Advanced Engineering Informatics 73 (2026) 104579

Contents lists available at ScienceDirect

Advanced Engineering Informatics

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

Full length article

Rethinking keys in data-driven short-term building energy predictions: Towards standardized methods for data preparation, model training and evaluation ☆

Cheng Fan [a,b,c], Enqi Shen [c], Da Yan [d], Jinhan Mo [a,b,*]

[a] State Key Laboratory of Subtropical Building and Urban Science, Shenzhen University, Shenzhen, China
[b] Key Laboratory of Coastal Urban Resilient Infrastructures, College of Civil and Transportation Engineering, Shenzhen University, Shenzhen, China
[c] Sino-Australia Joint Research Center in BIM and Smart Construction, Shenzhen University, Shenzhen, China
[d] Building Energy Research Center, School of Architecture, Tsinghua University, Beijing, China

ARTICLE INFO

Keywords:
Building energy prediction
Time series forecasting
Artificial neural network
Data-driven method
Data science

ABSTRACT

Accurately describing temporal dynamics in building energy patterns is crucial for optimizing real-time building operation performance. The wide availability of building operation data and the recent advances in data science have greatly encouraged the development of data-driven solutions for short-term building energy predictions on an hourly, daily or weekly basis. Despite the vast amount of research in the building field, distinguishing truly competitive ones is challenging due to varying characteristics of building data, inconsistent data-driven modeling procedures, and different evaluation measures used. Leveraging actual measurements collected from 32 buildings, this study explores the impact of various practices in developing data-driven models for 24-hour ahead building energy predictions. Data experiments have been designed to answer eight questions across three major aspects (i.e., data preparation, model training, and performance evaluation) and to illustrate potential pitfalls in common practices and biases in quantifying the practical value of data-driven solutions. This study aims to raise awareness among building researchers and practitioners of best practices for short-term building energy prediction tasks. The results obtained are helpful for standardizing analytical procedures and further enhancing the reference values of academic research in the building field.

1. Introduction

The urgent need for sustainable and low-carbon buildings necessitates more efficient and intelligent building operation methods. Modern buildings are typically equipped with comprehensive sensor networks that enable systematic data collection for building performance monitoring and controls. While high-frequency data at the second or minute level is increasingly available, sub-hourly or hourly data remains the most common resolution for building energy prediction tasks due to the alignment with utility metering practices, operational decision-making timescales, and the availability of public benchmark datasets [1,2].

As a prominent example, short-term building energy predictions, which typically describe energy patterns over the next few hours or days, have attracted considerable interest from both academic researchers and industrial practitioners, as they are closely related to daily building operations, such as real-time anomaly detection and control optimization [3,4]. Over the past decades, a broader paradigm shift has been observed in the building field, evolving from using small datasets for grey-box model parameterization to leveraging large datasets of building operations for complex machine learning model development [5,6]. Existing studies have proved the advantages of machine learning techniques in prediction accuracies [7,8]. For instance, ensemble models, which may adopt either bootstrapping or bagging for establishing base models, have been widely used due to their technical superiority on accuracies and robustness against overfitting [9,10]. Artificial neural network-based methods have been developed considering their great extensibility in model architectures and compatibility with various tasks, such as informative data sample identification through active learning [11], fault detection and diagnosis through semi-supervised graph convolutions [12], knowledge transfer and collaborative model training through model weight sharing and fine-

☆ This article is part of a special issue entitled: 'ADVEI_LLMs in Energy' published in Advanced Engineering Informatics.

* Corresponding author..
E-mail address: mojinhan@szu.edu.cn (J. Mo).

https://doi.org/10.1016/j.aei.2026.104579
Received 4 November 2025; Received in revised form 1 March 2026; Accepted 7 March 2026
Available online 10 March 2026

Nomenclature

MAE	Mean absolute error
RMSE	Root mean square error
CV-RMSE	Coefficient of variance of the root mean square error
RPG	Relative performance gap
GRU	Gated recurrent unit
CONV	One-dimensional convolutional neural network
TRAN	Transformer-based neural network
GBM	Gradient boosting machine
ARIMA	Autoregressive integrated moving average
ADF	Augmented Dickey-Fuller test

tuning [13,14], and synthetic data generation through generative learning [15,16]. Recent surveys have systematically examined the deep learning architectures for time series forecasting, highlighting the evolution from traditional statistical methods to advanced approaches using convolutional, recurrent, and transformer-based architectures [17,18]. The recent development of both model-agnostic and model-specific interpretation methods for neural networks further enhances the popularity of artificial neural network-based solutions in data-driven modeling [19,20]. Several studies have assessed evaluation methods for building energy predictions and identified common pitfalls in time-series modeling workflows [21,22]. Benchmark studies, including the seminal work from the ASHRAE Great Energy Predictor III competition, have provided valuable insights into model performance across diverse building portfolios [23]. Researchers have realized that the selection of appropriate baseline methods is the key to rigorous model performance comparison [24].

Despite these advances, it remains challenging to identify truly promising and competitive data-driven solutions for analyzing building energy data. The main reasons are two-fold. Firstly, most studies used one or two specific buildings for analysis and therefore, the conclusions drawn may not be generalizable as buildings may have unique operation patterns and different levels of intrinsic predictabilities [25]. For instance, predicting the energy patterns of buildings on a smaller scale can be more difficult as they are more sensitive to random changes in individual occupant schedules and activities. Secondly, there is a lack of standardized analytical pipelines in the building field [26]. As a result, different studies may adopt varying approaches in key tasks such as data preprocessing and model optimization, making it almost impossible to make direct comparisons. As an example, the performance of data-driven models is highly sensitive to their hyperparameters. Considering the varying efforts in model optimization, the results reported may not reflect the optimal performance of the methods proposed.

The abovementioned challenges can be resolved from two perspectives. The first is to evaluate the performance of data-driven models proposed using a fixed open dataset of many testing buildings [27]. Researchers have made efforts in collecting and preparing such datasets on building energy patterns [28,29], yet it is still at a preliminary stage, and most open datasets are collected in the U.S., U.K., Ireland, and Singapore [30]. The second is to correctly understand the impacts and consequences of various analytical procedures and thereby reach a consensus on the best data-driven modeling practices [31]. It should be mentioned that the rapid advancement of large language models (LLMs) has provided powerful tools for building energy research. LLMs have demonstrated great capabilities in automating complex engineering workflows, such as building energy model generation [32,33], energy management optimization [34], and fault detection and diagnosis [35]. In the context of data-driven energy predictions, LLMs are increasingly being used as intelligent agents that can autonomously design data preprocessing pipelines, select predictive modeling algorithms, tune data-driven model hyperparameters, and interpret prediction results in natural language [36,37]. A recent survey identified 13 distinct roles that LLMs can play in energy-related applications, ranging from data analysts and modelers to predictors and advisers [38]. However, the effectiveness of LLM-driven energy analytics fundamentally depends on the quality and consistency of the underlying data-driven modeling practices. If the data preparation, model training, and evaluation procedures fed into or followed by LLMs are not standardized, the results obtained may not be correct or justified. As an example, as we shall demonstrate in the latter part of this study, an LLM agent which defaults to random data partitioning will produce optimistic accuracy estimates without being aware of the underlying data leakage problem. Such concerns further motivate this study to empirically quantify how different methods affect the data-driven building energy prediction performance. The findings can serve as domain-specific knowledge to guide the process of LLM-assisted building energy prediction tasks.

While existing reviews have comprehensively surveyed data-driven building energy prediction methods, and benchmark competitions like the ASHRAE Great Energy Predictor III [23] have compared model performance across buildings, there remains a gap in quantifying the practical impacts of methodological choices in the building energy domain. The gap is becoming increasingly consequential as LLMs are being adopted to automate data-driven energy analytics pipelines and proper guidance should be provided to justify the validity of LLM-generated workflows and results. This study addresses this gap through comprehensive data experiments using hourly measurements from 32 buildings. The contributions are three-fold. Firstly, the impact of different data preparation, modeling, and evaluation practices will be quantified, providing useful references for analytical pipeline designs. Secondly, evidence-based recommendations for standardized analytical pipelines will be derived to ensure reproducibility and fair cross-study comparisons in the building field. Thirdly, a structured methodology framework is proposed to serve as a knowledge base for LLM-driven building energy analytics, enabling the adoption of best practices for automated building energy predictions.

The remaining contents are organized as follows. Section 2 introduces the theoretical basics and data experiment setups. The following three sections illustrate data experiment results on how different data preparation, modeling training and evaluation methods may affect the overall analysis. Sections 6 and 7 serve as discussions and conclusions, respectively.

2. Theoretical basics and data experiment setups

2.1. General procedures for pointwise building energy predictions

This study focuses on pointwise building energy predictions in a fixed window-wise context, where data-driven models are established to use historical measurements of the previous p timesteps (i.e., denoted as X_{t-p+1}, X_{t-p+2},…, X_{t-1}, X_t) to predict the values for the next f steps (i.e., denoted as Y_{t+1}, Y_{t+1},…, Y_{t+f}). The problem is defined as a one-step ahead prediction task when f equals one and a multi-step ahead task otherwise. In essence, such a context corresponds to a rolling-window setup, where the original multivariate time series needs to be segmented into window-wise data samples of length $p + f$. It should be noted that this study focuses on pointwise predictions, and data-driven methods to formulate prediction or confidence intervals are out of the scope of this study.

As shown in Fig. 1, the general data-driven procedures for building energy prediction can be summarized as follows. The building time series data are typically recorded as a two-dimensional matrix, where each row and column represent time steps and monitoring variables respectively. A format transformation is needed to prepare the original data into suitable data samples for predictive modeling. Expanding and rolling windows are two of the most widely used setups for time series forecasts, i.e., the former has an increasing input window size as more data become available, while the latter has a fixed input window size.

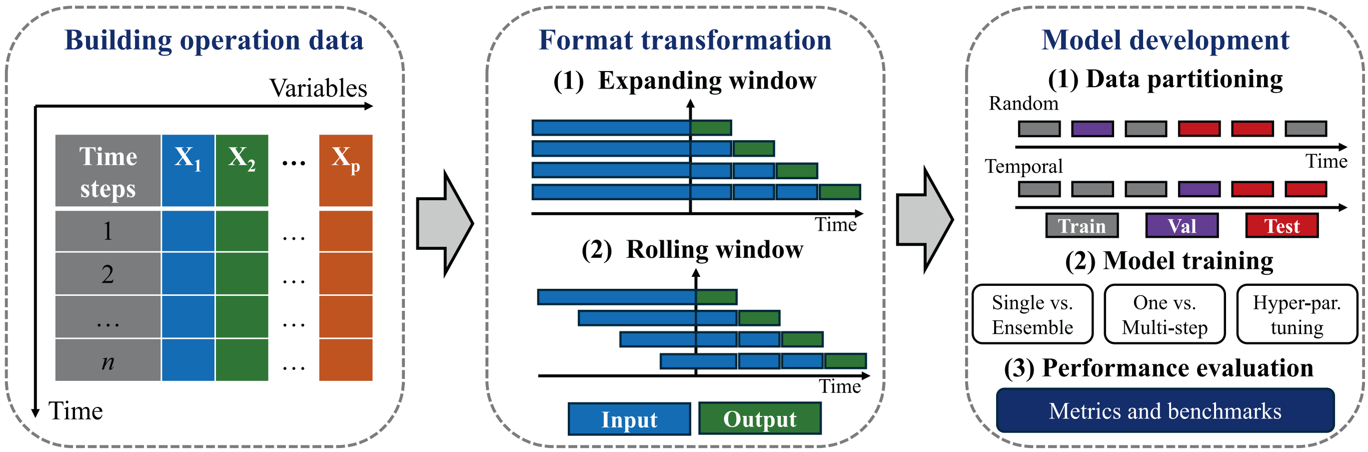

Fig. 1. General procedures for data-driven building energy predictions.

