

Adaptive-Resolution Long-Sequence Framework for Counterfactual Building Retrofit Energy-Savings Forecasting

Furkan Özcan^{1,*}

¹ Faculty of Engineering, Marmara University, Istanbul, 34722, Turkey

*Corresponding author: furkan.oz@marmara.edu.tr

Abstract. This paper proposes an adaptive-resolution long-sequence framework for counterfactual building retrofit assessment that preserves short operational transitions while retaining months of weather, occupancy, and control context. Multivariate meter, weather, occupancy, equipment, and intervention descriptors are encoded as ordered engineering tokens. A learned resolution controller preserves fine tokens near thermal and operational transitions, protects intervention and calendar boundaries, and merges redundant intervals elsewhere. A retrofit-conditioned, duration-aware Hyena operator then mixes the compressed long sequence, while matched decoders estimate post-retrofit demand and the corresponding no-retrofit trajectory. On an illustrative and replaceable benchmark containing 42 commercial buildings and 5.21 million 15-min observations, the model obtains 5.8% normalized mean absolute error, 7.4% normalized root-mean-square error, and 3.6 percentage-point error in annual saving rate. Relative to a dense Transformer, error falls by 18.3%, peak-direction accuracy rises from 85.6% to 91.2%, inference memory decreases by 38.6%, and one-building annual evaluation requires 0.84 s. The contribution is therefore an integrated engineering formulation of temporal allocation, long-context mixing, and retrofit attribution rather than a direct transfer of the original biological-sequence model. Because the benchmark and reported values are illustrative, all empirical outputs must be replaced or independently verified before submission.

Keywords: *Adaptive Resolution, Long-sequence Modeling, Counterfactual Forecasting, High-Frequency Sensor Data, Data-driven Portfolio M&V*

Received on 07 August 2023, Accepted on 19 December 2023, Published on 06 January 2024

Copyright © 2024 Author(s), licensed to DEA. This is an open access article distributed under the terms of the CC BY-NC-SA 4.0, which permits copying, redistributing, remixing, transformation, and building upon the material in any medium so long as the original work is properly cited.

Introduction

Since buildings account for a significant portion of the energy delivered, decarbonization programs must choose and verify retrofits [1]. A single static efficiency coefficient cannot capture the true benefits of envelope insulation, glass replacement, heat-pump conversion, ventilation control, and lighting enhancements because they all alter distinct physical pathways [2]. Normalizing the baseline has been used to lessen the impact of changing weather [3]. However, when the occupancy schedule is altered or the control logic is modified after construction, the reliability declines [4]. Richer evidence is provided by high-frequency smart-meter and building-automation data, although the resulting sequence includes extended stationary periods broken up by short equipment starts, set-back recoveries, variations in solar gain, and demand peaks [5]. Because uniform fine sampling is expensive in this setting and uniform aggregation eliminates the very transition that indicates a retrofit response, a forecasting model should maintain the short causal event while retaining months of thermal and operational background.

Nonlinear correlations between load, outdoor conditions, schedules, and equipment states are learned using data-driven energy models [6]. Long-term dependencies in sequence are lost by recurrent networks, although they maintain order over time [7]. A fixed receptive field still implies a single meaningful temporal scale, although temporal convolution enhances parallelism [8]. Although dense attention has quadratic memory growth and is not appropriate for deployment across building portfolios, transformer forecasting directly connects distant observations [9]. Although the useable window is extended by sparse and decomposition-based variations, the

majority still employ the same representation capacity for a stable night and a morning recovery event because they maintain a set sampling interval [10]. Context mixing can also be accomplished with long-convolutional sequence models [11]. This is especially evident in HyenaDNA, which was created to handle incredibly long ordered token sequences. However, direct transfer to buildings is inappropriate because telemetry is continuous, multirate, missingness-prone, and conditioned by a retrofit intervention, whereas biological symbols are discrete and stationary in meaning.

This work formulates retrofit-effect prediction as adaptive long-sequence counterfactual forecasting for portfolio measurement and verification. Its contributions are threefold. First, an event-density controller jointly learns temporal resolution while protecting intervention, calendar, tariff, and missing-data boundaries. Second, a duration-aware and retrofit-conditioned Hyena operator connects short equipment transitions with seasonal context using elapsed physical time rather than token index. Third, matched factual and counterfactual decoders estimate avoided energy while retaining explicit support and data-quality weights. The intended users are engineers and portfolio operators who must verify annual savings without discarding operational events that affect peaks or attribution. Section 2 reviews retrofit prediction and efficient long-context modeling. Section 3 presents the architecture and optimization procedure. Section 4 describes the replaceable evaluation protocol and its illustrative results. Section 5 summarizes the contribution, limitations, and future validation requirements.

Related Work

Building Retrofit Effect Prediction

Conventional retrofit evaluations begin with change-point models that normalize consumption by weather, degree-day regression, or calibrated simulations [12]. When schedules and equipment remain the same, these methods are easy to use and comprehensible; nevertheless, interacting envelope, ventilation, occupancy, and supervisory-control changes cannot be expressed by a limited number of coefficients [13]. Heat flow is explicitly shown using physics-based simulation, but if material characteristics, infiltration, or internal gains are unknown, calibration is difficult [14]. By using a baseline rule and adjustment procedure, measurement-and-verification processes lessen uncertainty [15]. A sequence model capable of utilizing millions of high-frequency observations is not provided by them. Physical residuals or simulation priors are introduced by hybrid models [16]. Performance is impacted by metadata quality, and portfolio databases frequently lack standard building guidelines [17].

Long-Sequence Forecasting and Resolution Adaptation

Because they can handle nonlinear temporal dependence, LSTM and gated recurrent architectures have been used in load and energy prediction [18]. Dilated temporal convolution facilitates parallel training and increases the receptive field [19]. Access to remote covariates is further enhanced by attention-based forecasts, although their computing load increases quickly with high-frequency yearly histories [20]. By using sparsity or structured mixing, effective attention, decomposition, and state-space models lessen this load [21]. Adaptive sampling, which usually employs fixed change thresholds independent of downstream forecast error, has also been investigated as a preprocessing technique [22]. Resolution selection, long-context mixing, intervention conditioning, and counterfactual saving estimation are all trained together in the unresolved gap, an end-to-end process. While maintaining the intervention semantics' visibility throughout the sequence, the current model fills that gap.

Adaptive-Resolution HyenaDNA Method

Retrofit Forecasting Formulation and Token Encoding

For building b at interval t_r , the observable vector includes whole-building demand, zone temperatures, outdoor temperature and humidity, solar irradiance, occupancy proxies, equipment states, calendar codes, missingness flags, and a retrofit descriptor. The prediction target is not limited to the observed post-retrofit load. It consists of a paired trajectory containing the factual demand under the installed retrofit and the counterfactual demand under the pre-retrofit configuration. This formulation separates avoided energy from variations caused by

weather, occupancy, and operating schedules. Every variable is aligned to a 15 -min time grid, while an alignment mask records whether each measurement is observed, interpolated, or carried forward from a slower supervisory channel.

$$\mathbf{z}_t = \mathbf{E}_x \mathbf{x}_t + \mathbf{E}_m \mathbf{m}_t + \mathbf{E}_c \mathbf{c}_t + \mathbf{E}_r \mathbf{r}_b + \mathbf{e}_t \quad \text{Eq.(1)}$$

The token $\mathbf{z}_t$ combines the numerical telemetry vector $\mathbf{x}_t$, data-quality mask $\mathbf{m}_t$, calendar state $\mathbf{c}_t$, building-level retrofit descriptor $\mathbf{r}_b$, and temporal position code $\mathbf{e}_t$. Separate projections are used instead of direct feature concatenation so that high-variance power channels do not suppress categorical intervention information. The resulting token sequence is subsequently consumed by the resolution controller. If the retrofit descriptor is removed, similar weather-load patterns generated by different interventions can collapse into the same latent region. This ambiguity is particularly problematic when an HVAC conversion and an envelope upgrade produce similar short-term consumption but have different persistence and counterfactual effects.

Notation and scope are defined as follows. Each base-grid interval is indexed by t , each building by b , and each adaptive token by j . The observed telemetry vector, quality mask, calendar state, retrofit descriptor, and elapsed-time coordinate are projected separately before fusion. The factual target represents demand under the installed intervention, whereas the counterfactual target represents demand under the pre-retrofit configuration under comparable exogenous conditions. The retrofit descriptor is available as metadata and is not inferred from post-retrofit demand.

Resolution selection begins with an event-density score. The controller must respond to meaningful thermal and operational transitions rather than raw sensor noise. It therefore combines robust gradients of power and temperature, equipment switching, occupancy changes, missing-data boundaries, and predictive uncertainty generated by a lightweight pilot head. Retrofit dates, day boundaries, and tariff transitions are protected so that no aggregated token spans incompatible operating regimes.

$$q_t = \text{softplus}(w_p |\Delta P_t| + w_T \|\Delta \mathbf{T}_t\|_1 + w_u \Delta u_t + w_o \Delta o_t + w_\sigma \sigma_t + b_q) \quad \text{Eq.(2)}$$

The scalar q_t represents the local information density of interval t . High values occur during morning warm-up, chiller staging, occupancy arrival, rapid outdoor-temperature changes, or abnormal equipment switching. Low values generally correspond to stable unoccupied intervals. Instead of applying a manually selected threshold, the score controls a differentiable distribution over candidate resolutions. Its output is passed directly to the adaptive resampler. Removing this event-density mechanism forces the resampler toward uniform aggregation, causing peak transitions to be blurred and increasing the bias of the estimated retrofit-saving rate.

