

Privacy-Preserving Federated Learning for Building Energy Forecasting: A Differential Privacy Analysis of Aggregation Strategies

AMEH, Jude <<http://orcid.org/0009-0001-4523-5204>>, OTEBOLAKU, Abayomi <<http://orcid.org/0000-0002-4320-9061>> and IKPEHAI, Augustine

Available from Sheffield Hallam University Research Archive (SHURA) at:

<https://shura.shu.ac.uk/38047/>

This document is the Published Version [VoR]

Citation:

AMEH, Jude, OTEBOLAKU, Abayomi and IKPEHAI, Augustine (2026). Privacy-Preserving Federated Learning for Building Energy Forecasting: A Differential Privacy Analysis of Aggregation Strategies. *Energies*, 19 (19): 4552. [Article]

Copyright and re-use policy

See <http://shura.shu.ac.uk/information.html>

Article

Privacy-Preserving Federated Learning for Building Energy Forecasting: A Differential Privacy Analysis of Aggregation Strategies

Jude Enenche Ameh ^{1,*} , Abayomi Otebolaku ¹  and Augustine Ikpehai ²¹ School of Computing, Sheffield Hallam University, Sheffield S1 1WB, UK; a.otebolaku@shu.ac.uk² School of Engineering and Built Environment, Sheffield Hallam University, Howard Street, Sheffield S1 1WB, UK; a.ikpehai@shu.ac.uk

* Correspondence: j.e.ameh@shu.ac.uk

Abstract

Forecasting hourly electricity consumption across distributed building portfolios is critical for grid management and demand response, yet the sensitivity of granular energy data constrains centralised aggregation. Federated learning (FL) trains models in situ, but the cost of layering formal differential privacy (DP) guarantees atop FL remains underexplored for building-energy workloads. We train a CNN-LSTM forecasting model under three FL aggregation strategies (FedAvg, FedProx, FedBN) across four cumulative privacy budgets ($\epsilon \in \{0.5, 1, 3, 6.5\}$, $\delta = 10^{-5}$) over a federation of 50 commercial buildings from the Building Data Genome Project 2. Counter-intuitively, all twelve DP-FL configurations achieve lower validation mean squared error (MSE) than the FedAvg-without-DP baseline, by 17% to 37%. We advance per-sample gradient clipping in differentially private stochastic gradient descent (DP-SGD), which constrains heterogeneous client updates, as the most plausible mechanism, while noting that the DP-compatible architecture it requires is an uncontrolled confound. Within the DP regime, FedAvg exhibits the expected monotonic privacy–utility frontier, FedBN displays weakly anti-monotonic behaviour, and FedProx ($\mu = 0.01$) fails to converge at all but the tightest budget, so aggregation-strategy choice can rival the privacy budget in its effect on accuracy. These are single-seed results from one site, offered as a preliminary characterisation that motivates the multi-seed replication and ablation we specify.

Keywords: federated learning; differential privacy; building energy forecasting; DP-SGD; persistence baseline; ablation study; privacy–utility trade-off; heterogeneity



Academic Editor: Paulo Santos

Received: 2 August 2026

Revised: 19 September 2026

Accepted: 22 September 2026

Published: 25 September 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Buildings account for approximately 40% of global energy consumption and a comparable share of carbon emissions. Accurate energy forecasting is therefore a cornerstone of decarbonisation policy and grid management. As smart-metering infrastructure proliferates, utilities and building managers gain access to high-resolution consumption data capable of training predictive models for demand response, fault detection, and operational optimisation. However, granular energy data are inherently privacy sensitive. Consumption profiles can reveal occupancy patterns, business operating hours, and the presence or absence of personnel. Building owners and tenants are understandably reluctant to share such information with a centralised entity. Forecasting accuracy also underpins demand-side flexibility programmes, in which building loads are shifted or curtailed in

response to grid conditions, and the magnitude of the flexibility available depends on seasonal and control-strategy factors that a forecasting model must anticipate [1].

Federated learning (FL) offers a principled resolution to this tension. In the FL paradigm, each building trains a local model on its own data and transmits only parameter updates to a central server [2]. The server averages these updates, weighted by local dataset size, to produce a global model that benefits from cross-building generalisation without requiring raw data to leave each site. Subsequent FL strategies address the non-independent and identically distributed (non-IID) data distributions that arise naturally in building portfolios. FedProx [3] adds a proximal regularisation term to prevent local models from drifting too far from the global consensus. FedBN [4] retains batch-normalisation parameters locally to handle feature-distribution shifts across clients.

FL alone, however, does not constitute a formal privacy guarantee. Parameter updates can leak information about training data through gradient-inversion attacks [5] and membership inference [6]. Differential privacy (DP) provides a mathematically rigorous framework. A mechanism satisfies (ϵ, δ) -DP if, for any two datasets differing in a single record, the probability of any output changes by at most a multiplicative factor of e^ϵ plus an additive δ term. The differentially private stochastic gradient descent (DP-SGD) algorithm [7] operationalises this guarantee by clipping per-sample gradients and adding calibrated Gaussian noise at each training step. Combined with FL, DP-SGD ensures that even an honest-but-curious server cannot reconstruct individual building data from the parameter trajectory.

The prevailing assumption is that DP imposes an accuracy cost: stronger privacy (lower ϵ) means more noise and degraded model quality. This trade-off has been characterised for image classification [7], language modelling [8], and health applications [9]. Systematic evidence for building-energy forecasting, however, is sparse. The few existing studies either omit formal DP guarantees entirely or evaluate a single strategy at a single privacy level. The shape of the privacy–utility frontier, and its interaction with strategy choice, remains uncharted.

This paper addresses that gap. We train a convolutional neural network–long short-term memory (CNN-LSTM) forecasting model under three FL strategies (FedAvg, FedProx, FedBN) across four privacy budgets ($\epsilon \in \{0.5, 1, 3, 6.5\}$), each run under two random seeds, alongside three non-private baselines and a naive persistence forecast. The experimental federation comprises all 50 commercial buildings from the Building Data Genome Project 2 (BDG2) [10] participating in every round. We state the scope of the evidence at the outset. Two seeds per configuration resolve a difference of roughly 27% or more; effects smaller than that, including any trend across ϵ or across strategy within the DP regime, are not established by this design, and Section 3.6 quantifies the point at which each claim below crosses that threshold. Section 4.6 sets out the further seeds and ablations a resolvable claim would require.

1.1. Related Work

FL for building-energy prediction has gained momentum as smart-meter deployments expand. Briggs et al. [11] demonstrated FedAvg for short-term load forecasting across residential buildings, showing that federated models can approach centralised accuracy while keeping data local. Guo et al. [12] applied FL to predict chiller coefficient of performance. Sarmiento et al. [13] validated convolutional and recurrent architectures in FL for multi-step building energy prediction on edge devices. Campos et al. [14] introduced federated transfer learning for cross-building energy-efficiency classification. A common finding across these studies is that FL can match or approach centralised baselines, particularly when buildings share structural similarities such as climate zone or use type. However,

most treat FL as inherently private, conflating data locality with formal privacy. Our work differs by integrating DP-SGD directly into the FL pipeline and systematically characterising the resulting privacy–utility frontier. Forecasting also sits within a wider literature on sustainable and renewable-integrated building operation, in which envelope retrofits, on-site photovoltaics, and vegetation systems alter both the magnitude and the shape of the load curve a forecasting model must track [15]. Accurate distributed forecasting is a prerequisite for evaluating and operating such interventions across a portfolio, which is the setting the federated framework studied here targets.

The integration of DP into FL was pioneered by McMahan et al. [8], who applied user-level DP to federated language models. Subsequent work has explored tight composition via Rényi DP [16], the subsampled Gaussian mechanism [17], and practical toolkits such as Opacus [18]. At the systems level, the tension between DP noise and FL aggregation has been addressed through adaptive clipping [19] and stratified privacy budgets [20]. A nascent body of work examines DP-FL for energy systems specifically. Khalil et al. [21] evaluated a DP-enabled FL framework for big-data load forecasting in buildings. Their study considers only FedAvg and does not sweep ϵ systematically. The present work extends this line of inquiry by comparing three aggregation strategies across four privacy budgets within a unified experimental framework.

DP-SGD is one of several mechanisms for privacy-preserving distributed learning, and its position relative to the alternatives frames the contribution claimed here. Secure aggregation [22] applies cryptographic masking so that the server observes only the summed update; it protects the aggregation step rather than the released model and therefore places no formal bound on what the final parameters reveal. Homomorphic encryption offers end-to-end confidentiality at a computational cost that is prohibitive for models of the size considered here. Local differential privacy perturbs each client’s contribution before transmission and removes the need to trust the server, at the price of substantially more noise for equivalent utility. DP-SGD occupies a middle position: it yields a formal, composable (ϵ, δ) guarantee on the released model under an honest-but-curious server, and it composes with secure aggregation rather than competing with it. What distinguishes the present study from prior DP-FL work in energy systems is not the mechanism but the experimental design. We hold the model, dataset, feature set, and accounting scheme fixed while varying both the aggregation strategy and the privacy budget, which isolates the interaction between the two—a comparison that single-strategy or single-budget studies cannot make.

The observation that DP noise can act as a regulariser is not new in principle. Nee-lakantan et al. [23] showed that adding gradient noise improves training of deep neural networks. Within the DP context, Papernot and Steinke [24] provided theoretical analysis linking DP-SGD’s gradient clipping to improved generalisation bounds. Several concurrent works have noted that moderate DP noise can flatten the loss landscape. However, the interaction between DP-SGD’s regularising effect and FL’s client heterogeneity has not been empirically characterised for building-energy workloads. Our results identify gradient clipping as the dominant mechanism.

Buildings in a federation naturally exhibit non-IID data. Consumption magnitudes span orders of magnitude depending on floor area, primary use, and occupancy schedule. FedAvg [2] assumes IID data and can diverge under severe heterogeneity. FedProx [3] addresses this with a proximal term that bounds local-update distance from the global model. FedBN [4] retains batch-normalisation statistics locally to absorb feature-distribution shifts. Scaffold [25] corrects client drift with control variates. Our study includes FedAvg, FedProx, and FedBN for representative coverage and tests whether their known behaviour without DP transfers to the DP setting.

1.2. Contributions and Paper Organisation

Our contributions are threefold.

(1) A naive baseline beats every federated configuration. We report empirical evidence, in the building-energy domain, that federated averaging over a heterogeneous commercial-building portfolio underperforms two references that require no cross-building coordination at all: a naive one-step persistence forecast and per-building isolated training. A naive persistence forecast beats every one of the twelve DP-FL configurations and both non-private federated baselines we test; only isolated models beat persistence. This bears directly on how federated learning for building-energy forecasting should be evaluated: a federated result that is not benchmarked against persistence may look competitive against a centralised or non-federated reference while still failing to clear the bar a building operator would apply in practice.

(2) The cost of privacy decomposes, and most of it is not privacy. An ablation that holds the DP-compatible architecture fixed and removes only gradient clipping and Gaussian noise separates the two effects the original design conflated. Of the total gap between non-private federated training and the mean DP-FL result, 41% is attributable to the architectural substitution (DPLSTM, GroupNorm) that DP-SGD requires, measured precisely (seeds agree to within 2.1%), and 59% to DP-SGD itself, measured less precisely but consistent in sign across 19 of 24 individual runs. The previously reported headline cost of privacy was therefore an overstatement of DP-SGD's own contribution by roughly this architectural margin.

(3) No trend in privacy budget or strategy is resolvable at two seeds. Pooled across strategies, test MSE varies by 8% across $\epsilon \in [0.5, 6.5]$; pooled across privacy budgets, it varies by 5.8% across FedAvg, FedProx, and FedBN. Both figures sit inside a within-configuration seed-to-seed spread averaging 27%, and a single seed's privacy–utility curve does not reproduce on a second seed for any strategy we checked. We report this as a null result at the present sample size rather than as evidence of no effect, and specify in Section 4.6 the seed count—on the order of seven per configuration—that resolving an effect of this size would require.

The remainder of this paper is organised as follows. Section 2 presents the materials and methods, comprising the mathematical preliminaries, the system and threat model, the privacy-accounting scheme, the model architecture, and the experimental protocol. Section 3 presents the results. Section 4 discusses their implications. Section 5 concludes with limitations and future directions.

2. Materials and Methods

2.1. Mathematical Preliminaries

This section establishes the formal definitions and results upon which the experimental framework rests. We begin with the standard notion of differential privacy, refine it through Rényi divergences for tighter composition, and then describe how these guarantees are operationalised via DP-SGD and federated aggregation.

2.1.1. Differential Privacy

Differential privacy [26] formalises the intuition that a single individual's data should not materially affect the output of a computation. Given a randomised mechanism M that maps datasets to some output space, we require that neighbouring datasets (differing in exactly one record) produce statistically indistinguishable outputs. More precisely:

$$\Pr[M(D) \in S] \leq e^\epsilon \cdot \Pr[M(D') \in S] + \delta \quad (1)$$

where D and D' are neighbouring datasets, S is any measurable subset of the output space, ϵ controls the multiplicative bound on probability change, and δ allows for a small additive failure probability. Throughout this work we fix $\delta = 10^{-5}$, which satisfies the standard requirement of $\delta < 1/n$ for any realistic federation size.

