Data Driven Chiller Sequencing for Reducing HVAC Electricity Consumption in Commercial Buildings

Zimu Zheng
The Hong Kong Polytechnic University

Qiong Chen
Huazhong University of Science and Technology

Cheng Fan
Shenzhen University

Nan Guan
The Hong Kong Polytechnic University

Arun Vishwanath
IBM Research Australia

Dan Wang
The Hong Kong Polytechnic University

Fangming Liu
Huazhong University of Science and Technology

ABSTRACT

It is well-known that the HVAC (heating, ventilation and air conditioning) dominates electricity consumption in commercial buildings. Alongside, electricity prices are increasing in several nations around the world, putting pressure on facility managers to reduce the electricity consumption incurred in operating their HVAC and buildings. In this paper, we focus on one of the core problems in building operation, namely *chiller sequencing* for reducing HVAC electricity consumption. Our contributions are threefold. First, we make a case for why it is important to quantify the performance profile of a chiller, namely coefficient of performance (COP), at *run-time*, by developing a data-driven COP estimation methodology. Second, we show that predicting COP accurately is a non-trivial problem, requiring considerable computation time. To overcome this barrier, we develop a dominant-graph based COP prediction technique and a time-constrained chiller sequencing algorithm integrating the COP predictions, which strikes a good balance between electricity consumption reduction and ease of use for real-world deployment. Finally, we evaluate the performance of our scheme by applying it to real-world data, spanning 4 years, obtained from multiple chillers across 3 large commercial buildings in Hong Kong. The results show that our solution is able to save on average 21 MWh of electricity consumption in each of the 3 buildings, which is an improvement of over 30% compared to the current mode of operation of the chillers in the buildings. We offer our data-driven chiller sequencing framework under time constraints as an effective and practical mechanism for reducing the electricity consumption associated with HVAC operation in commercial buildings.

CCS CONCEPTS

• **Information systems → Data analytics**; *Process control systems*; • **Hardware → Power estimation and optimization**; • **Applied computing →** *Decision analysis*;

e-Energy '18, June 12–15, 2018, Karlsruhe, Germany

ACM ISBN 978-1-4503-5767-8/18/06…$15.00
https://doi.org/10.1145/3208903.3208913

KEYWORDS

HVAC Operation, Chiller Sequencing, Applied Machine Learning

ACM Reference Format:
Zimu Zheng, Qiong Chen, Cheng Fan, Nan Guan, Arun Vishwanath, Dan Wang, and Fangming Liu. 2018. Data Driven Chiller Sequencing for Reducing HVAC Electricity Consumption in Commercial Buildings. In *Proceedings of e-Energy '18: International Conference on Future Energy Systems (e-Energy '18)*. ACM, Germany, 13 pages. https://doi.org/10.1145/3208903.3208913

1 INTRODUCTION

Centralized chilled water based HVAC plants are commonly used for cooling in large commercial buildings. These HVAC plants consume anywhere between 40% and 70% of a building's total electricity consumption [1, 2], a vast majority of which can be attributed to the chillers in the HVAC. In several nations around the world, the electricity bill paid by commercial buildings, which is dominated by the energy consumption of the HVAC, is often in the top-three list of an organization's operating expenses [3]. This trend is putting upward pressure on facility managers to improve the energy efficiency of their buildings by means of reducing the electricity consumption associated with HVAC operation.

Various techniques have been proposed in the literature for mitigating the energy impact of building HVAC. These include controlling the HVAC based on the spatio-temporal profile of occupancy inside a building [4], pre-cooling a building in advance of expected increase in occupancy [5], and incorporating renewables such as solar panels and battery storage into the energy mix [6].

While all of these approaches have merit, in this paper, we focus on the problem of chiller sequencing for reducing HVAC electricity consumption in commercial buildings. Chiller sequencing refers to operating the most efficient combination of chillers in a building at real-time to meet the time-varying cooling demand. For example, sequencing a building with two chillers [0.5, 0.7] implies that chiller 1 and chiller 2 are operating at 50% and 70% of their maximum rated capacity, respectively. Thus, the sequencing problem is to allocate the cooling load at any given time to the chillers in the most energy efficient manner so that the overall cooling demand of the building is satisfied while at the same time the electricity consumed by the chillers is kept at a minimum [7].

Most prior work on chiller sequencing has focused on developing techniques for predicting the cooling demand accurately. However, the efficacy of chiller sequencing control also relies heavily on the *run-time* performance profile of the chillers, namely the COP under different cooling load regimes. COP is a measure of the energy-efficiency of a chiller and captures the cooling power that it can output for a certain input power consumption. COP is typically greater than 1; larger values implying better efficiency [8].

Despite advances in chiller performance profiling, they have relied on developing fixed form thermodynamic models for obtaining the COP given cooling load [9]. These models have limited value in practice because COP is highly dependent on a variety of factors such as the operating conditions, configuration dynamics, varying cooling demands, degradation over time, weather and so on, making it extremely difficult to capture the impact of these parameters accurately in an analytical model. For e.g., it was recently shown that over 12 months, there was a 20% reduction in the chilled water flow rate, caused by excessive fouling that blocked the tubes in the chiller condenser [10]. Using the COP from these fixed form models therefore can introduce large errors, rendering them impractical for use in the real-world. In practice, facility managers often perform chiller sequencing using COP profiles obtained when the chillers are first tested and commissioned during installation in a building, called *initial profiles*. The initial profile considers cooling load as the sole parameter. For reasons mentioned above, and detailed in the rest of the paper, these initial profiles fail to capture the impact of other real-world parameters, and thus are not accurate. It is evident that robust estimation of the run-time COP of chillers is critical for the success of any chiller sequencing technique.

Inspired by the advent of IoT deployment in buildings, and the availability of IoT sensor data logged by modern building management systems (BMS), in this paper, we advocate a data-driven COP profiling approach to facilitate chiller sequencing. Our COP estimation relies on data collected by BMS. Specifically, our COP profiling techniques are underpinned by BMS data obtained at 30 minute intervals from 17 chillers, over four years, across three high-rise office buildings located in Hong Kong. We make three important observations. First, existing thermodynamic models for COP estimation can be inaccurate. For example, the run-time COP of water-cooled chillers with constant-speed primary pumps (like the ones considered in this paper) does not increase monotonically with the cooling load, as is typically assumed in practice and found in the initial profiles [11]. Second, there is a significant difference between the COP obtained from the data-driven approach and initial profiles for different cooling loads. Third, data-driven profiling increases the accuracy of chiller COP estimation, paving the way for energy-efficient chiller sequencing in practice. In this context, the contributions of this paper are:

- We demonstrate that there is a need for individualized COP chiller performance profiling at run-time, which when done effectively can be instrumental in reducing HVAC electricity consumption. As discussed above, the resulting COP values can vary substantially from that obtained via initial profiling. The latter is often used for sequencing chillers in practice today, undermining their energy efficiency considerably.
- We show that COP performance profiling using data-driven techniques is a challenging problem, in terms of computation time. And so we propose a dominant-graph based COP prediction technique along with a time-constrained chiller sequencing control algorithm. We highlight that it strikes a good balance between reducing electricity consumption for chiller operation and ease of use for real-world deployment.
- We comprehensively evaluate the efficacy of our approach by applying the solution on BMS data, spanning 4 years (2012-2015), obtained from multiple chillers across 3 high-rise office buildings in Hong Kong. The results show that our chiller sequencing approach is able to save on average 21 MWh of electricity consumption in each of the 3 buildings, which is an improvement of around 30% compared to the current mode of operation of the chillers in the buildings.

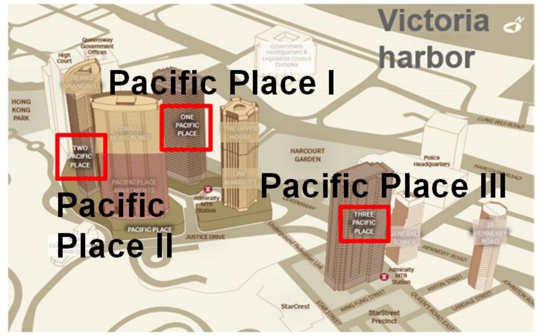


Figure 1: Towers of Pacific Place I, II and III in Hong Kong.

Our proposed data-driven COP estimation technique and chiller sequencing solution does not require any major capital expense and uses data readily available from any modern BMS. The solution recommends a chiller sequencing strategy that not only satisfies a building's cooling demand but also keeps the electricity consumption to a minimum. We offer our approach as an attractive mechanism for building facility managers to use who are on the look out for simple and low-cost means for reducing the energy and cost footprints of their buildings.

2 NEED FOR DATA DRIVEN COP PROFILING

2.1 Introduction to the Chiller Plants

Chiller plants are frequently used to generate cooling power for office buildings. For instance, in three office towers located in Hong Kong (Fig. 1), three chiller plants containing a total of 17 chillers serve more than ten thousand people. Four year data, spanning 2012 through 2015, at 30 minute intervals, for these different chiller models from Trane was collected from the BMS, as shown in Table 1.

Table 1: Chiller information in each building.