Afterwards, the data samples are partitioned into training, validation and testing datasets using either random or temporal strategies. Different methods can then be applied to train and optimize data-driven models, which may vary in their model complexity and temporal prediction capabilities. Finally, the generalization performance of data-driven models is evaluated based on accuracy metrics and comparisons with benchmark methods.

Unlike non-time series data analysis, the analytical procedures on time series data have more variants in data preparation, modeling and evaluation methods, making it difficult for fair comparisons across different solutions. Taking data partitioning as an example, both random and temporal data partitioning strategies can be applied. Even though existing studies have shown that both methods are valid from mathematical points [39], the former may result in more optimistic estimates on generalization performance due to the possible data leakage problem, i.e., future observations may be partitioned into the training dataset. Consequently, two equally good solutions may be quantified with varying prediction performance when testing datasets are partitioned in different ways. Similarly, a wide range of models is available for pointwise time series predictions, and many studies have claimed the superiority of their methods over others. However, as we shall demonstrate in the latter part of this study, the choices of benchmark, the intrinsic predictability of the time series, and the alignment between evaluation measures and model training objectives may all contribute to final conclusions.

2.2. Key questions to be answered

In general, the performance of data-driven prediction models may vary stemming from three main aspects, i.e., the methods used for data preparation, the techniques used for model training, and the measures adopted for performance evaluation. To raise the awareness of building professionals, a set of data experiments has been designed with the aim of quantifying the impacts of varying methods for short-term building energy predictions. In this study, eight general questions have been formulated from three aspects, i.e., data, model and evaluation. It should be mentioned that conducting exhaustive data experiments to address every conceivable detail in each aspect is virtually impossible. Therefore, this study tries to answer the following eight general questions using specific examples as summarized in Table 1. The former three questions are raised from the data perspective and discuss how different training data preparation methods may affect the prediction performance. The middle two questions are related to model training, and data examples are designed considering the use of artificial neural networks. The latter three questions address the importance of varying evaluation measures related to metrics, benchmarks and intrinsic data predictability.

Table 1
General questions to be answered through specific examples.

Aspects	Questions	Examples
Data	How will data partitioning methods affect the prediction performance evaluation?	Random vs. temporal partitioning
	What is the impact of training data with limited coverage on operating conditions?	Complete vs. incomplete training data coverage
	How does the amount of training data affect the model performance?	Training data samples with varying sliding step sizes
Model	Do more advanced modeling algorithms guarantee better prediction accuracies?	Dense, convolutional, recurrent neural networks
	What are the better practices in developing neural networks for building energy data?	Numerical and categorical input processing methods
Evaluation	What are the pros and cons for different pointwise evaluation metrics?	Scale-dependent and scale-free metrics
	Any competitive and easy-to-implement benchmarks for fair comparison?	Naïve model-free benchmark methods
	How to better quantify the values of data-driven methods beyond accuracies?	Measures on intrinsic time series predictability

2.3. Data experiment descriptions

2.3.1. Descriptions of building datasets

This study adopts actual measurements collected from 4 building clusters for data experiments [40]. Each building dataset contains hourly measurements of 1 to 2 years, and the monitoring variables include building energy consumptions, wind directions, wind speeds, outdoor dry-bulb and wet-bulb temperatures. As summarized in Table 2, there are 32 test buildings in total, which vary in their usage types and

Table 2
A summary of building meta characteristics.

Zone ID	Climate	Number of buildings	Usage types	Area (m^2)
1	Cold-humid continental	5	Education	[6320,14173]
2	Cold-humid continental	13	Education	[6000,80000]
3	Mixed-humid maritime	5	Hospital, Hotel	[3489,17993]
4	Mixed-humid maritime	9	Office	[1058,22569]

floor areas with gross floor areas ranging from 1,058 to 80,000 m^2. Each zone ID denotesa building cluster located in different geographic locations, where *Zones* 1 and 2 are located in Eastern North America (i.e., ASHRAE climate zone 6A) with a cold-humid continental climate, while *Zones* 3 and 4 are located in Europe (i.e., ASHRAE climate zone 4A) with a mixed-humid maritime climate. The buildings with a usage type of “Education” consist of university buildings serving as research facilities, offices, and teaching classrooms. The distributions of average hourly power consumptions are depicted in Fig. 2, indicating a wide range of energy patterns with minimum and maximum average hourly consumption of 2.44 kWh (Building 31) and 1491.72 kWh (Building 16) respectively. It is observed that different buildings exhibit significant variations in their hourly means, e.g., they are relatively small for Buildings No. 12, 14, 18, 27, 28 and 31, while significant variations are observed in Buildings No. 6, 7, 10, 11, 16, and 21.

To further visualize the temporal patterns across buildings, Fig. 3 shows a temporal heatmap of normalized energy use intensity (EUI) profiles over a 24-hour cycle. Each row represents an individual building, and each column corresponds to hours of the day. The color intensity encodes the max–min normalized EUI within each building, where darker and lighter hues indicate lower and higher consumptions respectively. Distinct temporal patterns can be observed in Fig. 3, where the energy peak hours may vary according to different occupancy schedules. In addition, Fig. 4 depicts the coefficient of variation (CV) of hourly energy consumptions against gross floor area for all 32 buildings. The CV is calculated as the ratio of the standard deviation to the mean values. Different colors are used to show the primary use types of each building, and the point size is set proportionally to the mean hourly energy consumption. It is observed that smaller buildings with an area below 5,000 m^2 exhibit a wider range of CV values, indicating that operational variability in compact structures is highly sensitive to occupancy schedules or equipment cycling. Such heterogeneous building datasets represent building energy consumption patterns across different environments, typologies and scales, thereby providing a relatively solid foundation to draw generalizable results through data experiments. It should be noted that the distribution of building types is unbalanced, with educational buildings accounting for most cases. Nevertheless, as shown in Figs. 3 and 4, the buildings exhibit diverse operational patterns and wide variability in building energy time-series data, which can help mitigate potential biases arising from uneven representation of building types.

2.3.2. *Descriptions of data experiments*

Considering building operations typically present weekly patterns, the prediction task is defined to predict 24-hour ahead building energy consumptions (i.e., denoted as values at timesteps $T+1$ to $T+24$) based on history measurements of energy consumptions, outdoor dry-bulb and wet-bulb temperatures during the last week (i.e., denoted as values at timesteps T-167 to T), the outdoor dry-bulb and wet-bulb temperatures of the next 24-hour, and the *Hour* and *Day Type* at time step $T+1$. In theory, the 24-hour ahead predictions can be generated in three ways [41]. The first is to develop a one-step ahead prediction model, which is used iteratively to generate 24-hour ahead predictions. The second is to design sequence-to-sequence models to enable multi-output predictions, while the third is to develop 24 sub-models, each for the prediction of a certain timestep. This study adopts the third approach considering its robustness against temporal error accumulation and the ease of implementation when using neural network architectures.

More specifically, this study firstly evaluates the impacts of data preparation, i.e., using random or temporal splitting strategies to create training and testing datasets, selecting training data with complete or incomplete seasonal coverage, and adopting training data with different amounts for modeling development. The relative differences in accuracy metrics are used to quantify their impacts on predictive modeling performance. Afterwards, the performance of three commonly used neural network architectures for analyzing time series data is compared, i.e., one-dimensional convolutional, recurrent and transformer-based neural networks, together with the use of different preprocessing techniques for numerical and categorical inputs. Thirdly, the pros and cons of evaluation measures on prediction performance and the intrinsic time series predictability are demonstrated.

3. Quantitative assessment on different data preparation methods

3.1. *Variations in training and testing data partitioning strategies*

As introduced in Section 2.1, random and temporal partitioning

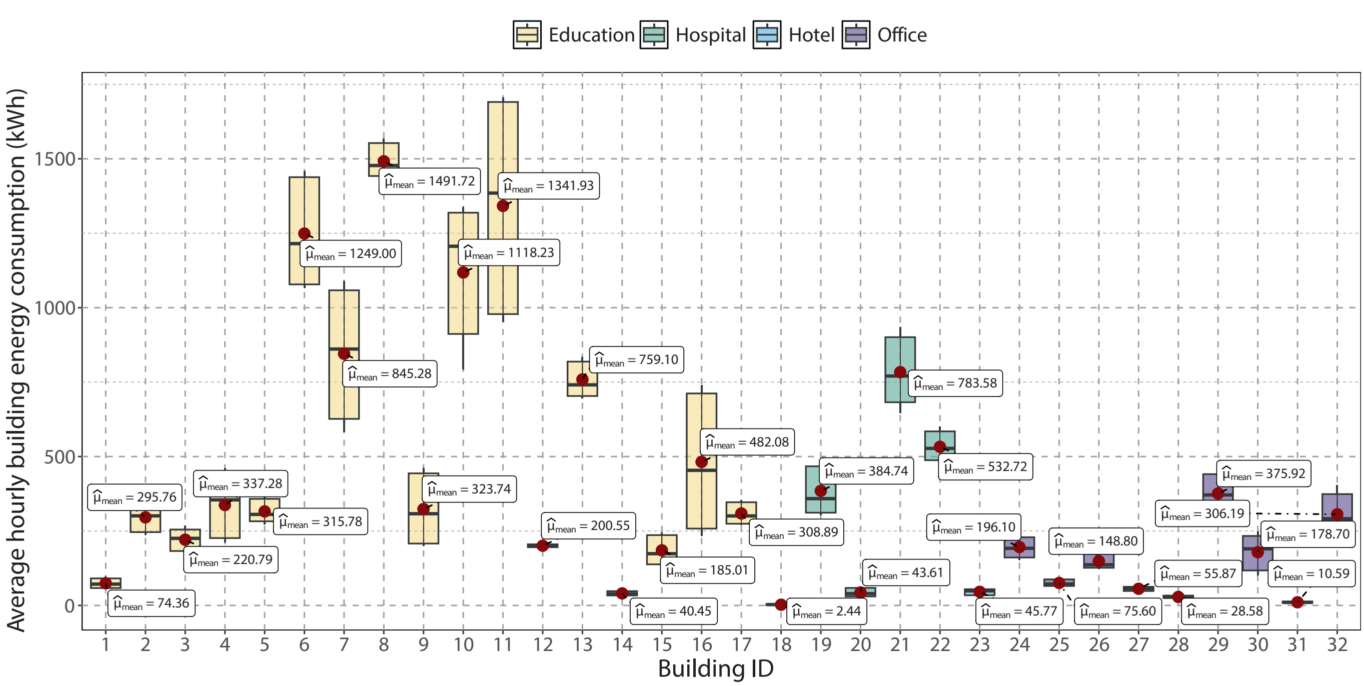

Fig. 2. Distribution of hourly power consumption for 32 buildings showing heterogeneous load magnitudes and variabilities.