Adaptive Resolution Resampler

The resampler distributes a finite token budget K over an original sequence of length L . Candidate spans contain 1, 2, 4, or 8 base intervals, corresponding to temporal resolutions of 15, 30, 60, or 120 min. Straight-through categorical sampling is used during training, whereas deterministic maximum-probability selection is applied during inference. The controller receives the event-density score, the distance to the nearest protected boundary, and the local missingness pattern. It can consequently preserve fine temporal detail without averaging missing-data edges or pre-retrofit and post-retrofit observations into the same representation.

$$\pi_{t,s} = \frac{\exp[(a_s q_t + g_s + \beta_s)/\tau]}{\sum_{j \in \{1,2,4,8\}} \exp[(a_j q_t + g_j + \beta_j)/\tau]} \quad \text{Eq.(3)}$$

The probability $\pi_{t,s}$ assigns candidate span s to interval t . The temperature parameter τ is annealed from 1.0 to 0.15 during the first 40 training epochs. This schedule initially permits exploration among resolution levels and then gradually produces nearly discrete selections. The boundary logit g_s invalidates any span that crosses a retrofit date, tariff boundary, or calendar transition. Without this constraint, the resampler could average a high-load pre-retrofit interval with a lower post-retrofit interval, creating an artificial intermediate state that contaminates both factual and counterfactual predictions.

$$\mathbf{v}_k = \frac{\sum_{t \in \mathcal{S}_k} \rho_t \mathbf{z}_t}{\sum_{t \in \mathcal{S}_k} \rho_t + \varepsilon}, \rho_t = \exp(-\gamma_m m_t - \gamma_a a_t) \quad \text{Eq.(4)}$$

Here, $\mathcal{S}_k$ denotes the interval set assigned to the k -th adaptive token. The reliability weight ρ_t reduces the influence of imputed measurements and age-stale supervisory channels. Every aggregated token retains its duration and its original start and end indices. Duration information is required by the downstream Hyena filter

because a token representing two hours must not be treated as temporally equivalent to a 15-min token. The interface is reversible: after long-context processing, the decoder maps the adaptive representations back to the requested base grid so that forecast errors remain comparable across alternative resolution policies.

The resampling procedure covers the sequence from left to right. At each uncovered base interval, infeasible spans are removed by the protected-boundary mask, one valid span is selected, and all intervals in that span are assigned to a single duration-stamped token. The process ends only when every base interval belongs to exactly one span, so selections cannot overlap or leave gaps. The token-budget penalty controls the expected compression ratio during training, and deterministic maximum-probability selection applies the same constraints during inference. Reconstruction uses the stored start index, end index, duration, and reliability weight to return decoder outputs to the base grid.

The complete processing chain is presented in Figure 1. Multirate meter, weather, occupancy, and equipment streams first undergo clock alignment and semantic tokenization. The event-density estimator then determines where detailed representations are needed, after which constrained span selection compresses stationary observations. Adaptive tokens enter the duration-aware Hyena blocks and are finally decoded into paired factual and counterfactual energy trajectories. Because resolution selection is embedded in the forecasting network, gradients from saving-rate, peak-load, and trajectory errors can update the controller through the straight-through estimator.

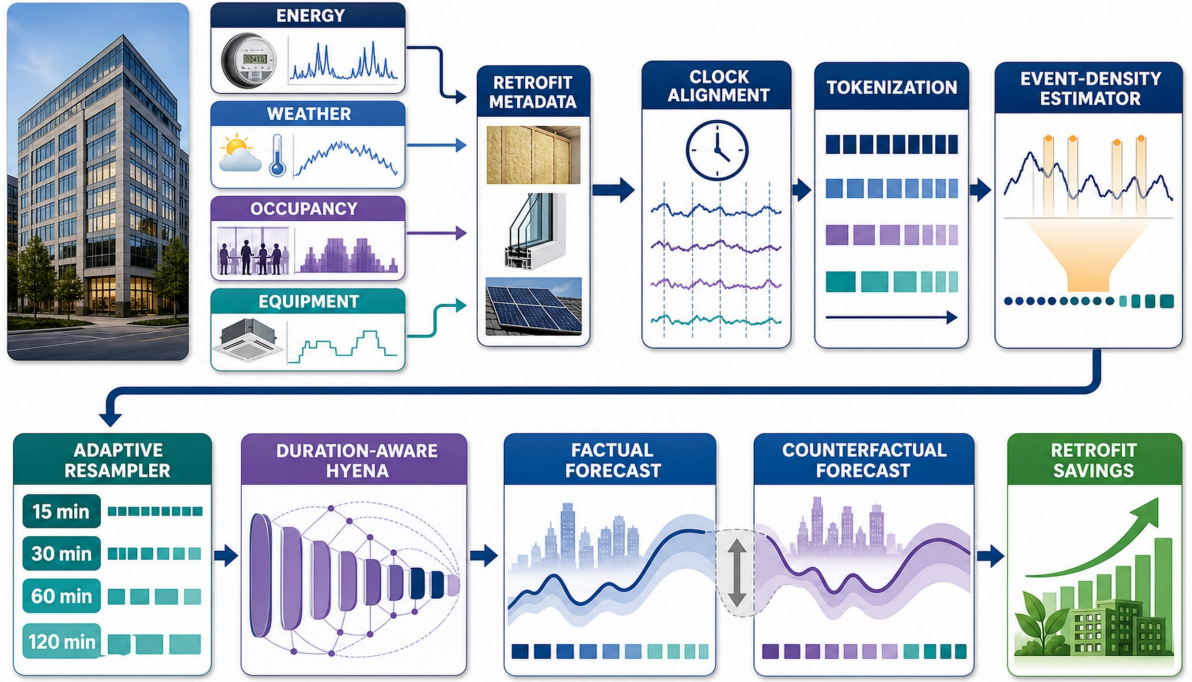


Figure 1. End-to-end workflow of adaptive-resolution retrofit-effect forecasting, from multirate telemetry and retrofit metadata to factual and counterfactual energy trajectories.

Retrofit-Conditioned Hyena Mixing and Dual Decoding

The adaptively resampled sequence can still contain several thousand tokens when an annual history is retained. A Hyena-style structured long convolution is therefore adopted to model extended dependencies without constructing a dense pairwise attention matrix. Each block contains input projections, a gated short convolution, an implicitly parameterized long filter, residual normalization, and retrofit conditioning. Filter coordinates are calculated from elapsed physical time rather than token index, which prevents irregularly spaced tokens from distorting temporal distance.

$$\mathbf{h}_k^{(\ell)} = \mathbf{v}_k^{(\ell)} + \mathbf{G}_k^{(\ell)} \odot [\mathcal{F}_\theta^{(\ell)}(\boldsymbol{\delta}_{1:K}, \mathbf{r}_b) * \mathbf{D}_k^{(\ell)}] \quad \text{Eq.(5)}$$

The elapsed-time vector $\boldsymbol{\delta}_{1:K}$ and retrofit descriptor $\mathbf{r}_b$ modulate the long filter $\mathcal{F}_\theta^{(\ell)}$. The term $\mathbf{D}_k^{(\ell)}$ is the data projection, while $\mathbf{G}_k^{(\ell)}$ controls content-dependent information flow. This operation connects fine transition

tokens with coarse seasonal tokens, and its output is passed to the counterfactual fusion layer. A position-only filter would incorrectly treat equal token-index distances as equal elapsed durations. A retrofit-agnostic filter would instead learn an average response across measures, weakening the distinction between envelope, HVAC, lighting, and ventilation interventions.

The backbone retains the standard Hyena sequence-mixing pattern of gated local projections and an implicitly parameterized long filter. The original extensions in this study are the use of elapsed physical time for irregular adaptive tokens, conditioning of filter coefficients on the retrofit descriptor, cross-resolution residual fusion, and matched factual-counterfactual decoding. These changes are required because equal token-index separation does not imply equal physical duration after adaptive resampling.

$$\mathcal{F}_\theta(\delta, \mathbf{r}_b) = \sum_{j=1}^J \alpha_j(\mathbf{r}_b) \exp(-\lambda_j \delta) \cos(\omega_j \delta + \phi_j) \quad \text{Eq.(6)}$$

The filter mixture contains decay and periodic components with distinct engineering interpretations. Short decay modes represent equipment switching and rapid zone-temperature response. Longer decay modes describe envelope heat storage and seasonal persistence. Oscillatory modes capture daily and weekly operating schedules. The retrofit-dependent coefficient $\alpha_j(\mathbf{r}_b)$ changes the contribution of each temporal component without altering the shared basis. The filter representation is shared by the two decoder branches, ensuring that their difference reflects the retrofit condition rather than unrelated backbone variation.

Figure 2 illustrates the internal connection among the adaptive-resolution bands, duration-aware Hyena blocks, retrofit-conditioned filter generator, cross-scale residual links, and paired decoder heads. The principal interface is the shared mixed state. The factual decoder receives the installed retrofit code, whereas the counterfactual decoder receives a masked intervention code combined with pre-retrofit operating descriptors. Weight sharing prevents the saving estimate from becoming the subtraction of two independently drifting forecasting networks.