Building Name	Regular Chiller	Backup Chiller	Vendor
Pacific Place I	4 × CVHG1100	2 × CVHE370	6 × Trane
Pacific Place II	3 × CDHG2250	2 × CVHG780	5 × Trane
Pacific Place III	4 × CVGF500	2 × CVGF500	6 × Trane
Total Number	11	6	17

In commercial buildings, sequencing of chillers is performed to keep the electricity consumed for meeting a certain cooling demand to a minimum. It follows two steps, i.e. Sequence Determination and Feedback Control, and they work as follows. When a cooling demand $\mathcal{D}$ arrives, the HVAC plant needs to determine the set of chillers that need to be active and the total cooling load $Q > \mathcal{D}$ to support the demand (Sequence Determination). The HVAC plant then needs to adjust the cooling load of each (active) chiller until the cumulative load of Q is attained (Feedback Control).

2.2 COP Computation

Chiller sequencing relies heavily on the energy-efficiency of the chillers. Clearly, electricity consumption increases as a function of the cooling load. Note that the amount of electricity consumed by a chiller is not only determined by Q but also by its energy-efficiency. Intuitively, if this efficiency is low (e.g. due to poor maintenance), then more electricity will be consumed to support a required cooling demand. It is therefore of paramount importance to quantify the energy-efficiency of a chiller, which is measured by its COP, determined as follows.

The cooling load of a HVAC plant at a given time is the sum of the cooling load Q_i over all chillers i, i.e., $Q = \sum_i Q_i$, where $Q_i = c_i \times m_i \times \Delta_{T\,i}$. Here, c_i is the thermal capacity of water (kJ/kg°C), m_i is the chilled water mass flow rate (kg/s) and $\Delta_{T\,i}$ is the temperature difference between the returned and supplied chilled water (°C) [12]. All these quantities are logged by our BMS.

The COP of chiller i to support cooling load Q_i is given by Q_i/E_i, where E_i is the electrical power consumed by chiller i to deliver the required amount of cooling. In practice, after a HVAC plant is installed in a building, there is a commissioning phase wherein a set of cooling loads is tested to ascertain the performance of the chillers. Following each test, values of Q and E are recorded which enable the facility manager to determine the corresponding COP profiles for the chillers. Once in production, certain statistical averaging techniques are used in the ensuing short period of operation to update the COP under different cooling loads [13, 14].

2.3 Observation of Significant COP Variation

Reliable chiller sequencing depends on the COP across all the loading conditions for chiller i. However, when we communicated with facility managers and applied the above computation on the historical data retrieved from the BMS, we learned and confirmed that not only does the COP degrade over time and raise after maintenance, which is well-known, but the COP fluctuates markedly over different cooling loads and environmental conditions.

To be specific, we first plot the average COPs as a function of the cooling loads in Fig. 2. We picked the first regular chiller from each building. It can be seen that these chillers often operate between 40% and 80% load, and their COP fluctuates somewhat randomly between zero and eight. The COP values for other loads are missing. For the same chillers, we plot in Fig. 3 the variation in their COP (i.e. difference between max and min) under different loading regimes, and note that there is a large fluctuation even for a given load. For e.g., COP for chiller 1 varies between 5.7 and 8.2 for 70% load, and between 1.8 to 8.3 for 60% load. This is because the chiller COP in practice is highly dependent on a variety of factors. We also observe from the data that if we classify the COPs at 5% cooling load increments, then more than 40% of the COPs are missing.

Chillers are complex systems. The COP fluctuation observed above is the result of thermodynamic processes under changing environmental conditions and configurations of the local building context, as well as the impact of other parameters such as chiller degradation and exposure to different seasons/weather. These factors are exceedingly difficult to capture within an analytical model.

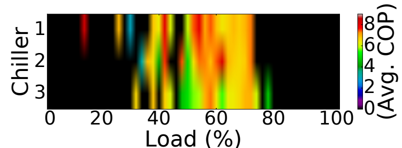


Figure 2: Average COP for different chillers and loads.

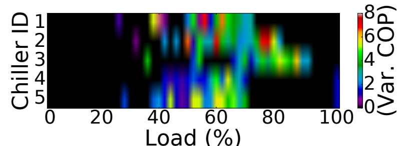


Figure 3: COP Variation for different chillers and loads.

Table 2: Example of a 10×5 updated profile on 2015.12.31 in Pacific Place II, with out-of-date entries discarded.

Load	Chiller 1	Chiller 2	Chiller 3	Chiller 4	Chiller 5
50%	7.4	-	-	7.6	7.0
55%	-	-	-	-	-
60%	6.4	7.1	4.7	6.9	-
65%	-	-	-	-	-
70%	7.3	5.6	7.2	-	-
75%	-	-	-	-	-
80%	-	6.9	5.4	-	-
85%	-	-	-	-	-
90%	-	-	7.0	-	6.6
95%	-	-	-	-	-

Data-driven techniques can thus play a crucial role in accurate COP prediction for improved chiller sequencing in the real world. For interested readers, a further discussion on the possible benefit of accurate COP prediction is available in Appendix A.

3 OVERVIEW OF TIME-CONSTRAINED DATA-DRIVEN CHILLER SEQUENCING (T-DCS) FRAMEWORK

In this section, we describe our solution framework for the data-driven chiller sequencing problem. The framework comprises three steps. An overview of each of these steps is described next.

3.1 T-DCS Problem Definition

1. The Data-driven COP Prediction (DPP) Sub-Problem. Our idea is to develop individualized COP for each chiller by applying machine learning techniques using historical chiller data. A private cloud is established to store the historical data from the BMS. When a cooling demand $\mathcal{D}$ arrives, the cloud can perform chiller sequencing assisted by our data-driven COP prediction schemes. To this end, we first introduce the Data-driven COP Prediction Sub-Problem.

Formal definition of DPP Sub-Problem: Given the current prediction task, infer the COP profile $\boldsymbol{COP}$ which minimizes the prediction loss, i.e., $1/T \sum_{t<T} ||F_t(\boldsymbol{X}_t, \boldsymbol{W}) - COP_t||_2$, where $\boldsymbol{X}$ denotes the features at time t of total time T; $\boldsymbol{COP_t}$ denotes the predicted COP for all chillers at time t; $F(\cdot)$ denotes the learning and prediction model and $\boldsymbol{W}$ denotes its parameters.

2. The T-DCS Problem. The next step is to determine the optimal sequencing of chillers. To do so, one needs to be wary of the following: (1) The cooling demand changes over time, so chiller

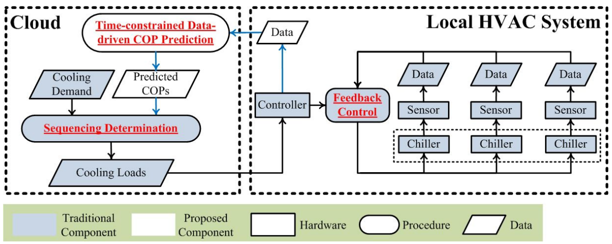


Figure 4: Framework of Data-driven Chiller Sequencing.

sequencing must be performed repeatedly in order to continuously meet the varying cooling demand. The common practice is to trigger chiller sequencing in a periodic manner [15]. (2) To ensure cooling performance, chiller sequencing needs time for feedback control until the system regains stability when switching from one sequence to another. There is also a minimum start-stop-start time (called *deadband*) for every chiller. (3) The chiller sequencing for each period must be completed before the start of the next sequencing period. Otherwise, the system can be unstable and return inaccurate data which can be detrimental to subsequent COP prediction and sequencing operations, as well as for the overall performance of the chillers. Due to the above issues, we define the term *deadline* T_D as the total time length of one chiller sequencing operation, including the computation time and the mechanical switching time, computed considering both the periodic interval t_P and mechanical switching time t_M, e.g., $T_D = \min(t_P, t_M)$.

Formal definition of T-DCS Problem: Given cooling demand $\mathcal{D}$, deadline T_D, historical COP labels $\boldsymbol{COP}$, the targeted and historical feature values $\boldsymbol{X_c}, \boldsymbol{X}$, our objective is to find a chiller sequence $\boldsymbol{Q} = \{Q_i\}$ which minimizes the total energy consumption E. The final solution should satisfy the cooling demand, i.e., $\sum_i Q_i > \mathcal{D}$ and the total needed time T is within the deadline T_D, i.e., $T \leq T_D$.

3.2 Solution Framework for T-DCS Problem

To solve the T-DCS problem, we propose the data-driven chiller sequencing framework, as shown in Fig. 4. The framework contains three main components: (1) Time-constrained Data-driven COP Prediction, (2) Sequencing Determination and (3) Feedback Control.

The Time-constrained Data-driven COP Prediction takes the historical data $\boldsymbol{X}$, historical COP labels $\boldsymbol{COP}$, deadline T_D as input and generates the COP prediction result within the deadline. Then, using the predicted COP and cooling demand $\mathcal{D}$, Sequencing Determination outputs the optimal chiller sequence and Feedback Control ensures the successful execution of the sequencing in the local HVAC plant. For simplicity, we omit the detailed designs of the Sequencing Determination and Feedback Control components. In the following sections, we focus on (1) Data-driven COP Prediction for the DPP Sub-Problem (Section 4) and (2) Time-constrained Data-driven COP Prediction for the T-DCS Problem (Section 5).

