

Parameter-performance scaling of hybrid quantum-classical policies for demand-charge-aware battery control

Eunsung Oh

esoh@gachon.ac.kr

Gachon University

Article

Keywords: behind-the-meter battery control, demand charge, hybrid quantum-classical model, parameter efficiency, ComStock, multi-objective control

Posted Date: September 4th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9933256/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Parameter-performance scaling of hybrid quantum-classical policies for demand-charge-aware battery control

Eunsung Oh^{1,*}

¹ Department of Electrical Engineering, Gachon University, Gyeonggi-do 13120, South Korea

* Corresponding author: esoh@gachon.ac.kr

ABSTRACT

Hybrid quantum-classical models are often proposed as compact alternatives to classical neural networks; however, the claim that they require fewer parameters has rarely been tested across full model size ranges against strong classical baselines. In this study, this claim was tested on a realistic energy-control task: 15-min, full-month control of behind-the-meter batteries that must minimize commercial electricity demand charges using 549 simulated office buildings under a standardized utility tariff. Three policy families—a classical neural network and two hybrid quantum-circuit variants sharing the same inputs and control head—were trained by behavioral cloning from a perfect-foresight optimizer and compared at four parameter budgets on a locked held-out test set. One compact hybrid design (~1,400 parameters) provided the best control performance among all tested model and modestly outperformed the size-matched classical network. However, the advantage is small, does not hold for other sizes, and depends on a task-specific control head rather than on the quantum circuit itself. The classical network remains a strong baseline and runs faster. Therefore, this study reports a narrow, bounded parameter-efficiency benefit rather than a general quantum advantage and outlines which hardware-level validation would still be required. All quantum circuits were evaluated using exact simulations.

Keywords: behind-the-meter battery control, demand charge, hybrid quantum-classical model, parameter efficiency, ComStock, multi-objective control

Introduction

Behind-the-meter battery storage can reduce commercial electricity bills by shifting grid imports under time-of-use energy prices and lowering monthly peak demand [1–4]. The latter is a long-horizon control problem in which the demand charge is determined by the highest short-interval import demand in the billing month; therefore, an apparently reasonable discharge can be harmful if the battery is depleted before the true monthly peak. Recent reinforcement-learning studies have begun to target demand-charge minimization explicitly [5–7] and combine bill reduction with battery state-of-health considerations [8, 9]. Consequently, demand-charge-aware storage control is now a well-formed setting for testing compact policies conditioned on state of charge, running peak, tariff state, time, building scale, and user preference between bill reduction and battery degradation.

Hybrid quantum-classical models are often used as compact function approximators in noisy intermediate-scale quantum settings [10–12]. Variational quantum-circuit (VQC) policies have been proposed for reinforcement learning more generally [13–18] and for building or industrial energy control specifically [19, 20]. However, for energy control, the direct scientific hypothesis is that a hybrid quantum policy will outperform all classical policies. A larger classical model may remain the strongest absolute regret controller. The relevant hypothesis is narrower: a hybrid quantum-classical policy family may reach a comparable control regime with fewer deployable actor parameters than well-matched classical policy families. This is a parameter-performance scaling question, not a single-model performance ranking.

Testing this hypothesis requires a locked model-family comparison. A compact hybrid checkpoint that performs similarly to a small multilayer perceptron (MLP) does not establish parameter efficiency. A valid test requires multiple actor parameter budgets, strong classical comparators, a validation-only threshold definition, a locked building-level test cohort, high-weight tail-risk diagnostics, and a resource audit that separates the actor parameter count from the inference latency and quantum circuit evaluations. As finite-shot sampling and hardware-noise effects can erode VQC trainability and inference fidelity [21, 22], the present study was restricted to an analytic-expectation simulation on a state-vector backend so that the parameter-budget comparison is not confounded by device-noise sources, implying that any hardware-execution claim would require a separate noise-aware validation pass.

In this paper, demand-charge-aware battery control is formulated as a target-native 15-min, full-month benchmark and parameter-performance scaling is evaluated across hybrid and classical policy families. The cohort was the ComStock 2025 Release 2 AMY2018 office buildings filtered to Los Angeles County, ASHRAE climate zone 3B, and vintage 1980 or later [23]. The PG&E B-10 tariff was imposed as a standardized summer demand-charge scenario [24]; the experiment does not claim that the buildings are native PG&E service-territory customers. The main split was at the building level, with 329 training buildings, 110 validation buildings, and 110 locked test buildings. The primary endpoint was the minimum deployable actor parameter count required to meet a validation-defined comparable-performance threshold for the locked test cohort. The results are interpreted using regime-balanced normalized regret, paired hybrid-versus-matched-MLP contrasts on a dense preference grid, high-weight timing-tail risk, and quantum/classical resource accounting.

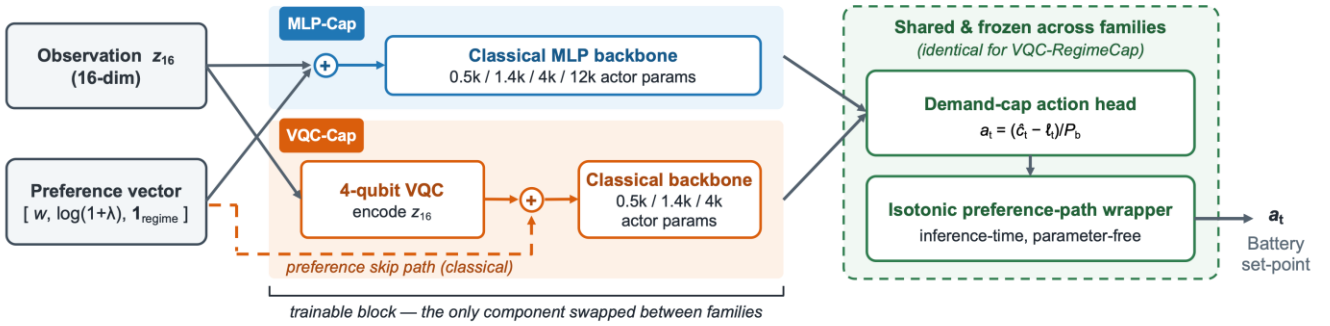
The main contributions of this study can be summarized as follows. (i) A frozen 549-building demand-charge benchmark with an explicit building-level train/validation/locked-test split, pre-registered comparator set, and SHA256-anchored finalist manifest; (ii) A same-action-projection parameter-performance comparison of two hybrid VQC families (skip-connected and soft preference-regime demand-caps) against a strong classical MLP family across four deployable actor-parameter budgets; (iii) A validation-derived performance-threshold ladder with explicit sensitivity bounds, replacing a single tuned threshold; and (iv) A resource and statistical audit that distinguishes deployable-parameter compactness from computational efficiency, reports Holm–Bonferroni-adjusted paired contrasts, and isolates an inference-time isotonic preference-path wrapper as a parameter-free post-processor.