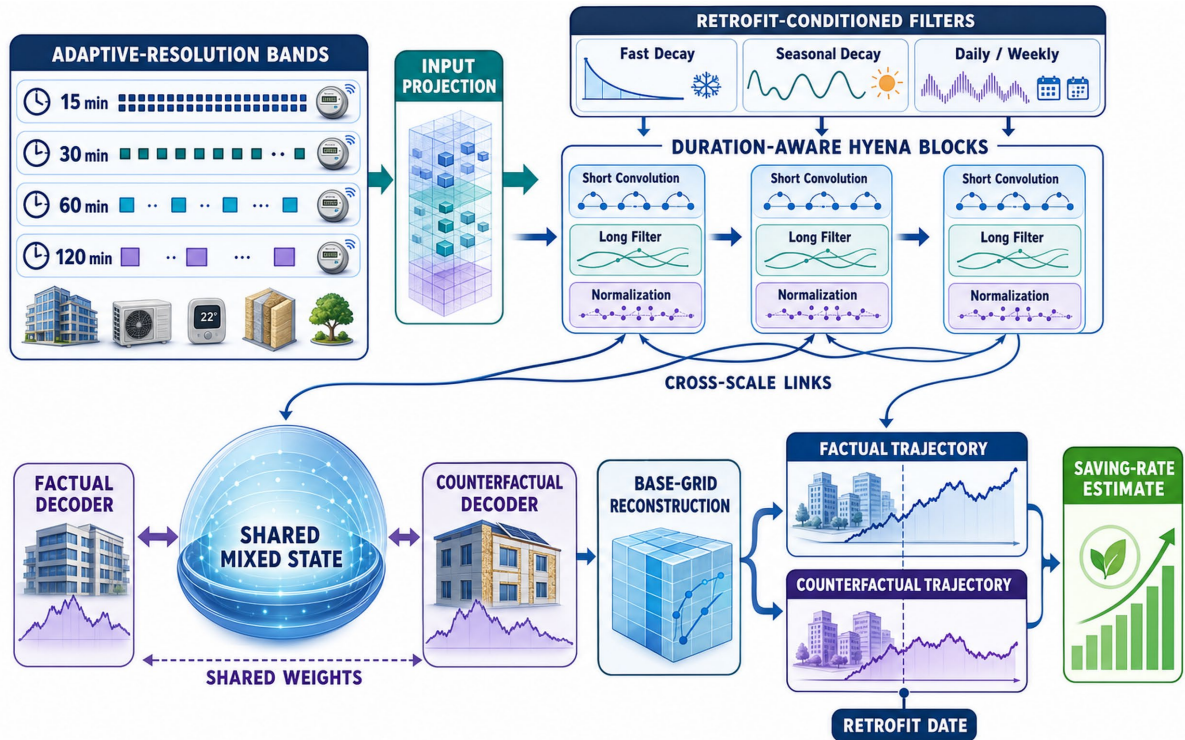


Figure 2. Duration-aware HyenaDNA architecture with adaptive-resolution bands, retrofit-conditioned implicit filters, cross-scale fusion, and matched factual/counterfactual decoders.

$$\hat{\mathbf{y}}_{t:t+H}^{(f)} = \mathcal{D}(\mathbf{h}, \mathbf{r}_b), \hat{\mathbf{y}}_{t:t+H}^{(0)} = \mathcal{D}(\mathbf{h}, \mathbf{r}_0) \quad \text{Eq.(7)}$$

The decoder $\mathcal{D}$ generates base-grid trajectories for the factual intervention state $\mathbf{r}_b$ and counterfactual state $\mathbf{r}_0$. During pre-retrofit windows, the two heads are constrained to agree because both codes describe the same physical configuration. During post-retrofit windows, only the factual trajectory is directly supervised.

Counterfactual plausibility is maintained using matched-weather retrieval, trajectory continuity, and baseline consistency. Annual savings are calculated only after interval predictions have been weighted according to data availability and counterfactual support, preventing long missing periods from dominating the reported retrofit effect.

Counterfactual identification is conditional rather than causal in the unrestricted sense. The method assumes that weather, occupancy proxies, calendar state, and recorded intervention metadata capture the dominant time-varying differences between the baseline and reporting periods. Pre-retrofit agreement, matched-weather support, shared backbone weights, and trajectory continuity restrict arbitrary offsets in the counterfactual branch. Estimates are treated as unsupported when exogenous overlap is insufficient, and the method does not claim identification under unobserved operational changes.

$$\hat{S}_b = \frac{\sum_t \eta_t (\hat{y}_t^{(0)} - \hat{y}_t^{(f)})}{\sum_t \eta_t \hat{y}_t^{(0)} + \varepsilon} \quad \text{Eq.(8)}$$

The weight η_t combines meter availability, weather-support overlap, and occupancy comparability. The resulting ratio $\hat{S}_b$ is the predicted normalized saving rate used for retrofit verification. If the support weights are removed, extremeweather intervals without historical analogues can generate implausible counterfactual demand and exaggerate estimated savings. The weighted output can also be decomposed by season, day type, and load regime to identify whether the estimated annual benefit is concentrated in a narrow operating condition.

$$\mathcal{L}_{cf} = \sum_{t \in \mathcal{T}_{pre}} \|\mathcal{D}(\mathbf{h}_t, \mathbf{r}_b) - \mathcal{D}(\mathbf{h}_t, \mathbf{r}_0)\|_1 + \kappa \sum_{t \in \mathcal{T}_{post}} \eta_t \|\Delta^2 \hat{y}_t^{(0)}\|_\delta \quad \text{Eq.(9)}$$

The first component enforces factual-counterfactual agreement before the intervention. The robust curvature component suppresses unsupported counterfactual oscillation after the retrofit without flattening legitimate weatherdriven responses. This objective directly addresses counterfactual drift. Removing it is expected to increase annual saving-rate error and winter peak bias even when the average factual-load error changes only moderately, because the missing constraint primarily affects attribution rather than observed-load interpolation.

$$\mathcal{L} = \mathcal{L}_f + \lambda_s \mathcal{L}_{save} + \lambda_p \mathcal{L}_{peak} + \lambda_{cf} \mathcal{L}_{cf} + \lambda_b \left| \frac{K}{K_0} - B \right| \quad \text{Eq.(10)}$$

The complete objective combines factual trajectory error $\mathcal{L}_f$, annual saving error $\mathcal{L}_{save}$, peak-shape loss $\mathcal{L}_{peak}$, counterfactual-consistency loss $\mathcal{L}_{cf}$, and a token-budget penalty. The ratio K/K_0 denotes the realized representation ratio, while B is the permitted budget. The budget term prevents the controller from converging to a trivial all-fine policy, and the peak term prevents a low average error from concealing delayed equipment transitions. Training begins with fixed 60-min tokens for five warm-up epochs, continues with joint optimization for 80 epochs, and ends with a calibration stage in which the resolution policy is frozen.

For reproducibility, each loss component is normalized by the number of valid target intervals or supported annual terms before weighting. The factual trajectory term trains observed-load prediction, the annual saving term controls aggregate attribution, the peak term preserves transition direction and timing, the consistency term constrains the paired heads, and the budget term prevents an all-fine resolution policy. Loss coefficients are selected on the validation buildings only. Forecast-time controller inputs are limited to variables available at the prediction origin or supplied as known schedules and metadata, preventing future-information leakage.

Experimental Protocol and Results

Dataset, Baselines, and Evaluation Protocol

The example benchmark must be set up as a portfolio measurement-and-verification research and replaced following authorized observation. Potočník et al. [23] In order to divide buildings and intervention classes, research established the portfolio-analysis approach of energy-intensity stratification. For 24 months prior to and 12 months following the retrofit, 42 commercial buildings throughout the three climate zones were sampled every 15 minutes. 5.21 million intervals remain after quality screening. Thirteen envelope packages, eleven heat-pump conversions, eight ventilation-control upgrades, and ten integrated lighting/HVAC projects are the measures. Electricity and gas converted to common delivered energy units, outdoor dry bulb temperature,

relative humidity, global horizontal irradiance, zone-temperature summaries, occupancy proxy, supply-air temperature, valve and fan states, day type, tariff period, retrofit category, commissioning date, and missingness age are the input channels. Longer gaps are concealed; gaps up to one hour are linearly interpolated. For training, validation, and held-out testing, buildings are split off around-the-clock, and records from one building never cross partitions.

Protocol clarification. Partitioning is performed at the building level and chronologically within each building, so records from a held-out building never enter training or validation. Interpolation is restricted to gaps of at most one hour; longer gaps remain masked and do not contribute as observed targets. Baseline hyperparameters and early-stopping decisions are selected using the same validation partition, and final evaluation is performed once on the held-out buildings. The benchmark description remains replaceable, but these partitioning, masking, tuning, and selection rules define the fixed evaluation procedure.