4 DATA-DRIVEN COP PREDICTION FOR DPP

To solve the DPP Sub-Problem outlined in Section 3.1, we develop an approach involving two steps: *Domain-assisted Feature Engineering* and *Clustered Multi-task Learning*, as described next.

4.1 Domain-assisted Feature Engineering

In industry domain, there is usually no luxury to have enormous data where a model can be trained to automatically eliminate irrelevant features. As such the first challenge is to select the proper feature set for chiller performance profiling. Our feature engineering uses domain knowledge to create features relevant to the problem at hand. The understanding includes the influence of external environmental conditions and the influence of inner mechanical factors associated with the chillers. We list our features in Table 3.

Temporal Features. First, we exploit the seasonality and the age of the chillers (in terms of days) as the temporal features. Intuitively, the chiller demands exhibit distinctive temporal characteristics: 1) cooling loads of chillers are different in different seasons, especially summer and winter, which leads to varying performance degradation; 2) as chillers age, its performance gradually degrades as well [10].

Meteorological Features. Second, we know that meteorological information such as temperature and weather drive the cooling demand imposed on the chillers. For example, a higher outdoor temperature requires more cooling power to ensure a comfortable room. This meteorological factor would affect the chiller mode and thus the chiller performance.

Mechanical Features. Finally, mechanical features are used to capture the chiller characteristics. The model type, building, operating power, water temperature difference, flow rate and the recent cooling load are important features. The cooling load is the amount of heat energy that would need to be removed from a space to maintain the temperature within an acceptable range. In practice, cooling loads are handled by air-conditioning equipment of chillers, which reflects the amount of work that chillers provide, and thus significantly impacts chiller degradation.

4.2 Clustered Multi-task Learning

Another challenge we face is the sparsity of the performance profile. Reliable sequencing requires the COP for all chiller loads. However, as shown in the Table 2, it is common for chillers to run on merely a small distinct set of loads, which leaves a sparse profile for training and prediction. The COP corresponding to the empty loads is difficult to infer reliably with little training data.

A natural way to solve this sparse problem is to infer the values using other non-empty ones. However, simply replacing with neighboring non-empty entries can cause significant errors. For example, in Table 2, we see that even for the same Chiller 3, replacing the COP of 80% with 90% leads to a relative error of 29.63%. That is because, at the time when these COPs are updated, external environment (e.g., meteorological factors) and inner conditions (e.g., temporal and mechanical factors) can be different, which leads to different thermodynamic processes of operation, resulting in different COPs. In other words, for different entries, different model parameters must be used to capture the underlying thermodynamic processes, i.e., we need to train one model for each entry, while at the same time exploit the benefits that come with (potentially) more information being present in larger training data sets.

Specifically, our idea here is to learn from not only the training data available for this single entry, but also learn from training data present in other pertinent contexts, e.g., cases with similar temporal, meteorological and mechanical conditions. To this end,

Table 3: The description of features.

Feature Type	Feature	Description
Temporal	Season	The season which the time interval is in
	Age of chiller	The number of days that the chillers have been working
Meteorological	Weather Condition	The description of weather condition in a time interval
	Outdoor Temperature	The outdoor temperature measured by Celsius in a time interval
Mechanical	Model Type	The model of the operating chiller
	Building	The building that the operating chiller is deployed in
	Operating Power	The power measured by kilowatts for the operating chiller
	Water Mass Flow Rate	The mass of water flowing per second, measured by kg/s
	Water Temperature Difference	The difference between the returned and supplied chilled water temperature
	Latest Cooling Load	The last recored cooling load assigned on this chiller

we apply *Clustered Multi-Task Learning (CMTL)* approach [16]. The COP prediction on an entry is called a *prediction task* in our paper. For each entry, the prediction task collects training data from similar contexts. CMTL is suitable for our multi-task and sparse condition, i.e., we not only need to develop different model parameters for each entry, but also need to share knowledge, e.g., training samples, among these entries. The prediction results of different learning methods are available in Appendix B.

5 TIME-CONSTRAINED DATA-DRIVEN COP PREDICTION FOR T-DCS

In the previous section, we described a solution for the Data-driven COP Prediction (DPP) sub-problem. However, the solution method cannot be directly applied in practice: (1) owing to the time-complexity associated with the prediction technique, which is computationally expensive. (2) This is exacerbated by the fact that data-driven prediction must be conducted in an on-line manner due to the sparsity issue, forcing the prediction model to be updated frequently to account for the time-varying nature of the cooling demand and change in operating conditions. To ensure high performance of our model, in this paper we update its parameters each time before sequencing is performed. (3) Finally, there is often a time constraint for each chiller sequencing operation, as mentioned earlier. Here is an example of the above issues: predicting the COP for each entry in Table 2 (i.e. for a given chiller and a cooling load) takes about 20 minutes using AdaBoost [500 trees] for a TB-level dataset on a private cloud with a 16-core CPU and 12 GB memory. So predicting the COP for the 50 entries in Table 2 would require $50 \times 20 = 1000$ minutes = 16.67 hours, which is significantly longer than the typical chiller sequencing period (of two hours [15]), and thus cannot be accomplished before the deadline.

A natural way to solve this problem is to provide higher computation capacity to finish the COP prediction before the deadline, but this will incur considerable costs. For our example, for a service provider of BMS in charge of 10 buildings, finishing our proposed DPP-based Chiller Sequencing (DCS) within the typical chiller sequencing period of two hours requires 87 m4.xlarge instances and 10-TB General Purpose SSD (gp2) volume on Amazon Web Service (AWS), for which the total annual price is $123,975 (m4.xlarge) +$14,746 (gp2)= $138,721.

In this paper, we propose to solve this problem by reducing the computation workload processed in each period, by only conducting COP prediction with the expensive learning model (AdaBoost

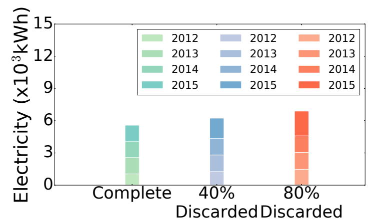


Figure 5: Electricity consumption with 1) all profiles, 2) 40% randomly discarded, and 3) 80% randomly discarded.

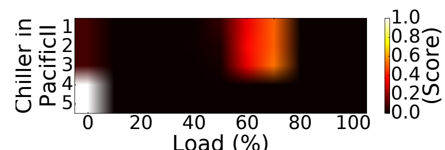


Figure 6: The probability of becoming best operation for different entries (Chiller 4 and 5 are backup chillers).

[500 Trees]) for a subset of the entries in Table 2, and using an efficient but less accurate learning model (Linear Regression) for the remaining entries. The question is how to decide which entries should be processed with the expensive learning model, in order to maximize the overall optimality of our DCS approach under the time constraint.

Fig. 5 shows the electricity consumption of data-driven sequencing with 40% and 80% randomly discarded entries to be predicted with the expensive learning model (while the remaining is predicted with the efficient model). We see a 13.21% and 25.63% higher electricity consumption, which eliminate 27.80% and 53.95% of our maximum potential saving, respectively. In the following, we will show that by carefully selecting the entries to be processed with the expensive learning model, the gain of our approach can be significantly improved under the time constraint.

5.1 Implementation Overview

The key observation of our approach is that, while in principle all COP entries may be selected to form the chiller sequencing, in practice only a small subset of them are frequently selected in the optimal sequence. The historical best operations can be computed with the sequencing optimization based on the ground truth of COP of 1460 days from 2012 to 2015. Then we can count the number of cases for each entry to be selected as the best operation and thus obtain the probability to become optimal. For example, if an entry is selected in 100 days as the best operation over the total 1460

Table 4: The description of factors to decide the priority.

Factor Type	Factor	Description
Mechanical	Cooling Load	The cooling power that an operation is supposed to provide
	Electricity Consumption	The amount of electricity that a sequencing would require for a period of time, e.g., one hour
Statistical	Past Success	The number of cases that a sequencing is selected in the optimal solution in the past
	Prediction Accuracy	The similarity between the predicted COP and the real COP for a given entry in the past

days, its probability to become optimal is computed as 100 / 1460 = 6.84%. Figure 6 shows the probability for different entries to be selected as the best operation in the whole year 2015 with a model trained with historical data of 2012-2014. The regular chillers often operate between load 50% and 80%, and sometimes operate on 0% load (switched off). The backup chillers 4 and 5 are seldom used and almost always remain on 0% load.

Intuitively, one should *prioritize* the entries, which is to order them according to the possibility for them to be selected as the optimal operation, and apply the expensive learning model to higher-priority entries first, until the end of the period. However, the possibility for an entry to be selected also changes over time, so the priority should also be adjusted dynamically. In Section 5.2, we will present how to decide this priority in detail.