Results

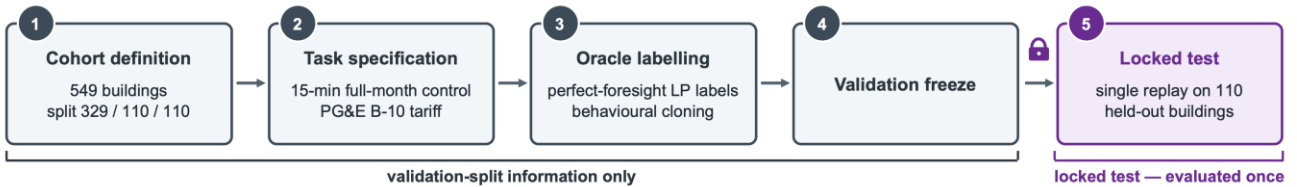
Locked 549-building test of parameter-performance scaling

The evidence path was designed to answer the following question: Does the hybrid policy family shift the compact parameter-performance frontier relative to strong classical alternatives? The model families, parameter budgets, seed count, reward definition, validation-selection rule, and statistical analysis were fixed before evaluating the locked test. The locked test was not used for the architecture choice, hyperparameter tuning, threshold selection, seed selection, or figure design. Figure 1 summarizes the comparison framework and evaluation workflow used throughout this study.

A Comparison framework — matched-budget policy families share every stage except the trainable block



B Locked evaluation workflow — selection is frozen on validation data before a single locked-test replay



Primary endpoint: the smallest deployable actor-parameter count reaching the validation-defined comparable-performance regime.

Figure 1. Comparison framework and evaluation workflow. **A.** The two retained policy families MLP-Cap (classical) and VQC-Cap (skip-connected hybrid) share the same observation vector, preference vector, demand-cap action head, and inference-time isotonic wrapper, and differ only in the trainable block; matched-budget pairing isolates the effect of swapping the classical block for the hybrid block. **B.** The five-stage workflow freezes all selection-relevant choices using validation-split information only (stages 1–4) before a single locked-test replay on the 110 held-out buildings (stage 5).

The two retained model families in Fig. 1A differ only in their training blocks. Both consume the same 16-dimensional observation vector z_{16} (normalized load, running peak, state of charge, tariff indicators, time-of-day, recent-load summaries, native peak, and battery power) and the same six-dimensional preference vector $[w, \log(1 + \lambda), \mathbf{1}_{\text{regime}}]$, and both map to an action through the same demand-cap action head $a_t = (\hat{c}_t - \ell_t)/P_b$, where $\hat{c}_t$ is the predicted import cap, ℓ_t is native load, and P_b is battery power, followed by the same inference-time implied-cap isotonic preference-path wrapper. The families differ only in the block placed between the observation and action head: MLP-Cap is a classical MLP that concatenates the two vectors at its input (495/1,393/3,985/11,929 actor parameters), whereas VQC-Cap is a skip-connected hybrid variational quantum circuit (4 qubits, 24 trainable quantum parameters; 499 / 1,393 / 3,993 actor parameters) in which only z_{16} is encoded into the circuit and the preference vector is routed through a classical skip path that rejoins the VQC measurement output before the classical backbone and demand-cap head. VQC-RegimeCap, which is not shown, follows the same template with a smooth preference-regime-gated demand-cap head. Holding the observation, action head, and wrapper constant across families and pairing families at matched parameter budgets isolates the architectural contribution of the hybrid block from the demand-cap engineering choice and from differences in observation or output handling; thus, the locked-test paired contrasts can be interpreted as block-level comparisons rather than head- or wrapper-level comparisons.

Figure 1B shows the five-stage workflow that produces locked-test evidence. Stages 1 and 2 fix the cohort (549 ComStock office buildings, split 329/110/110) and the control task (15-min full-month battery control under the PG&E B-10 summer demand-charge tariff); stage 3 generates perfect-foresight oracle labels for event-aware behavioral cloning; stage 4 freezes every selection-relevant choice—model family, parameter budget, seed, wrapper, and validation-derived threshold—using validation-split information only; finally, stage 5 is a single locked-test replay on the 110 held-out buildings with no further selection. The primary endpoint is the smallest deployable actor parameter count at which a family meets the validation-derived comparable-performance threshold in this locked test.

Table 1 presents the frozen experimental design. The primary cohort contains 549 buildings from ComStock 2025 Release 2 AMY2018 after filtering to Los Angeles County, ASHRAE climate zone 3B, office use, and vintage 1980 or later. The main experiment uses June–September only. As PG&E B-10 was applied as a standardized tariff scenario rather than a geographically native tariff assignment, tariff realism was handled by reporting the B-10 screening criteria, native peak distribution, and battery-size distribution rather than by claiming utility-territory representativeness. As shown in Supplementary Figure S1, the cohort distributions overlap closely across the train, validation, and locked-test splits.

Table 1. Frozen experimental design for the locked parameter-performance scaling study. Detailed source-data hashes, tariff constants, and battery parameters are tabulated in Supplementary Tables S4 and S5.

Design element	Frozen value
Dataset	ComStock 2025 Release 2, AMY2018; split manifest SHA256 prefix bb86e5e10df4.
Cohort	549 office buildings; LA County, ASHRAE 3B, vintage ≥ 1980 ; native summer peak 75.61–497.95 kW.
Split	Building-level 329/110/110 (train/validation/locked test); split seed 20260429.
Season and episodes	June–September only; 1,316/440/440 building-month episodes.
Tariff	PG&E B-10 (secondary bundled, effective 2026-03-01); demand charge 20.50 USD/kW-month; energy 0.339/0.278/0.245 USD/kWh (peak/part-peak/off-peak).
Battery	Behind-the-meter; $P_b = 0.25\max\ell$, $E_b = 2P_b$; $\eta = 0.96$; no export; initial SoC 0.5; degradation 0.0596 USD/kWh.
Training	Fixed-dose event-aware BC; two epochs; no early stopping or validation checkpoint search.
Model grid	MLP-Cap at 500/1400/4000/12000; VQC-Cap and VQC-RegimeCap at 500/1400/4000. Exact deployable counts: 495/1393/3985/11929; 499/1393/3993; 498/1393/3999.
Seeds	Ten seeds per retained cell; 100/100 locked-test cells completed, zero failures.
Primary endpoint	Minimum deployable parameters meeting a validation-defined regret threshold on the locked test.

Parameter efficiency defined via a validation-based performance threshold

The primary threshold is defined without locked-test information. Let $\tilde{r}_m(P)$ denote the dense-grid mean validation or test regret for model family m at actor parameter budget P . The high-capacity reference is the validation-selected 12k-parameter MLP-Cap policy; its validation performance defines a comparable-performance regime as follows:

$$\tau_R^{\text{val}} = \text{UCB}_{95}(\tilde{R}_{\text{MLP-Cap-12k}}^{\text{val}}), \quad (1)$$

where the upper confidence bound is computed using building-level bootstrap and then rounded upward according to a pre-specified rounding rule. A family reaches the comparable-performance regime if its locked-test mean regret is at or below τ_R^{val} :

$$P_m(\tau_R^{\text{val}}) = \min_p \{P: \tilde{R}_m^{\text{test}}(P) \leq \tau_R^{\text{val}}\}. \quad (2)$$

A secondary safety-aware version was pre-registered with the additional constraint that the binary high-weight tail-failure rate $T_m(P)$ remain at or below $\tau_T = 0.20$; however, this constraint is satisfied trivially by all retained locked-test rows (all rows have $T_m = 0.20$ exactly, the saturation value induced by the dense-grid coverage of $w \in \{0.8, 0.9\}$); therefore, the secondary threshold does not separate the retained families. The reported parameter-efficiency interpretation uses the primary threshold τ_R^{val} , and the continuous high-weight burden metrics in Supplementary Tables S25 and S26 are reported as the discriminating high-weight evidence. If a family does not reach the threshold within the tested grid, it is reported as “Not reached.”