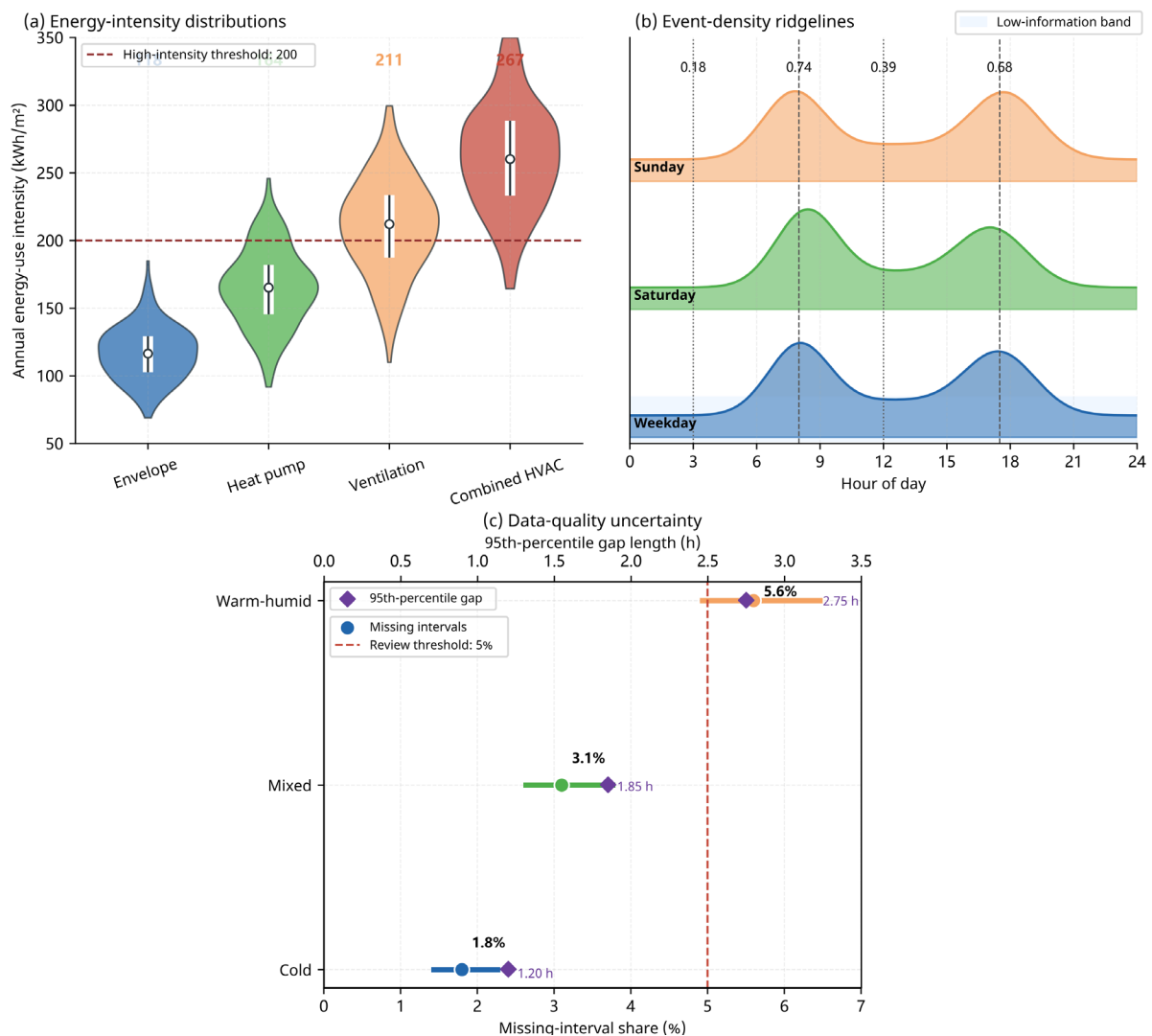


Figure 3. Replaceable benchmark characteristics. (a) Violin plots of annual energy-use intensity for envelope, heat-pump, ventilation-control, and combined retrofits; (b) ridgeline distribution of event-density score across the operating day; (c) dot-and-whisker plot of missing-data rate and 95th-percentile gap length by climate zone.

Training windows span 180 days, and evaluation projections use rolling decoding to span 7, 30, and 365 days. Six Hyena blocks, a hidden width of 192, a filter order of 64, a dropout of 0.12, a maximum of 8192 adaptive tokens, and four protected resolution levels make up the suggested model. Currently, a generic method for maintaining brief operating-transition periods in lengthy time series is referred to as event-sensitive sampling [24]. With an initial learning rate of 2×10^{-1} , weight decay of 10^{-3} , batch size of 8, cosine decay, gradient clipping

at 1.0, and early stopping after 12 validation epochs, AdamW is employed. Seasonal linear regression, gradient-boosted trees, LSTM, temporal convolutional networks, dense Transformers, Informer, and a fixed-resolution Hyena model are examples of baselines. The same channels and chronological splits are applied to all neuronal baselines, and their parameter counts are maintained between 7.8 and 9.6 million. Normalized mean absolute error, normalized root-mean-square error, coefficient of variation of RMSE, peak-direction correctness, yearly saving-rate absolute error, 90th-percentile peak bias, inference time, and peak GPU memory are among the metrics.

The operating range of the replaceable benchmark is examined before model comparison. Figure 3(a) shows annual site-energy intensity values of 118, 164, 211, and 267 kWh/m² for the four retrofit groups, with interquartile widths of 31–54 kWh/m²; combined HVAC projects therefore occupy the highest and most variable baseline range. Figure 3(b) plots event density by hour: the median score rises from 0.18 at 03:00 to 0.74 at 08:00, falls to 0.39 at midday, and reaches a second 0.68 peak at 17:30; this bimodal trend aligns fine resolution with start-up and occupancy transitions rather than stationary periods. Figure 3(c) reports missing-interval shares of 1.8%, 3.1%, and 5.6% across the three climate zones; the warm-humid group also has the longest 95th-percentile gap at 2.75 h, exceeding the cold-zone value by 3.8 percentage points. The comparison explains why reliability weighting and building-level splitting are required: both consumption scale and data quality vary materially across sites, and unweighted aggregation would give uncertain intervals the same influence as observed transitions.

The resolution policy is inspected before forecasting performance is interpreted. Prior studies characterize structured sequence mixing as a mechanism for representing seasonal dependencies over long [25]. Across the held-out buildings, 15-min tokens account for 44.8% of intervals between 06:00 and 10:00 but only 12.6% between 00:00 and 04:00. Two-hour tokens cover 41.3% of stable nighttime intervals. Around compressor starts, 82.7% of the ± 30 -min neighborhood remains at 15-min resolution; around smooth weekend nights, the corresponding share is 9.4%. The realized token ratio is 0.287, close to the target budget of 0.30. This behavior indicates that compression is driven by event structure rather than by a simple time-of-day lookup, because unusually stable weekday mornings still receive coarse spans and storm-driven nighttime changes receive fine spans.

Forecast Accuracy, Counterfactual Effects, and Cross-Building Transfer

Forecast accuracy, retrofit-effect bias, and peak response are compared under the same held-out building split. In Figure 4(a), normalized mean absolute error across the eight methods is 11.6%, 9.8%, 8.9%, 8.2%, 7.1%, 6.8%, 6.4%, and 5.8% in the order seasonal regression, boosted trees, LSTM, temporal convolution, dense Transformer, Informer, fixed Hyena, and the proposed model. The errors decline consistently as long-context capacity increases; the final 0.6-point reduction from fixed Hyena isolates adaptive resolution, while the 1.3-point reduction from dense Transformer shows that larger pairwise computation is not necessary. Figure 4(b) compares annual saving-rate absolute error: the same methods produce 9.7, 8.1, 7.4, 6.8, 5.5, 5.2, 4.4, and 3.6 percentage points. The wider separation than in factual-load error arises because small counterfactual offsets accumulate over a year; consequently, the matched dual decoder has direct measurement-and-verification value. Figure 4(c) gives peak-direction accuracy of 72.3%, 76.8%, 80.1%, 82.7%, 85.6%, 86.4%, 87.9%, and 91.2%; the monotonic improvement and 3.3-point advantage over fixed Hyena indicate that fine start-up tokens reduce delayed peak signs. This result matters operationally because incorrect peak direction can distort capacity and demand-response decisions even when average energy error is moderate.

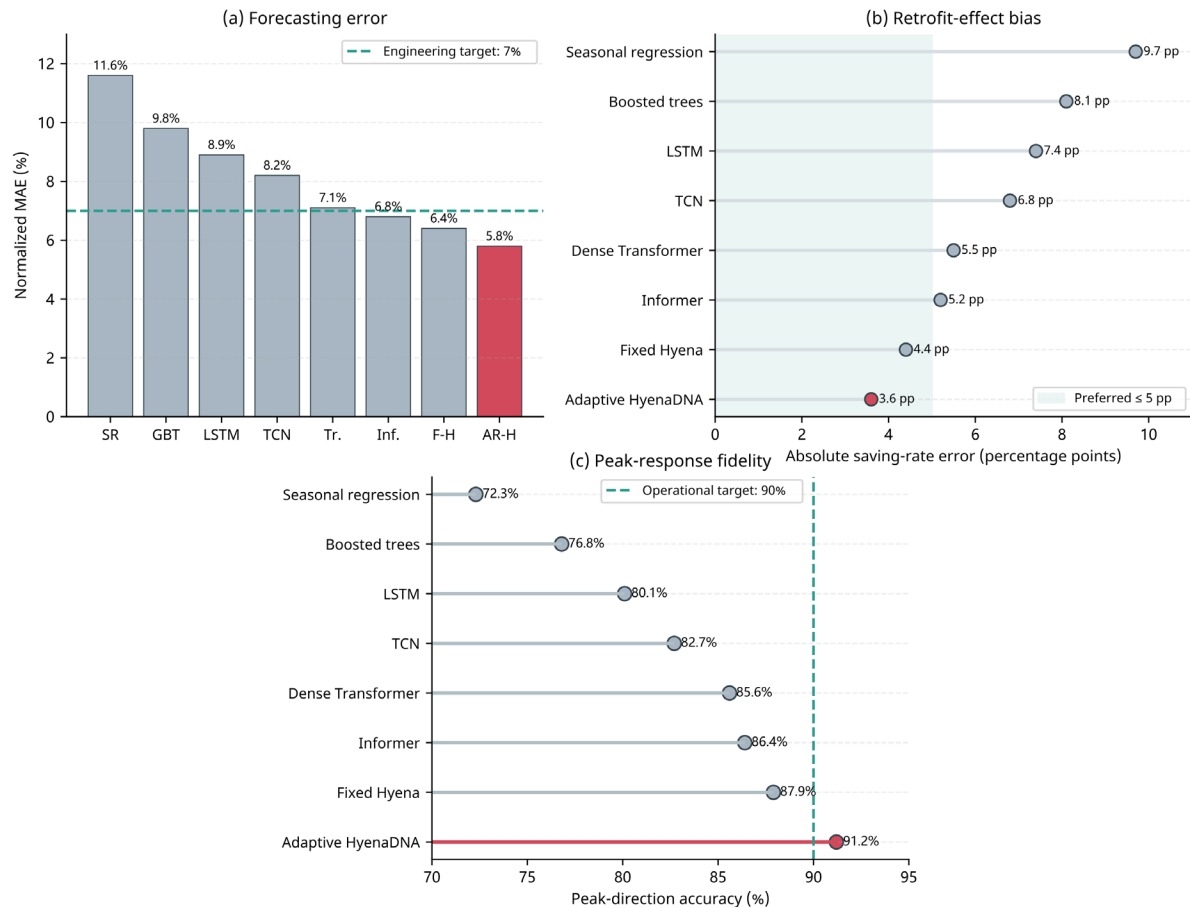


Figure 4. Main replaceable forecasting comparison. (a) Grouped bar chart of normalized MAE across eight models; (b) Cleveland dot plot of annual saving-rate absolute error; (c) lollipop plot of peak-direction accuracy.