Based on the above observations, we propose the Time-Constrained DCS (T-DCS) approach, the pseudo-code for which is shown in Algorithm 1. In each period, we first predict the COP of each entry by Linear Regression (line 2), used as their initial profiles. Second, we calculate the priority ordering of the entries in the current period jointly using multiple factors (line 3, which will be explained in detail in Section 5.2). Then we iteratively update the profiles of the entries using the expensive COP prediction model (line 5), according to the priority of the entries, until the available computation time expires. Finally, the optimal sequencing (line 7) is selected based on the up-to-date COP. Note that the stop condition of line 4 does not consider the computation time for selecting the optimal sequencing, because it can be ignored compared to the duration of the COP prediction procedure.

Algorithm 1 Time-Constrained DCS (T-DCS)

1: $\mathcal{S} \leftarrow$ possible sequencing, $\mathcal{D} \leftarrow$ cooling demand
2: $\mathcal{P} \leftarrow$ *LinearRegression*() ▷ Fill all entries with initial values.
3: $\mathcal{S}' \leftarrow$ *JointPriorityOrdering*($\mathcal{S}$)
4: **while** not yet reach the deadline **do**
5: $\quad\mathcal{P} \leftarrow$ *ExpensiveCOPPrediction*($\mathcal{S}$')
6: **end while**
7: $S \leftarrow$ *SelectOptimalSequencing*($\mathcal{D}, \mathcal{P}$)
8: **return** S

5.2 Joint Priority Ordering

To find the priority of entries for prediction, we first select *Factors to decide the entry priority*. With the *Historical priority determination* showing the priority of entries in the similar situations of the past, we develop the *Joint score and priority determination* to obtain the final priority of the entry.

Example: Referring to Table 2, assume we have two entries for Chiller 1 COP. The first corresponding to Chiller 1 running at 55% load (denoted as Entry 1) and the second corresponding to Chiller 1 running at 80% load (denoted as Entry 2). There could me multiple sequences that include Entry 1, for e.g. the loads for the five chillers being (55%, 60%, 60%, 60%, 60%), (55%, 70%, 70%, 75%, 70%), and so on. We denote these sequences as Sequence 1, Sequence 2, etc. Let Prob. 1 and Prob. 2 denote the probabilities of Sequences 1 and 2 being selected as the optimal sequence, respectively (probability computed as described in Section 5.1). The higher the probability is for a sequence, the more likely that it will be selected as the optimal sequence (in turn increasing the ranking or priority of the entries in the sequence). We denote the priority of Entry 1 as Pri. 1, which is computed as Pri. 1 = 1 - Prob. 1. Similarly for Pri. 2. The lower the priority score the higher is its importance. Assuming Prob. 1 = 0.2 and Prob. 2 = 0.8, then Entry 2 will be prioritized above Entry 1 because Pri. 2 = 0.2 < Pri. 1 = 0.8.

1. Factors to Decide the Entry Priority. To decide the entry priority, we select several important factors which help to indicate the probability of entries to be selected in the optimal solution, as summarized in Table 4. These factors fall into two categories: *Mechanical Factors* and *Statistical Factors*. In the following, we explain these factors in detail.

Mechanical Factors reflect the mechanical performance of a chiller operation. They directly affect the possibility for a chiller operation to be optimal. We consider two mechanical factors:

- **Cooling Load**, which is a clearly important factor as the total cooling load of the solution must meet the cooling demand.
- **Electricity Consumption** over a period under the corresponding cooling load and with the average COP value of the past period. This is an important factor because our objective is to minimize the electricity consumption.

Statistical factors reflect the performance of a chiller operation in the statistical view over the historical data. In particular, we consider the following two factors in this category:

- **Past Success** refers to the number of cases that an entry is selected in the optimal solution in the past. In general, if the chiller status, configuration, and environment do not change too much, a chiller operation that works well in the past will likely work well in the current period.
- **Prediction Accuracy** is the average similarity between the predicted performance profile and the real performance profile for a given operation in the past. This is also an important factor since an accurate prediction can reduce the possibility of failure in meeting the cooling demand, and thus decreases the probability to launch the energy-inefficient protecting process using backup chillers.

Note that the weight of these factors in the final decision of the entry priority may be different under different cooling demands. Due to page limitation, we put the discussion of the priority for each factor under different cooling demands in Appendix C.

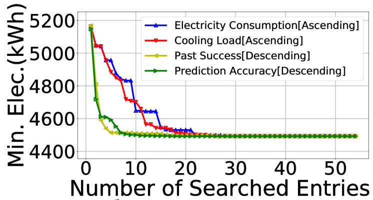


Figure 7: Minimum electricity consumption as a function of the number of searched entries ranked by different factors.

Figure 7 shows minimum electricity consumption with the historical data in the past four years, with respect to the number of entries visited in the searching procedure using the priority decided by each of the above four factors. We can see from the figure that prioritizing the entries by Past Success and Prediction Accuracy (in the descending order) reduces about 90% of the search space to reach the optimal solution, and prioritizing the entries by Cooling Load and Electricity Consumption (in the ascending order) reduces about 70% of the search space.

2. Historical Priority Determination. With above factors for each entry, we can learn the priority of entries from the past data. Basically, the same entry works well for similar conditions in the past can be useful in the present. In the following we start by deciding the priority of the values of each factor $\boldsymbol{r}_k^t$ for the past data, which are used to decide the priority in the present.

In general, the more important a sequence is, the more important its entries can be. With such an idea, the priority of the entries is then decided by the priority of the entries' sequences in which they are a part of (as given in the example above). Two types of relationships are used for priority decision.

- **Dominant Relationship** compares entries used in different sequences. These entries are clearly comparable because their priorities can be inferred according to the priority of their respective sequences, which can be obtained according to the electricity consumption objective and the cooling demand constraint with the past data.
- **Cooperative Relationship** compares entries used in the same sequence. We collect such cooperative entries in a set *layer*, where the entries may not be clearly comparable.

Accordingly, we apply the *Dominant Graph (DG)* [17] for the relationship modeling, which treats the two types of relationships differentially, i.e., the dominant relationship is more important for priority decision than the cooperative one. Two entries having a dominant relationship are ordered and directly linked in DG. The entries in the same layer can be ordered according to the number of cases when it is used for sequencing which meets cooling demand. Then we decide the best priority $\mathcal{COP}_{ec}^t$ of the entries with the *traveler algorithm* in Dominant Graph [17]. [1] Note that such a historical priority of entries should not be directly applied as the current priority, due to the changing COPs and cooling demands.

Finally, we obtain the historical priority $\boldsymbol{r}_k^t$ of the four factors for all historical time instances: $\boldsymbol{r}_k^t = [COP_i^t.r_k], \forall COP_i^t \in \mathcal{COP}_{ec}^t$, where $\boldsymbol{r}_k^t$ denote the historical priority of the factors for a past time instant t and $r_{ki}^t \in \boldsymbol{r}_k^t$ denote the ith value of $\boldsymbol{r}_k^t$. With the historical priority $\boldsymbol{r}_k^t$, the score s_k^j can be inferred for each factor k of entry j. Basically, the higher the priority of an entry is in the past, the higher is its score in the present. [2]

[1] An entry can be used by multiple candidate sequences and should not be updated repeatedly using the same model in the sequencing decision because the result will be the same. In one sequencing decision, we simply skip the update of an entry if it has been updated already.

3. Joint Score and Priority Determination. To jointly consider the dynamic priorities of different factors, we also compute the joint score of an entry using a weighted sum of these scores s_k^j. The joint scoring function $F(j)$ for entry j is defined as

$$F(j) = \sum_k^K M_k^j s_k^j.$$

where M_k^j denotes the weight of the kth factor of entry j.

The weight of each factor is proportional to the similarity between its current value and the most similar value in the past time instances $t' \in T'$. More specifically, for factor k, we compute the average minimum distance m_k^j between the current value $COP_j.r_k$ and the most similar value $r_{ki}^{t'}$ of the same factor for $t' \in T'$.

$$m_k^j = \frac{1}{|T'|} \sum_{t' \in T'} \text{MinDistance}_{i<|\boldsymbol{r}_k^{t'}|}(r_{ki}^{t'}, COP_j.r_k), \forall k.$$

To avoid the noise from the different value range of the factors, the final weight M_k^j is also normalized across the distance m_k^j of k factors, i.e., $M_k^j = \frac{\max(m^j)-m_k^j}{\max(m^j)-\min(m^j)}, \forall k$. Finally, the priority is determined by ranking the entries with the joint score.

For interested readers, a brief evaluation of joint priority ordering is available in Appendix D.

6 PERFORMANCE EVALUATION

Experimental Setting. The total data collected from the BMS is more than 1 TB. We configure a private cloud to process the data for our experiments, with 16 cores of 2.6GHz CPU and a total memory of 64GB. We train the models with three-year data and predict with one-year data, which is a common setting in time-series data mining [18] and multi-task learning [19].

Baselines. To make the COP prediction, we employ the following state-of-the-art models as baselines.

- **Initial Profiling Model (IPM)** predicts the COP using the initial profile of chillers under different loads.
- **Thermodynamic Model (TDM)** predicts the COP using pre-calibrated thermodynamic model. Thermodynamic models [9] capture the thermodynamic process of chillers and try to obtain the chiller COP with fixed form given the chiller loading.
- **Data-driven COP Prediction Model (DPP)** predicts the COP by the data-driven approach which learns the model with historical data samples, where **DPP-Ada** denotes the approach using AdaBoost Regression as learning model; **DPP-SVR** denotes the approach using SVM Regression.