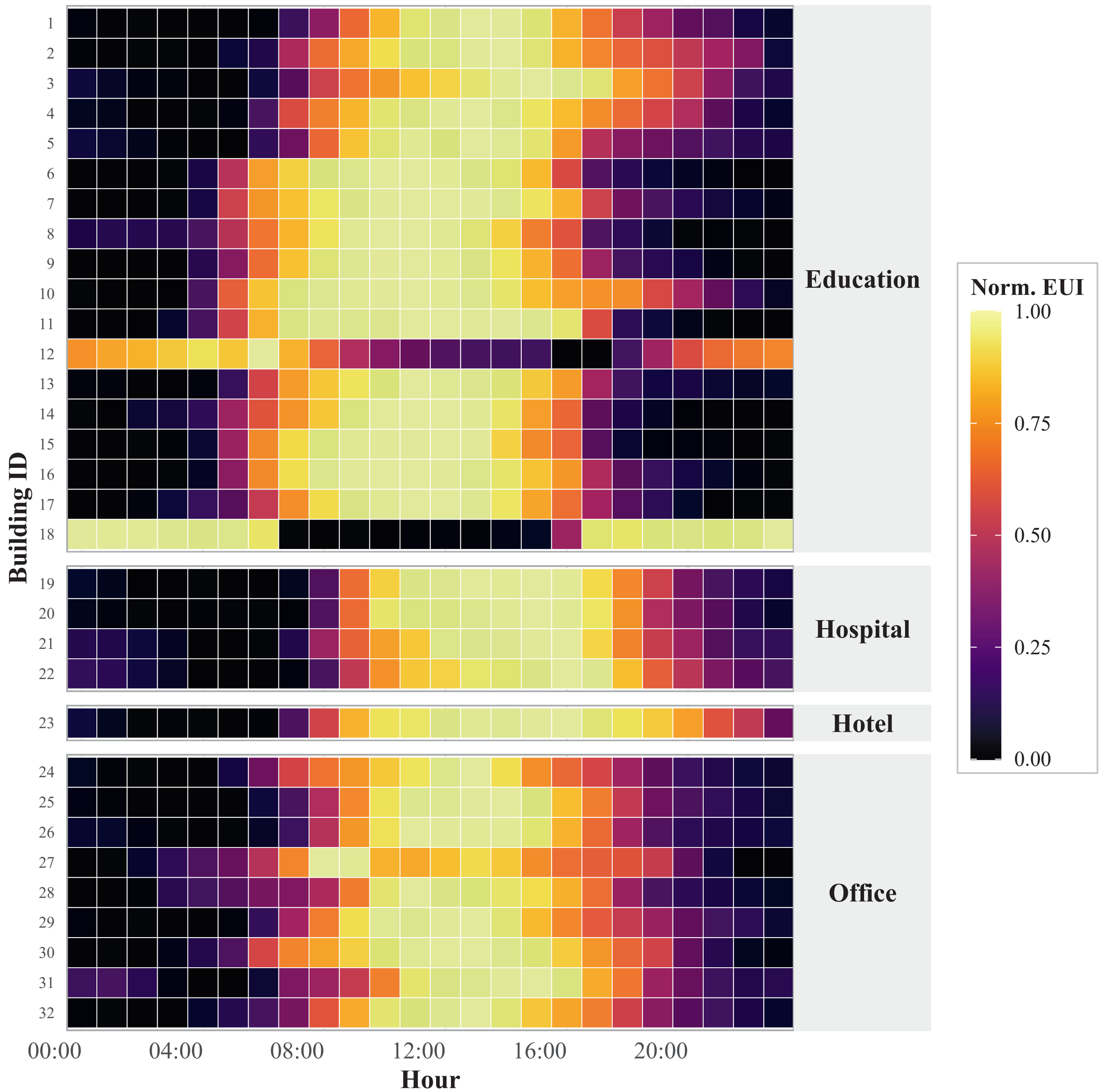

Fig. 3. Normalized 24-hour energy use intensity profiles revealing distinct diurnal patterns across building typologies.

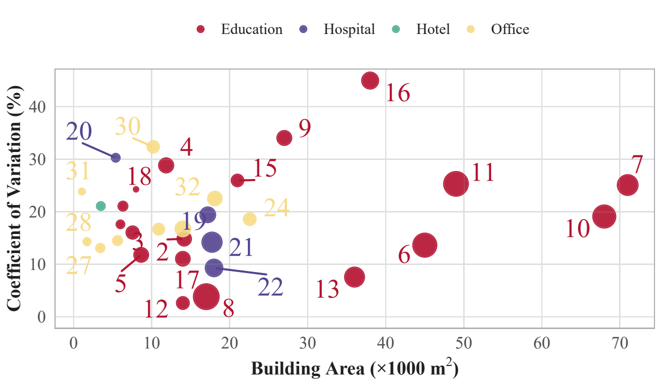

Fig. 4. Scatter plot of floor area versus consumption variability (CV), with point color indicating building type and size denoting mean energy consumption.

strategies differ in whether temporal order is preserved during data splitting. This section quantifies the practical impact of this choice on prediction accuracy across 32 buildings.

To illustrate this point, data experiments have been designed to assess the differences in accuracy metrics when training and testing data are sampled without and with the consideration of their temporal orders. More specifically, a fixed neural network architecture has been used for 24-hour ahead predictions, where one-dimensional convolutional layers are used for multivariate time-series data processing, standardization and embedding techniques are used for handling numerical and categorical input variables respectively. For each building scenario, the training and testing data ratios are fixed at 80% and 20%. If using a random partitioning strategy, all data subsequences are divided into training and testing datasets without considering their temporal orders. By contrast, the former 80% and the latter 20% of each month's data are selected for model training and testing using the temporal partitioning strategy, simulating the data environments in real practices.

Fig. 5 presents the performance comparison between temporal and random data partitioning strategies across different model types and buildings, where ΔCV-RMSE refers to the difference between CV-

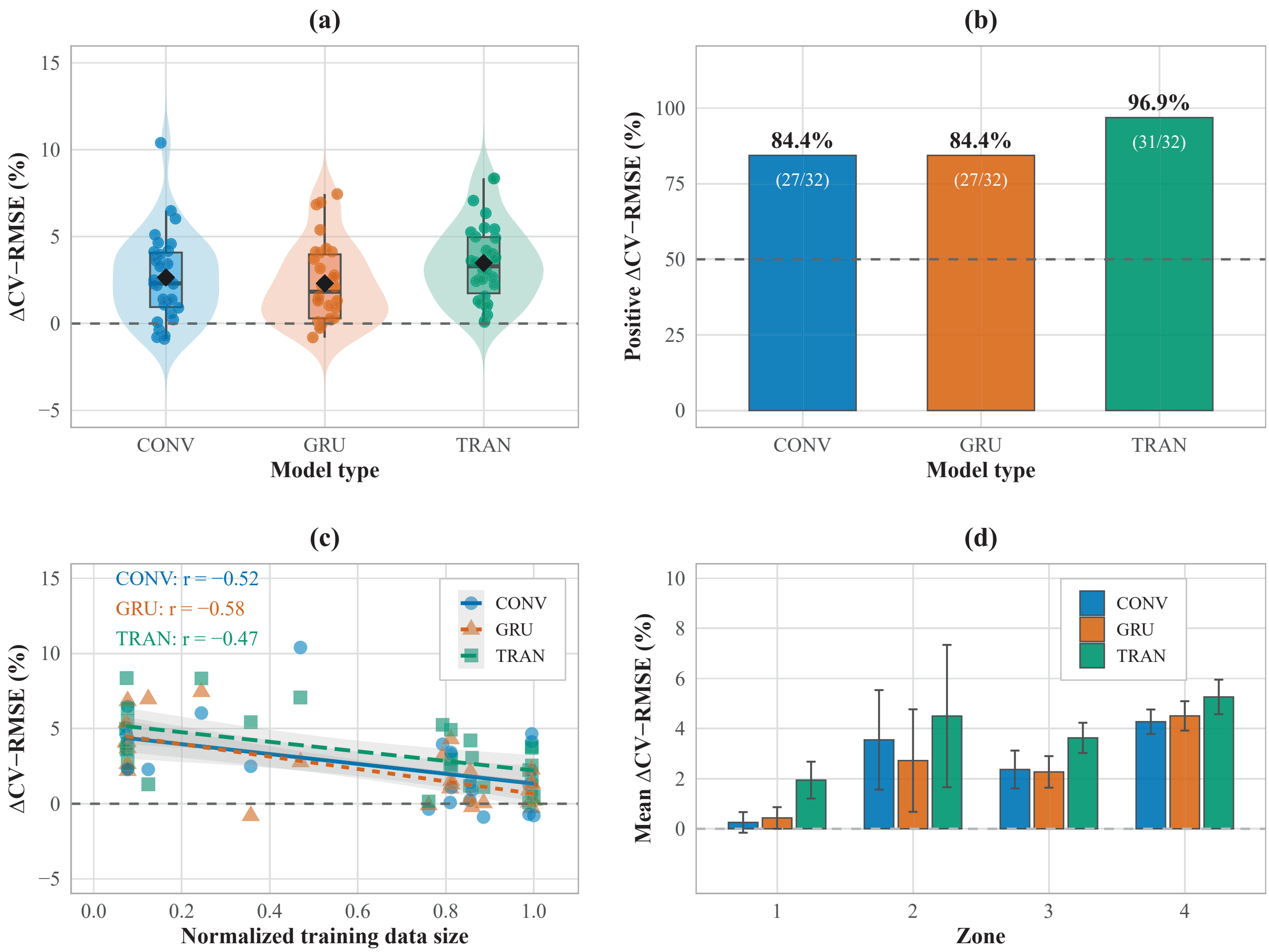

Fig. 5. Impact of data partitioning strategies on building energy prediction performance. (a) Distribution of ΔCV-RMSE across model types. (b) Proportion of cases showing positive ΔCV-RMSE. (c) Negative correlation between training data size and ΔCV-RMSE. (d) Mean ΔCV-RMSE for different zone buildings. Positive values indicate that random partitioning yields lower CV-RMSE estimates.

$\text{RMSE}_{\text{temporal}}$ and $\text{CV-RMSE}_{\text{random}}$. As shown in Fig. 5(a) and (b), the ΔCV-RMSE values are predominantly positive across all model types. It indicates that random partitioning consistently yields optimistic performance estimates with a mean CV-RMSE decrease of 3.34%. More specifically, the average decreases across 32 buildings are 3.05%, 2.79% and 4.18% for CONV, GRU and TRAN models respectively. Such results align with domain expertise, as random partitioning is more likely to cause data leakage, i.e., future observations are included in the training dataset. Consequently, two equally good solutions may be quantified as having different prediction performance when testing datasets are partitioned in different ways. It is shown that transformer-based models demonstrate the highest sensitivity to data partitioning strategies used, where convolutional and recurrent models show slightly greater robustness. This is in line with expectations, as transformers are more sensitive to data distribution shifts because they can learn very fine-grained patterns that may not generalize well. Fig. 5(c) indicates that a moderate negative correlation exists between normalized training data size and ΔCV-RMSE, suggesting that larger training datasets help to mitigate the discrepancy between partitioning strategies by capturing a broader range of building operating conditions. It is further observed in Fig. 5(d) that ΔCV-RMSE values in *Zone* 1 are relatively small with a mean of 0.88%. By contrast, the mean values of ΔCV-RMSE in the other three zones are much larger, i.e., 3.59%, 2.76%, and 4.68% for *Zones* 2 to 4 respectively.

3.2. Complete vs. Incomplete training data coverages over operational seasonalities

Prior to the development of data-driven models, practitioners should always ask themselves whether the training data at hand could provide enough information to ensure the model generalizability over various operating conditions. Taking the building energy modeling task as an example, building operations typically present unique patterns in different months considering variations in the outdoor environment and indoor needs. As a result, a full-year operational dataset would be desired to ensure the reliability of data-driven models. If the operation data of certain seasons is missing in the training data, it is highly likely that the model developed will not generalize well.

To illustrate this point, a data experiment has been designed to compare the differences in predictive accuracy when the operational data of certain months are artificially excluded from the training datasets. More specifically, training and testing data of each building are firstly partitioned within each month by preserving their temporal orders with proportions of 80% and 20% respectively. Afterwards, 12 unique training datasets are constructed for each building by removing one month of training data at a time. Data-driven models are then developed using the same architecture as introduced in Section 3.1. To investigate the impacts of such data absence, benchmark models have been developed using the same model architecture given the full 12-

month training datasets. Lastly, the data-driven models developed in each building are evaluated using the testing data of the missing month in the training datasets. In such a case, the results can be used as estimators of model generalizability under unseen operational conditions.

Fig. 6 presents the averaged CV-RMSEs across all building datasets given training data with complete and incomplete coverage. The green and purple lines represent the performance when data-driven models are trained using complete and incomplete data coverage respectively, while the grey bar shows the resulting CV-RMSE differences. It is observed that the intrinsic difficulties in building energy modeling vary according to different months. For instance, the benchmark CV-RMSEs of April and December are noticeably larger than the others. One possible explanation is that data-driven models developed in this study are essentially making predictions based on the past 7-day operation patterns. As April and December are typically the starting months of Summer and Winter, the operation patterns may undergo dramatic changes in thermal loads, and therefore make it more challenging to generate accurate predictions based on their recent operations, leading to an averaged CV-RMSE of 14.99% and 16.60%, respectively.