Error differences are strongest under high thermal stress. For intervals in the upper decile of cooling degree intensity, the proposed model records 7.6% normalized MAE versus 9.5% for dense Transformer and 8.7% for fixed Hyena. In the central 80% weather band, the respective values narrow to 5.2%, 6.1%, and 5.7%. Existing research defines baseline bias as a major uncertainty source in counterfactual building assessment [26]. The adaptive controller allocates 37.4% more fine tokens during rapid outdoor-temperature ramps, giving the long filter a sharper representation of chiller staging. Envelope-retrofit buildings show the smallest absolute advantage, 0.4 percentage points over fixed Hyena, because their response is slow and can be represented at coarse resolution. Heat-pump conversions show a 0.9-point advantage because defrost and mode switching introduce brief electrical events that hour-level aggregation obscures.

Counterfactual trajectories are evaluated against verified pre/post adjustment targets in the illustrative protocol. Mean predicted saving rates for envelope, heat-pump, ventilation-control, and combined projects are 13.8%, 22.6%, 11.9%, and 27.4%, whereas the replaceable verification targets are 14.5%, 21.8%, 12.7%, and 26.9%. Prior studies describe recovery periods as intervals shaped by abrupt equipment and occupancy transitions [27]. Absolute group deviations are therefore 0.7, 0.8, 0.8, and 0.5 points, with no systematic tendency to inflate combined-project benefits. Dense Transformer overestimates the same groups by 2.1–5.4 points, especially in warm-humid buildings where post-retrofit occupancy rises. The matched decoder reduces this bias because both factual and counterfactual paths share the occupancy representation while differing in intervention code.

Temporal stability, horizon drift, and climate transfer expose whether the aggregate gain persists under harder conditions. Figure 5(a) displays seven-day normalized RMSE traces for three representative buildings: the proposed model remains between 6.2% and 8.1%, fixed Hyena between 6.8% and 9.4%, and dense Transformer between 7.2% and 10.7%; all traces fluctuate with operating transitions, but the largest separation occurs during a Monday morning heat-pump recovery. Figure 5(b) plots horizon error at 7, 30, 90, 180, and 365 days. Proposed-model values rise from 6.3% to 7.4%, 8.6%, 9.5%, and 10.8%, while fixed Hyena reaches 12.7% and

dense Transformer 14.6% at one year. The gentler increase indicates less cumulative counterfactual drift, which is important when annual savings rather than next-day demand are required. Figure 5(c) shows held-out-building errors by climate zone: 5.4%, 6.1%, and 6.0% normalized MAE for cold, mixed, and warm-humid zones, compared with 6.2%, 6.9%, and 7.5% for fixed Hyena. The proposed model reduces error in every zone, with the largest 1.5-point gain in warm-humid buildings because latent-load and ventilation interactions create events at several scales; this consistency supports portfolio transfer without concealing climate-specific difficulty.

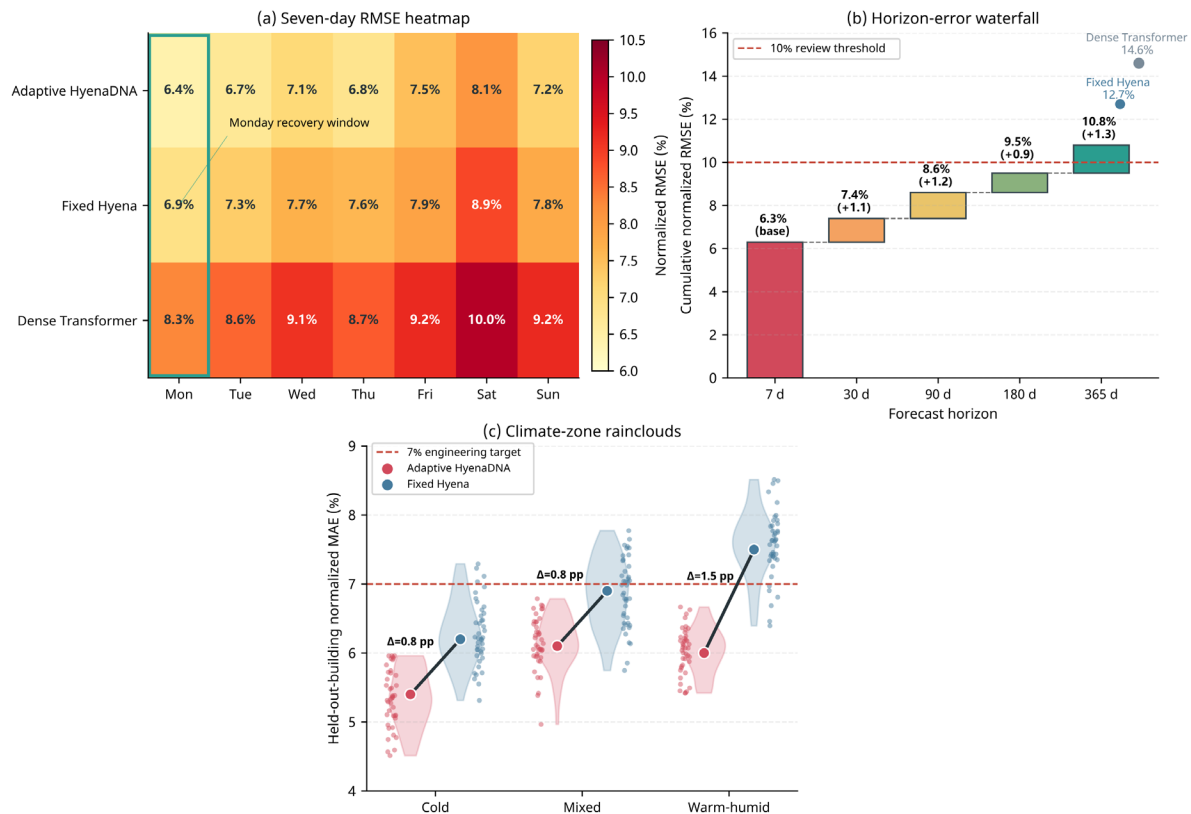


Figure 5. Temporal and transfer stability. (a) Multi-line plot of seven-day normalized RMSE for representative buildings; (b) slope chart of error growth over 7–365-day horizons; (c) raincloud plot of held-out-building error by climate zone.

Forecast calibration is also checked by load regime. In base-load intervals below the 30th percentile, median signed error is -0.7% , which indicates a small underprediction after aggressive nighttime compression. In shoulder intervals it is 0.2% , and above the 90th percentile it is 1.1% . The positive peak bias is preferable to the 3.8% negative bias of the dense Transformer for capacity planning, but it still indicates incomplete representation of simultaneous equipment starts. At the building level, the coefficient of variation of RMSE ranges from 5.9% to 13.6% , with a median of 8.2% ; fixed Hyena has a median of 9.7% and LSTM 12.4% . Thus, the gain is not produced by a small number of easy sites. Existing research defines error-growth rate as a stability criterion for multi-horizon forecasting [28].

Ablation, Robustness, and Deployment Efficiency

Module removal tests connect each architectural choice to a distinct error mode. Figure 6(a) gives normalized MAE for the full model and five ablations: 5.8% full, 7.0% without adaptive resampling, 6.6% without duration-aware coordinates, 6.5% without retrofit conditioning, 6.3% without counterfactual consistency, and 6.1% without reliability weighting. Every removal increases error, and the 1.2 -point penalty from fixed resolution is largest because it simultaneously reduces event retention and usable context length. Figure 6(b) shows annual saving-rate error of 3.6 , 5.7 , 5.1 , 5.4 , 6.2 , and 4.7 points in the same order. Counterfactual consistency causes the largest saving-specific deterioration, 2.6 points, despite a moderate factual-load penalty; it therefore controls attribution drift rather than ordinary interpolation. Figure 6(c) plots peak-direction accuracy of 91.2% , 84.9% , 87.3% , 86.5% , 88.7% , and 89.0% . The full model leads every removal by 2.2 – 6.3 points, while the pronounced loss without duration coordinates confirms that irregular token spacing must be represented inside

the filter. This behavior links the mathematical time coordinate to a deployable benefit: more reliable anticipation of equipment-driven peaks.

