, pp. 2900–2906. 11, 20

[177] M. Zuraimi, A. Pantazaras, K. Chaturvedi, J. Yang, K. Tham, and S. Lee, "Predicting occupancy counts using physical and statistical CO2-based modeling methodologies," *Building and Environment*, vol. 123, pp. 517–528, 2017. 18

Appendix A

Appendix

A.1 Performance of RL control agents

A.1.1 With the zone-level occupancy schedule

The data corresponding to Figure 4.4 can be found in Table A.1, which includes details when zone-level occupancy information is factored in. For a comparative analysis of the optimal trade-offs realized by each RL agent across various scenarios, refer to Tables A.2 that present data incorporating zone-level occupancy details.

A.1.2 With the building-level occupancy schedule

The data corresponding to Figure 4.4 can be found in Table A.1, which includes details when building-level occupancy information is factored in. For a comparative analysis of the optimal trade-offs realized by each RL agent across various scenarios, refer to Tables A.2 that present data incorporating zone-level occupancy details.

Table A.1: RL Agent performance results for different control scenarios using zone-level occupancy schedule.

		Blinds always open				**Single blind setpoint**				**Multiple blind setpoints**			
		No dimming		**With Dimming**		**No dimming**		**With Dimming**		**No dimming**		**With Dimming**	
		Jan.	**Jul.**	**Jan.**	**Jul.**	**Jan.**	**Jul.**	**Jan.**	**Jul.**	**Jan.**	**Jul.**	**Jan.**	**Jul.**
Energy (MWh)	Baseline	8.34	4.44	8.4	3.2	7.07	3.92	6.81	3.02	7.07	3.92	6.81	3.02
	BDQN	6.48	3.36	6.16	2.11	6.51	3.12	6.34	2.16	6.55	3.17	6.35	2.2
	SAC	***6.44***	3.58	***6.11***	2.27	***6.23***	3.21	***6.1***	2.36	***6.18***	3.33	***6.06***	2.43
	PPO	6.46	***3.33***	6.14	***2.04***	6.39	***3.05***	6.19	***2.05***	6.53	***2.93***	6.36	***2.06***
Thermal Comfort ($\|PMV\|$)	Baseline	***0.25***	***0.22***	***0.28***	***0.28***	***0.28***	***0.27***	0.32	0.32	***0.28***	***0.27***	0.32	0.32
	BDQN	0.31	0.28	0.37	0.31	0.28	0.31	***0.3***	***0.3***	0.28	0.27	***0.3***	***0.3***
	SAC	0.33	0.29	0.38	0.45	0.28	0.42	0.3	0.3	0.29	0.4	0.31	0.39
	PPO	0.32	0.32	0.38	0.39	0.28	0.49	0.3	0.34	0.28	0.31	0.31	0.59
Thermal Comfort Violation (%)	Baseline	***4.06***	***0.78***	***3.98***	***0.86***	4.92	***0.86***	5.47	***1.02***	4.92	***0.86***	5.47	***1.02***
	BDQN	5.62	8.78	9.04	7.35	3.21	7.14	***4.44***	8.39	***3.38***	5.32	***4.38***	7.41
	SAC	8.22	3.39	10.86*	4.55	***2.77***	1.94	4.93	3.29	3.8	14.61*	4.55	4.55
	PPO	5.98	10.64*	10.86*	15.46*	3.37	8.93	5.32	10.61*	3.66	10.8*	4.73	8.04
Visual Comfort Violation (%)	Baseline	70.16	96.33	70.16	96.33	70.16	37.11	70.16	37.11	70.16	37.11	70.16	37.11
	BDQN	70.16	96.33	70.16	96.33	11.28	2.11	13.26	3.4	12.07	1.39	15.62	2.57
	SAC	70.16	96.33	70.16	96.33	***9.19***	***0.17***	***8.45***	***0.26***	***8.77***	***0.17***	***9.38***	***0.83***
	PPO	70.16	96.33	70.16	96.33	9.38	5.51	10.59	7.77	20.5	4.32	11.03	2.67

* The value exceeds the 10% threshold for thermal comfort violation, which is suggested by ASHRAE.