In the final locked replay, the 12k MLP-Cap point is retained as the high-capacity classical reference, but is not the best locked-test MLP point, with a macro regret of 0.659 (95% confidence interval (CI), 0.615–0.711), compared with 0.596 (0.547–0.654) for the 1.4k MLP-Cap and 0.596 (0.566–0.632) for the 4k MLP-Cap. Therefore, the interpretation does not rely on any single tuned threshold; instead, Table 2 reports the minimum-parameter crossing under a pre-specified validation-derived threshold ladder. The ladder spans the strictest validation point (best validation mean across all MLP-Cap budgets, $\tau_R = 0.580$), the matched-budget compact reference ($\tau_R = 0.590$), the high-capacity reference mean ($\tau_R = 0.660$), and the upper 95% CI used as the primary U95 threshold ($\tau_R^{\text{val}} = 0.710$). The full parameter-performance frontier and matched-budget paired contrasts remain the primary interpretation, with the threshold ladder serving as a calibrated robustness check rather than a tunable selection rule.

Table 2. Validation-threshold sensitivity for parameter-budget crossing on the locked test. Thresholds are derived from the validation split only and applied without retuning to the locked-test mean regret. Each threshold is the unrounded validation source statistic rounded upward to the nearest 0.005 increment. “Not reached” indicates that no tested budget for the given family has mean regret at or below the threshold.

Threshold τ_R (source statistic; unrounded)	VQC-Cap	VQC-RegimeCap	MLP-Cap	Note
0.580 (best MLP val. mean, 0.5758)	Not reached	Not reached	Not reached	Strictest sensitivity.
0.590 (1.4k MLP val. mean, 0.5863)	1,393	Not reached	Not reached	Compact-reference; only VQC-Cap 1.4k crosses.
0.660 (12k MLP val. mean, 0.6594)	1,393	1,393	495	High-capacity reference mean.
0.710 (12k MLP val. U95, 0.7077)	1,393	1,393	495	Primary U95 threshold (broad).

Hybrid and classical frontier comparison across actor-parameter budgets

Figure 2 shows the locked-test parameter-performance frontier, which is the main result of this study. The frontier compares validation-selected checkpoints for each model family and actor parameter budget. Table 3 summarizes the retained main grid. These results support the hybrid parameter-efficiency hypothesis only if the hybrid family reaches the validation-defined comparable-performance regime with fewer deployable actor parameters than the strongest pre-specified classical comparator family. If the hybrid family is competitive only with the screened secondary comparators and not with MLP-Cap, the conclusion is that the tested hybrid design does not shift the strongest compact frontier. If the hybrid family does not reach the regime earlier than the classical comparators, the parameter-efficiency hypothesis is not supported for this benchmark.

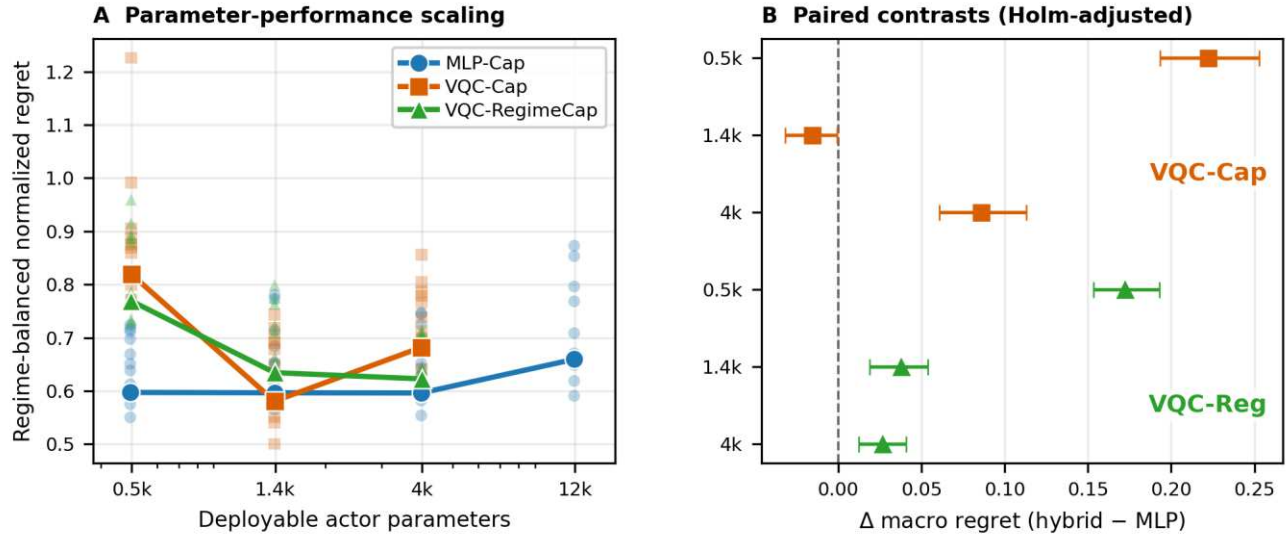


Figure 2. Locked-test parameter-performance frontier. **A.** Regime-balanced normalized regret vs deployable actor parameters. Bold lines connect ten-seed means; faint markers are per-seed locked-test cells. The hybrid frontier is non-monotone: VQC-Cap is best at 1.4k but worse at 0.5k and 4k; MLP-Cap is flat from 0.5k to 4k and rises at 12k under the fixed two-epoch dose. **B.** Forest plot of paired hybrid-minus-matched-MLP macro regret with building-level 95% bootstrap CI; family is indicated by marker shape and in-line labels. All six pre-specified contrasts are significant after Holm–Bonferroni adjustment, but only the 1.4k VQC-Cap row favors the hybrid (marginal, Holm $p=0.050$); see Table 5 for exact values.

Table 3. Model families and locked-test parameter grid. Main rows present the locked-test replay; deployable parameter counts are reported per cell. Screening rows are validation-only and described in the Supplementary Information.

Family (block + head)	0.5k	1.4k	4k	12k	Role
MLP-Cap (MLP + demand-cap)	495	1,393	3,985	11,929	Main classical comparator.
VQC-Cap (skip-VQC + demand-cap)	499	1,393	3,993	–	Main hybrid finalist.
VQC-RegimeCap (regime-VQC + demand-cap)	498	1,393	3,999	–	Main hybrid robustness family.
MLP direct/residual			Screening only		SI screening evidence.
Fourier / RBF / earlier VQC variants			Screening only		SI screening evidence.
Rule-based peak shaving			No trainable parameters		Sanity/floor reference.
No storage and oracle			Not applicable		Normalization floor and ceiling.