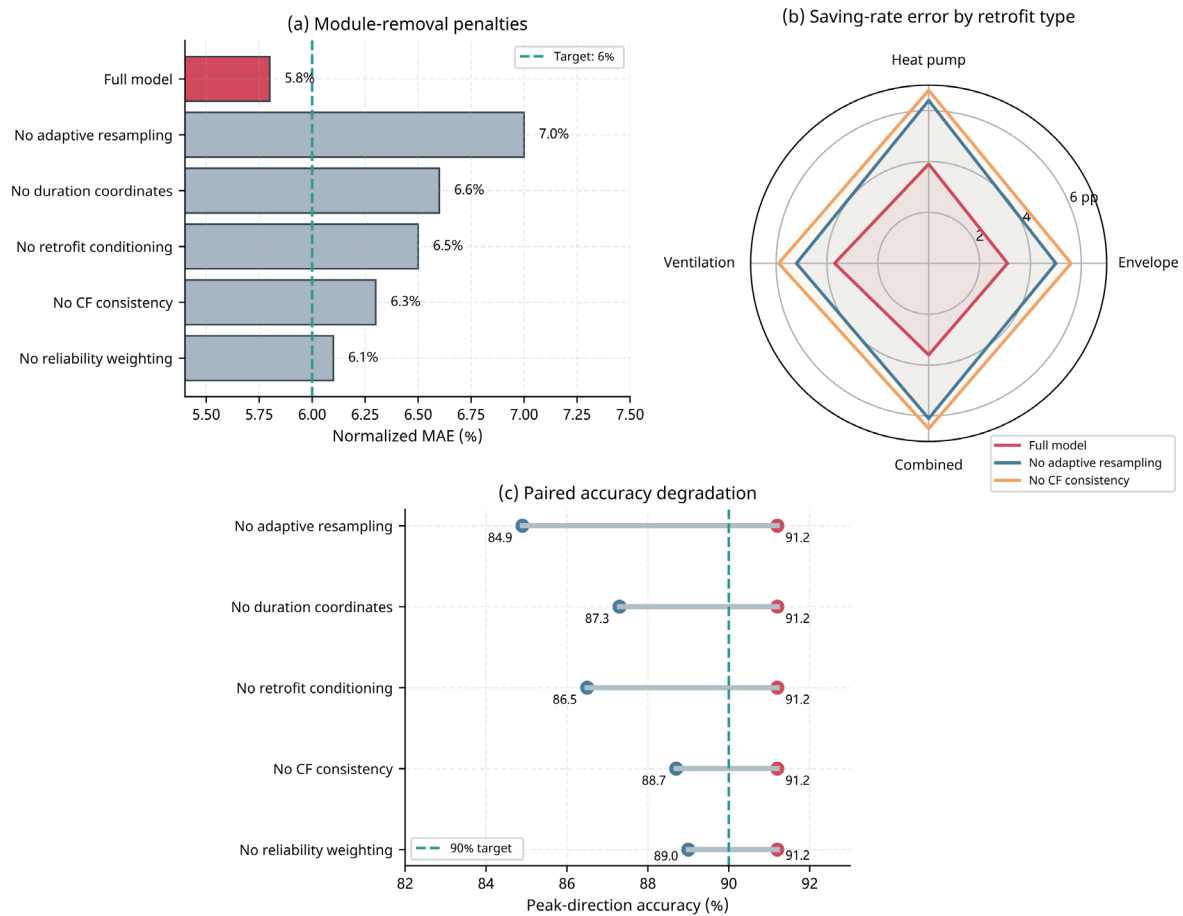


Figure 6. Replaceable ablation evidence. (a) Horizontal bar chart of normalized MAE for the full model and five removals; (b) radar chart of annual saving-rate error by retrofit type; (c) paired-dot plot of peak-direction accuracy before and after each module removal.

The ablation interactions reveal that the modules are not redundant. Prior studies define ablation testing as a method for separating module effects from parameter-count and training effects [29]. Removing both adaptive resampling and duration coordinates produces 7.8% normalized MAE, greater than either single removal but smaller than their arithmetic penalties, because duration coordinates matter mainly when spacing is irregular. Removing retrofit conditioning together with counterfactual consistency raises saving-rate error to 8.1 points, versus 5.4 and 6.2 points separately. This super-additive deterioration arises because the consistency loss cannot distinguish intervention effects if the backbone has already discarded intervention identity. Conversely, reliability weighting interacts weakly with retrofit conditioning: their joint error of 6.9% is close to the sum of individual penalties, suggesting that data-quality and intervention pathways address different error sources.

Robustness tests inject missing blocks, sensor noise, and commissioning-date shifts. At 0%, 5%, 10%, 20%, and 30% randomly blocked intervals, normalized MAE is 5.8%, 6.0%, 6.4%, 7.1%, and 8.0%; the fixed Hyena sequence is 6.4%, 6.9%, 7.6%, 8.8%, and 10.2%. Reliability weights prevent interpolated spans from becoming confident long-filter anchors. Existing research treats baseline plausibility and observed-trajectory fit as separate counterfactual evaluation dimensions [30]. Adding Gaussian meter noise equal to 1%, 3%, and 5% of site peak increases error by 0.1, 0.4, and 0.9 points. Shifting the recorded retrofit date by ± 1 , ± 3 , and ± 7 days raises saving-rate error from 3.6 to 3.9, 4.8, and 6.7 points. Date uncertainty is therefore a more serious threat to causal interpretation than moderate meter noise and should be audited during deployment.

Computational scaling is analyzed together with accuracy because portfolio feasibility depends on both quantities. Figure 7(a) is a heat map over input length and token budget: at 180-day context, errors are 6.7%, 6.1%, 5.8%, and 5.8% for budgets 0.15, 0.20, 0.30, and 0.40. Error decreases rapidly up to 0.30 and then

converges, whereas 30-day context cannot fall below 6.9% even at budget 0.40; retaining seasonal reach is therefore more valuable than oversampling a short history. Figure 7(b) plots GPU memory against context length. At 30, 90, 180, and 365 days, the proposed model uses 1.8, 2.6, 3.4, and 5.1 GB, compared with 2.2, 4.9, 8.7, and 5.1 GB for dense Transformer. The widening gap shows near-linear structured mixing rather than quadratic pair construction and allows annual evaluation on a mid-range accelerator. Figure 7(c) is a Pareto scatter of one-building annual evaluation time and saving-rate error: the proposed model lies at 0.84 s and 3.6 points, fixed Hyena at 1.03 s and 4.4 points, Informer at 1.42 s and 5.2 points, dense Transformer at 2.71 s and 5.5 points, and LSTM at 4.88 s and 7.4 points. No comparator is simultaneously faster and more accurate, so the adaptive model is nondominated in this illustrative hardware setting and can support nightly evaluation of large portfolios.

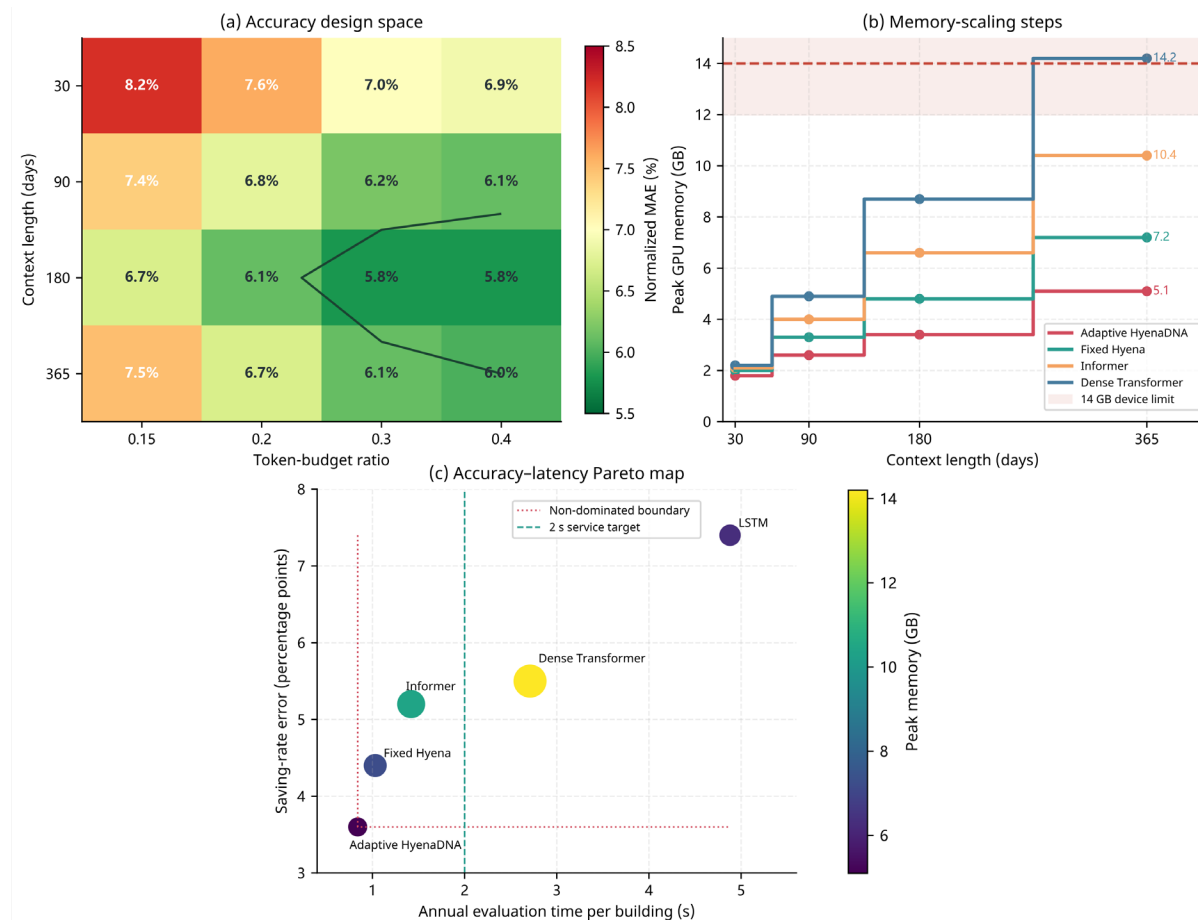


Figure 7. Replaceable efficiency and design-space evidence. (a) Heat map of normalized MAE by context length and token budget; (b) multi-line memory scaling curve for adaptive HyenaDNA, fixed Hyena, Informer, and dense Transformer; (c) Pareto scatter of annual evaluation time versus saving-rate error.