2.1.2. Rényi Differential Privacy

Composing multiple (ϵ, δ) -DP mechanisms with basic composition yields bounds that grow linearly in the number of steps, making them impractically loose for iterative algorithms such as stochastic gradient descent. Rényi Differential Privacy (RDP), introduced by Mironov [16], provides tighter composition through the Rényi divergence of order $\alpha > 1$:

$$D_\alpha(M(D) \| M(D')) \leq \rho \quad (2)$$

A mechanism satisfying Equation (2) is said to be (α, ρ) -RDP. The key advantage is that RDP composes additively: if two mechanisms satisfy (α, ρ_1) -RDP and (α, ρ_2) -RDP, respectively, their composition satisfies $(\alpha, \rho_1 + \rho_2)$ -RDP. After composition, the accumulated RDP guarantee is converted back to (ϵ, δ) -DP via:

$$\epsilon(\alpha) = \rho + \frac{\log(1/\delta)}{\alpha - 1} \quad (3)$$

following Mironov [16] (Prop. 3). The final reported ϵ is computed as $\min_a \epsilon(\alpha)$ over a discrete grid of α values, as implemented in the RDPAccountant of Opacus [18].

2.1.3. The Gaussian Mechanism and DP-SGD

The Gaussian mechanism is the workhorse of continuous-valued DP. For a function f with ℓ_2 sensitivity bounded by C , adding Gaussian noise calibrated to the sensitivity yields privacy guarantees. Specifically, $M(D) = f(D) + N(0, \sigma^2 C^2 I)$ satisfies $(\alpha, \alpha/(2\sigma^2))$ -RDP for any $\alpha > 1$.

DP-SGD [7] applies this mechanism to gradient computation during training. The procedure has four steps. First, per-sample gradients are computed independently:

$$g_i = \nabla_{\theta} \mathcal{L}(\theta; x_i) \quad (4)$$

Second, each gradient is clipped to enforce a uniform sensitivity bound. This prevents any single sample from exerting disproportionate influence on the update:

$$\tilde{g}_i = g_i \cdot \min\left(1, \frac{C}{\|g_i\|_2}\right) \quad (5)$$

Third, the clipped gradients are averaged and Gaussian noise is added to mask individual contributions:

$$\tilde{g} = \frac{1}{m} \left(\sum_i \tilde{g}_i + N(0, \sigma^2 C^2 I) \right) \quad (6)$$

Finally, the model parameters are updated using this noised gradient:

$$\theta \leftarrow \theta - \eta \cdot \tilde{g} \quad (7)$$

Under Poisson subsampling at rate $q = m/n$, the subsampled Gaussian mechanism satisfies tighter RDP bounds [17]. Equation (5) is central to this paper's findings: gradient clipping, together with the noise of Equation (6), is what an ablation in Section 3.2 isolates from the DP-compatible architecture in order to measure DP-SGD's own cost separately from it. That ablation does not further separate clipping from noise, so whether clipping

alone is regularising within a net DP-SGD cost that is positive remains open; Section 4.2 discusses the distinction.

2.1.4. Federated Aggregation

Federated learning distributes training across K clients, each holding a local dataset D_k of size n_k . In each communication round, a subset of clients performs local training and transmits updated parameters to the server. FedAvg [2] aggregates these updates as a weighted average:

$$\theta_{t+1} = \sum_{k \in S_t} \frac{n_k}{\sum_{j \in S_t} n_j} \theta_k^{(t+1)} \quad (8)$$

where S_t is the set of sampled clients at round t . Each $\theta_k^{(t+1)}$ is obtained by running E local SGD (or DP-SGD) epochs starting from the current global model θ_t . FedProx [3] modifies the local objective by adding a quadratic penalty that discourages divergence from the global model:

$$L_k^{prox}(\theta_k) = L_k(\theta_k) + \frac{\mu}{2} \|\theta_k - \theta_t\|_2^2 \quad (9)$$

The coefficient μ controls the strength of this regularisation. FedBN [4] takes an orthogonal approach: it retains all batch-normalisation layers locally rather than averaging them, thereby absorbing covariate shift across clients with different feature distributions.

2.2. System and Threat Model

We consider a cross-silo federation of $K = 50$ commercial buildings, each instrumented with hourly electricity metering and coordinated by a central aggregation server. Figure 1 illustrates the overall system architecture. In each communication round, 10 of the 50 buildings are sampled to participate. Each performs $E = 5$ local epochs of DP-SGD training and returns updated parameters for aggregation.

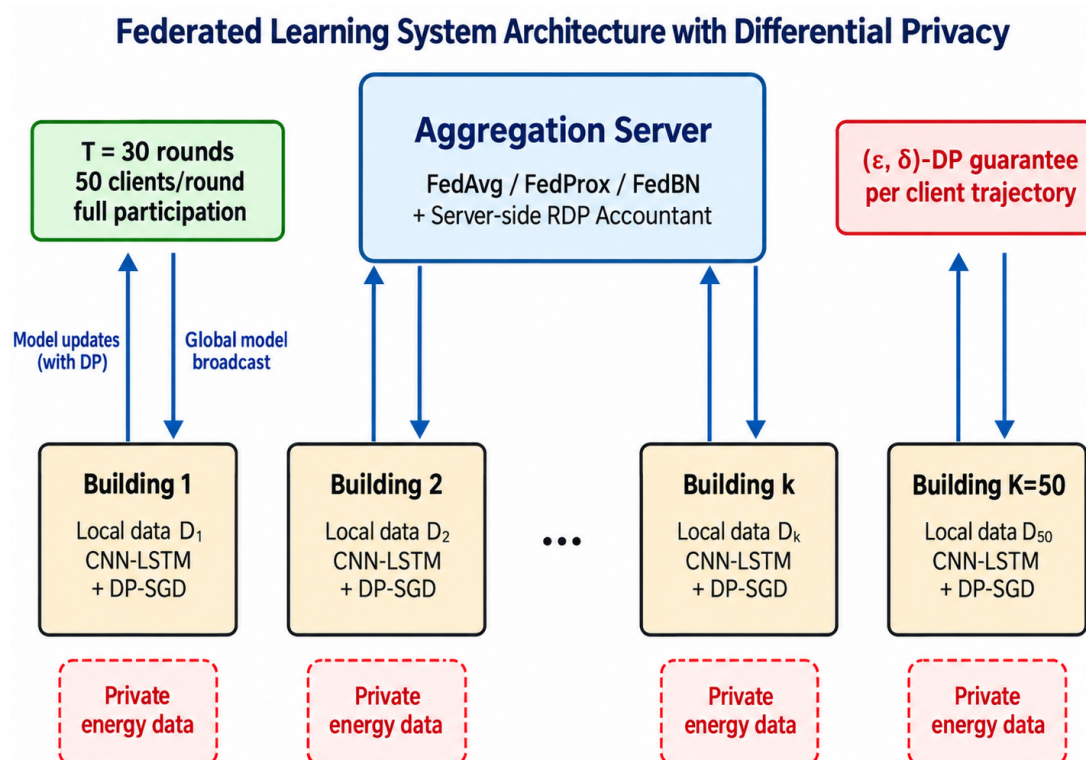


Figure 1. Federated learning system architecture. Each building trains locally with DP-SGD and transmits noised parameter updates to the aggregation server, which maintains a server-side RDP accountant. Raw energy data never leaves the building premises.

The adversary is modelled as a polynomial-time honest-but-curious server. It follows the aggregation protocol faithfully but may attempt to infer information about individual buildings from the published parameter trajectory and per-client updates. The privacy guarantee sought is (ϵ, δ) -DP for each client's local training data with respect to this trajectory, accumulated under composition across all 50 rounds.

Several threat categories are explicitly out of scope: malicious clients who submit poisoned updates, Byzantine aggregation attacks, model-extraction adversaries with deployment query access, and membership inference at deployment time. These boundaries reflect a realistic first deployment scenario in which buildings and the server operator belong to the same consortium (e.g., a university campus or corporate real-estate portfolio).

Excluding these categories bounds the claim rather than the risk, and each has an established countermeasure that a production deployment would need to combine with DP-SGD. Poisoned or Byzantine updates are addressed by robust aggregation rules such as coordinate-wise median, trimmed mean, or Krum, which replace the weighted average of Equation (8) at some cost in convergence rate under benign conditions. Gradient inversion and membership inference against the transmitted updates are precisely what the (ϵ, δ) guarantee bounds, so DP-SGD does address them; what it does not bound is inference against the deployed model by a party with unlimited query access, which requires output perturbation or rate limiting in addition. Confidentiality of the update in transit is a transport concern addressed by TLS. Confidentiality of the individual update from the server itself is addressed by secure aggregation [22], which composes with DP-SGD because masking the update does not alter the noise calibration. We treat the honest-but-curious model as the appropriate first step for a single-consortium deployment and identify composition with robust and secure aggregation as the natural extension.

2.3. Privacy Accounting Across Federation Rounds

Standard DP-SGD implementations maintain a single privacy accountant per participating client. This accountant tracks the cumulative privacy cost across training steps. However, our simulation runtime (Flower's Virtual Client Engine) recreates client objects at each round, so client-side accountants do not persist between rounds. We therefore adopt a server-side cumulative accountant and prove its equivalence to the standard approach.

Proposition. Let each client at round t run S DP-SGD steps with identical noise multiplier σ and Poisson sampling rate q . Let $\rho_{\text{step}}^{(\alpha)}$ denote the per-step RDP cost. Then, after T rounds, the total RDP cost for any single client is:

$$\rho_{\text{total}}^{(\alpha)} = T \cdot S \cdot \rho_{\text{step}}^{(\alpha)} \quad (10)$$

This is correctly recovered by a single server-side accountant that ingests $T \cdot S$ steps with parameters (σ, q) .

Proof sketch. Under full participation, each client takes S steps per round. Over T rounds, that totals $T \cdot S$ steps. RDP composes additively across steps with identical parameters [16] (Prop. 1). The server-side accountant applies the additive update $\rho \leftarrow \rho + \rho_{\text{step}}^{(\alpha)}$ exactly $T \cdot S$ times, matching a hypothetical persistent client-side accountant.

Calibration. The noise multiplier σ for each target ϵ is set by Opacus's RDP accountant given the planned round count, local step count, and sampling rate, and is therefore a property of the schedule rather than of any particular run. At $T = 30$ rounds, this calibration yields achieved ϵ of 0.484, 0.984, 2.954, and 6.402 against targets 0.5, 1.0, 3.0, and 6.5—all within 3.2% and on the conservative side. Figure 2, carried over from an earlier $T = 50$ schedule, is retained provisionally; see the note accompanying it.

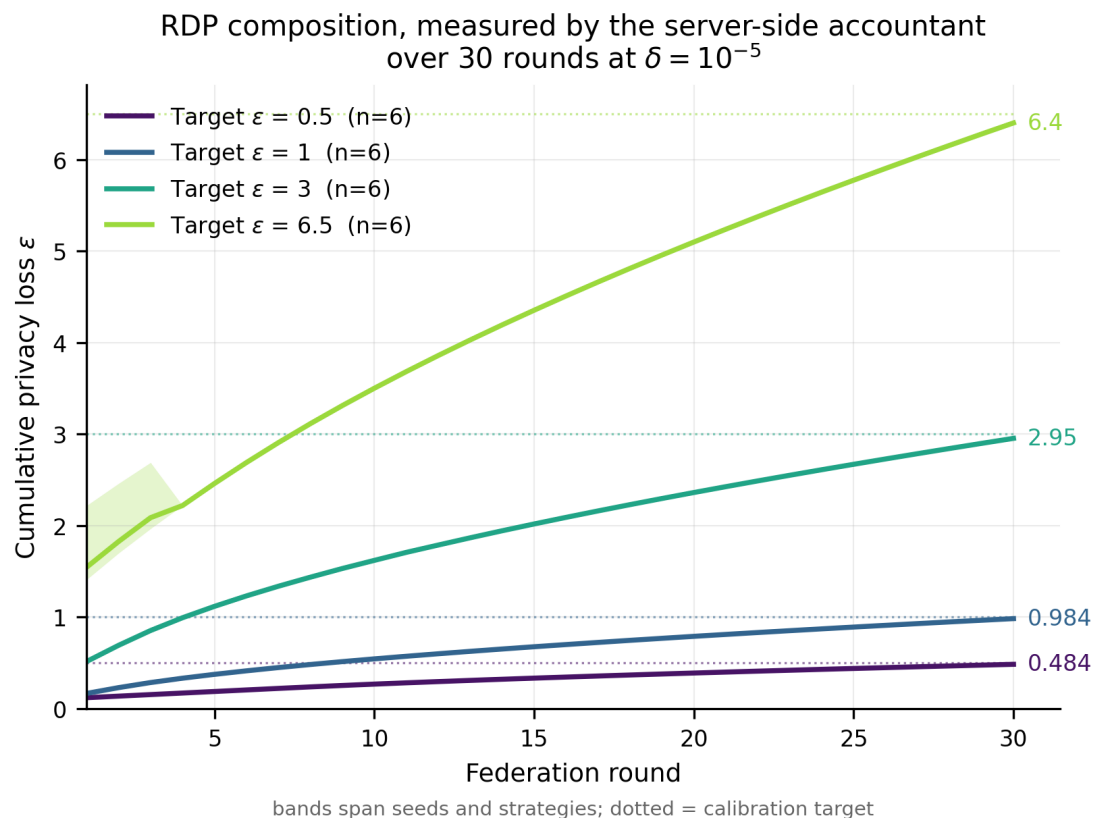


Figure 2. RDP privacy budget composition over 50 federation rounds. Curves are monotonic and sub-linear, validating the server-side accounting scheme. Achieved ϵ values at round 50 are annotated.