[2] Due to page limitations, the discussion of scoring under different cooling demands can be found in Appendix C.

- **Time-constrained Data-driven Profiling Model (T-DPP)** predicts the COP by data-driven approach under time constraints. As for the default approach, we adopt our proposed Joint Priority Ordering method to order entries.

To leverage the estimated COP, we employ the following state-of-the-art sequencing models as baselines.

(1) **Predefined Sequencing (PS)** conducts sequencing with predefined prediction model and starts backup chillers when it fails to meet the cooling demand [20, 21]. We adopt Thermodynamic Model as default predefined prediction model instead of Initial Profiling Model because it performs better.
(2) **Data-driven Chiller Sequencing (DCS)** conducts sequencing with DPP and predicts all profiles without considering any time-constraint. We adopt the most accurate DPP-Ada as default DPP model and use backup chiller for sequencing.
(3) **Time-constrained Data-driven Chiller Sequencing (T-DCS)** conducts chiller sequencing with DPP under time constraint. We adopt our Joint Priority Ordering method to select entries and start backup chillers when necessary.

Evaluation Metrics. For a sequencing method, the ability to provide credible energy saving is crucial to all stakeholders. Electricity is always the first concern, and we measure the Average Electricity Consumption (AvgEC), which measures the average electricity used by all sequencing operations on one day where all time instances in one day is denoted by T and each sequencing is conducted at time $t \in T$. Let L_i denote the maximum cooling capacity of chiller $i < n$. Formally,

$$AvgEC \quad = \quad \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n} L_i \cdot S_{i,t}/COP_{i,t},$$

where $COP_{i,t}$ denotes the real performance of chiller i at time t.

It is also significant that our decision should be accurate so that our final decision making can be reliable. Thus, we also measure the Accuracy, which indicates the similarity between our predicted COP and the real COP. Formally,

$$Accuracy \quad = \quad \frac{1}{T}\sum_{i<n}\frac{|COP_{i,t} - \hat{COP}_{i,t}|}{COP_{i,t}},$$

where $\hat{COP}_{i,t}$ denotes its predicted value.

Our decision should be conducted before the deadline, and we also measure the two metrics on time: Total Time and Run Time. Total Time indicates the total time over which sequencing was made, including the computation time and the mechanical switching time, to see whether the proposed operation can be done within time limitations. Run Time indicates the computation time of prediction models, and thus indicates the power of searching methods, to see whether all of the prediction tasks should be done. Formally,

$$Total\ Time = t_s - t_c, \quad Run\ Time = t_p - t_c,$$

where t_s denotes the time instant when the sequencing decision is made; t_p denotes the time instant when the predicted COP is known; t_c denotes the time when each experiment starts.

6.1 DPP and T-DPP Model

We compare the prediction results of our Data-driven COP Prediction (DPP-Ada and DPP-SVR) with that of Initial Profiling Model

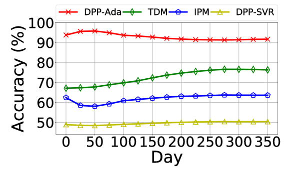


Figure 8: The accuracy as a function of day.

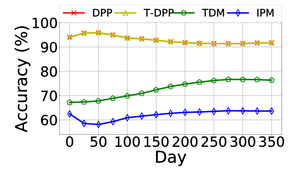


Figure 9: The accuracy of prediction models as a function of day.

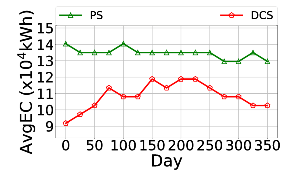


Figure 10: The average electricity consumption of days comparing DCS with PS.

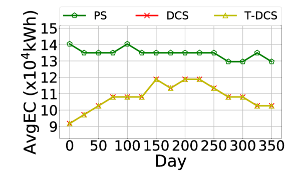


Figure 11: The average electricity consumption of days comparing T-DCS with DCS and PS.

(IPM) and Thermodynamic Model (TDM). Figure 8 shows that, DPP-Ada outperforms DPP-SVR, Thermodynamic Model, and Initial Profiling Model by 43.19%, 20.14%, and 30.77% respectively on average, which illustrates the prediction power of our approach. That is because, our data-driven approach is developed based on the runtime data in the real environment and leverages the ensemble technique to avoid overfitting in non-linear modeling, which can successfully capture the chiller local and dynamic performance.

In Fig. 9, we compare the Accuracy of Time-constrained DPP (T-DPP) with that of DPP (we adopt DPP-Ada as our default DPP approach due to its high accuracy), Initial Profiling Model and Thermodynamic Model for one year. Though T-DPP significantly reduces the computation time (will be shown later), we can see that our T-DPP is almost the same as DPP in terms of prediction accuracy, which outperforms Initial Profiling Model and Thermodynamic Model by 32.45% and 21.65%. That is because T-DPP reduces the computation time merely by selecting important tasks to conduct. For the selected prediction tasks, it leverages the state-of-the-art data-driven model, which ensures the accuracy and maintains the superiority of our data-driven techniques.

6.2 DCS and T-DCS Model

Result on Electricity Consumption. First, Figure 10 compares our Data-driven Chiller Sequencing (DCS) approach with Predefined Sequencing over days, in terms of Average Electricity Consumption. On average, our DCS outperforms Predefined Sequencing (PS) by 32.04%. In the day No.25, the improvement increases to 38.89%. That is because our data-driven method captures the performance dynamics of chillers and adjusts the cooling load in a smarter way. The Predefined Sequencing remains stable because its backup chiller mechanism is triggered frequently due to low prediction accuracy of COP and thus consistently provides more cooling power than needed.

Figure 11 shows Average Electricity Consumption as a function of day. Though our Time-constrained DCS (T-DCS) significantly reduces the computation time (will be shown later), it still performs almost the same as DCS, which always outperforms Predefined

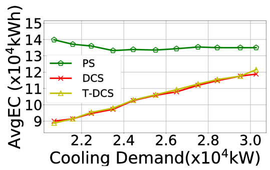


Figure 12: The average electricity consumption of cooling demands.

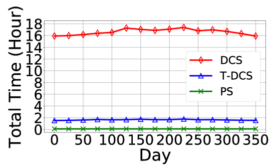


Figure 13: The total time as a function of day.

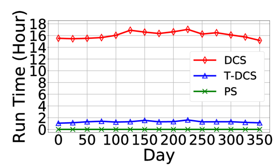


Figure 14: The run time as a function of day.

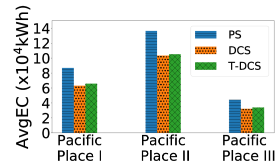


Figure 15: The average electricity consumption in multiple buildings.

Sequencing. That is because not all predictions on all operations are necessary. Our T-DCS captures the top important operations and still maintains the superiority of data-driven techniques.

From the cooling demand point of view, Figure 12 shows the changes in Average Electricity Consumption under different sequencing methods. We see that T-DCS is still almost the same as DCS and always outperforms Predefined Sequencing, which validates the performance of our approach as mentioned above. For example, T-DCS outperforms Predefined Sequencing by 24.56% at the average cooling demand of 26205.5 kW. Consistent with our intuition, as the cooling demand increases, the Average Electricity Consumption of DCS and T-DCS also gradually increases; while Predefined Sequencing still remains steadily high due to its limited prediction performance and backup chiller mechanism.

Result on Time We first compare the total time of the-state-of-the-art sequencing models. In Fig. 13, we compare the Total Time of T-DCS with that of Predefined Sequencing and DCS for one year under time limitation of 2 hours. We can see that T-DCS and Predefined Sequencing operations can be completed within the stipulated time except for DCS. It is mainly because DCS needs to update all the profile tables before making a sequencing decision, while our T-DCS only updates the most important sparse profile table, so it saves significant computation time.

To show the potential of saving time, we next compare the Run Time of T-DCS with that of Predefined Sequencing and DCS for one year. As we can see in Fig. 14, the average computation time of our T-DCS model is 1.6 hours, which is an improvement of 10 times over the DCS model. That is because, T-DCS uses Joint Priority Ordering to select the most important entries for prediction, unlike DCS. Though compared with Predefined Sequencing, it seems our T-DCS takes more time, but in return, we get a more accurate result as detailed in Section 6.1.

Result on Multiple Buildings In Fig. 15, we compare the Average Electricity Consumption of T-DCS with that of Predefined Sequencing and DCS for different buildings. As we can see, though T-DCS significantly reduces the computation time, the Electricity Consumption of T-DCS is still quite close to DCS. Compared with Predefined Sequencing, T-DCS saves 20980 kWh of Electricity Consumption in Pacific Place I, which is an improvement of 31.42%, respectively. In the remaining two buildings of Pacific Place II and Pacific Place III, the improvement is 31.63% and 30.98%, respectively. These results highlight the generality of our approach. When it comes to multiple buildings, our approach merely shares the similar training samples using multi-task learning and selects important entries under similar cooling demands, thus avoiding the noise when switching between different contexts and models.