The locked test supports a compact hybrid demand-cap point but not a simple monotone scaling law, such that larger is better. Table 4 presents the seed-averaged metrics with building-level bootstrap intervals. VQC-Cap at the 1.4k budget is the best tested macro-regret point, with regime-balanced normalized regret (lower is better; 0 = perfect-foresight oracle, 1 = no-storage baseline) 0.581 (95% CI, 0.538–0.632), mean bill saving 449.5 USD/month (407.7–494.1), and mean negative degradation cost -193.9 USD/month (-212.4 to -177.1). The matched 1.4k MLP-Cap has regret 0.596 (0.547–0.654), bill saving 440.5 USD/month (400.4–482.4), and negative degradation cost -193.0 USD/month (-211.6 to -175.5). The paired VQC-Cap minus MLP-Cap macro-regret difference at 1.4k is -0.016 (-0.032 to -0.001), computed from the unrounded per-building paired differences before any rounding rather than from the three-decimal point estimates above (the rounded inputs $0.581 - 0.596$ give -0.015 , which differs by 0.001 at the third decimal), with Cohen’s $d_z = -0.19$ (just below the conventional small-effect threshold of $|d| = 0.20$) and a sign-flip permutation $p = 0.050$, which remains at the conventional 0.05 boundary after Holm–Bonferroni adjustment across the six pre-specified hybrid-vs-MLP paired contrasts. The other five paired contrasts, summarized in Table 5, go in the opposite direction: VQC-Cap 0.5k and 4k and all three VQC-RegimeCap budgets are significantly worse than the matched MLP-Cap budget under the same Holm-adjusted permutation rule, with positive paired regret differences of 0.027 to 0.222 ($d_z = 0.35$ – 1.59 ; Holm-adjusted $p \leq 0.001$). The 0.5k hybrid points underfit, and VQC-Cap at 4k reverses relative to 1.4k. VQC-RegimeCap improves from 0.5k to 4k, reaching regret 0.623 (0.588–0.664) at 4k, but does not exceed the best VQC-Cap or MLP-Cap points. Thus, the evidence is strongest for a 1.4k

skip-connected demand-cap hybrid design rather than for monotone parameter scaling across all hybrid budgets, and the only paired-contrast advantage that survives multiplicity correction is the 1.4k VQC-Cap row.

Table 4. Primary locked-test metrics over the 110-building locked-test cohort and dense preference grid. Brackets indicate building-level bootstrap 95% CIs (10,000 resamples; seeds averaged within each family-budget-building unit). The binary high-w tail-failure rate (saturated at 0.20 for all rows) is reported in Supplementary Table S19.

Model and budget	Params	Macro regret	Bill saving	Neg. degradation
MLP-Cap, 0.5k	495	0.597 [0.552, 0.651]	433.2 [392.5, 475.2]	-182.7 [-200.2, -166.2]
MLP-Cap, 1.4k	1,393	0.596 [0.547, 0.654]	440.5 [400.4, 482.4]	-193.0 [-211.6, -175.5]
MLP-Cap, 4k	3,985	0.596 [0.566, 0.632]	397.6 [361.8, 434.1]	-187.2 [-204.3, -170.6]
MLP-Cap, 12k	11,929	0.659 [0.615, 0.711]	416.6 [376.0, 459.2]	-186.5 [-204.4, -169.9]
VQC-Cap, 0.5k	499	0.819 [0.750, 0.900]	402.8 [363.6, 443.9]	-214.9 [-235.5, -195.1]
VQC-Cap, 1.4k	1,393	0.581 [0.538, 0.632]	449.5 [407.7, 494.1]	-193.9 [-212.4, -177.1]
VQC-Cap, 4k	3,993	0.682 [0.632, 0.741]	435.0 [392.9, 477.9]	-212.7 [-232.7, -193.7]
VQC-RegimeCap, 0.5k	498	0.769 [0.711, 0.834]	403.0 [362.6, 446.1]	-206.5 [-226.1, -187.7]
VQC-RegimeCap, 1.4k	1,393	0.634 [0.596, 0.679]	411.0 [372.8, 451.8]	-193.5 [-211.8, -175.8]
VQC-RegimeCap, 4k	3,999	0.623 [0.588, 0.664]	392.1 [354.4, 431.2]	-185.7 [-203.6, -168.6]

Table 5. Pre-specified hybrid-vs-MLP paired contrasts (locked test, six contrasts). Paired differences are hybrid minus matched MLP; negative is better for the hybrid. Cohen’s d_z uses the paired-difference SD across the 110 locked-test buildings (seeds averaged within each family-budget-building unit). p -values are two-sided sign-flip permutations (10,000 permutations) with Holm–Bonferroni adjustment across the six contrasts.

Contrast	Budget	Δ regret [95% CI]	Cohen’s d_z	Holm p	Favored
VQC-Cap vs MLP-Cap	500	0.222 [0.194, 0.253]	1.385	0.0006	MLP
VQC-Cap vs MLP-Cap	1,400	-0.016 [-0.032, -0.001]	-0.190	0.0500	Hybrid (marginal)
VQC-Cap vs MLP-Cap	4,000	0.086 [0.061, 0.113]	0.607	0.0006	MLP
VQC-RegimeCap vs MLP-Cap	500	0.172 [0.153, 0.193]	1.594	0.0006	MLP
VQC-RegimeCap vs MLP-Cap	1,400	0.038 [0.019, 0.054]	0.401	0.0006	MLP
VQC-RegimeCap vs MLP-Cap	4,000	0.027 [0.013, 0.041]	0.352	0.0006	MLP

Preference-conditioned replay to test the compact frontier usability

The policies are conditioned on a bill-degradation preference weight w , following the broader preference-conditioned multi-objective control idea that one deployable policy should span a trade-off path rather than a scalar objective [25–30]. A model that performs well for one weight but collapses for another does not provide a usable frontier. Training for the active rescue/finalist path uses a sparse preference grid chosen to cover the bill-priority plateau, main oracle knee, and high-degradation-avoidance endpoint, $w \in \{0.00, 0.30, 0.40, 0.80, 0.90\}$, while final validation and locked-test replay use the dense preference grid $w \in \{0.0, 0.1, \dots, 0.9\}$. A dense replay results in regret, bill savings, degradation cost, and throughput across the grid.

As shown in Fig. 3, the retained policies form usable but imperfect preference-conditioned paths. The learned policies are far from a perfect-foresight oracle, particularly at high degradation-avoidance weights, in which the oracle can combine low cycling with a well-timed peak reduction. Among the learned policies, the 1.4k VQC-Cap and 1.4k MLP-Cap trace similar bill-degradation regions at low and knee weights, whereas VQC-RegimeCap 4k is more conservative in terms of bill savings. Regret increases sharply at the high- w endpoint for all learned policies, indicating that sparse demand-charge peak timing remains the dominant difficulty; the full dense-grid frontier across all retained family-budget cells and representative monthly traces are presented in Supplementary Figures S2 and S3. Therefore, frontier-shape diagnostics remain part of the interpretation: the locked-test result supports compact hybrid competitiveness, but not a perfectly monotone preference frontier.