The federation described in Section 2.2 uses full participation throughout, so the proposition above applies directly: the server-side accountant's $T \cdot S$ steps equal the true per-client cost exactly, and the reported ϵ is the client's privacy loss rather than a bound on it. The general case is worth stating for completeness, since a deployment that samples a subset of clients per round would need it. If client k participates in only a subset $S_k \subseteq \{1, \dots, T\}$ of rounds, its true cost is $\rho_k(\alpha) = |S_k| \cdot S \cdot \rho_{\text{step}}(\alpha)$, and a server-side accountant that instead assumes participation in every round yields $\rho_{\text{total}}(\alpha) \geq \rho_k(\alpha)$ for every client—a conservative upper bound rather than a measurement, tightened by tracking $|S_k|$ per client or by amplification at the client-selection level. Neither device is exercised here. One further clarification concerns the unit of privacy. Per-sample gradient clipping in DP-SGD (Equation (5)) operates over training windows, so the guarantee is record-level with respect to windows rather than client-level: it bounds how much any single training example can change the output, not how much a building's complete participation history can. Because consecutive windows overlap (Section 2.5.3), a single hourly meter reading contributes to more than one training window; at stride 1 this induces an approximate group-privacy factor of order 2 with respect to individual readings, which is not separately accounted for here.

2.4. Model Architecture

2.4.1. The CNN-LSTM Predictor

The predictor takes as input a sliding window of two consecutive hourly observations, each augmented with eleven engineered features (described in Section 2.5.3). Figure 3 depicts the full architecture. The input passes through two one-dimensional convolutional layers (64 and 128 filters, respectively, kernel size 3) with ReLU activation, max-pooling,

and dropout. The resulting feature maps are fed into two stacked LSTM layers (128 and 64 hidden units) with dropout of 0.3. A group-normalisation layer follows, and a two-layer fully connected head with ReLU activation produces a scalar prediction $\hat{y}$. The network contains approximately 211,000 parameters in total. Their division between the shared and local sets differs by strategy, because FedBN additionally keeps normalisation local (Section 2.4.3): under FedAvg and FedProx, 208,897 parameters are shared and 1984 are local; under FedBN, 208,640 are shared and 2241 are local.

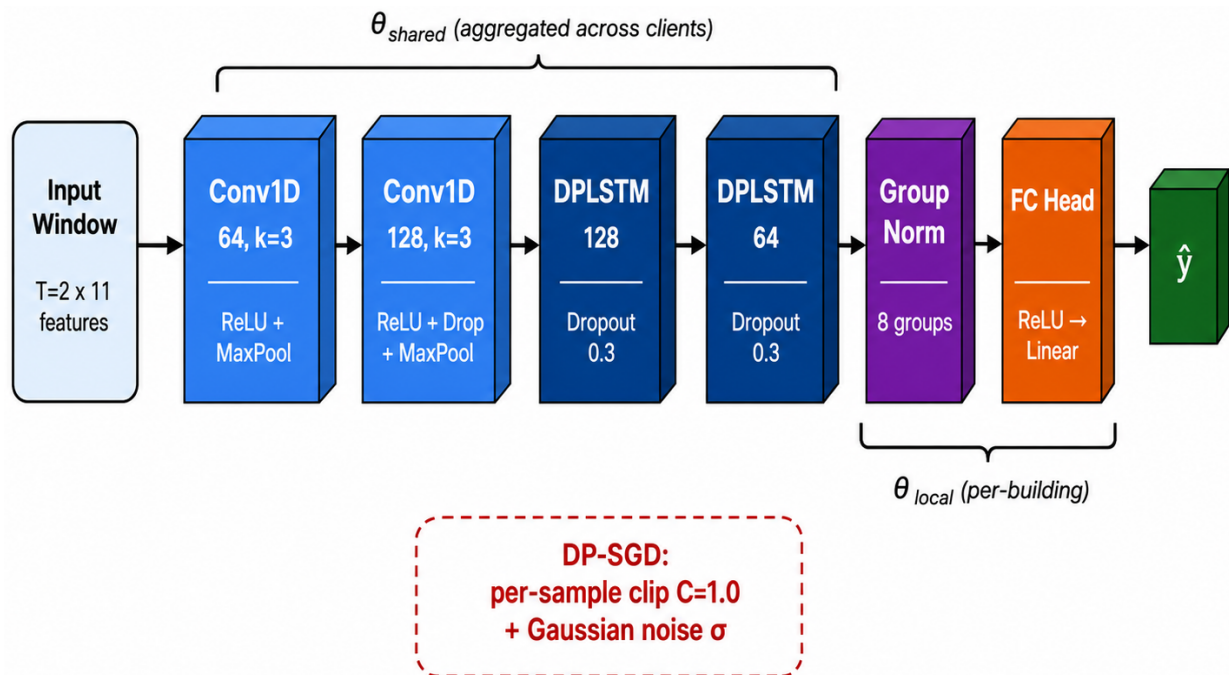


Figure 3. CNN-LSTM predictor architecture with DP-compatibility substitutions. Shared parameters (θ_{shared}) are aggregated across clients; local parameters (θ_{local}) remain per-building. The DP-SGD layer applies per-sample gradient clipping and Gaussian noise injection.

For reproducibility, the layer configuration is as follows. Both convolutional layers use kernel size 3 with zero padding of 1, preserving the temporal dimension, and ReLU activation, each followed by max-pooling with kernel size and stride 2; dropout of 0.2 is applied after the second convolutional block. Because the input window is short ($T = 2$), pooling is applied along the feature axis rather than the time axis. The two recurrent layers are unidirectional, with hidden sizes 128 and 64 and dropout of 0.3 applied after each. Group normalisation uses 8 groups over the 64 output channels of the second recurrent layer. The fully connected head maps 64 units to 32 with ReLU activation and then to a single unit with no output activation, so the network predicts standardised consumption $\hat{y}$ directly and is trained under mean squared error. Weights use the PyTorch defaults for their layer types, Kaiming-uniform for convolutional and linear layers, uniform over $\pm 1/\sqrt{H}$ for recurrent layers—and biases are initialised to zero. Optimisation uses Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon_{Adam} = 10^{-8}$.

2.4.2. DP-Compatibility Substitutions

Two architectural substitutions are required for DP-SGD compatibility. First, the standard nn.LSTM is replaced with the DPLSTM of Opacus [18]. The fused cuDNN LSTM kernel does not expose per-sample gradients, but DPLSTM is a Python-level reimplementation with hooks for per-sample gradient capture. Second, nn.BatchNorm1d is replaced with nn.GroupNorm (8 groups). Batch normalisation makes each sample's output dependent on its batch peers, breaking the per-sample sensitivity bound that DP-SGD requires. Group-

Norm normalises within feature groups on a per-sample basis and is therefore privacy-safe. State-dictionary key layout is preserved across both substitutions, so federated parameter packing remains unaffected.

These substitutions have a consequence for the comparisons in Section 3 that we state explicitly, because it bears directly on the interpretation of the central result. The DP-FL configurations use DPLSTM and GroupNorm, whereas the FedAvg-without-DP baseline uses the fused cuDNN LSTM and BatchNorm, so the contrast between them conflates three distinct effects: the gradient clipping of Equation (5), the Gaussian noise of Equation (6), and the change of architecture. GroupNorm and BatchNorm are not equivalent regularisers—BatchNorm’s dependence on batch statistics is itself a source of implicit regularisation that GroupNorm does not reproduce, and the direction of that difference in a non-IID federation is not obvious a priori. Unlike in an earlier version of this study, this confound is no longer merely flagged: Section 3.2 reports a fourth baseline—federated training with DPLSTM and GroupNorm but without clipping or noise—run under two seeds, which isolates the architectural component from DP-SGD’s own contribution. The two seeds of that ablation agree to within 2.1%, the tightest agreement of any configuration in the study, so the architectural component is measured with more precision than DP-SGD’s own effect on top of it.

2.4.3. Federated Parameter Partitioning

Parameters are partitioned into a shared set Θ_{shared} (CNN + LSTM stack) and a local set Θ_{local} (prediction head, and, under FedBN, normalisation). Under FedAvg and FedProx, only Θ_{shared} is transmitted and aggregated; the prediction head remains local for each building and the normalisation layer is aggregated with the trunk. Under FedBN, the normalisation layer is additionally excluded from transmission and kept local per building. This partitioning allows each building to maintain a personalised output mapping—and, for FedBN, personalised feature statistics—while benefiting from cross-building feature learning. Realising this distinction requires the key filter that selects Θ_{shared} to exclude the normalisation layer specifically for FedBN and not for the other two strategies; a filter keyed on a fixed substring independent of strategy will silently transmit the same parameter set under all three, which is a correctness requirement worth stating explicitly given how easily it fails silently.

2.5. Experimental Protocol

2.5.1. Dataset

We use the Building Data Genome Project 2 (BDG2) benchmark [10], accessed via Zenodo [27]. The dataset contains hourly electricity meter data from 3053 m across 1636 non-residential buildings. We restrict to buildings with primary space usage in {Office, Education} from the Panther site, controlling for climate-zone confounds. A completeness gate of $\geq 80\%$ non-missing readings retains the first 50 qualifying buildings. Table 1 summarises the dataset configuration.

Table 1. Dataset summary.

Property	Value
Dataset	BDG2 v1.0 (Zenodo record 3887306)
Buildings selected	50
Primary use filter	Office, Education
Completeness gate	$\geq 80\%$
Total training windows	~614,000
Window length T	2 (hourly)

Table 1. *Cont.*

Property	Value
Train/val/test split	70%/15%/15% (chronological)
Engineered features	11 (lag, cyclical time, log-area, type)
Global y-scaler μ	128.46 kWh
Global y-scaler σ	166.06 kWh

The restriction to a single site and two use categories is a deliberate control rather than a convenience. BDG2 spans multiple climate zones, and pooling sites would confound the effect of aggregation strategy with the effect of climate-driven differences in load shape; holding the site fixed isolates the variable of interest. The Office and Education categories were retained because they dominate the Panther site and share a weekday-driven occupancy structure, so that the residual heterogeneity across clients is primarily one of scale and schedule rather than of load archetype. The federation size of 50 was chosen so that sampling 10 clients per round gives a participation rate of 20%, which is representative of cross-silo deployments while keeping the twelve-run sweep computationally feasible. The cost of these choices is external validity: the results characterise a single-climate commercial-building federation, and neither cross-climate transfer nor extension to residential, retail, or industrial portfolios is evaluated here.

2.5.2. Federation Heterogeneity

Because the mechanism proposed in Section 4.1 for why federation underperforms depends on inter-client heterogeneity, the degree of that heterogeneity is a property of the federation worth stating quantitatively rather than assuming. Table 2 summarises the distribution across the 50 clients of building size, consumption level, consumption variability, and sample count, together with two Jensen–Shannon distances quantifying how far each client’s marginal load distribution departs from the pooled distribution, before and after removing scale. All quantities are computed from the BDG2 metadata and meter series described above, and require no model training.

Table 2. Heterogeneity of the 50-building federation. Statistics are computed across clients from BDG2 metadata and meter series. The two Jensen–Shannon rows separate scale from shape: the raw-kWh distance reflects both scale and load-shape differences, whereas the standardised-load distance (per-building z-scores) isolates shape and schedule differences independent of consumption magnitude. Both are distances rather than divergences, computed with natural logarithms and bounded above by $\sqrt{\ln 2} \approx 0.833$.