Table A.2: Total facility energy use for the best trade-off offered by each RL algorithm using zone-level occupancy information.

Control Scenario	**Baseline Number**	**Month**	**Baseline** (MWh)	**BDQN** (MWh)	**SAC** (MWh)	**PPO** (MWh)	**Best Agent** (improvement over baseline)
SAT setpoint Blinds always open	(1)	January	8.34	6.48 (22.3%)	6.44 (22.78%)	6.46 (22.54%)	SAC (22.78%)
		July	4.44	3.36 (24.32%)	3.58 (19.37%)	—	BDQN (24.32%)
SAT setpoint Blinds always open Auto dimming	(3)	January	8.4	6.16 (26.67%)	—	—	BDQN (26.67%)
		July	3.2	2.11 (34.06%)	2.27 (29.06%)	—	BDQN (34.06%)
SAT setpoint Single blind setpoint	(2)	January	7.07	6.51 (7.92%)	6.23 (11.88%)	6.39 (9.62%)	SAC (11.88%)
		July	3.92	3.12 (20.41%)	3.21 (18.11%)	3.05 (22.19%)	PPO (22.19%)
SAT setpoint Single blind setpoint Auto-dimming	(4)	January	6.81	6.34 (6.9%)	6.1 (10.43%)	6.19 (9.1%)	SAC (10.43%)
		July	3.02	2.16 (28.48%)	2.36 (21.85%)	—	BDQN (28.48%)
SAT setpoint Multiple blind setpoints	(2)	January	7.07	6.55 (7.36%)	6.18 (12.59%)	6.53 (7.64%)	SAC (12.59%)
		July	3.92	3.17 (19.13%)	—	—	BDQN (19.13%)
SAT setpoint Multiple blind setpoints Auto-dimming	(4)	January	6.81	6.35 (6.75%)	6.06 (11.01%)	6.36 (6.61%)	SAC (11.01%)
		July	3.02	2.2 (27.15%)	2.43 (19.54%)	2.06 (31.79%)	PPO (31.79%)

— The agent's thermal comfort violation rate exceeds the 10% threshold. Thus, the realized reduction in energy use is not reported.

Table A.3: RL Agent performance results for different control scenarios using building-level occupancy schedule.

		Blinds always open				**Single blind setpoint**				**Multiple blind setpoints**			
		No dimming		**With Dimming**		**No dimming**		**With Dimming**		**No dimming**		**With Dimming**	
		Jan.	**Jul.**	**Jan.**	**Jul.**	**Jan.**	**Jul.**	**Jan.**	**Jul.**	**Jan.**	**Jul.**	**Jan.**	**Jul.**
Energy (MWh)	Baseline	8.34	4.44	8.4	3.2	***7.07***	3.92	***6.81***	3.02	***7.07***	3.92	***6.81***	3.02
	BDQN	7.59	3.47	7.5	2.22	7.78	3.32	7.89	***2.35***	7.85	3.31	8.08	***2.35***
	SAC	***7.17***	***3.49***	***6.84***	2.26	7.85	***3.12***	7.83	2.37	7.72	***3.21***	7.69	2.42
	PPO	7.57	3.87	7.18	***2.18***	7.16	3.33	7.15	2.38	7.22	3.21	8.07	2.47
Thermal Comfort ($\|PMV\|$)	Baseline	***0.25***	***0.22***	***0.28***	***0.28***	0.28	***0.27***	0.32	0.32	0.28	***0.27***	0.32	0.32
	BDQN	0.29	0.28	0.31	0.34	0.23	0.3	0.25	0.3	0.23	0.27	0.24	***0.3***
	SAC	0.3	0.42	0.34	0.33	***0.22***	0.46	***0.23***	***0.24***	***0.22***	0.42	***0.23***	0.46
	PPO	0.29	0.56	0.33	0.38	0.26	0.8	0.27	0.97	0.26	0.29	0.26	0.56
Thermal Comfort Violation (%)	Baseline	***4.06***	***0.78***	***3.98***	***0.86***	4.92	***0.86***	5.47	***1.02***	4.92	***0.86***	5.47	***1.02***
	BDQN	7.19	8.59	6.72	7.19	3.75	8.83	4.14	9.38	3.67	7.27	3.28	9.45
	SAC	6.8	2.5	10.08*	3.59	***2.97***	3.2	***2.58***	3.44	***3.36***	3.36	***3.2***	3.28
	PPO	6.25	30.86*	8.75	15.47	5.55	8.98	4.53	8.67	5.62	9.45	4.45	30.16*
Visual Comfort Violation (%)	Baseline	70.16	96.33	70.16	96.33	70.16	37.11	70.16	37.11	70.16	37.11	70.16	37.11
	BDQN	70.16	96.33	70.16	96.33	12.5	3.52	11.8	4.45	10.86	4.14	10.78	5.31
	SAC	70.16	96.33	70.16	96.33	***8.05***	***0.0***	***8.05***	***0.23***	***8.28***	***0.08***	***8.05***	***1.64***
	PPO	70.16	96.33	70.16	96.33	12.19	5.16	9.22	5.47	15.62	11.95	12.11	7.19