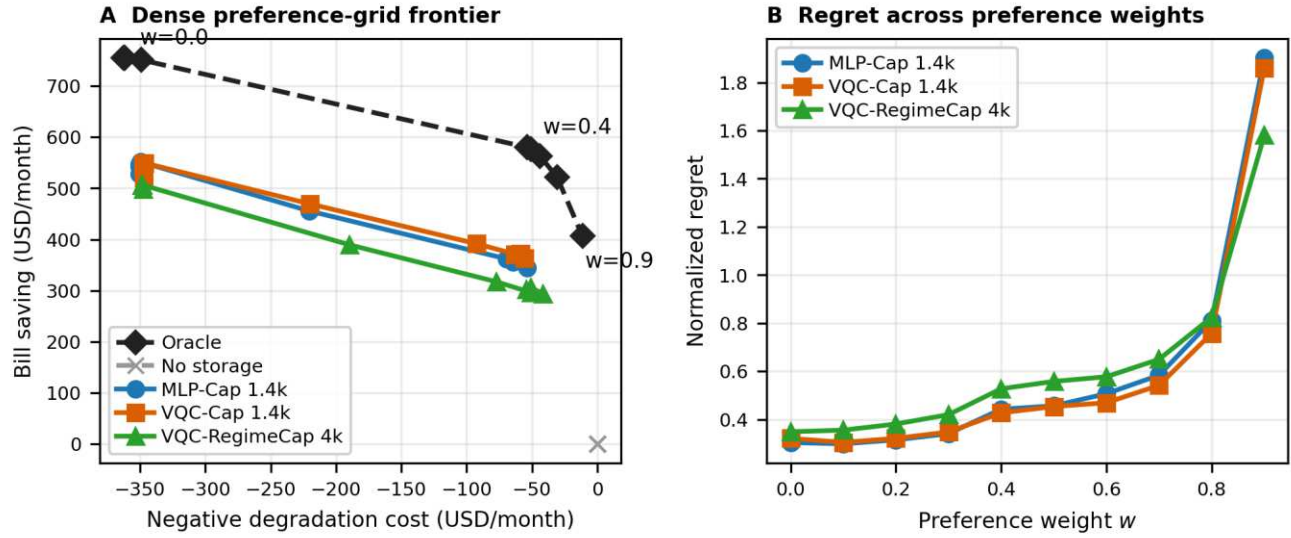


Figure 3. Dense preference-grid evaluation results. **A.** Bill-saving vs negative-degradation-cost frontier for the oracle, no-storage reference and representative locked-test policies; numeric labels indicate selected preference weights. **B.** Oracle-relative normalized regret across the dense preference grid, showing the sharp increase at high- w for all learned policies.

High-weight tail risk to constrain practical parameter efficiency

Demand-charge control is vulnerable to rare timing failures. A compact policy that reduces the average regret but depletes the battery before high-cost monthly peak events is not practically parameter-efficient. Therefore, in Fig. 4A, $w \in \{0.8, 0.9\}$ is treated as a high-weight operational stress test and continuous high-weight normalized regret is reported instead of relying on a binary failure label alone. Per-weight locked-test rankings, locked-test row-level bill savings, and negative degradation costs are tabulated in Supplementary Note 9 (Tables S25 and S26). A representative bill-priority monthly trace at $w = 0$ is provided in Fig. S2 to illustrate the demand-cap policy's cycling pattern, with the understanding that the high-weight stress test is characterized by per-weight metric tables rather than by this single bill-priority trace. At $w = 0.9$, the locked-test ranking invert-MLP-Cap 4k attains the lowest mean normalized regret among retained cells (1.561) and no hybrid configuration beats it, whereas at $w = 0.8$, VQC-Cap 1.4k is the best (0.758), with the next-best three rows being MLP-Cap budgets between 0.774 and 0.811; the full per-weight ranking is reported in Supplementary Table S25. A binary site-month tail indicator such as $R^\pi(e, w) > 0.05$ is retained as a sensitivity diagnostic in the Supplementary Information; Supplementary Tables S19 and S24 respectively present the binary rate along the continuous metrics and catalogue the candidate failure-case taxonomy (early depletion, late response, over-conservative cycling avoidance, no-export/projection clipping, forecast-blind event, and terminal-correction artifact) that motivates the per-weight diagnostic. As $w \in \{0.8, 0.9\}$ are both trained sparse-grid weights, the high-weight diagnostic characterizes intra-training-grid tail behavior rather than generalization to unseen high weights.

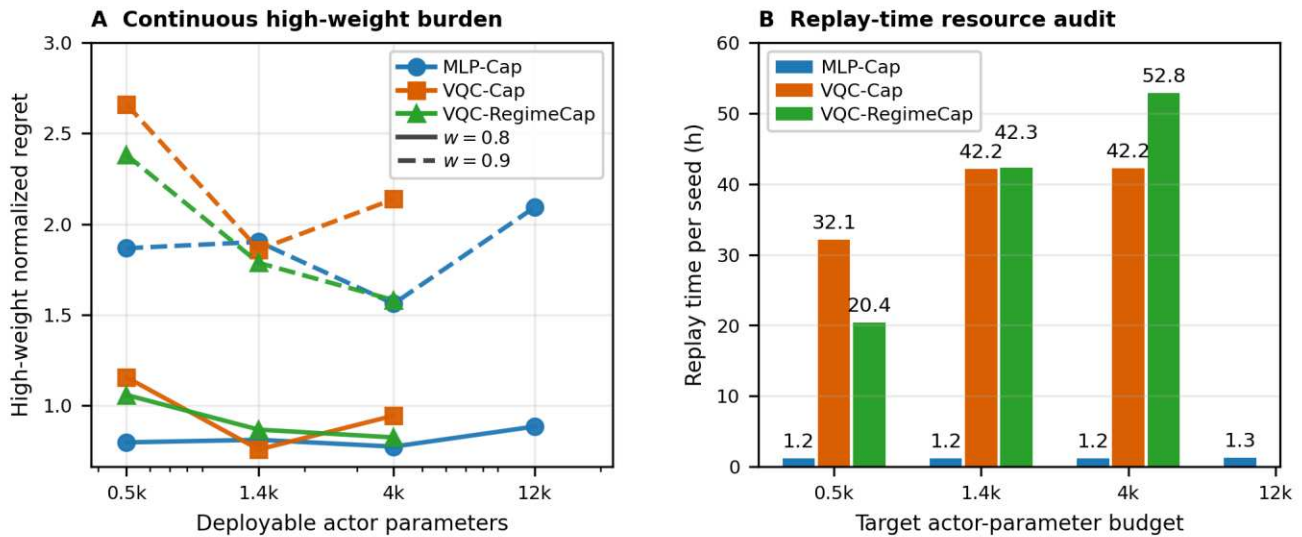


Figure 4. High-weight and resource diagnostics. **A.** Continuous high-weight normalized regret vs deployable actor parameters at $w = 0.8$ (solid) and 0.9 (dashed); at $w = 0.9$, MLP-Cap 4k beats all hybrid configurations. **B.** Scalar locked-test replay time per seed, showing that compactness is not computationally efficient.

Using the claim ladder, the current result is a positive but bounded parameter-compactness result. The tested VQC-Cap family provides the best locked-test macro-regret point at the compact 1.4k budget and improves the matched MLP-Cap in the paired building-level contrast. However, the MLP-Cap frontier is close to 0.5k–4k, and is much cheaper to replay. Therefore, the appropriate assertion is that the skip-connected demand-cap hybrid VQC can shift the deployable-parameter frontier in this benchmark and not that hybrid quantum-classical policies are generally superior or computationally faster.

Resource accounting to distinguish parameter compactness from computational efficiency

Deployable actor parameters are primary resource variables, but they are not the only variables. A hybrid VQC policy may use fewer trainable parameters while requiring more circuit evaluations per action and higher inference latency; Fig. 4B and Table 6 quantify this gap on the locked-test scalar replay path. Therefore, the final claim distinguishes the deployable parameter compactness from computational efficiency.

A resource audit separates the positive parameter results from the deployment cost. The seed-averaged scalar replay time on the CPU-based scalar replay path is approximately 1.18, 42.18, and 42.31 h for 1.4k MLP-Cap, 1.4k VQC-Cap, and 1.4k VQC-RegimeCap, respectively. Therefore, the compact hybrid result is not computationally efficient. It is meaningful only under a storage, communication, or deployable-parameter constraint or under future execution settings in which the quantum-circuit evaluation cost is reduced.

Table 6. Resource and selection audit. Per-seed replay times are obtained from the scalar locked-test path on a 32-core Threadripper CPU; quantum circuit details are presented in Supplementary Table S12.