The resulting monthly CV-RMSE differences range from 4.36% to 12.94%. The top 5 largest differences occur when July, June, September, December and April training data are excluded, leading to 12.94%, 9.79%, 9.71%, 8.74% and 8.48% increases in CV-RMSEs over the benchmark performance respectively. The 95% confidence intervals (i.e., CI) shown in Table 3 indicate that complete and incomplete training data coverage will cause statistically significant differences in CV-RMSEs. It indicates that the operational data of these months have rather unique characteristics, and the information conveyed can be hardly represented by data collected from other months. By contrast, the absence of the middle month of four seasons (i.e., February, May, August and November) will cause much smaller degradations in CV-RMSEs, as their operational characteristics could be better described by their neighboring months.

3.3. The impact of training data amounts on predictive performance

It is generally believed that the performance of data-driven models is positively correlated with the number of training data samples. Considering that training data amounts may vary according to how data subsequences are generated for time-series modeling, data experiments have been designed to quantitatively assess their impacts on prediction accuracies. More specifically, training and testing data are partitioned following their temporal orders in each month with ratios of 80% and 20% respectively. While the testing data is fixed for each building, the training data has been further randomly sampled across 10 possible proportions (i.e., ranging from 10% to 100% with an increment of 10%) to simulate different training data availabilities. In addition, sliding step sizes have been defined from one to five for generating time-series subsequences. The model architecture follows the same setting as introduced before.

Table 3
Statistics on CV-RMSEs given complete and incomplete data coverage.

Month	Coverage	Mean CV-RMSE (%)	SD	95% CI of Δ CV-RMSE
1	Complete	7.16	3.50	[2.50%, 8.34%]
	Incomplete	12.58	7.41	
2	Complete	9.38	5.30	[1.56%, 7.16%]
	Incomplete	13.74	5.89	
3	Complete	10.57	5.50	[1.48%, 9.29%]
	Incomplete	15.96	9.54	
4	Complete	14.99	22.03	[-6.65%, 23.60%]
	Incomplete	23.47	36.48	
5	Complete	12.68	5.22	[2.86%, 10.14%]
	Incomplete	19.18	8.81	
6	Complete	11.23	10.43	[2.44%, 17.13%]
	Incomplete	21.01	17.86	
7	Complete	9.32	5.67	[7.16%, 18.71%]
	Incomplete	22.26	15.12	
8	Complete	10.28	4.44	[3.63%, 11.07%]
	Incomplete	17.63	9.46	
9	Complete	8.04	2.69	[4.94%, 14.48%]
	Incomplete	17.75	13.00	
10	Complete	9.14	4.60	[2.14%, 9.71%]
	Incomplete	15.06	9.58	
11	Complete	9.01	5.64	[1.24%, 8.09%]
	Incomplete	13.68	7.87	
12	Complete	16.60	8.32	[3.36%, 14.12%]
	Incomplete	25.34	12.69	

Fig. 7 presents the differences between 1 and normalized CV-RMSEs across different training data ratios, where the training data are generated with a sliding step size of 1 for all 32 buildings. Note that normalized CV-RMSEs are calculated for each building using their maximum and minimum CV-RMSEs of 10 unique training data ratios. Normalized CV-RMSEs are typically the maximum when the training data ratio is the minimum, i.e., 0.1. A clear increasing trend is observed when more training data become available for model development, especially at the early stage when the training data ratios increase from 0.1 to 0.3. The prediction accuracies tend to become saturated when the training data ratio exceeds 0.5. In this study, the number of training data samples for each building ranges from around 6800 to 13,800 when generated using a sliding step size of one, indicating that around 50% of these values would be sufficient to ensure the modeling performance given hourly data sampled from all-year datasets.

It is worth mentioning that the number of training data will vary

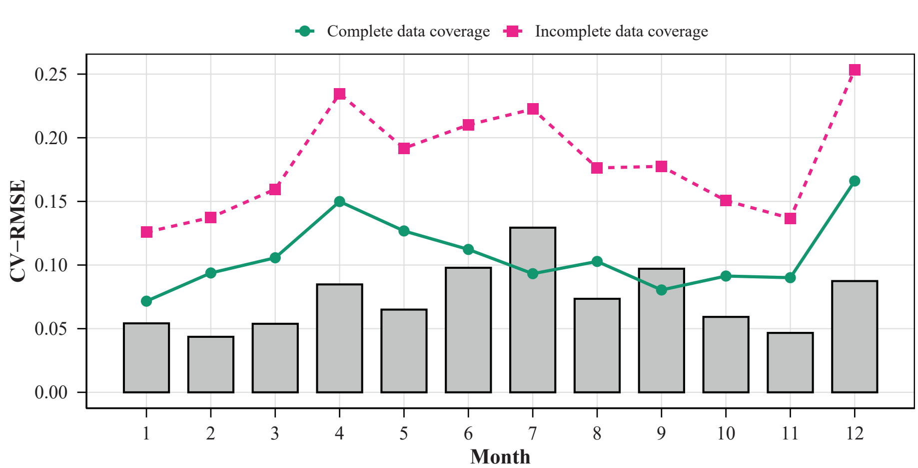

Fig. 6. Comparison of CV-RMSE across months under complete and incomplete training data coverage. Lines indicate CV-RMSEs when trained with complete (green) and incomplete (purple) data coverage, while grey bars show their differences.

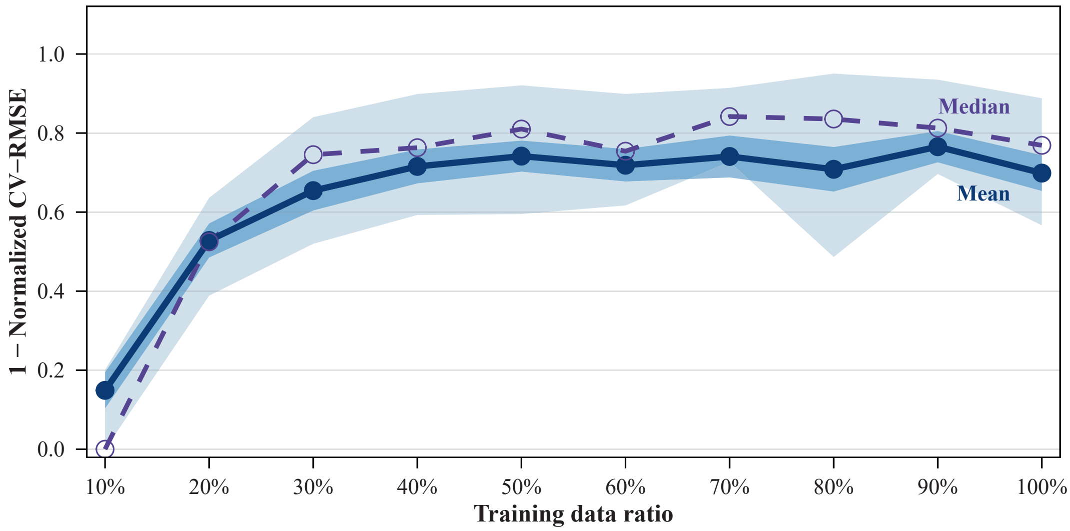

Fig. 7. Effect of training data ratio on predictive model performance. Solid and filled circles represent the mean, dashed line and open circles represent the median. Dark shaded region indicates +/-1 standard error of the mean, while light shaded region indicates +/-1 standard deviation.

significantly when using varying sliding windows to generate data subsequences. For instance, given a univariate timeseries with hourly measurements, a one-year operation will generate 8760 values. The number of 24-hour subsequences will drop from 8737 to 1748 when the sliding step size increases from 1 to 5. Coupled with variations in training data ratios, such differences in data generation settings may also affect the performance of data-driven models. To visualize such coupled effects, Fig. 8 presents the mean of CV-RMSEs across 32 buildings given different training data ratios and sliding step sizes. A clearly decreasing trend in CV-RMSEs is observed with increases in training data ratios, indicating that more training data will lead to better generalization performance. Data samples generated using a sliding step size of 1 (i.e., finest temporal resolution) achieve the lowest and most stable CV-RMSEs across all training ratios, while a step size of 5 (i.e., coarsest temporal resolution) exhibits substantially higher CV-RMSE and greater inter-building variability, particularly at training ratios below 30%. All step sizes converge to comparable performance when training data exceeds 40%. In practice, using smaller sliding step sizes will create highly similar subsequences, which may lead to possible biased conclusions if training and testing data are divided randomly without preserving their temporal orders, i.e., possible data leakage as two highly similar subsequences may fall into training and testing datasets respectively.

4. Quantitative assessment on different model training techniques

4.1. General performance interpretations based on Shapley values

The advances in computer science have provided great variations in data-driven modeling approaches in terms of data processing techniques, model architectures, and training schemes [42,43]. Unlike complicated modeling tasks in the field of computer vision and natural language processing, building energy analysis is typically of less complexity due to its smaller data amounts and more traceable physics-related operation patterns. A natural question may emerge as whether the adoption of the state-of-the-art data modeling procedures or techniques, if migrated to building energy data analysis, will bring cost-effective differences in prediction performance. To answer these questions, data experiments have been designed from three perspectives, i.e.,

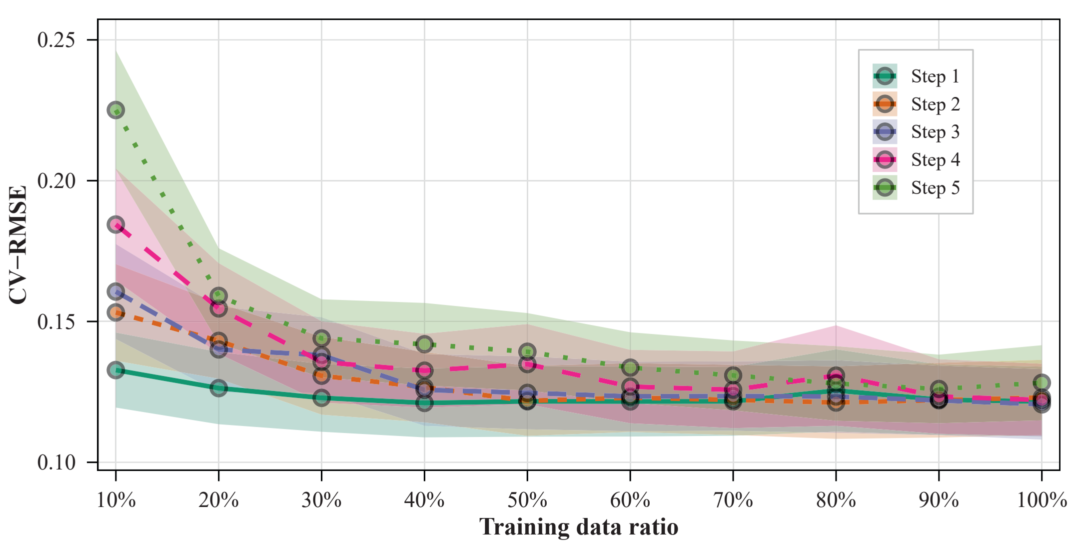

Fig. 8. Model CV-RMSEs given different training data ratios and sliding step sizes. Lines and filled circles represent mean values, while shaded bands indicate +/-1 standard error across 32 buildings.

different model architectures for time series modeling, numerical and categorical data preprocessing methods.

Building energy predictions are typically treated as a short-term time series modeling task. In theory, the model architecture used should be capable of deriving correlations between historical building measurements and future operation conditions. In general, three types of neural networks are compatible with time series data analysis, i.e., fully connected, convolutional and recurrent neural networks. In this study, variations of the abovementioned architectures have been adopted for performance comparisons. More specifically, the first relies on one-dimensional convolutional operations to extract temporal information from historical and future time series input data. The second adopts recurrent neural networks to iteratively update the hidden state of input time-series sequences to model temporal relationships. The third is a variation of transformers, which combines the use of fully connected layers with multi-head attentions for time series modeling. It should be mentioned that the neural network models are designed with similar complexity in terms of their total parameter numbers to ensure fairness in performance comparison.