* The value exceeds the 10% threshold for thermal comfort violation, which is suggested by ASHRAE.

Table A.4: Total facility energy use for the best trade-off offered by each RL algorithm using building-level occupancy information.

Control Scenario	**Baseline Number**	**Month**	**Baseline** (MWh)	**BDQN** (MWh)	**SAC** (MWh)	**PPO** (MWh)	**Best Agent** (improvement over baseline)
SAT setpoint Blinds always open	(1)	January	8.34	7.59 (8.99%)	7.17 (14.03%)	7.57 (9.23%)	SAC (14.03%)
		July	4.44	3.47 (21.85%)	3.49 (21.4%)	—	BDQN (21.85%)
SAT setpoint Blinds always open Auto dimming	(3)	January	8.4	7.5 (10.71%)	—	7.18 (14.52%)	PPO (14.52%)
		July	3.2	2.22 (30.63%)	2.26 (29.38%)	2.18 (31.88%)	PPO (31.88%)
SAT setpoint Single blind setpoint	(2)	January	7.07	7.78 (-10.04%)	7.85 (-11.03%)	7.16 (-1.27%)	PPO (-1.27%)
		July	3.92	3.32 (15.31%)	3.12 (20.41%)	3.33 (15.05%)	SAC (20.41%)
SAT setpoint Single blind setpoint Auto-dimming	(4)	January	6.81	7.89 (-15.86%)	7.83 (-14.98%)	7.15 (-4.99%)	PPO (-4.99%)
		July	3.02	2.35 (22.19%)	2.37 (21.52%)	2.38 (21.19%)	BDQN (22.19%)
SAT setpoint Multiple blind setpoints	(2)	January	7.07	7.85 (-11.03%)	7.72 (-9.19%)	7.22 (-2.12%)	PPO (-2.12%)
		July	3.92	3.31 (15.56%)	3.21 (18.11%)	3.21 (18.11%)	SAC (18.11%)
SAT setpoint Multiple blind setpoints Auto-dimming	(4)	January	6.81	8.08 (-18.65%)	7.69 (-12.92%)	8.07 (-18.5%)	SAC (-12.92%)
		July	3.02	2.35 (22.19%)	2.42 (19.87%)	—	BDQN (22.19%)

— The agent's thermal comfort violation rate exceeds the 10% threshold. Thus, the realized reduction in energy use is not reported.