Family / audit item	Actor parameters	Per-seed replay time	Role
MLP-Cap	495–11,929	1.18–1.26 h	Main classical comparator.
VQC-Cap	499–3,993	32.1–42.2 h	Main hybrid finalist (best compact point at 1.4k).
VQC-RegimeCap	498–3,999	20.4–52.8 h	Hybrid robustness family with smooth regime gates.
Screened classical families	Screening only	Validation-only	MLP-direct/residual, Fourier and RBF (SI evidence).
Screened VQC families	Screening only	Validation-only	Direct/residual/one-hot regime VQCs (SI evidence).
Selection audit	–	–	No early stopping or best-checkpoint search; final BC checkpoint replayed.

Discussion

In this study, a narrow hypothesis was tested: Hybrid quantum-classical policies may not outperform large classical policies but may be more deployable-parameter efficient in a demand-charge-aware control task. Experiments were designed accordingly. Parameter-performance frontiers, rather than single checkpoints, were compared and comparable performance is defined using validation-only information before the locked-test evaluation.

The locked-test outcome supports the narrow parameter-performance claim of the hybrid design. VQC-Cap at 1.4k deployable parameters is the best tested macro-regret point and modestly improves on the matched 1.4k MLP-Cap point estimate. VQC-RegimeCap improves with parameter budget, but remains below the best VQC-Cap and MLP-Cap rows. The binary high-weight tail diagnostic is not discriminative in the final replay because all retained rows have the same rate; therefore, the dense frontier, regret-by-weight path, and monthly economics carry more interpretive weight. Computationally, the hybrid replay is far slower than the MLP replay; therefore, the assertion is deployable-parameter compactness rather than speed.

The hybrid advantage in this study is also conditional on the demand-cap action head, rather than being intrinsic to the VQC block, with the caveat that the ablation evidence is preliminary because of its limited seed coverage and residual architecture. Supplementary Table S28 presents a matched-1.4k-budget cap-head ablation under the same observation, BC training, and isotonic wrapper as the locked-test path: the cap-head baselines use the locked-test 10-seed rows (MLP-Cap at 0.5863 and the skip-connected hybrid VQC variant h5 at 0.5606), while the direct-head ablation rows are two-seed validation cells (MLP direct at 0.7020, seed standard deviation 0.0003; VQC direct variant h1 at 0.7649, seed standard deviation 0.1092); therefore, the ablation rows have wide and asymmetric seed-level uncertainty and should be read as a preliminary, not a definitive, head-isolation diagnostic. With this caveat, removing the cap head worsens regret for both blocks ($\Delta = +0.116$ and $+0.204$ for MLP and VQC, respectively, both computed from the unrounded source values in Supplementary Table S28 and reported here at three-decimal precision for consistency). The within-block hybrid-minus-MLP contrast at the cap-head baseline is -0.026 (0.5606 vs 0.5863, 10 vs 10 seeds), whereas the contrast at the direct head is $+0.063$ (0.7649 vs 0.7020, 2 vs 2 seeds, with the wide 0.109 VQC seed standard deviation entering directly). The $+0.063$ direct-head contrast additionally swaps the VQC variant (skip-connected h5 in the cap-head baseline and direct h1 in the ablation), which confounds the head change with a change in the VQC architecture. The combined evidence, therefore, supports the narrower interpretation that the deployable-parameter compactness assertion is a property of the combined demand-cap + skip-connected VQC design rather than of a generic VQC block; a fully head-isolated, seed-matched (10-seed) ablation using the h5 architecture with a direct head was identified as a confirmatory follow-up.

The 12k MLP-Cap row deserves a brief mechanistic note. Its locked-test macro regret (0.659) is worse than that of the 1.4k and 4k MLP-Cap rows (both at 0.596); therefore, the frontier is non-monotone in the actor-parameter count for MLP-Cap and VQC-Cap. Supplementary Table S15 presents a pre-specified training-dose sensitivity check that doubled the supervised dose from two to four epochs for the 4k and 12k MLP-Cap rows under the same observation, head, isotonic wrapper and validation protocol; doubling the dose worsened the mean validation-split regret at both budgets (4k: 0.5973 $\rightarrow$ 0.6402; 12k: 0.6594 $\rightarrow$ 0.7366, where the unrounded validation 12k value 0.6594 happens to round to the same 0.659 as the locked-test 12k value). Therefore, the 12k anomaly is not an undertraining artifact and is more consistent with a capacity-data mismatch in the fixed training cohort. Consequently, the 1.4k sweet spot is treated as a genuine feature of the frontier rather than a transient consequence of the chosen training dose, and the threshold-sensitivity ladder presented in Table 2 indicates parameter-budget crossings under both the 12k and best-MLP validation means to make this interpretation explicit.

Mapping the locked-test evidence onto the pre-specified interpretive grid yielded a mixed answer. The pre-specified rubric was that the parameter-efficiency hypothesis would be supported if a hybrid family reached the validation-defined comparable-performance regime with fewer deployable actor parameters than MLP-Cap, that hybrid competitiveness without parameter advantage would be reported if a hybrid family matched MLP-Cap at similar budgets, and that no evidence of a hybrid parameter-efficiency advantage would be reported if MLP-Cap reached the regime with equal or fewer parameters. Under this rubric, only the 1.4k VQC-Cap row satisfies the first clause, which is the best macro-regret point in the locked test, and crosses the validation-defined threshold ladder at a smaller deployable-parameter count than the nearest MLP-Cap row that crosses the same ladder, with a paired hybrid-minus-MLP regret difference of -0.016 (Holm $p = 0.050$). The remaining VQC-Cap budgets (0.5k and 4k) and all three VQC-RegimeCap budgets fall under the third clause: the matched MLP-Cap point is either at the same parameter count with a smaller paired regret or at a larger parameter count with a non-discriminative gap, and the paired contrasts favor MLP-Cap with Holm $p \leq 0.001$. The screened Fourier transform, RBF, and earlier VQC variants were retained as validation-only screening evidence and did not change the overall mapping. Therefore, the conclusion is deliberately narrow: a bounded parameter-compact advantage for the 1.4k skip-connected demand-cap hybrid, no general parameter efficiency claim across hybrid budgets, and no quantum-advantage claim in the broader sense, which was not required by the original hypothesis.

A high-weight analysis is essential for interpreting compactness. Demand charges reward correct timing near monthly peaks, and rare timing errors or unnecessary cycling can dominate economic outcomes. A compact policy that reaches the average regret threshold but carries a higher high-weight degradation burden, lower bill savings, misplaced state of charge, or more binary timing failures should be described as parameter compact but operationally fragile. Conversely, a hybrid policy that matches the MLP-Cap frontier, while maintaining a high-weight operational burden, provides stronger evidence of practical parameter efficiency.

A resource audit limits the deployment interpretation. The primary endpoint was the deployable actor parameters. This is relevant when the model size, policy storage, or communication are constrained. This is not identical to computational efficiency. A hybrid VQC policy can perform more circuit evaluations per action than a classical policy, which performs matrix operations. Therefore, the final claim should use the phrase deployable-parameter compact unless the latency audit also supports computational efficiency. The compactness claim is also conditioned on the simulation backend, and all VQC results in this study used analytic state-vector expectations; therefore, finite-shot sampling noise and hardware decoherence [21, 31] would need to be retested before any equivalent hardware-execution statement can be made. Supplementary Table S30 reports a single-seed, 30-building shot-noise sensitivity diagnostic on the 1.4k VQC-Cap. Under ideal binomial shot sampling, the 1.4k VQC-Cap macro regret moves by at most 0.001 between the analytic backend (0.5575) and $n_{\text{shots}} \in \{1,024,8,192\}$ settings (0.5568 and 0.5573), which is well below the locked-test bootstrap uncertainty band. The diagnostic does not recompute the matched 1.4k MLP-Cap row on the same 30-building subset; therefore, the supported claim is the VQC-side stability of the macro-regret estimate under shot noise rather than the preservation of the VQC-vs-MLP paired contrast, which would require a matched-MLP rerun (this should be explored in future work). The diagnostic also does not model coherent gate errors or device decoherence; therefore, it bounds the sampling-noise sensitivity rather than confirming hardware-execution robustness. Recent reinforcement-learning evaluations have also emphasized that compact-policy claims should be supported by multi-seed locked-test protocols [32, 33], which is the role of the ten-seed, building-level paired bootstrap used in the present study.