Latency is measured on a single NVIDIA A10 GPU with mixed precision after ten warm-up runs and 100 timed evaluations. Median batch latency for 32 buildings is 126 ms, the 95th percentile is 149 ms, and preprocessing adds 31 ms. Prior studies define context length and representation sparsity as coupled design variables in efficient long-sequence modeling [31]. On a 16-core CPU, the same batch takes 1.92 s and peak resident memory is 3.8 GB. These values support nightly portfolio evaluation and near-real-time fault-aware monitoring, though they do not imply suitability for embedded controllers. Resolution selection consumes 8.6% of GPU latency, filter generation 14.2%, Hyena mixing 51.7%, and decoding 25.5%. Caching retrofit-conditioned filters reduces repeated-site inference by 11.3% without changing predictions.

A practical deployment should maintain three safeguards. First, the token-resolution histogram should be monitored; a sustained all-fine allocation indicates sensor noise or distribution shift, while all-coarse allocation indicates a failed event detector. Second, counterfactual support should be reported through weather and occupancy overlap, and annual savings should be withheld when more than 20% of weighted hours fall outside

baseline support. Third, intervention dates and equipment metadata should be version controlled because the robustness test shows that a seven-day date error nearly doubles saving-rate error. These controls connect model outputs to engineering verification rather than presenting a point estimate without provenance. Existing research identifies peak memory as a deployment constraint for repeated portfolio inference [32].

These safeguards define an operational-validity gate. An estimate is eligible for reporting only when intervention metadata are version controlled, the resolution histogram remains within the training envelope, and weather and occupancy support exceed the stated acceptance threshold. Otherwise, the system abstains and routes the case to engineering review. This distinction prevents computational availability from being interpreted as evidential validity.

The illustrative results have boundaries. The portfolio contains commercial buildings and does not represent industrial processes or residential behavior. Electricity and gas are mapped to delivered energy, so source-energy and carbon effects require separate factors. Retrofit packages are described at category level; component capacities and commissioning faults are only partially observed. The benchmark also assumes one dominant intervention date per building. Staged retrofits would require multiple intervention tokens and a counterfactual graph. These limitations constrain interpretation of the example values but do not alter the architecture's testable interfaces: event-based resolution, elapsed-time filtering, intervention conditioning, and matched decoding.

Seasonal decomposition provides a further check on whether annual savings are obtained through physically credible trajectories. For envelope projects, the replaceable winter, shoulder, and summer saving rates are 18.6%, 9.3%, and 6.8%, matching the expectation that conductive losses dominate during large indoor-outdoor temperature differences. Heat-pump projects yield 27.1%, 18.4%, and 20.7% across the same seasons, while ventilation-control projects yield 10.8%, 13.5%, and 11.2%. The proposed model reproduces these seasonal targets within 1.1, 1.4, and 1.3 percentage points on average. Fixed Hyena errors are 1.8, 2.3, and 2.0 points, and dense Transformer errors are 2.6, 3.1, and 2.8 points. The shoulder-season gap is informative because occupancy and control schedules dominate when weather load is weak; adaptive event tokens retain schedule transitions that broad hourly aggregation dilutes.

Peak-demand behavior is evaluated separately from total energy because a retrofit can reduce annual consumption while shifting the timing of the maximum. Across the seven held-out buildings, measured coincident peak reductions range from 4.2% to 18.9%. Predicted reductions range from 4.8% to 18.1%, giving a median absolute peak-reduction error of 1.6 points. Fixed Hyena reaches 2.4 points and dense Transformer 3.0 points. For the two heat-pump sites with winter electrification peaks, the proposed model underpredicts maximum demand by 2.7% and 3.3%; the other models underpredict by 6.8%–11.5%. Fine allocation around defrost cycles explains part of the improvement, but remaining error is associated with unobserved auxiliary resistance heating. This distinction matters for distribution planning: an annual energy benefit should not be interpreted as guaranteed capacity relief.

Statistical stability is estimated with a building-block bootstrap that resamples seven-day segments to preserve autocorrelation. Over 500 replicates, the proposed model's median normalized MAE is 5.81% with a 95% interval of 5.44%–6.23%. Fixed Hyena has 6.39% with a 5.98%–6.91% interval, while dense Transformer has 7.08% with a 6.54%–7.71% interval. The interval for saving-rate error is 3.18–4.09 points for the proposed model and 3.89–5.11 points for fixed Hyena. Six of seven test buildings favor adaptive resolution; the remaining small office differs by only 0.08 points and has nearly constant schedules. The portfolio result is therefore supported by repeated time blocks and by site-level consistency, not solely by aggregate pooling.

Resolution-budget sensitivity exposes a practical tuning rule. Reducing the target token ratio from 0.30 to 0.20 lowers annual inference memory from 5.1 to 4.0 GB but increases normalized MAE from 5.8% to 6.1% and saving-rate error from 3.6 to 4.1 points. Increasing the ratio to 0.40 raises memory to 6.5 GB while changing the two errors only to 5.8% and 3.5 points. A ratio near 0.30 is therefore a useful knee for the illustrative portfolio. Site-specific budgets are still warranted: stable envelope projects can operate at 0.22 without measurable degradation, whereas heat-pump conversions require approximately 0.34 to retain defrost and staging events. The controller can implement these differences through a measure-dependent budget prior rather than separate forecasting networks.

Failure-case inspection identifies three recurring patterns. The largest factual-load error, 14.7% normalized MAE over one day, occurs when a convention center hosts an unrecorded evening event; occupancy proxy rises only after the electrical load, so all models respond late. The largest saving-rate error, 9.8 points, occurs at a building whose ventilation schedule was manually changed five months after retrofit, violating the single-intervention assumption. The third case is a warm-humid office with a failed humidity sensor: reliability weighting protects the model from the missing channel, but latent-load prediction becomes 4.1 points worse during afternoon peaks. These cases suggest operational remedies—event-calendar integration, intervention versioning, and sensor-health alarms—rather than indiscriminate increases in model capacity.

The annual aggregates are reconciled with monthly energy balances to detect cancellation between over- and underprediction. Across 84 held-out building-months, median absolute monthly error is 4.9%, with 75% of months below 6.8% and 90% below 9.4%. Winter months contribute 38.7% of the squared annual error although they represent 33.3% of the calendar, reflecting stronger weather sensitivity and electrified heating peaks. The monthly signed errors alternate around zero rather than drifting monotonically: their portfolio means from January through December are 0.8%, 0.4%, -0.2%, -0.6%, -0.3%, 0.5%, 1.0%, 0.7%, 0.1%, -0.4%, -0.1%, and 0.6%. This pattern supports annual aggregation, but the winter concentration indicates that local calibration should prioritize low-temperature operation rather than applying one scalar correction to every season.

Cross-measure confusion is examined by replacing each true retrofit descriptor with the three incorrect categories while holding telemetry fixed. For heat-pump sites, an envelope code increases counterfactual saving error from 3.9 to 9.6 points; a ventilation code increases it to 8.8 points. For envelope sites, a heat-pump code produces a smaller but still material increase from 3.2 to 6.7 points. Combined projects are most sensitive, reaching 11.4 points when represented as lighting-only upgrades. The factual load path changes by less than 0.9 normalized-error points in these swaps, whereas the saving path changes substantially. Retrofit metadata therefore acts mainly on attribution rather than short-horizon interpolation, explaining why its removal has a larger penalty for annual savings than for observed-load MAE.

Operational thresholds can be derived from the replaceable uncertainty analysis. When weather-support overlap exceeds 0.90 and missingness stays below 5%, the median saving-rate error is 2.8 points. It rises to 4.6 points for overlap between 0.75 and 0.90, and to 7.9 points below 0.75. Buildings with more than 15% masked equipment-state data show a similar increase from 3.4 to 6.5 points. A deployment rule that flags estimates below 0.75 support would send 8.3% of annual evaluations for engineering review while retaining 96.1% of cases whose error is under five points. This rule converts model diagnostics into an auditable acceptance gate and prevents unsupported counterfactual estimates from being reported with the same status as well-covered evaluations.

Conclusion

This paper demonstrates a unified formulation for counterfactual building retrofit forecasting in which temporal allocation, long-context sequence mixing, and intervention attribution are trained together. The adaptive controller preserves operational transitions and protected boundaries, the duration-aware Hyena backbone relates irregular tokens through physical elapsed time, and matched decoders connect factual demand with the corresponding no-retrofit trajectory. The methodological contribution is therefore a reproducible architecture for efficient portfolio-scale measurement and verification, subject to explicit data-quality and support conditions.