Besides variations in model architectures, two data preprocessing techniques have been adopted for handling numerical and categorical input variables respectively, i.e., normalization or standardization for numerical input variables, and one-hot encoding or embedding for categorical input variables. Similar to the settings used in previous sections, the ratios of training and testing data are set to 80% and 20% respectively, and their temporal orders within each month are preserved to avoid possible data leakages. The resulting prediction accuracies on testing data are then used as a proxy to illustrate the values of the abovementioned modeling approaches through SHAP-based interpretable machine learning methods [44]. It should be noted that the following results may not represent the optimal performance of each approach, yet they could serve as a fair foundation for performance comparison as data-driven models are developed using the same training protocols, i.e., all models have similar complexity in terms of parameter numbers and are developed using the same learning rates, batch sizes, and early stopping scheme.

Fig. 9 presents the distribution of Shapley values extracted from the LightGBM model developed, where the average CV-RMSEs for each building across all twelve months are used as the model output, and the model types, preprocessing methods for numerical and categorical inputs are used as model inputs. It is observed that models using one-dimensional convolutions generally have the smallest prediction errors, while GRU models perform the worst. Admittedly, recurrent operations and transformers are theoretically more powerful in describing temporal relationships. However, special efforts are needed to realize their potential in the building field due to the limited complexity of building operations and the higher sensitivity to hyperparameter tuning. The second subplot in Fig. 9 shows that max–min normalization could lead to slightly better prediction performance than standardization. One possible explanation is that max–min normalization is distribution-agnostic and preserves the original distribution shape after data transformation. By contrast, while effective for Gaussian-distributed data, standardization may be less appropriate for building operation data where such distributional assumptions are often violated. As illustrated in the right subplot of Fig. 9, embedding is more efficient than conventional one-hot encoding for transforming categorical variables. Such an observation aligns with domain expertise, as one-hot encoding may introduce unnecessary sparsity in the input data, making it more difficult to develop reliable neural networks.

4.2. Quantitative impacts of varying techniques on building energy predictions

Given fixed settings on numerical and categorical input preprocessing methods, Fig. 10 presents the average of relative differences in CV-RMSEs together with their 95% confidence intervals when using different model architectures, e.g., GRU-CONV is calculated as $\frac{(CV-RMSE_{GRU}-CV-RMSE_{CONV})}{CV-RMSE_{CONV}}$. The GRU-CONV comparison shown in red shows consistently higher ΔCV-RMSE than the TRAN-CONV comparison shown in blue, indicating that the one-dimensional convolutional neural network outperforms the GRU-based model by a larger margin than it outperforms the Transformer-based model. Both comparisons remain positive throughout all months, demonstrating robust superiority of convolutional neural networks across seasonal variations. Compared with one-dimensional convolutional neural networks, using recurrent operations will increase the prediction error from 9.71% to 18.33%, while adopting a transformer architecture will slightly increase the error from 4.53% to 9.36%. Again, these results do not reflect the optimal performance of each model type, but they indicate that further optimization is required to fully realize the value of recurrent and transformer-based models.

Table 4 summarizes the average relative differences in CV-RMSEs given different numerical and categorical input preprocessing methods. It is shown that normalization and embedding are the superior choices for preprocessing numerical and categorical input variables respectively. The relative differences in CV-RMSEs given different numerical data preprocessing methods (i.e., ranging from −0.18% to 4.91%) are much smaller than those of using varying categorical data

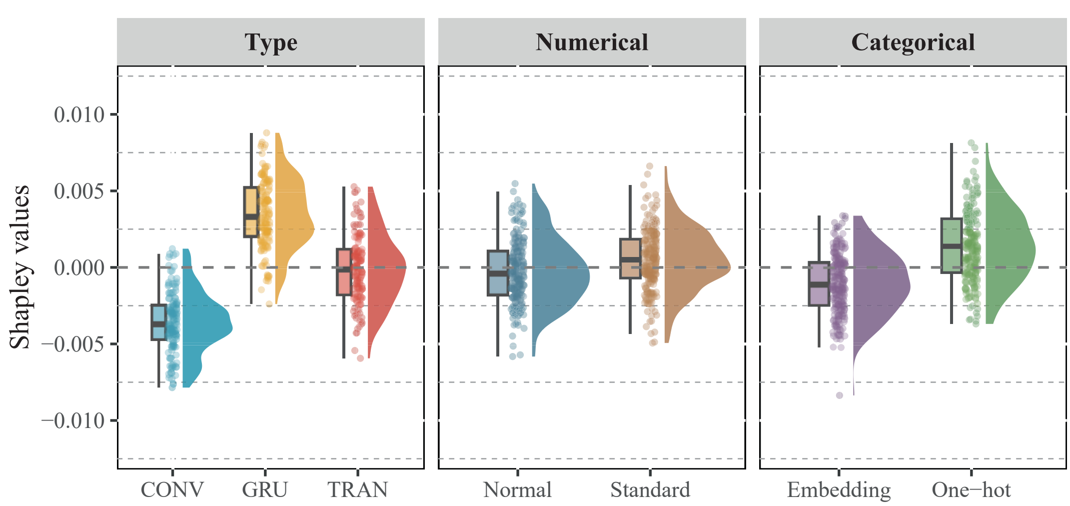

Fig. 9. Distribution of Shapley values for different variables on CV-RMSEs. Each panel displays kernel density estimates using a half-violin plot, while the boxplot shows medians and interquartile ranges.

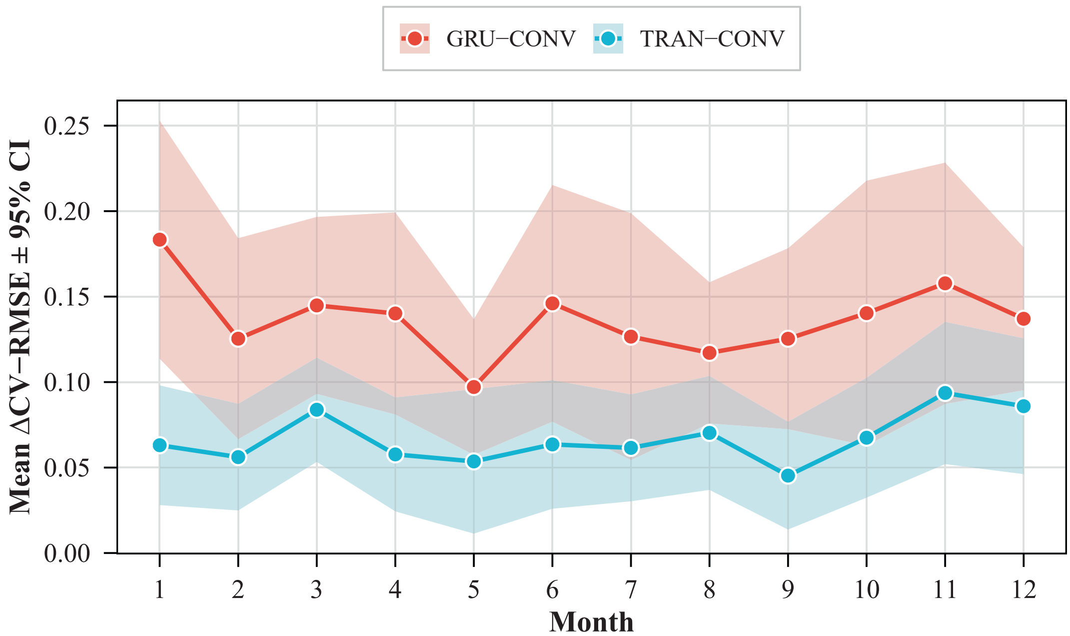

Fig. 10. Monthly comparison of predictive performance given varying model architectures. Lines represent the means of CV-RMSE differences, while shaded regions denote their 95% confidence intervals.

Table 4
Relative CV-RMSE differences given varying input preprocessing methods.

Month	Relative differences in CV-RMSEs (%)	
	Standardization vs. Normalization	One-hot encoding vs. Embedding
1	−0.18	13.24
2	3.70	10.36
3	0.66	8.22
4	4.28	10.28
5	3.89	6.65
6	4.91	7.37
7	3.98	6.98
8	0.83	4.22
9	0.46	5.38
10	4.98	11.75
11	1.53	8.25
12	0.51	3.93

preprocessing methods (i.e., ranging from 3.93% to 13.24%). One possible explanation is that normalization does not assume the original data are normally distributed, and it will not change their distribution. By contrast, standardization assumes the original data follow a normal distribution, which may not be the case for building operation data. It is further observed that the variations in the latter will result in similar CV-RMSE differences shown in Table 4, indicating that input preprocessing methods can be as important as model architectures in building energy prediction tasks.

5. Evaluation measures and performance benchmarks

5.1. *Scale-dependent and scale-free evaluation measures for pointwise predictions*

Many evaluation measures exist for assessing the prediction performance of point-wise time series predictions. In general, evaluation measures can be divided into two groups, i.e., scale-dependent and scale-free measures, and we direct interested readers to [45,46] for a detailed summary on point-wise time series evaluation measures. Scale-dependent measures quantify prediction errors using the same scale and unit of the target time series, e.g., mean absolute error (MAE), mean squared error (MSE) and root mean squared error (RMSE). One should always be cautious when using scale-dependent measures to compare data-driven model performance across different building energy time series, as their energy scales can be quite different. For instance, a data-driven model with an MSE of 15 is quite good if the mean of the target time series is around 1500, while such a model is almost meaningless if the mean is around 15. Scale-free measures are designed to avoid such problems through data scaling. For instance, CV-RMSE can be used to quantify the relative error by scaling RMSE based on the mean value, and MAPE is obtained when actual values are used as the denominator. In general, scale-free metrics may become troublesome when applied to time series with small values, e.g., ranging from −1 to 1, as the metrics may become quite large given a denominator close to zero.

Another point to note is that different measures may yield contradictory conclusions in practice. The reason behind is that evaluation measures are targeted towards optimizing for a specific statistic of the distribution [45]. For instance, MSE and RMSE will optimize towards the mean values, while MAE will optimize towards the median values. As time series prediction models are typically trained using L1 or L2 losses as the optimization function, the selection of different evaluation measures will also introduce additional decision biases. As an example, models trained using L1 losses tend to behave better when evaluated using MAE, while models trained with L2 loss functions will perform better if evaluated using squared-based measures such as RMSE or MSE.

To further illustrate this point, the relative orders between the pairwise monthly testing performance for each building are examined using MAE and RMSE respectively. All data-driven models are developed using MSE as the objective function. For each building, such pairwise comparisons of two specific months will result in 66 groups (i.e., C_{12}^2) of comparisons in total, and their ranking orders quantified by MAE and RMSE are extracted for consistency comparison. As shown in Fig. 11, it is found that the mismatched ratios range from 25.8% to 78.8% across all 384 data experiments (i.e., considering 32 buildings, 3 model types and 2 methods for numerical and categorical input preprocessing methods each). The average mismatch ratio is 54.7%, indicating that more than half of the relative orders assessed by MAE and RMSE are different. It suggests the necessity of adopting various measures for training and evaluating data-driven models.