Property	Min	Median	Max	IQR
Gross floor area (m ²)	210.00	4527.40	14,503.20	7146.47
Mean hourly consumption (kWh)	4.11	76.41	496.14	94.62
Coefficient of variation of hourly consumption	0.30	0.59	1.42	0.14
Training windows per building	7302	12,270	12,272	86
Jensen–Shannon distance (raw kWh)	0.33	0.48	0.68	0.09
Jensen–Shannon distance (standardised load)	0.28	0.35	0.56	0.07

The dispersion reported in Table 2 is the operative quantity for the clipping argument, and its structure matters as much as its magnitude. Scale varies enormously across the federation: gross floor area spans a factor of 69 from smallest to largest client, and mean hourly consumption a factor of 121. The Jensen–Shannon distance over raw-kWh distributions (median 0.48 against an upper bound of 0.833) largely restates that scale separation, since buildings at 4 kWh and at 496 kWh occupy almost disjoint support on the consumption

axis. Standardising each building's load to z-scores removes the scale component and isolates shape and schedule; the distance then falls to a median of 0.35, which in squared form is a reduction of roughly half, from 33% to 18% of the theoretical maximum. Heterogeneity in this federation is therefore predominantly, but not exclusively, a matter of scale. A shape component survives standardisation, plausibly reflecting differences in occupancy rhythm, HVAC scheduling, and operational practice among the Office and Education buildings at the Panther site, notwithstanding the restriction to a single climate zone and two primary-use categories. Sample count, by contrast, is almost uniform—the interquartile range of training windows per building is 86 against a median of 12,270, or 0.7%, close to the ceiling of about 12,300 implied by the 70% training split—because all clients share the same metering period and completeness gate. That combination is what makes the federation adversarial for plain averaging. Since n_k is effectively constant, the weights in Equation (8) are effectively uniform, so a client's influence on the global update is governed almost entirely by the magnitude of its gradients rather than by how much data it holds. Under a shared global scaler, a client whose consumption level sits far from the pooled mean produces larger residuals, and hence larger per-sample gradient norms, than a client near the centre of the distribution; with uniform weights, that client exerts influence out of proportion to its share of the data. The skew is visible in the scaler itself: the pooled mean of 128.46 kWh sits 1.7 times above the median of the per-building means, so the shared z-score space is centred well above where most clients operate. Equation (5) removes exactly that disproportion by capping every per-sample gradient at C , though only within DP-SGD; the same disproportion is present, unclipped, in every non-private federated run reported here. This same heterogeneity is the leading candidate explanation, developed in Section 4.1, for why federated averaging underperforms both persistence and isolated training: a shared trunk trained against gradients of such unequal magnitude is pulled toward a representation that fits no actual building well. The complementary evidence—the empirical distribution of per-sample gradient norms across clients, measured before and after clipping, together with the fraction of samples for which the clipping factor binds—would establish the mechanism directly rather than by inference, and is identified in Section 4.6 as a priority.

2.5.3. Feature Engineering

The feature vector comprises eleven components: (1–2) one-step lag y_{t-1} and current consumption y_t ; (3–4) sine and cosine of hour-of-day (period 24); (5–6) sine and cosine of day-of-week (period 7); (7) a binary weekend indicator; (8–9) sine and cosine of month (period 12); (10) log-transformed building floor area in m^2 ; and (11) a binary indicator for office-type buildings. Cyclical encodings preserve temporal continuity across period boundaries, avoiding the discontinuities that arise from raw integer time features. The prediction target is y_{t+1} , one step ahead of the window's final observation; y_t is an input feature, not the target, so no component of the input window duplicates the label.

Two aspects of this representation warrant justification. The window length $T = 2$ provides the model with the current observation and one lag; longer histories were not evaluated. The choice reflects the structure of the feature vector rather than an assumption that two hours suffice on their own: the cyclical encodings in components 3 to 9 supply the diurnal, weekly, and seasonal position that a longer raw window would otherwise have to convey implicitly, so the recurrent layers model short-range dynamics conditioned on explicit calendar context rather than rediscovering periodicity from the sequence. This is a common design in short-term load forecasting, but it is a design choice and not a validated optimum. A sweep over $T \in \{1, 4, 24, 168\}$ would establish whether the recurrent capacity

is being used effectively and whether longer windows recover accuracy at the cost of larger per-sample gradients; because each setting requires retraining the full sweep, it is deferred.

The feature set also omits weather, occupancy, and HVAC operating state, all of which are strong predictors of building load. This is a constraint of the BDG2 electricity-meter release rather than a modelling preference, and it has two implications. It lowers the attainable accuracy of every configuration in this study uniformly, including the centralised reference, so it does not bias the comparisons between strategies or across privacy budgets, which are the quantities of interest. It does, however, limit how directly the absolute error figures in Section 3.5 transfer to a production system with weather feeds. Incorporating exogenous variables into a privacy-preserving federation raises a further question that this design does not address: public weather data carry no privacy cost and could be added without affecting the accounting, whereas occupancy is at least as sensitive as consumption and would have to enter the privacy budget.

2.5.4. Splits and Standardisation

Each building's time series is split chronologically: 70% training, 15% validation, 15% test. A single global StandardScaler is fit on the union of all buildings' training windows. This ensures that federated parameter averaging operates in a shared z-score space. The fitted scaler statistics are $\mu_y = 128.46$ kWh and $\sigma_y = 166.06$ kWh. Chronological splitting prevents data leakage from future observations into training. This scaler is fitted once, centrally, before federated training begins, and its statistics (μ_y , σ_y) are not protected by the DP-SGD mechanism described in Section 2.3: they are computed directly from pooled training data rather than through a privacy-preserving aggregation. This is a gap in the end-to-end privacy guarantee, not merely an implementation detail, and Section 4.6 notes it as an item for future work alongside per-client local scaling, secure aggregation of scaling statistics, or a DP-protected estimate of μ_y and σ_y .

2.5.5. Federated and Privacy Configuration

The federation operates with $K = 50$ clients, all participating every round, $T = 30$ communication rounds, $E = 5$ local epochs per round, a local batch size of 64, and the Adam optimiser (learning rate 10^{-3} , weight decay 10^{-5}). Aggregation follows the weighted-average rule in Equation (8). The FedProx proximal coefficient is $\mu = 0.01$. Each round involves an average of 959 DP-SGD steps, totalling 28,758 steps across the full training run. Per-sample gradient clipping is set at $C = 1.0$. Table 3 reports the resulting noise multiplier and achieved ϵ for each target privacy budget.

Table 3. Noise multipliers calibrated via Opacus to hit each ϵ target at round 30. Total steps are the median across buildings (Table 2 reports the range).

ϵ Target	ϵ Actual	σ (Calibrated)	δ	Sample Rate q	Total Steps
0.5	0.484	7.03	10^{-5}	0.00533	28,758
1.0	0.984	3.71	10^{-5}	0.00533	28,758
3.0	2.954	1.51	10^{-5}	0.00533	28,758
6.5	6.402	0.945	10^{-5}	0.00533	28,758

2.5.6. Baselines

Three non-private baselines anchor the comparison. The Isolated baseline trains one CNN-LSTM per building on that building's data only, representing the accuracy attainable when no information is shared. The Centralised baseline trains a single CNN-LSTM on pooled data from all 50 buildings; it is a non-private centralised reference, in the sense that it requires raw data to leave each site and carries no formal guarantee, and it bounds what

federation costs in accuracy rather than representing an achievable privacy-preserving deployment. The FedAvg-without-DP baseline uses federated training with no privacy mechanism and is the comparator against which the marginal effect of adding DP is measured, subject to the architectural caveat in Section 2.4.2.

All baselines and all DP-FL configurations share the same experimental conditions: the same 50 buildings, the same chronological 70/15/15 split at identical timestamps, the same global StandardScaler fitted on the union of training windows, the same eleven engineered features, the same CNN-LSTM topology and parameter count, the same loss function, and the same seed for weight initialisation. Validation and test metrics are computed over the same building-hours in every case. The only differences across configurations are the aggregation rule, the presence and calibration of the DP mechanism, and the DPLSTM and GroupNorm substitutions that DP-SGD requires. A fourth baseline—non-private federated training with the DP-compatible architecture—would close the remaining gap in this design and was not run.

2.5.7. Evaluation Metrics

We report MSE, root mean squared error (RMSE), and mean absolute error (MAE) in the standardised (z-score) space for cross-building comparability. Operational kWh-unit equivalents are provided for interpretability. Two temporal granularities are reported: native hourly and daily (computed on 24-h averages). Because the standardisation in Section 2.5.4 uses the pooled standard deviation σ_y , nRMSE is reported as $\sqrt{\text{MSE}_z}$, RMSE expressed as a fraction of σ_y . Mean absolute percentage error (MAPE) is computed on unscaled kWh, since standardised values are centred near zero and a ratio taken there is meaningless, and as a mean over buildings rather than over points, with actuals below 1 kWh excluded (the ratio is undefined at zero and unstable just above it; this excludes 2.5% of test points). Symmetric MAPE (sMAPE), $2|\hat{y} - y| / (|\hat{y}| + |y|)$ averaged over every point with no exclusion, is reported alongside MAPE because it has a bounded denominator and is not dominated by the smallest buildings the way MAPE is. MAPE and nRMSE are retained for comparability with the forecasting literature; RMSE in kWh and sMAPE are the more informative operational figures on this dataset.

2.5.8. Reproducibility

The implementation uses Python 3.12, PyTorch 2.x, Flower 1.x, and Opacus 1.4 [18]. Random seeds are fixed at module load. Per-round checkpointing enables resume after interruption. The complete experimental pipeline, including all code, configuration, and result-processing scripts, is openly released as FedBuild [28]. Total compute on a single NVIDIA RTX-class laptop GPU is approximately 150 h for the full pipeline, of which the Phase 3 DP sweep (12 runs) accounts for approximately 50 h.

3. Results

All errors below are test-set mean squared error (MSE) in standardised (z-score) units unless kWh is stated; the conversion is $\text{RMSE}_{\text{kWh}} = \sqrt{\text{MSE}_z} \times \sigma_y$, with $\sigma_y = 166.06$ kWh (Section 2.5.4). Every DP-FL run is trained under two random seeds; both are reported.

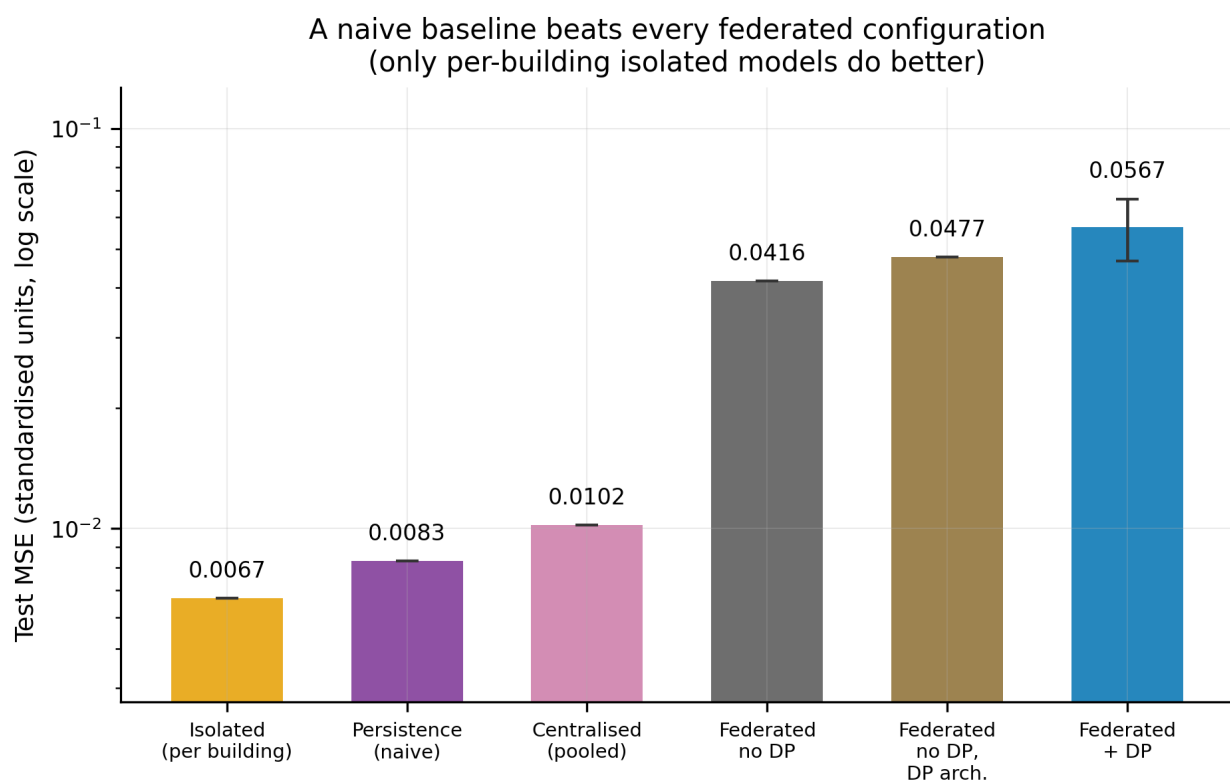
3.1. Where Federated Learning Sits

Table 4 places every method on a common scale. A naive one-step persistence forecast ($\hat{y}_{t+1} = y_t$), which requires no training at all, is beaten only by per-building isolated models; it beats centralised training and every federated configuration we test, private or not.