Despite its many advantages, this study had certain limitations. For instance, the cohort comprised office buildings in one climate-zone/geography filter, the tariff was a standardized PG&E B-10 scenario, only summer was considered, the battery sizing rule was fixed, and the model families were finite. Accordingly, the results should not be generalized to all building archetypes, tariffs, seasons, degradation models, or quantum architectures. A causal model-predictive-control comparator with a fitted load forecast was considered during protocol freeze and excluded because the validation-tuned deterministic rule-based controller (the closest pre-registered non-learning baseline) achieved a mean normalized regret of 0.953 against a no-storage regret of 1.0 on the validation split. Adding a stronger MPC baseline would constitute a separate study rather than a fair comparator under the locked actor-parameter accounting used in this study.

Overall, the appropriate conclusion is that hybrid quantum-classical battery control is generally superior. The locked S8 replay indicates that the tested skip-connected demand-cap hybrid VQC shifts the compact deployable-parameter frontier in this 549-building demand-charge-aware benchmark, while the strong MLP-Cap comparator and significantly larger hybrid inference cost keep the conclusion deliberately narrow.

Methods

Comparison framework

In this study, three retained model families (MLP-Cap, VQC-Cap, and VQC-RegimeCap) in a frozen building-level locked-test cohort were compared using an identical observation vector, action projection, oracle target source, and inference-time post-processor. The primary endpoint was the minimum deployable actor-parameter count needed to satisfy a validation-derived performance threshold on the locked test, with the sensitivity ladder reported in Table 2. The secondary endpoint was the paired hybrid-versus-MLP macro-regret contrast at each matched budget, as reported in Table 4. The split was fixed before training; the finalist manifest with SHA256 checkpoint hashes is frozen before the locked-test replay, and no architecture, hyperparameter, threshold, seed, or figure choice is informed by the locked-test outcomes. The remaining subsections specify the cohort, tariff, battery, observations, model families, training protocols, baselines, statistics, and resource accounting that epitomize this framework.

For traceability, the validation-only pipeline that produced the frozen finalist manifest is organized into internal stages whose codes appear in the source-data and run-ledger references throughout this manuscript: S7-A (five-seed screening of classical and hybrid candidates on the validation split), S7-B (10-seed full-validation finalist replay), S7-C (sparse-grid full-train refit and isotonic-wrapper development at the matched 0.5k, 1.4k and 4k budgets), S7-D (12k budget-scaling delta on the MLP-Cap arm), S7r_rescue (preference-conditioned rescue path that introduced the demand-cap action head and produced the dense-grid frontier targets), and S8 (one-shot locked-test replay on the 110 held-out buildings using the validation-frozen

finalist manifest). The locked-test results reported in the main text come entirely from S8; all model development decisions are upstream of S8 and use only the training and validation splits.

ComStock cohort and split

The benchmark used ComStock 2025 Release 2 AMY2018 load profiles. The public ComStock dataset describes the time series and annual energy consumption of the U.S. commercial building stock at end-use resolution, and the public release table lists the 2025_2 AMY2018 release. The ComStock 2025 Release 2 software stack was documented using OpenStudio 3.8.0, EnergyPlus 24.1.0, and Buildstockbatch 2024.11.0. The final cohort consisted of 549 buildings after filtering for Los Angeles County, ASHRAE climate zone 3B, office building type, vintage 1980 or later, and the study-specific PG&E B-10 screening rule.

The split unit is a building. All June–September months from a given building remained in the same split. The primary split comprised 329 training, 110 validation, and 110 locked-test buildings, yielding 1,316, 440, and 440 summer building-month episodes, respectively. The split is stratified according to the native summer peak demand and total summer energy use, with optional balancing by floor area and load factor if sparse strata do not arise. The split manifest was frozen before model training.

Standardized PG&E B-10 tariff scenario

The PG&E Schedule B-10 was imposed as a standardized commercial demand-charge scenario. This tariff framing isolates parameter-performance scaling under a fixed demand-charge structure and does not claim that Los Angeles County buildings are native PG&E service territory customers. The primary experiment used only the summer months (June–September). The tariff implementation uses a secondary bundled B-10 rate snapshot effective 2026-03-01. The summer energy prices were 0.33947, 0.27778, and 0.24522 USD/kWh for the peak, part-peak, and off-peak periods, respectively. The summer peak window was 16:00–21:00 Pacific standard time, part-peak windows were 14:00–16:00 and 21:00–23:00, and all other intervals were off-peak. The monthly demand charge was 20.50 USD/kW-month based on the highest 15-min average import demand in the billing month, with a running-peak reset at the start of each month. Peak Day Pricing events and fixed customer charges were excluded from the objective because they are either event-specific or policy-invariant in this benchmark. Source load timestamps were recorded in Eastern standard time and shifted by three hours to Pacific standard time for the summer tariff assignment.

Battery model and action projection

The battery is a behind-the-meter device with a fixed sizing rule, power limits, energy capacity, one-way efficiency, state-of-charge limits, and no export projection. For building b , the battery power is

$$P_b = 0.25 \max_{t \in \text{summer}} \ell_{b,t}, \quad E_b = 2P_b, \quad (3)$$

Therefore, every building uses a 2-h battery sized to one-quarter of its native summer peak. Let ℓ_t be the native load, u_t be the battery action, g_t be the projected grid import, and s_t be the battery state of charge. The projection enforces

$$g_t \geq 0, \quad s_{\min} \leq s_t \leq s_{\max}, \quad (4)$$

with $s_{\min} = 0$, $s_{\max} = 1$, an initial SoC of 0.5, and one-way charge and discharge efficiencies of 0.96. The same projection was applied to learned policies, rule-based controllers, and oracle settlements. Month-end SoCs are handled by an economic terminal settlement rather than by imposing a hard terminal equality constraint. The degradation proxy used 0.0596 USD/kWh, following convex per-throughput cycle-degradation formulations [9, 34]. Under the signed-discharge convention used in the implementation, the interval degradation energy is $q_t^{\text{deg}} = q_t^{\text{chg}} + q_t^{\text{dis+}}/\eta_{\text{dis}}$, where $q_t^{\text{chg}} \geq 0$ is the charged energy and $q_t^{\text{dis+}} = \max(q_t^{\text{dis}}, 0)$ is the positive discharged-energy portion of the signed discharge q_t^{dis} , with $\eta_{\text{dis}} = 0.96$.

Objective and regret

Each policy is conditioned on a scalar preference weight $w \in [0,1]$. Lower weights emphasize bill reduction, whereas higher weights emphasize degradation avoidance. For building-month episode e , policy π , and preference w , the objective is

$$J^\pi(e, w) = (1 - w)C_{\text{bill}}^\pi(e, w) + wC_{\text{deg}}^\pi(e, w), \quad (5)$$

with lower values indicating better performance. The objective components are monthly USD costs and are not standardized before scalarization; the bounded variant used for regret has the same form, with no-storage degradation fixed at zero.