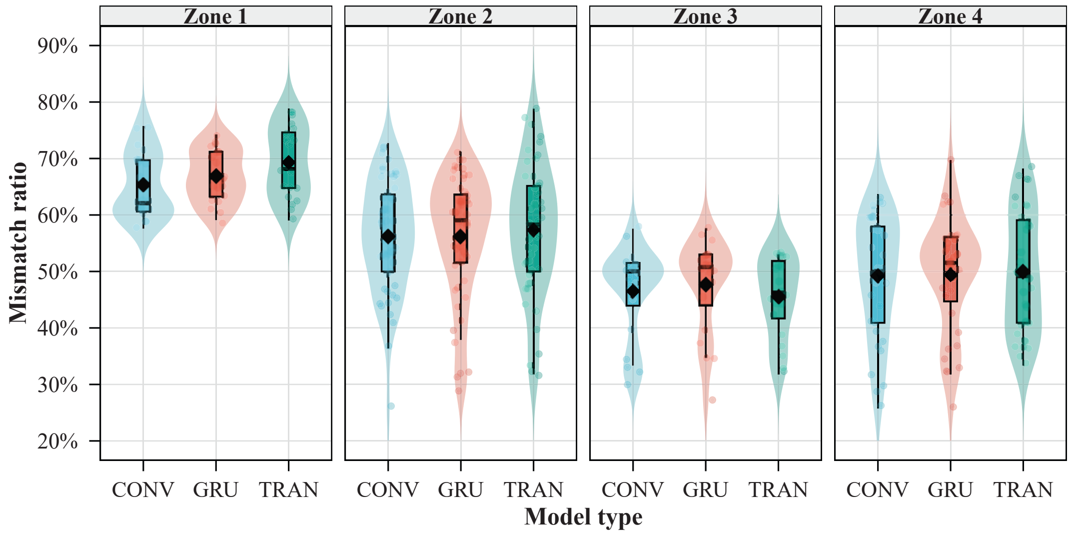

Fig. 11. Distribution of mismatch ratios quantifying disagreement between two accuracy metrics (MAE and RMSE) in ranking pairwise monthly prediction performance.

5.2. Choices of performance benchmarks

Considering the efforts and complexity in training data-driven models, it is essential to justify their cost-effectiveness over benchmark methods. As building operations typically present evident seasonalities, naïve model-free methods, which simply use historical measurements as predictions, can be used as competitive benchmarks [47]. As one of the simplest naïve methods, the persistence or no-change method takes the measurement of time step *T-1* as the prediction for time step *T*. Similarly, considering that building operations also have repeating weekly and daily patterns, historical measurements of the same day type in the last week or last day can be used as naïve benchmarks for next-day predictions.

Besides naïve model-free methods, other model-based benchmarks can be formulated to further reveal the advantage of models proposed over off-the-shelf predictive modeling algorithms, ranging from traditional ARIMA models to advanced deep learning variants such as Informer and ensemble models using boosting or bagging techniques [48]. As machine learning models are typically sensitive to their hyperparameter settings, rigorous model optimizations should be conducted for both the proposed and benchmark models to ensure fairness in performance comparison. In practice, it can be quite challenging to claim that researchers have put equal effort into optimizing all data-driven models. To enhance the values of related research in the field of building energy predictions, it is suggested to at least include naïve model-free benchmarks to illustrate the value of the modeling methods proposed. As we shall demonstrate in the following subsection, naïve benchmarks are helpful to quantify the intrinsic predictability or the difficulty of the prediction task at hand. For instance, if the naïve benchmarks have already provided a low prediction error, it indicates that the building energy patterns exhibit less variation or randomness and can be well explained based on their seasonalities. In such cases, developing complex data-driven models may not be entirely justified considering their computational costs, overfitting risks, and difficulties in real-time implementations.

5.3. Evaluation on the intrinsic predictability of building energy time-series data

The predictability of time series data quantifies the intrinsic degree to which future observations can be explained and estimated from past observations [49]. It helps to provide a general understanding of the best achievable accuracy for a given time series prediction task. Unlike modeling tasks in other fields such as image recognition, the upper bound of prediction accuracy for building energy time-series data can be rather ambiguous due to intrinsic uncertainty in data measurements and randomness in building operations. As a result, researchers and practitioners may not be capable of answering the question of whether the data-driven model developed has reached the performance limit. It is argued that reporting conventional accuracy metrics alone can be less meaningful without knowing the intrinsic predictability of the building energy time series being analyzed.

As basic solutions for evaluating time series predictability, statistical tests have been developed based on classical time series assumptions. For instance, the augmented Dickey-Fuller (i.e., ADF) test has been widely used to determine whether a time series is stationary or not, as a time series is more predictable if it is stationary [50]. Fig. 12 presents the ADF test statistics for all 32 buildings analyzed in this study. It shows that all statistics are negative and much smaller than the 99% statistic threshold (i.e., around -3.96 in this study), indicating that building energy time series to be modeled are rather stationary and conventional preprocessing methods such as differencing are not necessary. It should be mentioned that such measures are designed for testing the whole time series and therefore, may not be fully compatible with the sequence-to-sequence prediction paradigm used in the building energy prediction field.

As alternatives, recent studies have also proposed direct answers to quantify the intrinsic predictability of time series data. For example, researchers at Tsinghua University formulated an accuracy law that defines the upper bound on the accuracy of sequence-to-sequence prediction models based on the window-wise pattern complexity [51]. As shown in Eq. (1), the window-wise complexity is calculated as the total variance of amplitude spectrum distribution, which quantifies the spread of the amplitude spectra distribution among all time windows of length $P + F$, where P and F represent the lengths of the past observations and prediction horizon respectively in a single forecast window, N denotes the number of divided time windows, A_i denotes the spectrum amplitudes of the i-th window extracted by Fast Fourier Transformation, and $\overline{A}$ represents the sample mean of $\{A_i\}$. The accuracy law proposed states that for any time series x with $Complexity(x) \in (C_{min}, C_{max})$, the lowest mean squared error achieved by all feasible deep models exhibits an exponential relation with the complexity as described in Eq. (2), where α equals to 0.0054, C_{min} and C_{max} are 0 and 309 respectively. It should be noted that such law is derived from empirical analysis and designed for univariate forecasting tasks only. Such measures are useful for evaluating the intrinsic difficulties of the prediction tasks and help

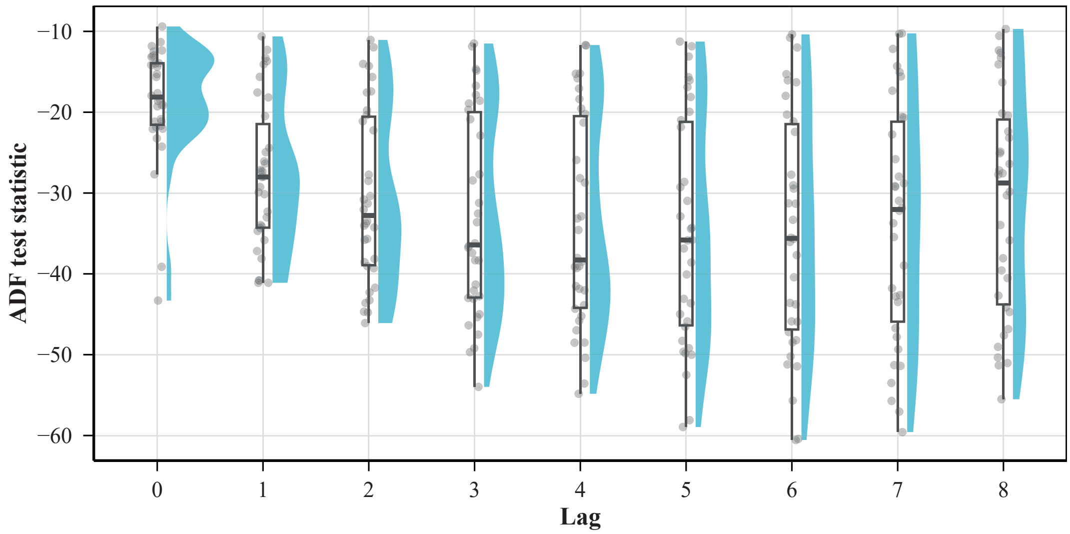

Fig. 12. ADF test statistics across time lags. Half-violin plots show the distributions, where boxplots indicate medians and interquartile ranges, grey points represent individual buildings, and the shaded green area shows the relative distribution density.

practitioners to better understand the improvement potential of their models.

$$Complexity(x) = \frac{1}{N}\sum_{1 \le i \le N} \|A_i - \overline{A}\| \tag{1}$$

$$MSE \approx \exp(\alpha \bullet Complexity(x)) - 1 \tag{2}$$

$$Relativeperformancegap = \frac{MSE_{benchmark} - MSE_{law}}{MSE_{benchmark}} \tag{3}$$

To illustrate, the building energy time series of all 32 testing buildings have been analyzed using the abovementioned approach, based on which the performance gaps of two naïve benchmark methods have been quantified. More specifically, the energy data has been firstly normalized using the max–min normalization method, and the window-wise complexity is then calculated considering a 24-hour ahead prediction task with inputs of the last 7 days, i.e., P and F are 168 and 24 respectively. The MSEs of two naïve benchmark methods, i.e., denoted as *Last Day* and *Last Week*, have been used for all 32 buildings. Fig. 13 presents the scatter plot showing the relationship between MSEs on normalized data using the naïve *Last Week* method and the window-wise complexity calculated. It is shown that 25 out of 32 buildings have a complexity of less than 309, which meets the prerequisite for implementing the accuracy law. Afterwards, as shown in Eq. (3), the relative performance gap (RPG), defined as the relative difference between the benchmark methods' MSEs and those calculated using the accuracy law, is computed for these 25 buildings. As shown in Fig. 14, it is evident that these two benchmark methods do have improvement potentials in terms of MSEs, as most of the RPGs calculated are positive, indicating that the MSE of benchmark methods is larger than the lower error bound calculated through the law. It is observed that using the same day in the last week as predictions is more accurate than using the last day measurements, as the means of RPG are 55.52% and 28.36% respectively. Such results indicate that using the same day in the last week can serve as a competitive performance benchmark, yielding an average MSE reduction of around 28% when developing data-driven models. It is worth mentioning that this study uses the accuracy law proposed in [51] as an example to show the potential benefits of knowing the intrinsic predictability of building energy time series and does not serve as a validation of this law in the building field.

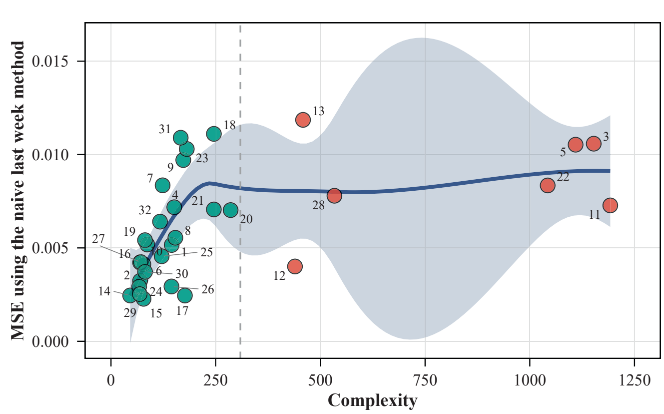

Fig. 13. Relationships between the window-wise complexity and prediction MSE on normalized data using the naïve weekly method. The vertical grey line indicates the complexity threshold of 309, while the blue line represents a LOESS fit with 95% confidence intervals. Green and red points represent building data that fall below and above the complexity threshold.

6. Discussions

6.1. Standardized procedures on pointwise building energy predictions

Based on the data experiment results shown above, it is argued that a standardized analytical pipeline is needed to better guide the research in the building field. As shown in Fig. 15, six actionable steps across three phases have been formulated, each accompanied by specific recommended practices or quantitative results derived from data experiments. From the data perspective, the first step is to check the training data coverage over different building operation seasonalities, as the absence of certain month's operational data can lead to significant prediction degradation, particularly for summer and winter transition months. The second step is to select an appropriate data partitioning strategy. Temporal data partitioning, which preserves the chronological data orders when dividing training, validation and testing datasets, is always recommended. Compared with random partitioning, temporal partitioning can prevent data leakages and avoid optimistic estimates on generalization performance, as random partitioning can lead to biased and smaller CV-RMSEs. Given rather short time series, random partitioning can be used, yet the sliding step size should be set relatively large to avoid generating highly similar training and testing data subsequences in the data sample generation step.