Table 4. Test-set error by method, on a common scale. n is the number of independent runs contributing to each row.

Method	Test MSE (z)	RMSE (kWh)	n
Isolated (per building)	0.0067	13.6	50 buildings
Persistence ($\hat{y}_{t+1} = y_t$)	0.0083	15.1	no training
Centralised (pooled)	0.0102	16.8	1
Federated, no DP	0.0416	33.9	2 seeds
Federated, no DP, DP architecture	0.0477	36.3	2 seeds
Federated + DP (all configurations)	0.0567	39.5	24 runs

The ordering, shown on a log scale in Figure 4, is the central result of this study and it is unflattering to federation as implemented here. Hourly commercial-building consumption is strongly autocorrelated, so a last-value carry-forward is a genuinely strong forecaster at this horizon—not a weak straw man. Federation, with or without differential privacy, does not clear it: the best federated configuration is worse than persistence by a factor of five, and federation costs a $6.2\times$ increase in test MSE relative to training a separate model per building. Reporting persistence correctly matters to this conclusion: an earlier implementation of the baseline indexed three steps behind the target rather than one, giving 0.0310—a value that would have placed persistence between centralised and federated training and made federation look mediocre rather than beaten. That error is corrected throughout this revision (Section 2.5.3).

**Figure 4.** The method hierarchy on a log axis. Only per-building isolated models beat the persistence forecast; every federated configuration, private or not, is worse than persistence.

Operational Units

Squared error in standardised units is the right metric for comparing configurations but says little to a building manager. Table 5 restates every result in kWh and in percentage terms: RMSE and MAE in kWh; nRMSE (RMSE against the scaler's own σ_y , as defined in Section 2.5.7); MAPE, a per-building mean over unscaled kWh with actuals below 1 kWh

excluded (2.5% of test points on this dataset); and sMAPE, which is bounded and needs no exclusion.

Table 5. Test-set error in operational units. MAE, MAPE, and sMAPE are not meaningful for the three non-federated references and are omitted for them (Section 2.5.7).

Configuration	RMSE (kWh)	MAE (kWh)	nRMSE	MAPE	sMAPE
Isolated	13.6	—	0.082	—	—
Persistence	15.1	—	0.091	—	—
Centralised	16.8	—	0.101	—	—
Federated, no DP	33.8	22.6	0.204	29%	36%
Federated, no DP, DP architecture	36.3	23.8	0.218	39%	52%
FedAvg + DP, $\epsilon = 0.5$	40.9	26.9	0.246	41%	47%
FedAvg + DP, $\epsilon = 1.0$	39.7	25.8	0.239	38%	41%
FedAvg + DP, $\epsilon = 3.0$	35.8	24.0	0.216	33%	35%
FedAvg + DP, $\epsilon = 6.5$	38.9	25.1	0.234	37%	46%
FedBN + DP, $\epsilon = 0.5$	38.5	26.2	0.232	42%	47%
FedBN + DP, $\epsilon = 1.0$	39.6	26.0	0.238	37%	43%
FedBN + DP, $\epsilon = 3.0$	40.0	24.8	0.241	35%	42%
FedBN + DP, $\epsilon = 6.5$	42.2	26.1	0.254	37%	41%
FedProx + DP, $\epsilon = 0.5$	37.5	25.7	0.226	44%	50%
FedProx + DP, $\epsilon = 1.0$	39.5	24.4	0.238	34%	39%
FedProx + DP, $\epsilon = 3.0$	40.1	25.5	0.241	38%	39%
FedProx + DP, $\epsilon = 6.5$	39.8	25.7	0.240	38%	42%

Against a mean hourly consumption of 128.46 kWh, non-private federation predicts to about 34 kWh RMSE—roughly a quarter of mean load—and DP configurations to 36–42 kWh. Persistence reaches 15.1 kWh and isolated models 13.6 kWh, so the operational picture matches the standardised one in Table 4.

MAPE and sMAPE diverge because MAPE weights the smallest buildings most heavily: a fixed absolute error is a larger fraction of a 4 kWh load than of a 496 kWh one, so buildings at the low end of Table 2’s consumption range dominate the mean. sMAPE’s bounded denominator avoids this and is the more informative percentage figure here. The two also disagree on how large the architectural term is: substituting GroupNorm and DPLSTM costs +14.8% in MSE (Table 6) but moves sMAPE from 36% to 52%, a 44% relative increase—percentage error is dominated by the small buildings, and that the substitution hurts more there than squared error suggests BatchNorm’s per-batch statistics were doing useful work specifically for low-consumption clients. This is offered as an observation, not a tested mechanism.

Table 6. Decomposing the cost of privacy. Each step is relative to the row above it; the cumulative column is relative to federated training with the standard architecture.

Configuration	Test MSE (z)	Step	Cumulative
Federated, no DP (BatchNorm + cuDNN LSTM)	0.04157	—	—
Federated, no DP, DP architecture (GroupNorm + DPLSTM)	0.04772	+14.8%	+14.8%
Federated + DP, mean over 24 runs	0.05665	+18.7%	+36.3%

3.2. Decomposing the Cost of Privacy

The DP-compatible architectural substitutions described in Section 2.4.2 (DPLSTM, GroupNorm) are required for DP-SGD but are not themselves privacy mechanisms. An ablation—the DP-compatible architecture trained federated, with no gradient clipping and no noise—separates their cost from DP-SGD’s own.

Of the total absolute gap between non-private federated training and the mean DP-FL result, shown in Figure 5, 41% is architectural and 59% is DP-SGD’s own contribution; the two do not need to be added as relative percentages, which would give the wrong total, but do add correctly as absolute increments (0.00615 and 0.00893 out of a total 0.01508).

A comparison that did not hold the architecture fixed would report the cost of privacy as +36%; correctly attributed, it is roughly +19%. The remainder is the price of using layers that admit per-sample gradients at all, independent of whether any privacy budget is actually spent.

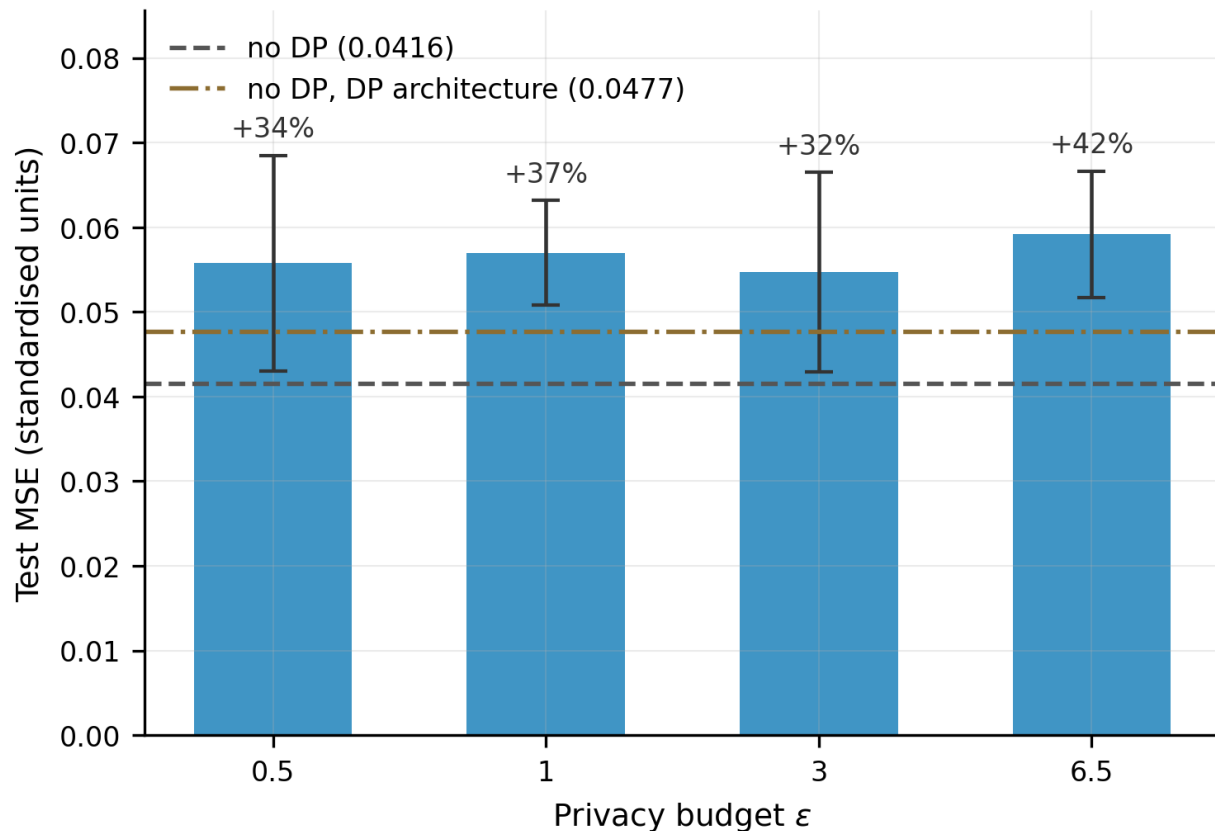


Figure 5. Cost of differential privacy, pooled over the three strategies ($n = 6$ per bar; whiskers are one standard deviation). The gap to the dash-dot reference line is DP-SGD’s own contribution; the gap below it, to the dashed no-DP line, is architectural.

Two further observations support treating this decomposition as reliable even though the DP-SGD term itself is not precisely measured. The ablation’s own two seeds differ by only 2.1%, the tightest agreement of any configuration in the study, so the architectural term is measured with unusual precision. And 19 of the 24 individual DP runs, and 11 of the 12 configuration means, sit above the ablation line, so DP-SGD’s cost is consistent in sign across the sweep even where its exact magnitude is not resolvable (Section 3.6).

3.3. Per-Configuration Results

Table 7 reports both seeds for every strategy $\times \epsilon$ configuration, together with their spread as a percentage of their mean, and the selected round for each seed.

Table 7. Test MSE at the validation-selected round, both seeds. Spread is $|seed1 - seed2|$ as a percentage of the mean. The final two rows (italicised) are the non-private references from Table 6, repeated here so the vs-no-DP and vs-ablation columns (italicised) have a visible base.

Strategy	ϵ	Seed 1	Seed 2	Mean	Spread	RMSE (kWh)	vs. No-DP	vs. Abl.	Rounds
FedAvg	0.5	0.08077	0.04356	0.06217	59.9%	41.4	+49.6%	+30.3%	19, 5
FedAvg	1.0	0.06222	0.05221	0.05722	17.5%	39.7	+37.6%	+19.9%	6, 6
FedAvg	3.0	0.05691	0.03705	0.04698	42.3%	36.0	+13.0%	−1.6%	5, 4
FedAvg	6.5	0.05595	0.05383	0.05489	3.9%	38.9	+32.0%	+15.0%	4, 5

Table 7. Cont.

Strategy	ϵ	Seed 1	Seed 2	Mean	Spread	RMSE (kWh)	vs. No-DP	vs. Abl.	Rounds
FedBN	0.5	0.04964	0.05778	0.05371	15.1%	38.5	+29.2%	+12.6%	7, 6
FedBN	1.0	0.06232	0.05148	0.05690	19.1%	39.6	+36.9%	+19.2%	4, 4
FedBN	3.0	0.07221	0.04526	0.05873	45.9%	40.2	+41.3%	+23.1%	7, 5
FedBN	6.5	0.06649	0.06292	0.06470	5.5%	42.2	+55.7%	+35.6%	1, 4
FedProx	0.5	0.05897	0.04371	0.05134	29.7%	37.6	+23.5%	+7.6%	6, 5
FedProx	1.0	0.04903	0.06454	0.05679	27.3%	39.6	+36.6%	+19.0%	5, 30
FedProx	3.0	0.06513	0.05174	0.05844	22.9%	40.1	+40.6%	+22.5%	5, 4
FedProx	6.5	0.06853	0.04735	0.05794	36.6%	40.0	+39.4%	+21.4%	5, 3
no DP	∞	0.04412	0.03902	0.04157	12.3%	33.9	—	−12.9%	4, 3
no DP, DP architecture	∞	0.04822	0.04721	0.04772	2.1%	36.3	+14.8%	—	10, 10

Only FedAvg at $\epsilon = 3.0$ falls below the ablation, shown in Figure 6, and by just 1.6%, well inside its own 42.3% seed spread, so this is not a real exception to the pattern above. No configuration in this table should be read as differing meaningfully from its neighbours without the confidence intervals that Section 4.6 specifies.