7 RELATED WORK

Energy Intelligent Buildings is a widely studied smart-building research topic [4–6], especially for the major energy consumer of HVAC system. Works include controlling the HVAC based on the spatio-temporal profile of occupancy inside a building [4], cooling a building in advance of expected increase in occupancy, also known as pre-cooling [5], and incorporating renewables such as solar panels and battery storage into the energy mix [6]. Recently, data-driven techniques are also introduced to capture occupant and device behavior in this research topic [4, 22–26]. In [4], rich sensor data is leveraged to predict occupancy and benefit setpoint control in HVAC system to save energy. Occupant feedback is also further considered in [22], where a joint model was developed for feedback and individually comfort learning. These scenarios usually assume the data to be sufficient for training.

Chiller Sequencing refers to operating the most efficient combination of chillers in a building at (near) real-time to meet the time-varying cooling demand. Previous studies mainly focused on developing reliable and robust sequencing according to instantaneous building cooling load [15, 27, 28]. These studies mainly set the chiller cooling capacity as constant (i.e., the same as the rated cooling capacity) [29]. Considering that the cooling capacity may vary under different operating conditions (e.g., different chiller evaporating and condensing temperatures), such approaches may fail to provide enough cooling energy or lead to extra energy usage [30]. As a solution, first, thermal energy storage was leveraged to improve the COP of multiple chiller plants [31]. Second, physical and grey box have been used to capture the variation in maximum cooling capacity given different operating conditions [30, 32]. General model is also proposed calibrated using real data from water plants [33–35] and centrifugal chillers [36]. However, the actual performance of these sequence control strategies is subject to the accuracy of these models because they are general purpose models and do not capture the practicalities that come with deployments in different building conditions and time-varying cooling demands. Worse still, when conducting data-driven sequencing, they also do not consider the time limitations, e.g., minimum start-stop-start time of chillers. For the first time, we introduce a novel data-driven chiller sequencing framework that also captures the need to perform chiller sequencing under time constraints. It provides dynamic cooling performance estimation given a set of possible cooling loads, other configurations and environmental settings, and conducts sequencing under practical time limitations.

Time issue is always an important problem in traditional complex system research, and now may need to be re-thinked when time-costly machine learning introduced in decision making [37, 38], especially on the edge of the network [39]. An example is SenseNetworks [40], a recent U.S.-based startup company, which uses millions of GPS estimates sourced from mobile phones within a city to predict the place where the people would be interested. In such a case, the computation time on the phone should be tackled carefully to meet user requirement. In practice, there are also quite a few works trying to reduce the data-driven computation time using distributed machine learning [37, 41, 42]. Admittedly, by increasing the total computation power, such methods work well in reducing data-driven computation time. However, they can still suffer from problems like higher computation cost, higher data transmission and sharing cost, or even data privacy concerns. There are also methods re-designing the algorithm for speed up, most of which are conducted with trade-off between computation time and accuracy [38]. Our idea, in this paper, is based on the observation that not all operations are valuable for prediction. Focusing on electricity minimization, we show that it is possible to cut down the computation time by reducing the number of less-important prediction tasks, rather than sacrificing prediction accuracy or incurring high computation cost, which sheds some new light on time-aware machine learning in decision making.

8 DISCUSSION

Accuracy and Energy Consumption Prediction accuracy generally affects energy consumption in our T-DCS sequencing, because additional operations, e.g., backup chillers, should be launched to fix the problem to meet the required cooling demand. Theoretically, it seems that using backup chiller does not necessarily lead to increase in overall energy consumption, when they are used to satisfy the exact cooling demand. However, generally, starting additional backup chillers in practice usually increases the energy consumption under perfect prediction due to the following reasons: 1) Using additional chillers and pumps in general leads to more energy consumption, because launching and maintaining additional chillers usually takes more energy than increasing the load on already-operating chillers when meeting the same cooling load. 2) Backup chillers are practically run in an over-provisioned manner, e.g., on the highest load, in order to ensure the required cooling demand and avoiding further adjustment, instead of meeting exactly the demand. Such over-provisioning leads to energy waste.

However, there are also other cases when inaccurate prediction may not lead to inefficient operation. For example, as mentioned in our paper, some operations may never be used in chiller sequencing optimization. The prediction accuracy of such unimportant operations may not affect the final decision. In this paper, we accordingly propose the dominant graph techniques. It would be an interesting future work to further investigate on the difference of data-driven techniques between industrial (focusing on industrial objective) and traditional (focus on merely accuracy objective) prediction.

Chiller Type In this paper, we focus on merely water-cooled chillers. As for water-cooled chillers, we first take into account the inlet and outlet water temperature. We also leverage the dry-bulb temperature (DBT) which reflects the cooling demand and usage for performance prediction on the chillers. Future work can include the feature design for different types of chillers. For example, the DBT can also be quite related for air-cooled chillers due to their heat transfer with outdoor air.

Sensor Missing The sensing data required by the COP computation are all accessible in our system. However, such an assumption may not be true in other systems where some of the required sensors, e.g., mass flow rates sensors, may not be not available. In that case, cooling load can be estimated indirectly by inverse physical models based on the power consumption of chiller motors, see [43].

Learning and Prediction Frequency The data samples of all operations are likely to be sparse, the prediction model should be updated frequently to adapt to the accumulated samples under demands, configurations and degradation overtime. To maintain high performance of our model, in this paper, we update the parameters each time before a chiller sequencing operation. It would be interesting future work to investigate the optimal learning and prediction frequency in data-driven industrial operations.

Multiple-time Chiller Sequencing In this paper, we focused on single-time chiller sequencing, i.e., without the consideration and cooperation of other sequencing at different times. Possible future work can include the consideration of minimum annual runtime of chillers for multiple-time sequencing, or minimum chiller utility for each single-time sequencing operation.

9 CONCLUSIONS

Developing energy efficient buildings has long been an important research topic as facility managers grapple with the problem of reducing their building’s electricity bills. In this paper, we focused on one of the core problems in building operation, namely HVAC chiller sequencing, and made the following contributions.

First, we demonstrated that using chiller COP values from initial profiles can be detrimental from the point of view of HVAC electricity consumption. We subsequently stressed the need for quantifying COP at run-time.

Second, we showed that predicting COP accurately is a challenging problem, requiring considerable computation time and hardware resources. To provide a practical solution, we developed a sequencing framework alongside a time-constrained approach for COP prediction, which opens the doors for HVAC electricity consumption reduction while enabling ease of use of the scheme for real-world deployment.

Finally, we evaluated the performance of our solution by applying it to BMS data, spanning 4 years, obtained from multiple chillers across 3 large commercial buildings in Hong Kong. We showed that our solution can save over 30% of HVAC electricity consumption compared to the current mode of chiller operation in the buildings. We believe that sequencing chillers using a data-driven approach for COP prediction offers a simple and effective mechanism for reducing the electricity consumption associated with operating the HVAC in large commercial buildings.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Abraham Lam and Adrian from Mega automation Ltd. for their valuable discussion and feedback. Parts of the work were supported by the Hong Kong Polytechnic University under Grant No.: 1-BBYX, and by the School of Computer Science and Technology, HUST, with the National Key Research & Development (R&D) Plan under Grant No.: 2017YFB1001703, and the National Natural Science Foundation of China under Grant No.: 61761136014 and 61722206.

APPENDIX

A POSSIBLE BENEFIT OF ACCURATE COP PREDICTION

Currently, in practice, facility managers often perform COP estimation using initial profile. There are two issues with current COP estimation schemes: 1) not all loads have been tested in the initial profiling period, and so COPs at these loads will be missing. A consequence is that these loads will never be used in the sequencing algorithm; 2) for the COPs with data, a simple averaging approach is often used. A consequence is that these COPs may be largely inaccurate and should not be used for making sequencing decisions.

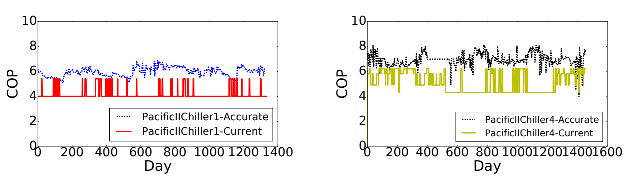
In this section, we demonstrate that there can be a substantial reduction in chiller electricity consumption when sequencing is performed with accurate COP prediction.

We now compute the electricity consumption using chiller sequencing under the current COP estimation scheme, which is based on the initial profiles, and compare it against a scheme that estimates COP accurately assuming there exists an oracle that can determine these values. Such an oracle can be obtained by computing COP using historical data (Section 2.2), and can be regarded as ground truth. Clearly, this is not a fair comparison, but it shows the benefit that comes with improving COP prediction.

Since the current chiller sequencing mechanism uses COPs that are inaccurate, it is possible that the cooling load Q provided by the chillers fails to satisfy the actual cooling demand $\mathcal{D}$. In practice, this is usually addressed by starting backup chillers immediately. [3]

A case study is conducted based on Pacific Place II, which contains the most complete chiller data amongst the three buildings. A recent COP profile for Pacific Place II is shown in Table 2. The COP matrix is indeed sparse, confirming that each chiller is routinely operated at only a few distinct loads.