From the model training perspective, selecting the most complicated

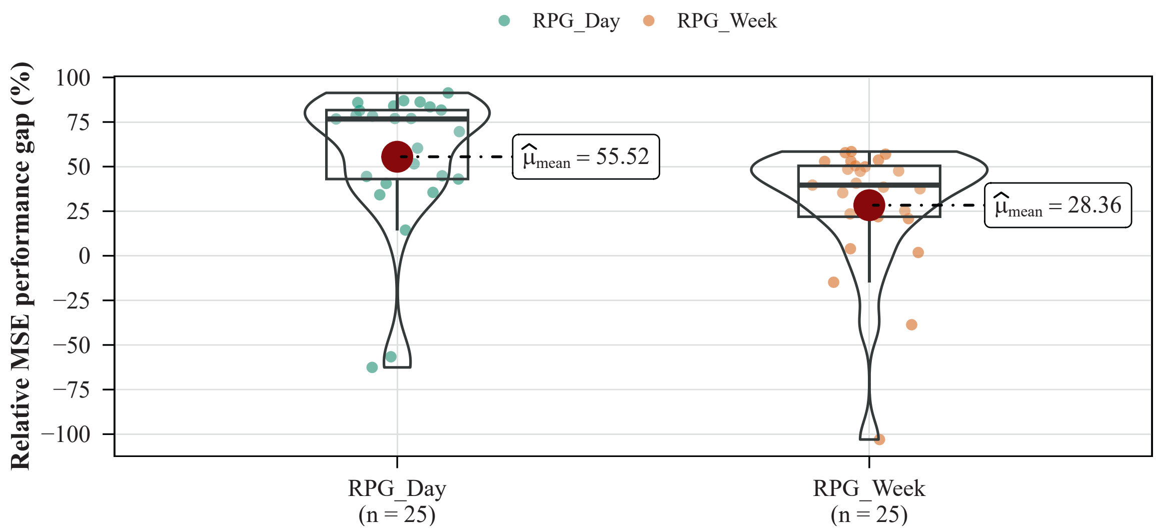

Fig. 14. Relative MSE performance gaps of naïve prediction methods compared with the lower accuracy bounds derived from the accuracy law proposed in [51]. Boxplots show medians and interquartile ranges, while red points represent mean values.

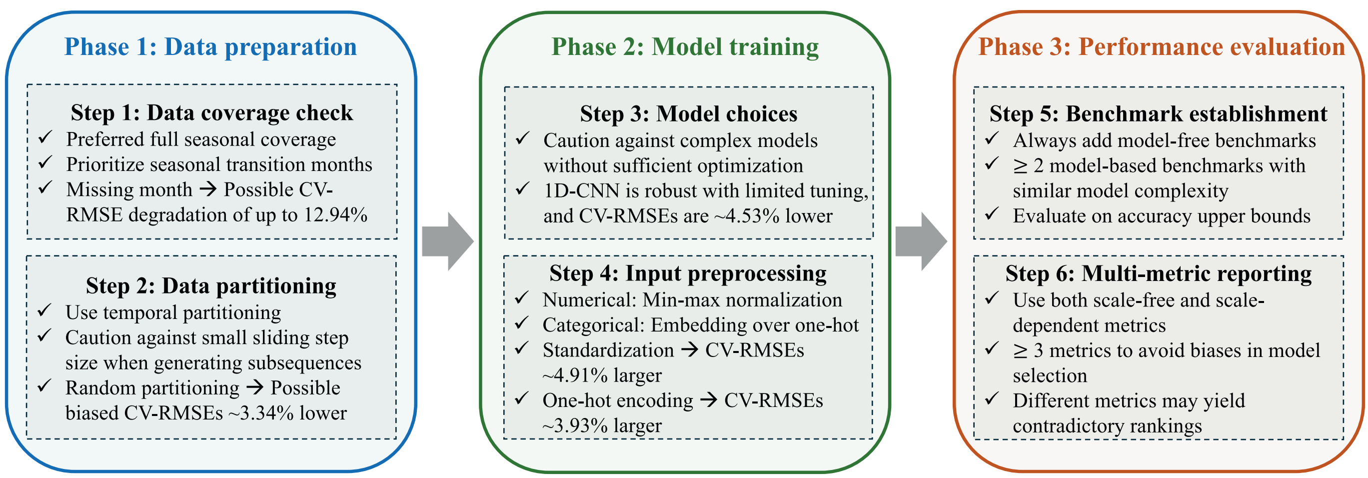

Fig. 15. Evidence-based standardized pipeline for short-term building energy predictions. Each step includes recommended practices, quantitative thresholds or cautions derived from the data experiments.

algorithms may not ensure the predictive modeling performance without sufficient optimization effort. Taking the neural network-based experiments conducted in this study as an example, the results suggest that practitioners should exercise caution against defaulting to complex model architectures without adequate hyper-parameter optimization, model regularization, and over-fitting controls. In addition, the choice of input data preprocessing methods can be equally important. For numerical variables, min–max normalization is preferred over standardization, as standardization assumes normally distributed data, which may not hold for building operation data. For categorical variables, embedding is preferred over one-hot encoding, as one-hot encoding may introduce unnecessary input sparsity. From the performance evaluation perspective, the first priority is to establish proper benchmarks to contextualize the prediction accuracy. Practitioners are recommended to always report model-free benchmarks as they are easy to compute and can help building professionals to understand the underlying daily or weekly seasonalities. In addition, at least two model-based benchmarks with similar model complexity should be reported to claim the competitiveness of the methods proposed. It is also advised to evaluate the intrinsic predictability of the building energy data to quantify the upper bound of prediction accuracy achievable, although such an evaluation can be challenging and calls for more research efforts. The second priority is to adopt a multi-metric reporting strategy. Practitioners should use both scale-free and scale-dependent metrics, with at least three metrics reported to avoid biases in model selection. It should be noted that different metrics may yield contradictory rankings in practice, as evaluation measures optimize towards different statistics of the distribution (e.g., MSE and RMSE towards the mean, MAE towards the median). Consequently, the selection of the loss function during model training should also be considered in conjunction with the evaluation metrics used.

It is worth noting that the 32 buildings used in this study cover four primary usage types (i.e., education, office, hospital, and hotel) with varying floor areas and load profiles. Although the distribution of building types is unbalanced, the key experimental findings demonstrate consistent patterns across these diverse building categories. For example, the data partitioning bias (i.e., random partitioning leading to optimistic accuracy estimates) is observed as predominantly positive across all 32 buildings regardless of their usage types, with *Zones* 1 and 2 buildings (education, ASHRAE 6A) and *Zones* 3 and 4 buildings (office, hospital, hotel, ASHRAE 4A) both exhibiting this effect. Similarly, the sensitivity to incomplete seasonal coverage and the model complexity-performance trade-offs are derived as averaged results across buildings spanning different types, climate zones, and scales. These observations suggest that the standardized procedures proposed are fundamentally methodology-oriented rather than building type-specific, as they address analytical pitfalls inherent in the data-driven modeling process itself. Nevertheless, future studies should further validate these findings on a broader building portfolio that includes more balanced representation of building types, particularly for underrepresented categories such as lodging and healthcare facilities.

6.2. Possible research directions towards a wider use of data-driven solutions

From the authors' perspectives, the research on pointwise building energy predictions should no longer be treated as an algorithm-level issue. Instead, future research could focus more on improving transferability, interoperability, and the ease of practical implementation across diverse building contexts. As an example, collecting training data with full coverage on all possible operating conditions could be time-consuming at the individual building level. In such a case, it is appealing to develop novel paradigms (e.g., transfer learning or federated learning) to enable a data or model sharing mechanism among buildings. To further enhance the generalization performance of data-driven models, physics-informed machine learning techniques can be adopted by formulating input variables through physical equations, combining physical constraints into the objective loss functions, or designing model architectures guided by physical principles [52]. All the above may help building practitioners to establish reliable data-driven models given limited data and time resources.

Beyond conventional machine learning approaches, LLMs have emerged as a transformative technology for building energy research. Recent studies have demonstrated the feasibility of using LLMs to automate various building energy tasks, such as generating EnergyPlus-based simulation models [36], energy predictions [53] and intelligent fault diagnosis through domain-specific fine-tuning [54]. The results obtained in this study could contribute to the reliability of LLM-based building energy prediction tasks from several aspects. First, the empirical evidence on data partitioning biases can provide guidance for designing LLM-based analytical pipelines. For instance, if an LLM-generated pipeline adopts random data partitioning on time-series data without explicitly addressing data leakage, the system can automatically flag a methodological risk. Second, the sensitivity analysis on training data coverage shown in Section 3.2 has direct relevance to LLM-based transfer learning and few-shot learning paradigms. The results suggest that when LLMs are used to generate or augment training datasets, the completeness of seasonal and operational coverage should be validated as a quality assurance step. Third, the comparison of neural network architectures shown in Section 4 indicates that more complex models do not necessarily lead to superior performance without careful hyperparameter tuning. Such findings help to caution against LLM pipelines that default to the most complex algorithms or model architectures without adequate optimization effort, and support the inclusion of model complexity-performance trade-off analysis as a standard step in LLM-based workflows. Fourth, the evaluation methods discussed in Section 5 can provide structured templates which LLMs can follow when generating performance reports. It is particularly useful as LLMs can synthesize multi-metric evaluation results and present them in natural language for building practitioners without extensive data science expertise.

7. Conclusive remarks

Accurate building energy predictions are valuable for efficient and effective controls over building energy systems. This study discusses key issues in ensuring the reliability of data-driven models and enhancing the generalizability of related studies from three main aspects, i.e., data preparation methods, model training techniques and evaluation measures. Data experiments have been designed to quantify the impacts of different approaches on building energy predictions. The main findings and answers to the questions related to data preparation, model training, and evaluation measures proposed in Section 2.2 are summarized below:

(1) Improper data preparation may lead to undesirable decision biases or suboptimal models. Compared with rigorous data partitioning methods preserving temporal orders, random data splits will lead to optimistic estimates on model generalization performance with an averaged CV-RMSE decrease of 3.34%. Given that building operations exhibit significant seasonal patterns, training data with incomplete month coverage will result in increases in CV-RMSEs ranging from 4.36% to 12.94%. In addition, the increase of sliding step sizes will decrease the size of window-wise data samples, which in turn will affect model accuracy.
(2) Customizations on modeling techniques are needed considering building energy data characteristics. It is shown that without sufficient efforts on hyper-parameter optimizations, theoretically more powerful algorithms such as recurrent neural networks may not be cost-effective in describing building energy patterns. Distribution-agnostic data processing techniques, such as max–min normalization and embedding, are shown to be more compatible for handling numerical and categorical variables respectively in building energy prediction tasks.
(3) Choosing the right evaluation measures and benchmarks is of great significance to enhance the value of related research. It is shown that different scale-dependent and scale-free measures can lead to contradictory conclusions, whereas naïve model-free methods can serve as competitive benchmarks for performance evaluation. Furthermore, designing methods to quantify the intrinsic predictability of building energy time series is a promising research direction, as it can help building professionals to know the accuracy limit and justify the actual value of data-driven models proposed.

This research aims to raise the awareness of building researchers and practitioners regarding the best practices and possible pitfalls associated with different data preparation, modeling and evaluation methods. The quantitative results obtained from data experiments can serve as actionable benchmarks for researchers to contextualize their own studies. In addition, the results obtained can serve as a structural knowledge base to guide the development of domain-specific LLMs in building energy analytics. The quantitative thresholds, decision rules, and recommended practices identified through data experiments can be encoded into LLM systems through several mechanisms, i.e., (1) as domain-specific instructions or system prompts that constrain LLM behavior when designing the analytical pipeline; (2) as structured entries in retrieval-augmented generation (RAG) knowledge base for LLM agents to query when making methodological decisions; (3) as fine-tuning examples to teach LLMs to distinguish between sound and flawed analytical practices in building energy contexts. Such integration may help to develop truly automated, reliable and reproducible building energy analytics at scale.

Several limitations should be acknowledged. First, the building datasets used may not fully represent all possible building scenarios. In particular, the distribution of building types across the 32 case buildings is unbalanced, with educational buildings accounting for most cases. As a result, some of the observed patterns may be more strongly supported for such buildings. Future work should further validate these findings using a more balanced building portfolio with broader representation of underrepresented categories such as lodging, healthcare, and industrial facilities. Second, this study focuses on neural network architectures and the generalizability of research findings to other machine learning methods may need further investigation. Third, the standardized procedures proposed are derived from hourly building energy prediction tasks and may need customization when applied to building operational data with higher time resolutions. Fourth, while this study provides empirical results which LLM-based systems can leverage, future work is needed to evaluate how effectively LLMs can adapt when designing building energy prediction pipelines in practice.