A.2 Space allocation

A.2.1 `BestFit-Energy` algorithm

Algorithm 3 BestFit-Energy Algorithm

Require: $\mathcal{T}, \mathcal{X}^*, G, C, \theta, f(\cdot), G_k, p_k(\cdot), \theta$
Ensure: $i', t_{i'}$ ▷ Assigned zone and temperature setpoint

$\mathcal{S} \leftarrow \emptyset$
for i **in** $1, \ldots, |\mathcal{T}|$ **do**
  $occupancy \leftarrow \sum_{j \in \mathcal{M}} \mathcal{X}^*_{i,j} G_j$
  if $C_i - occupancy \geq G_k$ **then** ▷ If there is enough capacity
    $\Delta e \leftarrow f_i(\mathcal{T}_i, G_k + occupancy) - f_i(\mathcal{T}_i, occupancy)$
    $entry.zone \leftarrow i$
    $entry.occupied \leftarrow 0$ **if** $occupancy = 0$ **else** 1
    $entry.energy \leftarrow \Delta e$
    $\mathcal{S} \leftarrow \mathcal{S} \cup entry$
  end if
end for

if $\mathcal{S} = \emptyset$ **then** ▷ If $\mathcal{S}$ is empty, then no feasible solution
  reject k
end if

sort $\mathcal{S}$ **by** $value.occupied, value.energy$ **ascending**

for $entry$ **in** $\mathcal{S}$ **do**
  $i \leftarrow entry.zone$
  if $\sum_{j \in \mathcal{M}} \mathcal{X}^*_{i,j} G_j = 0$ **then** ▷ If the zone is vacant
    $i' \leftarrow i$
    $t_{i'} \leftarrow \text{argmin}_{t_i \in [t^{lb}, t^{ub}]} f_i(t_i, G_k)$
  else if $p_k(\mathcal{T}_i) \geq \theta$ **then**
    $i' \leftarrow i$
    $t_{i'} \leftarrow \mathcal{T}_i$
  end if
  terminate if i' **exists**
end for

for *entry* **in** $\mathcal{S}$ **do**
  $i \leftarrow entry.zone$
  $\mathbb{T} \leftarrow \mu_k \pm \sqrt{-2\sigma_k^2 \log\theta}$ ▷ Range of feasible setpoints
  for j **in** $1, \ldots, |G|$ **do**
    if $\mathcal{X}^*_{i,j} = 1$ **then**
      $\mathbb{T} \leftarrow \mathbb{T} \cap \left[\mu_j - \sqrt{-2\sigma_j^2 \log\theta},\ \ \mu_j + \sqrt{-2\sigma_j^2 \log\theta}\right]$
    end if
  end for

  if $\mathbb{T} \neq \emptyset$ **then**
    $i' \leftarrow i$
    $t_{i'} \leftarrow \text{argmin}_{t_i \in \mathbb{T}} f_i(t_i, G_k)$
    terminate
  end if
end for
reject k

A.2.2 `BestFit-Space` algorithm

Algorithm 4 Diversifying Temperature Setpoints

Require: $\mathcal{T}, \mathcal{X}^*, G, C, t^{lb}, \theta, \bar{\sigma}$
Ensure: $\mathcal{T}^*$
$\Delta \leftarrow \sqrt{-2\bar{\sigma}^2 \log\theta}$ ▷ Acceptable deviation from setpoint
$\mathcal{T}^* \leftarrow \emptyset, \mathcal{T}^{existing} \leftarrow \{26\}, t \leftarrow t^{lb}$
for i **in** $1, \ldots, |\mathcal{T}|$ **do**
  $occupancy \leftarrow \sum_{j \in \mathcal{M}} \mathcal{X}^*_{i,j} G_j$
  if $0 \leq occupancy \leq C_i$ **then**
    $\mathcal{T}^{existing} \leftarrow \mathcal{T}^{existing} \cup \{\mathcal{T}_i - \Delta\}$
    ▷ Get lower limit of acceptable temperature range
  end if
end for
sort $\mathcal{T}^{existing}$ **ascending**
for t_i **in** $\mathcal{T}^{existing}$ **do**
  if $t < t_i$ **then**
    $num \leftarrow \lceil (t_i - t)/(2 \times \Delta) \rceil$
    $\delta \leftarrow (t_i - t)/(num + 1)$
    for i **in** $1, \ldots, num$ **do**
      $\mathcal{T}^* \leftarrow \mathcal{T}^* \cup \{t + i \times \delta\}$
    end for
  end if
  $t \leftarrow t_i + 2 \times \Delta$
end for