To conduct sequencing in Pacific Place II, COP estimation is needed for all the entries in the table. In Fig. 16, we start by comparing the predicted COP in the current mode of operation (bottom curve) against the accurate oracle scheme (top curve). The curves depict chiller 1 (left) and chiller 4 (right) in Pacific Place II. We see that the current mode of operation often under estimates the COP. There is a high estimation error of over 35% and has little correlation with the top curve.



(a) (b)

Figure 16: The accurate COP and estimated COP with present scheme on different days in Pacific Place II for two types of chillers (a) Chiller 1; (b) Chiller 4.

In Fig. 17, we compare the electricity consumption between 2012 and 2015 for each of the two chillers. Stacked bars on the right indicate the current mode of operation while bars on the left show what the consumption would have been if COP had been predicted accurately. We see that the latter could have resulted in lowering electricity consumption by over 45% on average, with nearly 60% reduction coming in 2015. These results demonstrate that there is a significant room to reduce chiller electricity consumption when more accurate, and robust, COP prediction schemes are used.

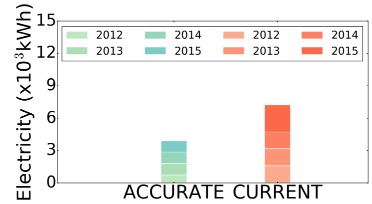


Figure 17: Electricity consumption of sequencing with ACCURATE and CURRENT schemes, from 2012 to 2015.

Table 5: Accuracy of Learning Methods.

Learning Model	Learning Method	Accuracy
AdaBoost [500 Trees]	Single Task	54.22%
	Independent Multi-task	73.15%
	Clustered Multi-task	94.03%

B A BRIEF EVALUATION ON PREDICTION

With the clustered training samples, a learning model can then be trained and predicted. We train our model with the first-three-year data and predict the COP of the last year with the given feature. We apply the evaluation metric of *Accuracy*, i.e., $1/T \sum_{t<T} ||F(\boldsymbol{X}_t, \boldsymbol{W}) - \boldsymbol{COP}_t||_2$, which is the average similarity between the ground truth and the estimated value in the last year.

To evaluate the performance of CMTL, we compare with Single Task Learning which learns a single model by pooling together data from all entries, and Independent Multi-task Learning which learns each entry independently without sharing any instance or knowledge. In CMTL, AdaBoost under 500-Trees setting in Python scikit-learn is adopted as the default learning model due to its high accuracy, which will be also shown in our experiments next. The result with different learning methods is shown in Table 5. CMTL approach significantly outperforms all other methods. That is because (1) it can better capture the different thermodynamic models of different entries than a Single Task Learning method, and (2) it enables the knowledge sharing among entries and reduces the negative effects due to the little training data in most entries, other than Independent Multi-task Learning.

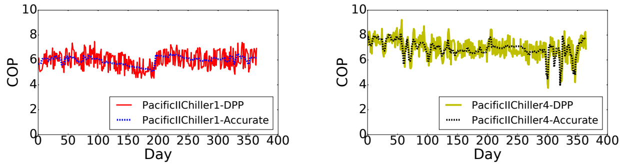
In Fig. 18, we show both the measured COP and the estimated COP with our DPP. The results are shown for one chiller from each type in Pacific Place II. We see that our approach well matches the value and the trend of the real COP, with an average RMSE of 0.52.

We also show result with different learning models $F(\cdot)$ in Table 6. We found that the ensemble approach like AdaBoost (1) can better capture the non-linearity than linear regression, and (2) are less likely to become over-fitted other than support vector regression on large datasets, due to the model combination nature of AdaBoost.

C DISCUSSION ON PRIORITY OF FACTORS

Priority under Different Cooling Demands. To order the entries and obtain the priority, we need to decide the four priorities for

[3] It is possible to over-provision the cooling loads by a certain degree. There is a trade-off between over-provisioning and the fail-over for backup chillers. This problem is orthogonal to ours and we do not consider it in this paper.



(a) (b)

Figure 18: The accurate COP and the estimated COP with DPP scheme on different days in Pacific Place II for two types of chillers of (a) Chiller 1; (b) Chiller 4.

Table 6: Learning Models: Computation Time and Accuracy.

Learning Model for Prediction	Time per Entry	Accuracy
Linear Regression	< 1 second	46.22%
Support Vector Regression	1 minute	51.67%
AdaBoost [200 Trees]	3 minutes	82.14%
AdaBoost [400 Trees]	11 minutes	90.14%
AdaBoost [500 Trees]	20 minutes	94.03%

the four factors. An important observation here is that the priority for the factors changes with different cooling demands. Figure 19 shows that the cooling demand varies considerably from 20058 kW to 30483 kW over time, mainly due to the meteorological status. When the cooling demand is high, an entry with a low cooling load may not meet the cooling demand and thus is not likely to be selected in the optimal solution, while it may be suitable when the cooling demand is low because of the low electricity consumption.

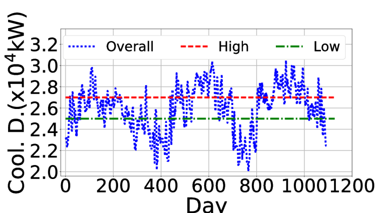
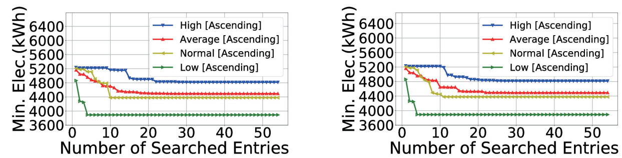


Figure 19: The cooling demand as a function of day.

Under different cooling demands, Figure 7 shows minimum electricity consumption with the historical data in the past four years, with respect to the number of entries visited in searching procedure using the priority decided by Electricity Consumption and Cooling Load. We can see from the figure that the entry priority is changed under different cooling demands. For example, searching with Electricity Consumption (in the ascending order) under low cooling demands (bottom 33% of all cases) further reduces 87.50% of the total search space under high cooling demands (top 33% of all cases); prioritizing the entries by Cooling Load (in the ascending order) under low cooling demands reduces 85.71% of the search space under high cooling demands. The Past Success and Prediction Accuracy are independent of the cooling demand and thus do not need to be treated differently under different cooling demands.

Because the priority is influenced by the cooling demand, it will be decided according to the past priority with similar cooling demands.

One possible solution is to divide the historical data into several coarse-grained subsets, e.g., High, Normal and Low cooling demand,



(a) (b)

Figure 20: Under High, Normal, Average and Low cooling demands: minimum electricity consumption under the number of searched entries, ranked by (a) Electricity Consumption and (b) Cooling Load.

and use the historical data in the corresponding subset to decide the current priority. However, this can be very biased for entries close to the boundaries of the subsets. For example, suppose we define a cooling demand above 27095.8 kW to be of High Cooling Demand. When the current cooling demand is 27100 kW, it will be assigned with a similar subset of High Cooling Demand. However, a past time instant in the Normal Cooling Demand, e.g., with cooling demand of 27000 kW, can be more similar with the current time instant of 27100 kW, other than another instant in the High Cooling Demand, e.g., 30000 kW.

The factors of electricity consumption and cooling load can vary significantly with the given cooling demand. Let $\boldsymbol{T}$ denote the set of time instants in the history. The priority is obtained using merely a subset of time instances $\boldsymbol{T'} \in \boldsymbol{T}$ with similar cooling demands. Let $\mathcal{D}^{\boldsymbol{T}}$ denote the cooling demands at past time instances in $\boldsymbol{T}$.

To tackle bias problem mentioned above, a subset of time instances $\boldsymbol{T'}$ can be computed by clustering algorithms such as *k Nearest Neighbors* (kNN) [44], i.e., $\boldsymbol{T'} = kNN(\mathcal{D}, \mathcal{D}^{\boldsymbol{T}}, \boldsymbol{T})$. For example, when we are clustering for a time instant of cooling demand 27500 kW, the resulting time instances in the cluster will be around 27500 kW.

Then, the priority can be inferred from similar past time instances $t' \in \boldsymbol{T'}$. Let $P_j.r_k$ denote the value of entry j considering the kth factor at the present time point. Let $f^j_{kt'}$ denote the rank using the kth factor in entry j at a past time instant t'. Then, we have the priority computation at a past time instant t':

$$f^j_{kt'} = \arg\min_i (P_j.r_k - r^{t'}_{ki}).$$

Let $j \in J$ denote an entry. Then, for each factor k of entry j, we turn its value into a score s^j_k by considering the ranking of the value. The s^j_k is computed by:

$$s^j_k = \frac{1}{|\boldsymbol{T'}|} \sum_{t' \in \boldsymbol{T'}} (|\boldsymbol{r}^t_k| - f^j_{kt'})/|\boldsymbol{r}^t_k|.$$

Note that the scores computed above are normalized and have values in $[0, 1]$.

D A BRIEF EVALUATION OF JOINT PRIORITY ORDERING

With the joint priority, we can decide the order of entries to conduct prediction and sequencing. Within the deadline of two hours, different orderings provide different probability to output the optimal solution. Such a probability can be obtained by comparing the output sequencing and the optimal sequencing for a period of time.

Table 7: The probability to find optimal solution using different factors within deadline of 2 hours.