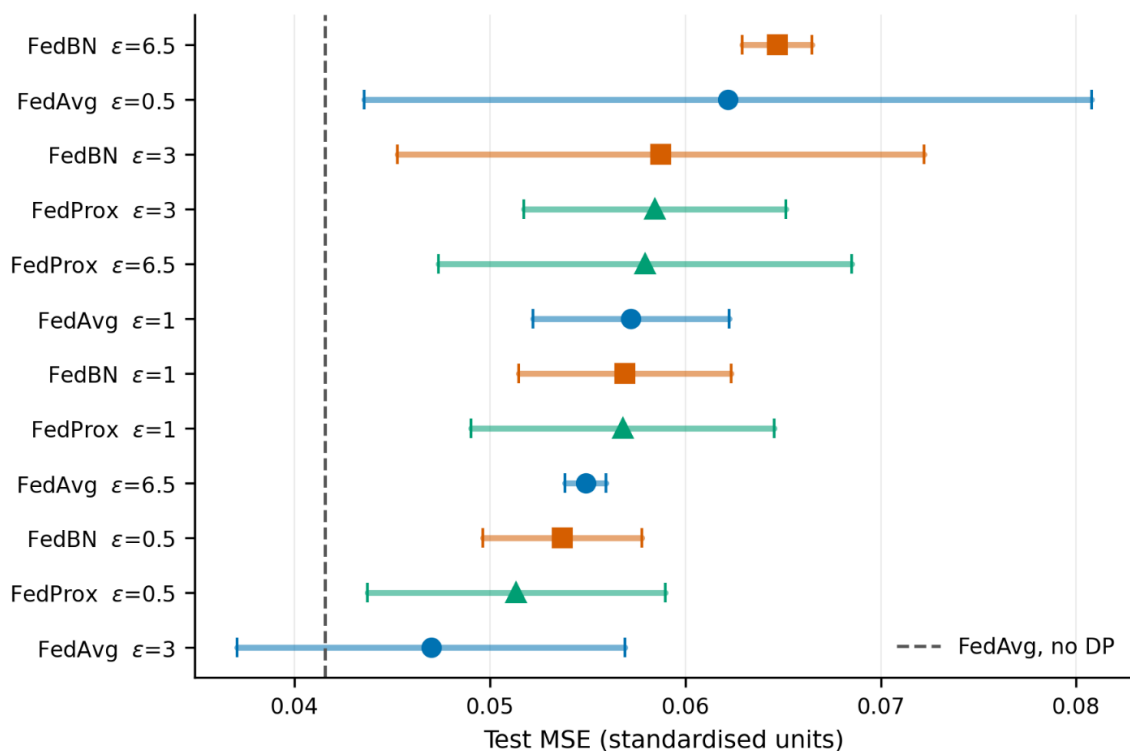


Figure 6. Seed-to-seed range within each configuration, ordered by mean. All 12 configuration means exceed the no-DP baseline; one configuration's range (FedAvg, $\epsilon = 3$) crosses below it.

3.4. Effect of the Privacy Budget

Pooling across the three strategies, six runs contribute to each ϵ level, reported in Table 8.

Table 8. Test MSE pooled across strategies, by privacy budget.

ϵ	Mean	SD	Range	vs. No-DP	vs. Abl.
0.5	0.05574	0.01272	0.0436–0.0808	+34.1%	+16.8%
1.0	0.05697	0.00618	0.0490–0.0645	+37.0%	+19.4%
3.0	0.05472	0.01176	0.0371–0.0722	+31.6%	+14.7%
6.5	0.05918	0.00746	0.0474–0.0685	+42.4%	+24.0%

The four means span 8% of their own range, against standard deviations within each level of 11–23%; the ranges overlap heavily. Per strategy, the directions disagree outright, as Table 9 shows:

Table 9. Test MSE by strategy and ϵ , showing the pattern each strategy traces across the privacy-budget sweep.

Strategy	$\epsilon = 0.5$	$\epsilon = 1.0$	$\epsilon = 3.0$	$\epsilon = 6.5$	Pattern
FedAvg	0.06217	0.05722	0.04698	0.05489	improves, then reverses
FedBN	0.05371	0.05690	0.05873	0.06470	monotonic, reversed
FedProx	0.05134	0.05679	0.05844	0.05794	worsens, then flattens

As Figure 7 shows, no two strategies agree, and FedBN runs monotonically opposite to the direction that a privacy–utility trade-off predicts. Isolating a single strategy does not rescue the trend: seed 1 alone gives a clean monotonic FedAvg curve (0.0808, 0.0622, 0.0569, 0.0560), but seed 2 does not reproduce it. That non-reproduction, more than any single number in this table, is the reason a single-seed privacy–utility curve should not be trusted at this federation size.

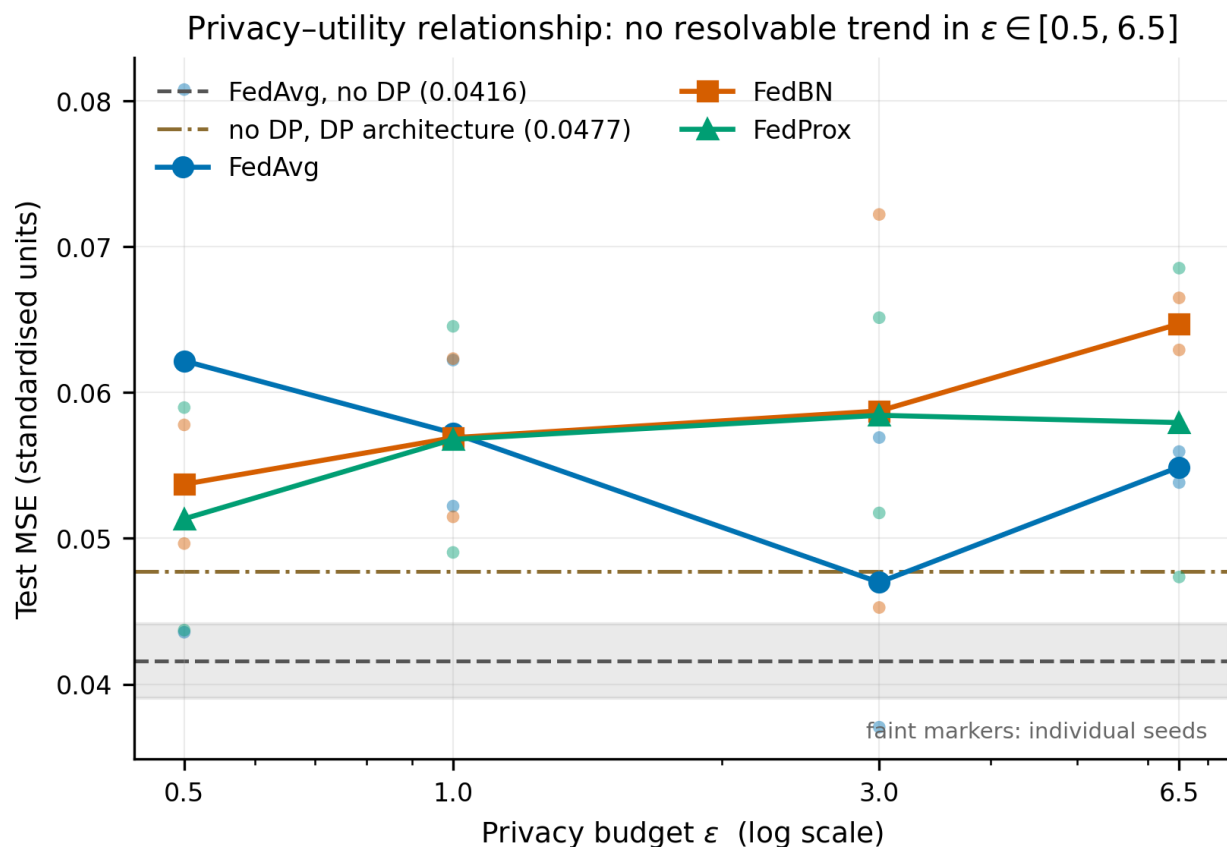


Figure 7. Privacy–utility relationship: no resolvable trend across $\epsilon \in [0.5, 6.5]$. Faint markers are individual seeds; solid markers are the per-strategy mean at each budget.

3.5. Effect of Aggregation Strategy

Pooling across ϵ , eight runs contribute to each strategy, reported in Table 10.

Table 10. Test MSE pooled across privacy budgets, by strategy.

Strategy	Mean	SD	vs. No-DP	vs. Abl.
FedAvg	0.05531	0.01216	+33.1%	+15.9%
FedProx	0.05612	0.00876	+35.0%	+17.6%
FedBN	0.05851	0.00858	+40.8%	+22.6%

The three strategies separate by 5.8% of the grand mean against a within-configuration seed spread averaging 27%; this ordering is not resolvable at the present sample size. FedProx and FedBN do show lower dispersion (SD 0.0088, 0.0086) than FedAvg (0.0122), consistent with the proximal term and local normalisation each damping client divergence—but with $n = 8$ runs per strategy this is an observation to flag for future replication, not a finding.

3.6. Variance

Every comparison above is governed by the quantities in Table 11.

Table 11. Sources of variance across the sweep.

Quantity	Value
Within-configuration seed spread, mean	27.1%
Within-configuration seed spread, max	59.9% (FedAvg, $\epsilon = 0.5$)
Spread of all 12 configuration means	31.3%
Coefficient of variation, 24 DP runs	17.8%
Ablation seed spread	2.1%

Run-to-run variance under identical settings is comparable to the total variation across every condition tested in this study. Differences smaller than roughly 27% cannot be attributed to ϵ or to strategy on this evidence; the differences actually observed in Tables 7–9 are 6–8%. Three effects nonetheless survive scrutiny, because each is consistent in sign across every run rather than merely large on average: DP versus no-DP (+36%), the architectural term (+14.8%, measured to 2.1%), and the gap between federation and every non-federated reference. Resolving a ~10% effect against this level of noise would need on the order of seven seeds per configuration—84 runs, roughly 200 GPU-hours—with no guarantee the effect exists once measured.

Where the Variance Comes from

Selected rounds cluster early, as Figure 8 shows: median round 5 of 30, mean 6.2, range 1–30.

Because the validation optimum arrives within the first few rounds, the selected model comes from the period when the federated aggregate is least settled: few aggregation steps, large client divergence, and each round still contributing substantially to the shared trunk. Small perturbations—floating-point reduction order, dropout sampling—change which round wins, and different rounds differ materially in the model they hand to evaluation. Two illustrations: FedAvg at $\epsilon = 0.5$ selected round 19 on seed 1 and round 5 on seed 2, giving 0.0808 and 0.0436, respectively; FedProx at $\epsilon = 1.0$ selected round 30 on seed 2 against round 5 on seed 1. The ablation is the counterexample that supports this reading: both of its seeds selected round 10 and landed within 2.1% of each other, the tightest agreement in the study, consistent with a configuration whose optimum is not being caught mid-transient.

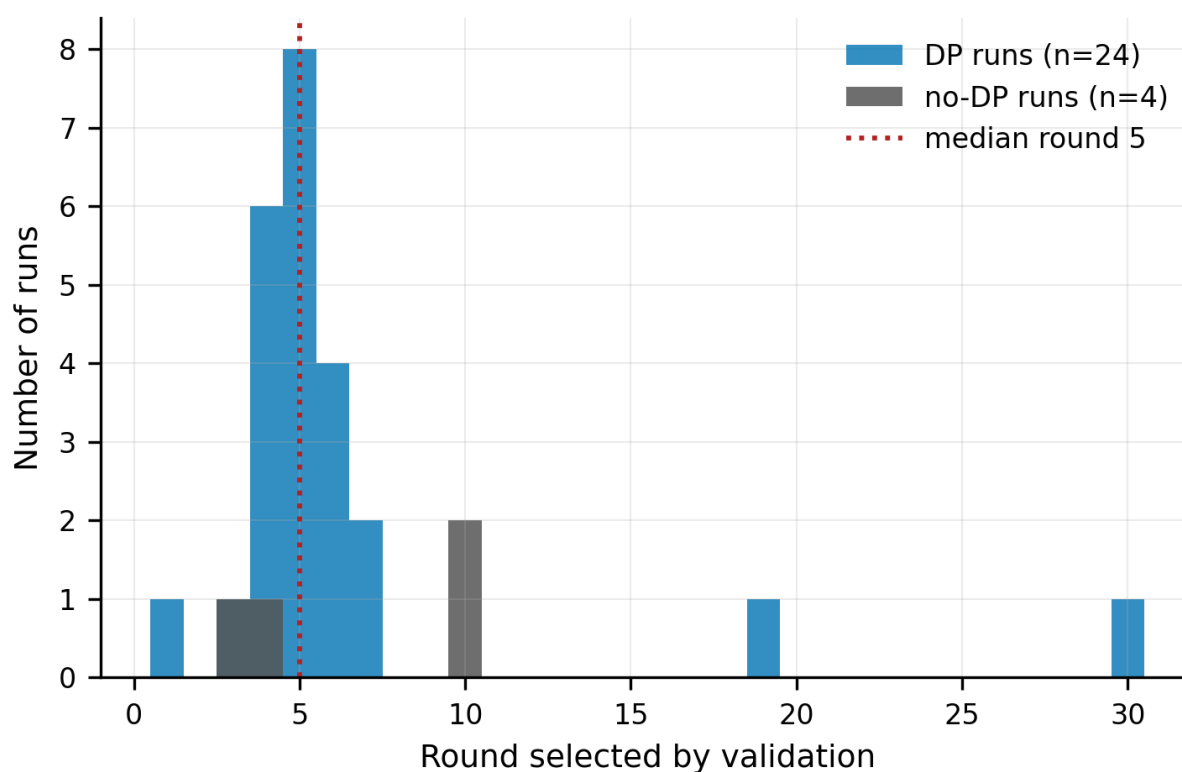


Figure 8. Validation optimum arrives early, with or without DP. Round selected by validation, DP runs ($n = 24$) against no-DP runs ($n = 4$).