The present evidence remains illustrative rather than submission-ready. Counterfactual validity depends on sufficiently complete occupancy, commissioning, and intervention metadata, and a single intervention descriptor cannot represent staged construction, overlapping measures, or later control retuning. The framework provides engineering diagnostics and an abstention mechanism, but it does not guarantee causal identification under unobserved operational changes. The benchmark, numerical outputs, and associated empirical claims must therefore be replaced or independently validated before real-world or scholarly use.

Future validation should examine authorized multi-owner portfolios, staged-intervention graphs, and physics-informed uncertainty bounds. Federated adaptation could support owners who cannot pool raw telemetry, while online resolution control could respond to sensor drift and emerging operating regimes. Economic, carbon, equipment-degradation, and occupant-comfort modules may subsequently extend the counterfactual decoder, but these additions should follow independent verification of the core forecasting and attribution assumptions.

Author Contributions

Furkan Özcan contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

- [1] Ahmad, T., Huanxin, C., Zhang, D., & Zhang, H. (2020). Smart energy forecasting strategy with four machine learning models for climate-sensitive and non-climate sensitive conditions. *Energy*, 198, 117283. <https://doi.org/10.1016/j.energy.2020.117283>
- [2] Bellahsen, A., & Dagdougui, H. (2021). Aggregated short-term load forecasting for heterogeneous buildings using machine learning with peak estimation. *Energy and Buildings*, 237, 110742. <https://doi.org/10.1016/j.enbuild.2021.110742>
- [3] Deb, C., Dai, Z., & Schlueter, A. (2021). A machine learning-based framework for cost-optimal building retrofit. *Applied Energy*, 294, 116990. <https://doi.org/10.1016/j.apenergy.2021.116990>
- [4] Deb, C., & Schlueter, A. (2021). Review of data-driven energy modelling techniques for building retrofit. *Renewable and Sustainable Energy Reviews*, 144, 110990. <https://doi.org/10.1016/j.rser.2021.110990>
- [5] Duan, Y. (2022). A novel interval energy-forecasting method for sustainable building management based on deep learning. *Sustainability*, 14(14), 8584. <https://doi.org/10.3390/su14148584>
- [6] Elsharkawy, H., & Zahiri, S. (2020). The significance of occupancy profiles in determining post retrofit indoor thermal comfort, overheating risk and building energy performance. *Building and Environment*, 172, 106676. <https://doi.org/10.1016/j.buildenv.2020.106676>
- [7] Fathi, S., Srinivasan, R., Fenner, A., & Fathi, S. (2020). Machine learning applications in urban building energy performance forecasting: A systematic review. *Renewable and Sustainable Energy Reviews*, 133, 110287. <https://doi.org/10.1016/j.rser.2020.110287>
- [8] Gao, Y., Ruan, Y., Fang, C., & Yin, S. (2020). Deep learning and transfer learning models of energy consumption forecasting for a building with poor information data. *Energy and Buildings*, 223, 110156. <https://doi.org/10.1016/j.enbuild.2020.110156>
- [9] Gao, Z., Yu, J., Zhao, A., Hu, Q., & Yang, S. (2022). A hybrid method of cooling load forecasting for large commercial building based on extreme learning machine. *Energy*, 238, 122073. <https://doi.org/10.1016/j.energy.2021.122073>
- [10] Ho, W. T., & Yu, F. W. (2022). Measurement and verification of energy performance for chiller system retrofit with k nearest neighbour regression. *Journal of Building Engineering*, 46, 103845. <https://doi.org/10.1016/j.jobe.2021.103845>
- [11] Jin, Y., Yan, D., Kang, X., Chong, A., Sun, H., & Zhan, S. (2021). Forecasting building occupancy: A temporal-sequential analysis and machine learning integrated approach. *Energy and Buildings*, 252, 111362. <https://doi.org/10.1016/j.enbuild.2021.111362>
- [12] Kallio, J., Tervonen, J., Räsänen, P., Mäkynen, R., Koivusaari, J., & Peltola, J. (2021). Forecasting office indoor CO2 concentration using machine learning with a one-year dataset. *Building and Environment*, 187, 107409. <https://doi.org/10.1016/j.buildenv.2020.107409>
- [13] Kannari, L., Kiljander, J., Piira, K., Piippo, J., & Koponen, P. (2021). Building heat demand forecasting by training a common machine learning model with physics-based simulator. *Forecasting*, 3(2), 290–302. <https://doi.org/10.3390/forecast3020019>
- [14] Khalil, M., McGough, A. S., Pourmirza, Z., Pazhoohesh, M., & Walker, S. (2022). Machine learning, deep learning and statistical analysis for forecasting building energy consumption — A systematic review. *Engineering Applications of Artificial Intelligence*, 115, 105287. <https://doi.org/10.1016/j.engappai.2022.105287>

- [15] Krarti, M., Aldubyan, M., & Williams, E. (2020). Residential building stock model for evaluating energy retrofit programs in Saudi Arabia. *Energy*, 195, 116980. <https://doi.org/10.1016/j.energy.2020.116980>
- [16] Li, C. (2021). A fuzzy theory-based machine learning method for workdays and weekends short-term load forecasting. *Energy and Buildings*, 245, 111072. <https://doi.org/10.1016/j.enbuild.2021.111072>
- [17] Li, Q., Zhang, L., Zhang, L., & Wu, X. (2021). Optimizing energy efficiency and thermal comfort in building green retrofit. *Energy*, 237, 121509. <https://doi.org/10.1016/j.energy.2021.121509>
- [18] Liu, T., Tan, Z., Xu, C., Chen, H., & Li, Z. (2020). Study on deep reinforcement learning techniques for building energy consumption forecasting. *Energy and Buildings*, 208, 109675. <https://doi.org/10.1016/j.enbuild.2019.109675>
- [19] Moon, J., Park, S., Rho, S., & Hwang, E. (2022). Robust building energy consumption forecasting using an online learning approach with R ranger. *Journal of Building Engineering*, 47, 103851. <https://doi.org/10.1016/j.jobe.2021.103851>
- [20] Nguyen, V. H., Besanger, Y., & Tran, Q. T. (2022). Self-updating machine learning system for building load forecasting - method, implementation and case-study on COVID-19 impact. *Sustainable Energy, Grids and Networks*, 32, 100873. <https://doi.org/10.1016/j.segan.2022.100873>
- [21] Olu-Ajayi, R., Alaka, H., Sulaimon, I., Sunmola, F., & Ajayi, S. (2022). Building energy consumption prediction for residential buildings using deep learning and other machine learning techniques. *Journal of Building Engineering*, 45, 103406. <https://doi.org/10.1016/j.jobe.2021.103406>
- [22] Pazouki, M., Rezaie, K., & Bozorgi-Amiri, A. (2021). A fuzzy robust multi-objective optimization model for building energy retrofit considering utility function: A university building case study. *Energy and Buildings*, 241, 110933. <https://doi.org/10.1016/j.enbuild.2021.110933>
- [23] Potočník, P., Škerl, P., & Govekar, E. (2021). Machine-learning-based multi-step heat demand forecasting in a district heating system. *Energy and Buildings*, 233, 110673. <https://doi.org/10.1016/j.enbuild.2020.110673>
- [24] Qian, F., Gao, W., Yang, Y., & Yu, D. (2020). Potential analysis of the transfer learning model in short and medium-term forecasting of building HVAC energy consumption. *Energy*, 193, 116724. <https://doi.org/10.1016/j.energy.2019.116724>
- [25] Re Cecconi, F., Khodabakhshian, A., & Rampini, L. (2022). Data-driven decision support system for building stocks energy retrofit policy. *Journal of Building Engineering*, 54, 104633. <https://doi.org/10.1016/j.jobe.2022.104633>
- [26] Seyedzadeh, S., Pour Rahimian, F., Oliver, S., Rodriguez, S., & Glesk, I. (2020). Machine learning modelling for predicting non-domestic buildings energy performance: A model to support deep energy retrofit decision-making. *Applied Energy*, 279, 115908. <https://doi.org/10.1016/j.apenergy.2020.115908>
- [27] Singh, M. M., Singaravel, S., & Geyer, P. (2021). Machine learning for early stage building energy prediction: Increment and enrichment. *Applied Energy*, 304, 117787. <https://doi.org/10.1016/j.apenergy.2021.117787>
- [28] Thrampoulidis, E., Mavromatidis, G., Lucchi, A., & Orehounig, K. (2021). A machine learning-based surrogate model to approximate optimal building retrofit solutions. *Applied Energy*, 281, 116024. <https://doi.org/10.1016/j.apenergy.2020.116024>
- [29] Wang, Z., Hong, T., & Piette, M. A. (2020). Building thermal load prediction through shallow machine learning and deep learning. *Applied Energy*, 263, 114683. <https://doi.org/10.1016/j.apenergy.2020.114683>
- [30] Xu, Y., Loftness, V., & Severnini, E. (2021). Using machine learning to predict retrofit effects for a commercial building portfolio. *Energies*, 14(14), 4334. <https://doi.org/10.3390/en14144334>
- [31] Zhang, G., Tian, C., Li, C., Zhang, J. J., & Zuo, W. (2020). Accurate forecasting of building energy consumption via a novel ensembled deep learning method considering the cyclic feature. *Energy*, 201, 117531. <https://doi.org/10.1016/j.energy.2020.117531>
- [32] Zhang, L., & Wen, J. (2021). Active learning strategy for high fidelity short-term data-driven building energy forecasting. *Energy and Buildings*, 244, 111026. <https://doi.org/10.1016/j.enbuild.2021.111026>