Factor	Prob. of Finding Optimal
Electricity Consumption	23.4%
Cooling Load	23.6%
Past Success	53.1%
Prediction Accuracy	52.2%
Joint Priority	94.5%

We train our model using data from 2012 to 2014 and evaluate such a probability with data in the year 2015. Table 7 shows the result. We clearly see that (1) our joint priority ordering raises the probability of finding the optimal by about 70% than using just Electricity Consumption and Cooling Demand; (2) it also raises the probability by 40% than using Past Success and Prediction Accuracy. That is because our joint priority ordering captures the dynamic priorities of different factors, especially under different cooling demands.

In Fig. 21, we compare the average electricity consumption of sequencing using these single factors. We see that (1) our proposed joint priority method outperforms the ordering using factors of Electricity Consumption and Cooling Load by over 20%; (2) it also outperforms the ordering using factors of Past Success and Prediction Accuracy by about 10%. In the 150th day, the improvements raise to about 40% and 25%. That is because our joint priority ordering benefits the optimal solution searching within the deadline.

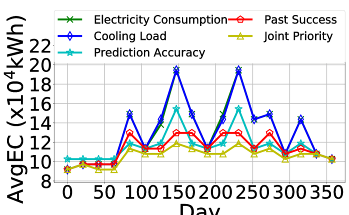


Figure 21: The average electricity consumption ordering with different factors under days.

REFERENCES

[1] Heating, Ventilation and Air Conditioning. http://eex.gov.au/technologies/heating-ventilation-and-air-conditioning/, 2018.
[2] Luis Pérez-Lombard, José Ortiz, and Christine Pout. A review on buildings energy consumption information. *Energy and buildings*, 40(3):394–398, 2008.
[3] Property Council of Australia. Survey of Operating Costs: Office Buildings. https://goo.gl/7oT1W7, 2018.
[4] Wilhelm Kleiminger et al. Smart heating control with occupancy prediction: How much can one save? In *Proc. ACM UbiComp'14*, pages 947–954.
[5] Arun Vishwanath et al. A data driven pre-cooling framework for energy cost optimization in commercial buildings. In *Proc. ACM e-Energy '17*, pages 157–167.
[6] Nair et al. Battery energy storage systems: Assessment for small-scale renewable energy integration. *Energy and Buildings*, 42(11):2124–2130, 2010.
[7] Zhaohui Liu, Hongwei Tan, Duo Luo, Guobao Yu, Jin Li, and Zhenyu Li. Optimal chiller sequencing control in an office building considering the variation of chiller maximum cooling capacity. *Energy and Buildings*, 140:430–442, 2017.
[8] Wikipedia. Coefficient of performance. https://en.wikipedia.org/wiki/Coefficient_of_performance, 2018.
[9] Kody M Powell, Wesley J Cole, et al. Optimal chiller loading in a district cooling system with thermal energy storage. *Energy*, 50:445–453, 2013.
[10] Nofirman Firdaus et al. Chiller: Performance deterioration and maintenance. *Energy Engineering*, 113(4):55–80, 2016.
[11] FW Yu and KT Chan. Optimization of water-cooled chiller system with load-based speed control. *Applied Energy*, 85(10):931–950, 2008.
[12] Yundan Liao, Yongjun Sun, and Gongsheng Huang. Robustness analysis of chiller sequencing control. *Energy Conversion and Management*, 103:180–190, 2015.
[13] Thomas Hartman. All-variable speed centrifugal chiller plants. *ASHRAE Journal*, 56(6):68–79, 2014.
[14] A Michopoulos et al. Three-years operation experience of a ground source heat pump system in northern greece. *Energy and Buildings*, 39(3):328–334, 2007.
[15] Yongjun Sun, Shengwei Wang, and Fu Xiao. In situ performance comparison and evaluation of three chiller sequencing control strategies in a super high-rise building. *Energy and buildings*, 61:333–343, 2013.
[16] Laurent Jacob, Jean-philippe Vert, and Francis R Bach. Clustered multi-task learning: A convex formulation. In *Advances in neural information processing systems (NIPS)*, pages 745–752, 2009.
[17] Lei Zou and Lei Chen. Dominant graph: An efficient indexing structure to answer top-k queries. In *Proc. IEEE ICDE'08*, pages 536–545.
[18] Yongxin Tong, Yuqiang Chen, Zimu Zhou, Lei Chen, Jie Wang, Qiang Yang, Jieping Ye, and Weifeng Lv. The simpler the better: a unified approach to predicting original taxi demands based on large-scale online platforms. In *Proc. ACM SIGKDD'17*, pages 1653–1662.
[19] Jaime Carbonell Keerthiram Murugesan. Self-paced multitask learning with shared knowledge. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 2522–2528, 2017.
[20] FW Yu and KT Chan. Economic benefits of optimal control for water-cooled chiller systems serving hotels in a subtropical climate. *Energy and Buildings*, 42(2):203–209, 2010.
[21] Jijun Zhou, Guanghua Wei, W Dan Turner, Song Deng, David E Claridge, and Oscar Contreras. Control optimization for a chilled water thermal storage system under a complicated time-of-use electricity rate schedule. 2005.
[22] Abraham Hang-yat Lam, Yi Yuan, and Dan Wang. An occupant-participatory approach for thermal comfort enhancement and energy conservation in buildings. In *Proc. ACM e-Energy'14*, pages 133–143.
[23] Tuan Anh Nguyen and Marco Aiello. Energy intelligent buildings based on user activity: A survey. *Energy and buildings*, 56:244–257, 2013.
[24] Luis M Candanedo and Véronique Feldheim. Accurate occupancy detection of an office room from light, temperature, humidity and co 2 measurements using statistical learning models. *Energy and Buildings*, 112:28–39, 2016.
[25] Dwi Rahayu, Balakrishnan Narayanaswamy, Shonali Krishnaswamy, Cyril Labbé, and Deva P Seetharam. Learning to be energy-wise: Discriminative methods for load disaggregation. In *Proc. ACM e-Energy'12*, page 10.
[26] Hoang Dung Vu et al. Data driven chiller plant energy optimization with domain knowledge. In *Proc. ACM CIKM'17*, pages 1309–1317.
[27] Yundan Liao et al. Uncertainty analysis for chiller sequencing control. *Energy and Buildings*, 85:187–198, 2014.
[28] Sen Huang, Wangda Zuo, and Michael D Sohn. Amelioration of the cooling load based chiller sequencing control. *Applied Energy*, 168:204–215, 2016.
[29] Pe Marchall Seymore. Simplified chiller sequencing for a primary/secondary variable chilled water flow system. *ASHRAE Journal*, 56:24–32, 2014.
[30] Yongjun Sun et al. Chiller sequencing control with enhanced robustness for energy efficient operation. *Energy and buildings*, 41(11):1246–1255, 2009.
[31] Madhur Behl, Truong X Nghiem, et al. Green scheduling for energy-efficient operation of multiple chiller plants. In *IEEE RTSS'12*, pages 195–204.
[32] Wei Jiang et al. General methodology combining engineering optimization of primary hvac&r plants with decision analysis methods - part ii: Uncertainty and decision analysis. *HVAC&R Research*, 13(1):119–140, 2007.
[33] Mark Hydeman, Nick Webb, Priya Sreedharan, and Steve Blanc. Development and testing of a reformulated regression-based electric chiller model/discussion. *ASHRAE Transactions*, 108:1118, 2002.
[34] Mark Hydeman et al. Tools and techniques to calibrate electric chiller component models/discussion. *ASHRAE transactions*, 108:733, 2002.
[35] Danielle Monfet and Radu Zmeureanu. Identification of the electric chiller model for the energyplus program using monitored data in an existing cooling plant. In *Proceedings of the international IBPSA conference. Sidney, Australia: International Building Performance Simulation Association*, 2011.
[36] JM Gordon et al. Centrifugal chillers: thermodynamic modelling and a diagnostic case study. *International Journal of refrigeration*, 18(4):253–257, 1995.
[37] Mu Li, David G Andersen, Alexander J Smola, and Kai Yu. Communication efficient distributed machine learning with the parameter server. In *Advances in Neural Information Processing Systems*, pages 19–27, 2014.
[38] Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2-3):209–232, 2002.
[39] Abbas Kiani and Nirwan Ansari. Toward hierarchical mobile edge computing: An auction-based profit maximization approach. *IEEE Internet of Things Journal*, 4(6):2082–2091, 2017.
[40] Panu Korpipaa et al. Managing context information in mobile devices. *IEEE pervasive computing*, 2(3):42–51, 2003.
[41] Evan R Sparks, Ameet Talwalkar, et al. Mli: An API for distributed machine learning. In *Proc. IEEE ICDM'13*, pages 1187–1192.
[42] Yucheng Low et al. Distributed graphlab: a framework for machine learning and data mining in the cloud. *Proc. the VLDB Endowment*, 5(8):716–727, 2012.
[43] Gongsheng Huang, Shengwei Wang, Fu Xiao, and Yongjun Sun. A data fusion scheme for building automation systems of building central chilling plants. *Automation in Construction*, 18(3):302–309, 2009.
[44] Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In *ICDT*, pages 217–235. Springer, 1999.