CRediT authorship contribution statement

Cheng Fan: Writing – original draft, Software, Methodology, Investigation, Conceptualization, Formal analysis, Visualization, Writing – review & editing. **Enqi Shen:** Visualization, Investigation, Formal analysis. **Da Yan:** Writing – review & editing, Investigation, Conceptualization. **Jinhan Mo:** Writing – review & editing, Supervision, Investigation, Conceptualization.

Declaration of competing interest

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

Acknowledgement

The authors gratefully acknowledge the support of this research by the National Natural Science Foundation of China (No. 52325801, 52278117, 52225801), Guangdong Basic and Applied Basic Research Foundation (2024A1515011549) and Shenzhen Science and Technology Program (20240813143330039).

Data availability

The authors do not have permission to share data.

References

[1] Z. Xing, Y. Pan, Y. Yang, X. Yuan, Y. Liang, Z. Huang, Transfer learning integrating similarity analysis for short-term and long-term building energy consumption prediction, Appl. Energy 365 (2024) 123276.
[2] H. Guo, T. Song, Y. Liu, A decomposition-prediction cooperative model with multi-strategy enhanced black-winged kite algorithm for multi-time-granularity short-term building load forecasting, Journal of Building Engineering 119 (2026) 115239.
[3] X.J. Luo, L.O. Oyedele, A.O. Ajayi, C.G. Monyei, O.O. Akinade, L.A. Akanbi, Development of an IoT-based big data platform for day-ahead building heating and cooling demands, Adv. Eng. Inf. 41 (2019) 100926.
[4] Y. Sun, F. Haghigaht, C.M. Fung, A review of the-state-of-the-art in data-driven approaches for building energy prediction, Energ. Buildings 221 (2020) 110022.
[5] S. Singaravel, J. Suykens, P. Geyer, Deep-learning neural-network architectures and methods: using component-based models in building-design energy prediction, Adv. Eng. Inf. 38 (2018) 81–90.
[6] P. Klanatsky, F. Veynandt, C. Heschl, Grey-box model for model predictive control of buildings, Energ. Buildings 300 (2023) 113624.
[7] H. Yesilyurt, Y. Dokuz, A.S. Dokuz, Data-driven energy consumption prediction of a university office building using machine learning algorithms, Energy 310 (2024) 133242.
[8] J. Runge, R. Zmeureanu, Deep learning forecasting for electric demand applications of cooling systems in buildings, Adv. Eng. Inf. 53 (2022) 101674.
[9] H. Yuan, M. Zhang, Z. Wang, Unveiling the impact of base model selection in heterogeneous ensemble learning for building energy prediction, Energy 332 (2025) 137162.
[10] M. Neshat, M. Thilakaratne, M. El-Abd, S. Mirjalili, A. Gandomi, J. Boland, Smart buildings energy consumption forecasting using adaptive evolutionary bagging extra tree learning models, Energy 333 (2025) 137130.
[11] C. Fan, Q. Wu, Y. Zhao, L. Mo, Integrating active learning and semi-supervised learning for improved data-driven HVAC fault diagnosis performance, Appl. Energy 356 (2024) 122356.
[12] C. Fan, Y. Lin, M.S. Piscitelli, et al., Leveraging graph convolutional neural networks for semi-supervised fault diagnosis of HVAC systems in data scarce contexts, Build. Simul. 16 (2023) 1499–1517.
[13] G. Li, Y. Wu, C. Yan, X. Fang, et al., An improved transfer learning strategy for short-term cross-building energy prediction using data incremental, Build. Simul. 17 (2024) 165–183.
[14] I.C. Akyol, Y.M. Karadag, S. Ucar, I. Talaz, F.E. Gursoy, I.G. Dino, S. Kalkan, Transfer learning and parameter-efficient fine-tuning for heating energy consumption prediction using urban building energy models (UBEM), Adv. Eng. Inf. 68 (Part A) (2025) 103576.
[15] M.M. Singh, K. Santer, J. Quesada-Allerhand, I.F.C. Smith, Enrichment of building energy models using error domain constrained generative machine learning, Adv. Eng. Inf. 69 (2026). Part C):104018.
[16] T. Hong, L. Zhang, AI for building energy modeling: a transformation, Build. Simul. 18 (2025) 2219–2225.
[17] D. Gao, X. Zhang, Y. Zhang, Y. Gao, W. Zou, K. Shan, Dataset preprocessing and optimization for machine learning-based building load prediction: a review, Energy 344 (2026) 139992.
[18] Y. Li, F. Arellano-Espitia, R. Aler, L. Igualada, C. Corchero, Data-driven methods and their applications to building HVAC energy consumption prediction: a review, Journal of Building Engineering 116 (2025) 114612.
[19] Y. Zhang, F. Tao, B. Qiu, et al., Interpretable data-driven fault diagnosis method for data centers with composite air conditioning system, Build. Simul. 17 (2024) 965–981.
[20] G. Li, Z. Yao, L. Chen, T. Li, C. Xu, An interpretable graph convolutional neural network based fault diagnosis method for building energy systems, Build. Simul. 17 (2024) 1113–1136.
[21] C. Bergmeir, Common pitfalls and better practices in forecast evaluation for data scientists, The International Journal of Applied Forecasting 70 (2023).
[22] M.A. Lones, Avoiding Common Machine Learning Pitfalls. Patterns 5 (10) (2024) 101046.
[23] C. Miller, P. Arjunan, A. Kathirgamanathan, et al., The ASHRAE Great Energy Predictor III competition: Overview and results, Sci. Technol. Built Environ. 26 (2020) 1427–1447.
[24] Ahir R.K., Delinchant B., Easwaran A. Time-series clustering: A benchmark study on energy data with insights into demand response. Engineering Applications of Artificial Intelligence, 2026, 163 (Part 1):112892.
[25] J.Y. Park, X. Yang, C. Miller, P. Arjunan, Z. Nagy, Apples or oranges? Identification of fundamental load shape profiles for benchmarking buildings using a large and diverse dataset, Appl. Energy 236 (2019) 1280–1295.
[26] M.M. Saad, U. Eicker, Investigating the reliability of building energy models: Comparative analysis of the impact of data pipelines and model complexities, Journal of Building Engineering 71 (2023) 106511.
[27] C. Tang, Z. Lu, Use of publicly available data to create a dataset for data-driven urban commercial building energy intensity classification: Model accuracy, interpretation, and implications of an open data framework in Hong Kong, Sustain. Cities Soc. 100 (2024) 105063.
[28] P. Emami, P.B. Graf, A large-scale dataset of 900k buildings and benchmark for short-term load forecasting. Open Energy Data Initiative (OEDI). National Renewable Energy, Laboratory (2018).
[29] W. Liao, X. Jin, Y. Ran, F. Xiao, W. Gao, Y. Li, A twenty-year dataset of hourly energy generation and consumption from district campus building energy systems, Sci. Data 11 (2024) 1400.
[30] X.Y. Jin, C. Zhang, F. Xiao, A. Li, C. Miller, A review and reflection on open datasets of city-level building energy use and their applications, Energ. Buildings 285 (2023) 112911.
[31] K.S. Skeie, L.S. Rokseth, C. Lausselet, A. Gustavsen, Characterizing and structuring open datasets for assessment of residential building energy use on the neighborhood and urban scale, Energ. Buildings 341 (2025) 115842.
[32] L. Zhang, V. Ford, Z. Chen, J. Chen, Automatic building energy model development and debugging using large language models agentic workflow, Energ. Buildings 327 (2025) 115116.
[33] L. Zhang, X. Fu, Y. Li, J. Chen, Large language model-based agent schema and library for automated building energy analysis and modeling, Autom. Constr. 176 (2025) 106244.
[34] T. Xiao, P. Xu, Exploring automated energy optimization with unstructured building data: a multi-agent based framework leveraging large language models, Energ. Buildings 322 (2024) 114691.
[35] M. Liu, L. Zhang, J. Chen, W. Chen, Z. Yang, L. Lo, J. Wen, Z. O'Neill, Large language models for building energy applications: Opportunities and challenges, Build. Simul. 18 (2) (2025) 225–234.
[36] G. Jiang, Z. Ma, L. Zhang, J. Chen, EPlus-LLM: a large language model-based computing platform for automated building energy modeling, Appl. Energy 367 (2024) 123431.
[37] M. Arslan, S. Munawar, Large language models in building energy applications: a survey, Energ. Buildings 352 (2026) 116800.
[38] C. Zhang, J. Zhang, J. Lu, Y. Zhao, Large language models meet energy systems: Opportunities, challenges, and future perspectives, Appl. Energy 403 (Part A) (2026) 127076.
[39] C. Bergmeir, R.J. Hyndman, B. Koo, A note on the validity of cross-validation for evaluating autoregressive time series prediction, Comput. Stat. Data Anal. 120 (2018) 70–83.
[40] C. Miller, A. Kathiragamanathan, B. Picchetti, et al., The Building Data Genome Project 2, energy meter data from the ASHRAE Great Energy Predictor III competition, Sci. Data 7 (2020) 368.
[41] C. Fan, D. Yan, F. Xiao, A. Li, J.J. An, X.Y. Kang, Advanced data analytics for enhancing building performances: from data-driven to big data-driven approaches, Build. Simul. 14 (2021) 3–24.
[42] J. Kim, H. Kim, H. Kim, D. Lee, S. Yoon, A comprehensive survey of deep learning for time series forecasting; Architectural diversity and open challenges, Artif. Intell. Rev. 58 (2025) 216.
[43] X. Kong, Z. Chen, W. Liu, K. Ning, et al., Deep learning for time series forecasting: a survey, Int. J. Mach. Learn. Cybern. 16 (2025) 5079–5112.
[44] A. Messalas, Y. Kanellopoulos, C. Makris, in: Model-Agnostic Interpretability with Shapley Values, Patras, Greece, 2019, pp. 1–7.

[45] H. Hewamalage, K. Ackermann, C. Bergmeir, Forecast evaluation for data scientists: common pitfalls and best practices, Data Min. Knowl. Disc. 37 (2023) 788–832.
[46] O. Rainio, J. Teuho, R. Klen, Evaluation metrics and statistical tests for machine learning, Sci. Rep. 14 (2024) 6086.
[47] N. Beck, J. Dovern, S. Vogl, Mind the naïve forecast! a rigorous evaluation of forecasting models for time series with low predictability, Appl. Intell. 55 (2025) 395.
[48] W. Li, K.L.E. Law, Deep learning models for time series forecasting: a review, IEEE Access 12 (2024) 92306–92327.
[49] G.E.P. Box, G.M. Jenkins, G.C. Reinsel, G.M. Ljung, Time series analysis: forecasting and control, John Wiley & Sons, 2015.
[50] J.H. Lopez, The power of the ADF test, Econ. Lett. 57 (1) (1997) 5–10.
[51] Wang Y., Wu H., Ma Y., Fang Y., Zhang Z., Liu Y., Wang S., Ye Z., Xiang Y., Wang J., Long M. Accuracy law for the future of deep time series forecasting. 2025, arXiv: 2510.02729v1.
[52] Z. Ma, G. Jiang, Y. Hu, J. Chen, A review of physics-informed machine learning for building energy modeling, Appl. Energy 381 (2025) 125169.
[53] Y. Zhang, D. Wang, G. Wang, P. Xu, Y. Zhu, Data-driven building load prediction and large language models: Comprehensive overview, Energ. Buildings 326 (2025) 115001.
[54] J. Zhang, C. Zhang, J. Lu, Y. Zhao, Domain-specific large language models for fault diagnosis of heating, ventilation, and air conditioning systems by labeled-data-supervised fine-tuning, Appl. Energy 337 (Part A) (2025) 124378.