3.7. Training Dynamics

Validation loss does not converge under any configuration tested, as Figure 9 shows. It falls for a few rounds, reaches a minimum, then rises while training loss keeps falling. Table 12 shows the no-DP baseline over an extended 60-round probe run to observe where this settles.

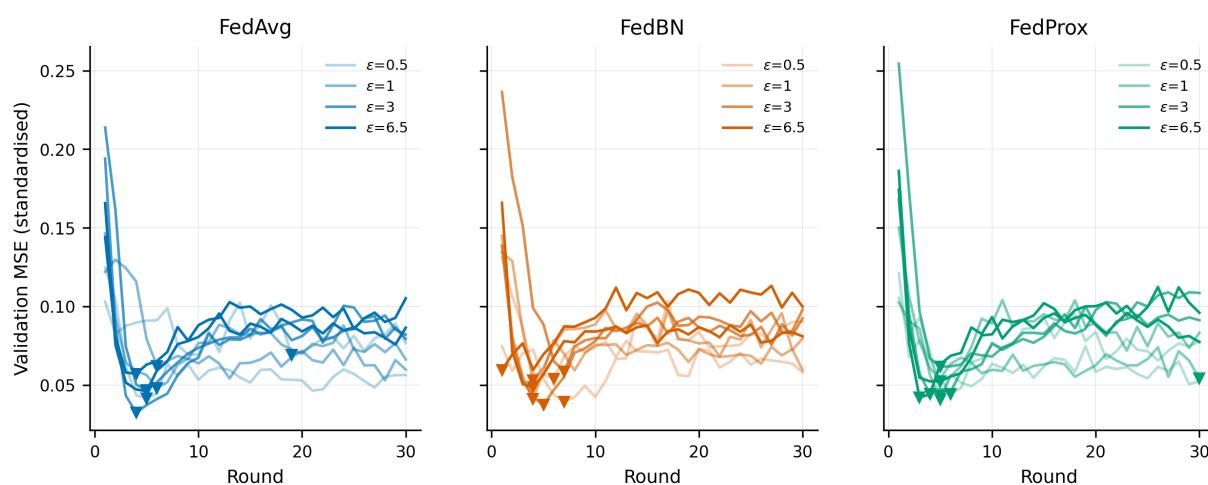


Figure 9. Validation trajectories by strategy: an early optimum, then steady degradation. Markers show the round each curve was selected at.

Table 12. No-DP federated baseline: validation MSE by round, over a 60-round probe (the sweep proper uses $T = 30$).

Round	1	3	10	20	40	60
Validation MSE	0.358	0.031	0.088	~0.06	~0.06	0.065

Training loss falls monotonically to 0.0048 over the same run, shown in Figure 10. The pattern is identical with and without DP, and with and without the DP-compatible architecture; since neither no-DP configuration contains privacy noise, the early peak cannot be attributed to DP-SGD. It is a property of federated training on this heterogeneous portfolio: the shared trunk increasingly fits an average client resembling no actual building, while each local head continues to fit its own (Section 4.1).

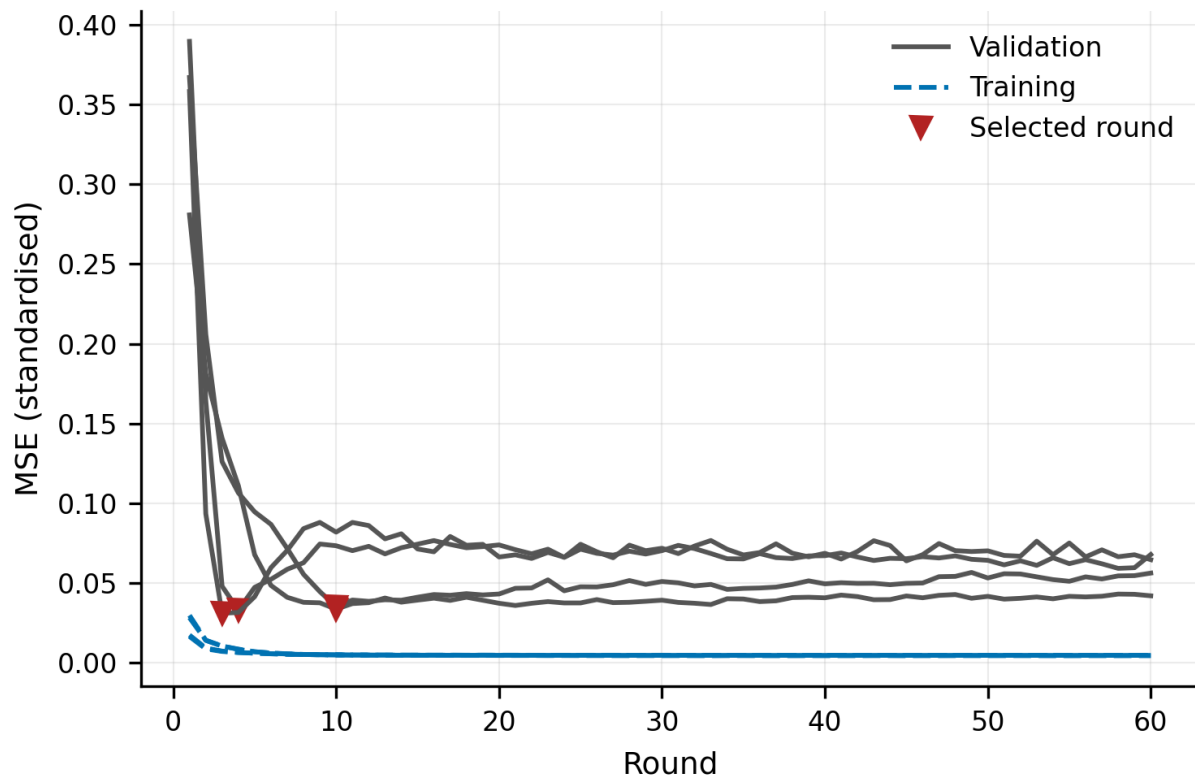


Figure 10. No-DP federation over the 60-round probe: training keeps improving while validation turns. The early optimum is not caused by DP noise, since none is present here.

3.8. Validation Against Test

The correlation between the validation loss at the selected round and the corresponding test MSE is $r = 0.898$ across all 28 runs; test exceeds validation by 17.4% on average. The correlation indicates that validation-based round selection transfers to held-out data rather than merely selecting validation-specific noise. The 17.4% gap is the expected optimism of a quantity used for selection, and is the reason every result in this paper is reported on the test set rather than on the validation loss used to select the round.

4. Discussion

4.1. Why Federation Underperforms Untrained Baselines

Persistence and isolated training both share one property that federation, as implemented here, does not: each prediction depends only on one building's own recent history. Federated averaging asks a single shared trunk to represent all 50 buildings simultaneously, and Table 2 quantifies just how different those buildings are—a $69\times$ range in floor area, a $121\times$ range in mean consumption, and a Jensen–Shannon distance that remains substantial even after standardising away scale. We advance one candidate mechanism, offered as a hypothesis rather than a demonstrated fact. A shared global scaler places every building in a common z-score space (Section 2.5.4), and because sample counts are near-uniform across clients (Section 2.5.2), the weighted average of Equation (8) is effectively unweighted: a

client's influence on each round's update is governed by the magnitude of its gradients rather than by how much data it holds. A building far from the pooled mean produces larger residuals, and hence larger gradients, than one near the centre of the distribution, and under near-uniform weighting that building pulls the shared trunk toward it out of proportion to its share of the data. The personalised head (Θ_{local}) can in principle correct for a building's own level once the trunk has settled, but if the trunk itself is repeatedly dragged toward whichever clients dominated recent rounds, the head is always chasing a moving target. This would also explain the early-peaking validation dynamic of Section 3.7: the trunk is closest to a workable shared representation in the first few rounds, before successive rounds of heterogeneous updates pull it toward no building in particular.

This account is consistent with the evidence assembled here—the heterogeneity characterisation of Table 2, the near-uniform sample counts, and the training dynamics of Section 3.7—but it is not established by it. The direct test would instrument per-client gradient norms before and after clipping and relate them to each round's movement in the shared trunk, which we did not do and identify in Section 4.6 as a priority.

4.2. What the Ablation Establishes, and What It Does Not

Section 3.2's ablation is, in our view, the most useful single addition to this revision. It answers a question the original submission could not: of the cost historically attributed to differential privacy, how much is privacy and how much is merely the architecture that privacy requires? The answer—41% architecture, 59% DP-SGD, with the architectural term measured to within 2.1%—means that a reader comparing this federation's DP and non-DP results should not attribute the full gap to ϵ having been spent.

What the ablation does not do is separate gradient clipping from Gaussian noise within DP-SGD's own 59% share. Both operations are removed together in the ablation and both are present together in every DP-FL run; a design that clipped without adding noise, or added noise without clipping, would be needed to divide that remaining share further. We regard this as a real limitation rather than a rhetorical one, because the two operations plausibly pull in different directions for the mechanism proposed in Section 4.1: clipping bounds the disproportionate influence of outlying clients regardless of privacy, and could in principle be mildly protective even where the noise it is paired with is costly. The present design cannot tell the two apart, and Section 4.6 specifies the further ablation that would.

4.3. Absence of a Resolvable Privacy–Utility or Strategy Trend

Neither Table 8 (by ϵ) nor Table 10 (by strategy) shows an effect that clears the 27%-spread threshold established in Section 3.6. This is worth dwelling on because the original submission of this manuscript reported a clean monotonic FedAvg frontier from a single seed, and that frontier does not reproduce on a second seed (Section 3.4). Single-seed federated-learning results at this scale of client heterogeneity should be treated with real caution: a trend that looks clean in one run can be an artefact of which round validation happened to select, not a property of the method.

We do not read this as evidence that ϵ and strategy have no effect, only that this design cannot detect one of the size actually observed (6–18%). The practical implication for a reader is the reverse of the original submission's: rather than choosing ϵ or an aggregation strategy on the basis of a privacy–utility curve, a practitioner working with a federation of this size and heterogeneity should not expect such a curve to be trustworthy without multi-seed replication, and should prioritise fixing the much larger federation-versus-persistence gap of Section 3.1 before optimising a budget that costs 6–18% relative to a baseline the whole system is currently losing to by $5\times$.

4.4. The Early-Peaking Validation Dynamic

Validation loss peaking within the first five rounds and degrading thereafter, identically with and without DP (Section 3.7), rules out privacy noise as the cause and points to the federation dynamic described in Section 4.1. One practical consequence is that the round budget itself may be poorly matched to this federation: if the shared trunk is at its most useful in the first few rounds, a protocol that personalises the head more aggressively relative to the trunk, or that stops updating the trunk earlier, may outperform one that continues federating regardless of the round count. We did not test a modified protocol along these lines and flag it as a direction rather than a finding.

4.5. Computational Cost

Table 13 summarises the wall-clock cost of the experimental pipeline reported in this paper, run on a single consumer-class GPU. The DP sweep alone—12 configurations $\times$ 2 seeds, plus 2 no-DP and 2 ablation seeds, 28 runs in total—took several weeks of elapsed time to complete.

Table 13. Computational cost of the experimental pipeline. Hardware: a single consumer-class GPU. Per-run hours are approximate.

Phase	Description	Wall-Time
Isolated + centralised baselines	50 buildings $\times$ CNN-LSTM, plus 1 pooled run	~1 h
Persistence baseline	No training required	<1 min
Federated, no DP (probe)	60 rounds, 2 seeds	~1–2 h
Ablation (DP architecture, no DP)	60 rounds, 2 seeds	~1–2 h
Single DP-FL run	30 rounds, full DP-SGD	~5–15 h
Full DP sweep	12 configurations $\times$ 2 seeds	several weeks

The single-run DP-SGD estimate is scaled down from the $T = 50$ schedule of the original submission in proportion to the reduced round count and is not independently timed; it is flagged for the authors' confirmation against their own logs. This cost is the practical constraint behind Section 4.6's two-seed design and behind deferring the further seeds and ablations specified there.

4.6. Limitations and Required Validation

The evidence in this paper is bounded in specific, quantifiable ways, and we set those boundaries out together with the work that would move them.