Algorithm 5 BestFit-Space Algorithm

Require: $\mathcal{T}, \mathcal{X}^*, G, C, \theta, f(\cdot), G_k, p_k(\cdot), \theta$
Ensure: $i', t_{i'}$

$\mathcal{S}, \mathcal{S}^{vacant} \leftarrow \emptyset$
$\Delta \leftarrow \sqrt{-2\bar{\sigma}^2 \log \theta}$

for i **in** $1, \ldots, |\mathcal{T}|$ **do**
  $occupancy \leftarrow \sum_{j \in \mathcal{M}} \mathcal{X}^*_{i,j} G_j$
  if $0 < occupancy \leq C_i - G_k$ **then**
    $c \leftarrow C_i - G_k$ ▷ Find effective capacity
    for j **in** $1, \ldots, |\mathcal{T}|$ **do**
      $occupancy' \leftarrow \sum_{k \in \mathcal{M}} \mathcal{X}^*_{j,k} G_k$
      if $occupancy' = 0$ **or** $2 \times \Delta \leq |\mathcal{T}_i - \mathcal{T}_j|$ **then**
        continue
      end if
      $c \leftarrow c + (1 - |\mathcal{T}_i - \mathcal{T}_j| / (2 \times \Delta)) * (C_j - occupancy')$
    end for

    $\Delta e \leftarrow f_i(\mathcal{T}_i, G_k + occupancy) - f_i(\mathcal{T}_i, occupancy)$
    $entry \leftarrow \{capacity : c, zone : i, energy : \Delta e\}$
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{entry\}$

  else if $occupancy = 0$ **then**
    $\mathcal{S}^{vacant} \leftarrow \mathcal{S}^{vacant} \cup \{i\}$
  end if

end for
if $\mathcal{S} \neq \emptyset$ **then** ▷ Occupied zones can accommodate G_k
  sort $\mathcal{S}$ **by** $value.capacity, value.energy$ **ascending**
  $i' \leftarrow \mathcal{S}_0.zone$
  $t_{i'} \leftarrow \mathcal{T}_{i'}$
  terminate
end if

$\mathcal{T}^* = \texttt{diversitySetpoints}(\mathcal{T}, \mathcal{X}, G, C, \Delta)$ ▷ Algorithm 4

for i **in** $1, \ldots, |\mathcal{T}|$ **do**
  if $t \in \mu_k \pm \sqrt{-2\sigma_k^2 \log \theta}$ **then**
    $c \leftarrow 0$ ▷ Find effective capacity
    for j **in** $1, \ldots, |\mathcal{T}|$ **do**
      $occupancy' \leftarrow \sum_{k \in \mathcal{M}} \mathcal{X}^*_{j,k} G_k$
      if $occupancy' = 0$ **or** $2 \times \Delta \leq |\mathcal{T}_i - \mathcal{T}_j|$ **then**
        continue
      end if
      $c \leftarrow c + (1 - |t - \mathcal{T}_j| / (2 \times \Delta)) * (C_j - occupancy')$
    end for

    $entry.capacity \leftarrow c$
    $entry.setpoint \leftarrow t$
    $entry.zone \leftarrow \text{argmin}_{i \in \mathcal{S}^{vacant}} (f_i(t, G_k) - f_i(t, 0))$
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{entry\}$

  end if
end for
sort $\mathcal{S}$ **by** $value.capacity$ **ascending**

if $\mathcal{S} = \emptyset$ **then** ▷ If $\mathcal{S}$ is empty, then no feasible solution
  reject k
end if

$i' \leftarrow \mathcal{S}_0.zone$
$t_{i'} \leftarrow \mathcal{S}_0.setpoint$