Seed count. Every comparison in Sections 3.4 and 3.5 sits inside a within-configuration seed spread averaging 27% (Section 3.6). Two seeds resolve only effects at or above roughly that size; the between-configuration differences actually observed (6–18%) are smaller. Reaching a resolution of about 10% would need on the order of seven seeds per configuration, 84 runs beyond what is reported here, at the per-run cost in Table 13—with no guarantee the underlying effect survives measurement at that resolution.

Clipping versus noise. Section 3.2's ablation separates the DP-compatible architecture from DP-SGD but not clipping from noise within DP-SGD's own share (Section 4.2). A design with four arms at fixed seed and fixed ϵ —standard architecture with no DP; DP-compatible architecture with no DP; DP-compatible architecture with clipping only; full DP-SGD, would complete the decomposition. The second and third arms are the two this paper's ablation does not yet run.

Clipping threshold. The clipping threshold was fixed at $C = 1.0$ throughout. If clipping bounds the disproportionate-gradient mechanism proposed in Section 4.1, its effect should vary systematically with C ; a sweep over $C \in \{0.1, 0.5, 1.0, 5.0\}$ at fixed ϵ would supply that

dose–response evidence and indicate whether $C = 1.0$ is a reasonable default or an accident of common practice.

Mechanism instrumentation. Direct evidence for the mechanism proposed in Section 4.1, the distribution of per-client gradient norms before and after clipping, and how each round’s movement in the shared trunk relates to it—was not collected and would require instrumenting the training loop and repeating the sweep.

Scope of the federation. The federation is 50 Office and Education buildings from a single BDG2 site, at hourly resolution, over $T = 30$ rounds with a two-hour input window and no weather or occupancy inputs. None of these choices is varied here, and each plausibly interacts with the heterogeneity structure this paper’s central argument depends on; a longer input window in particular is the most obvious lever on absolute accuracy given that even persistence is competitive.

Untuned hyperparameters. The FedProx coefficient $\mu = 0.01$ and $E = 5$ local epochs were taken from common practice, not tuned. Five local epochs under full participation drive substantial client divergence before each aggregation step and plausibly interacts with the early-peaking dynamic of Section 3.7; this interaction is not evaluated.

Privacy accounting. The privacy accounting in Section 2.3 is exact under the full participation used throughout this study; the general bound for partial participation, stated there for completeness, is not exercised by any result in this paper.

Normalisation statistics. The global StandardScaler of Section 2.5.4 is fitted centrally on pooled training data and its statistics are not protected by DP-SGD; this is a gap in the end-to-end guarantee rather than a component of it. Per-client local scaling, secure aggregation of scaling statistics, or a DP-protected estimate of μ_y and σ_y would close it and are left for future work.

Taken together, two experiments would most efficiently narrow the gap between what is reported here and a settled result: the seed count needed to resolve the ϵ and strategy comparisons, and the clipping-versus-noise arm that would complete Section 3.2’s decomposition. We regard the results in this paper as motivating that work rather than substituting for it.

5. Conclusions

This paper set out to characterise the privacy–utility trade-off in differentially private federated learning for building-energy forecasting. In the course of a corrected re-run—which fixed a local-head persistence bug, an inoperative privacy accountant, an unimplemented FedBN partition, and a mis-indexed persistence baseline, each described in Section 2 and disclosed in the accompanying response to reviewers—the central finding changed. Rather than characterising how differential privacy affects an already-competitive federated forecaster, this paper now reports that federated averaging over this heterogeneous 50-building portfolio does not clear a naive persistence baseline, private or not, and it characterises that gap and the additional, smaller cost of privacy on top of it.

Supported. Three findings are supported by the evidence presented:

A naive one-step persistence forecast beats every federated configuration tested—all 24 DP-FL runs and both non-private federated baselines. Only per-building isolated training beats persistence. Federation costs a $6.2\times$ increase in test MSE relative to isolated training.

An ablation separating the DP-compatible architecture from DP-SGD decomposes the historically reported +36% cost of privacy into an architectural component (+14.8%, agreeing to within 2.1% across seeds) and a DP-SGD component (+18.7%, consistent in sign across 19 of 24 individual runs). The architectural component is not a privacy cost and should not be reported as one.

The validation optimum arrives early (median round 5 of 30) and degrades thereafter, identically with and without DP and with and without the DP-compatible architecture, ruling out privacy noise as the cause. Validation-selected rounds correlate well with test performance ($r = 0.898$), which is why test error is reported rather than the best-validation figures.

Not supported at this sample size. Two claims are not supported at the sample size used here:

Any ordering of test error by ϵ within $[0.5, 6.5]$. The original submission's single-seed monotonic frontier for FedAvg does not reproduce on a second seed.

Any ordering among FedAvg, FedBN, and FedProx, or any interaction between strategy and privacy budget. Between-configuration differences of 6–18% sit inside a within-configuration seed spread averaging 27%.

Not tested. Four questions remain genuinely open and are not addressed by any result in this paper:

Privacy budgets outside $[0.5, 6.5]$; the absence of a gradient within this window does not imply the absence of one at $\epsilon = 0.1$ or $\epsilon = 50$.

Federation sizes other than 50 buildings, or sites other than the single BDG2 climate zone used throughout.

Local epoch counts other than $E = 5$, which plausibly interact with the early-peaking dynamic of Section 3.7 given how much client divergence five local epochs permits before each aggregation step.

Input windows longer than $T = 2$ h; given that even a one-step persistence forecast is competitive with the best federated result, window length is the most obvious lever on absolute accuracy that this study leaves untried.

The validation programme in Section 4.6—principally, a multi-seed repetition sufficient to resolve the ϵ and strategy comparisons, and a clipping-only ablation arm completing the decomposition in Section 3.2—is what separates the findings above from a settled characterisation of privacy-preserving federated learning for building-energy forecasting. We regard the negative result reported here—that federation as implemented does not yet beat a naive baseline—as itself a useful and, we believe, underreported class of finding in this literature, and release the corrected FedBuild pipeline [28] so that the further experiments specified above can be run against the same, now-verified, codebase.

Author Contributions: Conceptualization, J.E.A.; Validation, A.I.; Formal analysis, J.E.A.; Writing—original draft, J.E.A.; Writing—review & editing, A.I.; Supervision, A.O.; Project administration, A.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are openly available in Zenodo (v1.0.0), <https://doi.org/10.5281/zenodo.20527682>.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Li, W.; Zhu, J.; Sui, W.; Cheng, L.; Ji, Y.; Guo, Y. Quantifying Seasonal Demand-Side Flexibility in Residential Air Conditioning under Diverse Control Strategies. *Energy Build.* **2026**, *352*, 116764. [CrossRef]
2. McMahan, H.B.; Moore, E.; Ramage, D.; Hampson, S.; Agüera y Arcas, B. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), Fort Lauderdale, FL, USA, 20–22 April 2017*; PMLR: New York, NY, USA, 2017; Volume 54, pp. 1273–1282.
3. Li, T.; Sahu, A.K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; Smith, V. Federated Optimization in Heterogeneous Networks. In *Proceedings of the Machine Learning and Systems (MLSys), Austin, TX, USA, 2–4 March 2020*; Volume 2, pp. 429–450.

4. Li, X.; Jiang, M.; Zhang, X.; Kamp, M.; Dou, Q. FedBN: Federated Learning on Non-IID Features via Local Batch Normalization. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 3–7 May 2021.
5. Zhu, L.; Liu, Z.; Han, S. Deep Leakage from Gradients. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019; Volume 32.
6. Nasr, M.; Shokri, R.; Houmansadr, A. Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-Box Inference Attacks against Centralized and Federated Learning. In Proceedings of the IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 19–23 May 2019; pp. 739–753. [\[CrossRef\]](#)
7. Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H.B.; Mironov, I.; Talwar, K.; Zhang, L. Deep Learning with Differential Privacy. In Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS), Vienna, Austria, 24–28 October 2016; pp. 308–318. [\[CrossRef\]](#)
8. McMahan, H.B.; Ramage, D.; Talwar, K.; Zhang, L. Learning Differentially Private Recurrent Language Models. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
9. Choudhury, O.; Gkoulalas-Divanis, A.; Salonidis, T.; Sylla, I.; Park, Y.; Hsu, G.; Das, A. Differential Privacy-Enabled Federated Learning for Sensitive Health Data. *arXiv* **2019**, arXiv:1910.02578.
10. Miller, C.; Kathirgamanathan, A.; Picchetti, B.; Arjunan, P.; Park, J.Y.; Nagy, Z.; Raftery, P.; Hobson, B.W.; Shi, Z.; Meggers, F. The Building Data Genome Project 2, Energy Meter Data from the ASHRAE Great Energy Predictor III Competition. *Sci. Data* **2020**, *7*, 368. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Briggs, C.; Fan, Z.; Andras, P. Federated Learning for Short-Term Residential Energy Demand Forecasting. *Energy Build.* **2022**, *9*, 573–583. [\[CrossRef\]](#)
12. Cho, S.; Park, C.S. Federated learning for enhancing extrapolation ability of HVAC models: Case study on two real-life DOAS units. *J. Build. Perform. Simul.* **2026**, *19*, 622–637. [\[CrossRef\]](#)
13. Xu, C.; Chen, G.; Li, C. Federated learning for interpretable short-term residential load forecasting in edge computing network. *Neural Comput. Applic* **2023**, *35*, 8561–8574. [\[CrossRef\]](#)
14. Mármol Campos, E.; González Vidal, A.; Hernández-Ramos, J.L.; Skarmeta, A. Federated transfer learning for energy efficiency in smart buildings. In *Proceedings of the IEEE INFOCOM 2023—IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*; IEEE: Piscataway, NJ, USA, 2023; pp. 1–6. [\[CrossRef\]](#)
15. Ibrahim, M.M.; William, M.A.; Zidane, I.F.; Hanafy, A.A. Enviro-Economic Assessment of Vegetation–PV Envelope Retrofits for Nearly Zero Energy Buildings in Hot-Humid Climates. *Sustainability* **2026**, *18*, 4526. [\[CrossRef\]](#)
16. Mironov, I. Rényi Differential Privacy. In Proceedings of the IEEE 30th Computer Security Foundations Symposium (CSF), Santa Barbara, CA, USA, 21–25 August 2017; pp. 263–275. [\[CrossRef\]](#)
17. Mironov, I.; Talwar, K.; Zhang, L. Rényi Differential Privacy of the Sampled Gaussian Mechanism. *arXiv* **2019**, arXiv:1908.10530.
18. Yousefpour, A.; Shilov, I.; Sablayrolles, A.; Testuggine, D.; Prasad, K.; Malek, M.; Nguyen, J.; Ghosh, S.; Bharadwaj, A.; Zhao, J.; et al. Opacus: User-Friendly Differential Privacy Library in PyTorch. *arXiv* **2021**, arXiv:2109.12298.
19. Andrew, G.; Thakkar, O.; McMahan, H.B.; Ramaswamy, S. Differentially Private Learning with Adaptive Clipping. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Virtual, 6–14 December 2021; Volume 34.
20. Geyer, R.C.; Klein, T.; Nabi, M. Differentially Private Federated Learning: A Client Level Perspective. *arXiv* **2017**, arXiv:1712.07557.
21. Khalil, M.; Esseghir, M.; Merghem-Boulahia, L. Privacy-Preserving Federated Learning: An Application for Big Data Load Forecast in Buildings. *Comput. Secur.* **2023**, *127*, 103094. [\[CrossRef\]](#)
22. Bonawitz, K.; Ivanov, V.; Kreuter, B.; Marcedone, A.; McMahan, H.B.; Patel, S.; Ramage, D.; Segal, A.; Seth, K. Practical Secure Aggregation for Privacy-Preserving Machine Learning. In Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS), Dallas, TX, USA, 30 October–3 November 2017; pp. 1175–1191. [\[CrossRef\]](#)
23. Neelakantan, A.; Vilnis, L.; Le, Q.V.; Sutskever, I.; Kaiser, L.; Kurach, K.; Martens, J. Adding Gradient Noise Improves Learning for Very Deep Networks. *arXiv* **2015**, arXiv:1511.06807.
24. Papernot, N.; Steinke, T. Hyperparameter Tuning with Rényi Differential Privacy. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 25–29 April 2022.
25. Karimireddy, S.P.; Kale, S.; Mohri, M.; Reddi, S.; Stich, S.; Suresh, A.T. SCAFFOLD: Stochastic Controlled Averaging for Federated Learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, Virtual, 13–18 July 2020; PMLR: New York, NY, USA, 2020; Volume 119, pp. 5132–5143.
26. Dwork, C.; Roth, A. The Algorithmic Foundations of Differential Privacy. *Found. Trends Theor. Comput. Sci.* **2014**, *9*, 211–407. [\[CrossRef\]](#)

27. Building Data Genome Project 2. Available online: <https://zenodo.org/records/3887306> (accessed on 30 July 2026). [[CrossRef](#)]
28. Ameh, J.E. *FedBuild: A Privacy-Preserving Federated Learning Pipeline for Building Energy Forecasting*; Zenodo: Geneva, Switzerland, 2026. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.