A no-storage policy π^0 and perfect-foresight oracle π^* define the oracle-relative regret as follows:

$$R^\pi(e, w) = \frac{J^\pi(e, w) - J^*(e, w)}{J^0(e, w) - J^*(e, w) + \epsilon}. \quad (6)$$

Scalarization was performed in monthly USD without pre-scaling the bill and degradation components. To summarize regret and selection, the implementation uses the bounded objective $J_{\text{bd}} = (1 - w)C_{\text{bill}} + wC_{\text{deg}}$ with the no-storage objective defined as $(1 - w)C_{\text{bill}}^0$ as the no-storage degradation component is zero.

The denominator is audited to identify negligible opportunity episodes. The opportunity threshold is $\epsilon = 1$ USD; episodes with $J^0 - J^* \leq \epsilon$ are excluded from normalized-regret aggregation. Dense-grid frontiers are computed from bill-saving and negative-degradation-cost coordinates over $w \in \{0.0, 0.1, \dots, 0.9\}$, where both plotted axes are the-higher-the-better. The dominant-point and monotonicity diagnostics use a fixed tolerance of 1 USD.

The headline scalar used throughout this paper is regime-balanced normalized regret. The dense preference grid is partitioned into four preference regimes—plateau ($w \in \{0.0, 0.1, 0.2, 0.3\}$), knee ($w = 0.4$), balanced ($w \in \{0.5, 0.6, 0.7\}$) and high- w ($w \in \{0.8, 0.9\}$)—so that bill-priority, transition, mid-trade-off, and degradation-priority operating points are represented. Within each regime, the bounded normalized regret is averaged across all opportunity-valid episodes, and the per-regime means are then averaged with equal weights to obtain a regime-balanced scalar. This balanced aggregation prevents the eight non-high- w weights (plateau, knee, and balanced) from dominating the two-weight high- w tail, which would otherwise hide the high- w failure modes explicitly tracked in this study.

Exact incremental demand-charge reward

The training reward uses an exact per-interval decomposition of the monthly demand charge. Let $p_t = \max(g_t, 0)$ be the import demand, $m_t = \max(m_{t-1}, p_t)$ be the running monthly peak. With a monthly reset of $m_0 = 0$, the interval demand-charge increment is

$$c_t^{\text{dem}} = D(m_t - m_{t-1}), \quad (7)$$

where D is the demand-charge rate. As $\sum_t D(m_t - m_{t-1}) = Dm_T$, this is an exact telescoping decomposition of the billing-period peak charge and not an approximate shaping term. The per-step cost used for reinforcement learning is

$$c_t = (1 - w)[\pi_t g_t \Delta t + D(m_t - m_{t-1})] + w c_t^{\text{deg}}, \quad (8)$$

where π_t is the energy price and $\Delta t = 0.25$ h.

Policy observations

The active compact observation vector is the train-only scaled z16 feature set, which contains exactly sixteen features, namely, the normalized load, normalized running monthly peak, normalized load-minus-running-peak, state-of-charge fraction, hour sine, hour cosine, day-of-week sine, day-of-week cosine, month-progress fraction, peak and part-peak tariff indicators, remaining-days fraction, recent 1- and 4-h mean normalized load, normalized native summer peak, and normalized battery power. Continuous features are robust-standardized with training-split medians and interquartile ranges, cyclic and binary features are left unscaled, and scaled continuous values are clipped to $[-8, 8]$. Power features are formed from the native summer peak or battery power ratings before robust standardization. The scalar preference weight w enters the network as a separate six-dimensional preference vector $[w, \log(1 + \lambda), \mathbf{1}_{\text{regime}}]$, where $\lambda = w/(1 - w)$ is the dual ratio and $\mathbf{1}_{\text{regime}}$ is a one-hot indicator over the four preference regimes (plateau, knee, balanced, high- w). For MLP-Cap, the preference vector is concatenated with the z16 observation at the input of the classical backbone. For VQC-Cap and VQC-RegimeCap, only the z16 observation is encoded in the quantum circuit (the data rotations in the VQC card encode the 16 z16 features, not the preference), whereas the preference vector is routed through a classical skip path and concatenated with the VQC measurement output before the classical backbone and demand-cap head. All reported deployable parameter counts include the additional input width contributed by the six-dimensional preference vector on the classical side.

Model families

The hybrid VQC family combines classical pre-processing, variational quantum circuits, and classical readout heads. The active validation finalists are VQC-Cap and VQC-RegimeCap, which are introduced after the earlier direct, residual, and one-hot regime-gated VQC candidates fail to form a usable dense preference-conditioned frontier. The retained VQC block uses four qubits and two data-reuploading layers. Each layer applies RY/RZ data rotations from clipped normalized inputs,

trainable RX/RZ/RZ rotations on each qubit, and ring CNOT entanglement, and then reads out the Pauli-Z and -X expectations. Thus, the quantum block has 24 trainable quantum parameters, an eight-dimensional expectation readout, 48 estimated gates per action, analytic expectations on PennyLane default qubit, and Torch backpropagation through the state-vector simulator following the parameter-shift gradient formulation [35]. This interpretation is resource-aware because variational quantum circuits can be sensitive to circuit expressibility, trainability, barren-plateau effects, and training instability under finite shots or hardware noise [21, 22, 36–38].

The active classical comparator is MLP-Cap, which uses the same observation vector, action projection, oracle target source, fixed-dose training rule, and validation replay protocol as the hybrid family. The widths are chosen to match the target deployable actor-parameter budgets. MLP direct/residual, Fourier-feature, RBF, and earlier VQC variants are retained as validation screening or supplementary evidence, but they are not the main active comparators after the budget-scaling review.

At inference time, all retained families (MLP-Cap, VQC-Cap, and VQC-RegimeCap) apply the same implied-cap isotonic preference-path wrapper: at every action step the policy is queried on the dense preference grid $\{0.0, 0.1, \dots, 0.9\}$, the implied import caps are projected onto the non-decreasing cone in w via the pool-adjacent-violators (PAV) algorithm, and the cap at the requested w is converted back to an action. The wrapper is parameter-free (PAV is recomputed per step from the length-ten cap sequence and stores no fitted state), so the deployable actor-parameter counts reported herein reflect the policy module only; however, the wrapper does multiply the per-action policy forward-pass count by ten and is included in the scalar replay times presented in Table 6.

The demand-cap action head used by MLP-Cap and VQC-Cap predicts a normalized import-demand cap and converts it into a feasible normalized battery action. If $\hat{c}_t$ is the predicted cap, ℓ_t is the native load, and P_b is the battery power rating, the unclipped normalized action is

$$a_t = \frac{\hat{c}_t - \ell_t}{P_b}. \quad (9)$$

It is then clipped to the actor action range before a common battery projection is applied. VQC-RegimeCap uses the same cap-to-action conversion, but interpolates among three demand-cap heads with smooth gates centered on low-, knee-, and high-degradation preference regions. The high-weight sparsity regularizer penalizes non-zero actions on high- w intervals where the oracle target is near zero, such that degradation-avoidant preferences do not inherit unnecessary cycling from the bill-priority behavior.

Training protocol

All active family-budget-seed cells use a fixed-dose behavior-cloning protocol before locked-test access. Target actions are perfect-foresight oracle actions generated on the training split. Oracle actions are used only as supervised targets, whereas all learned policies use causal observations at inference. Training examples are sampled with event-aware stratification such that sparse peak intervals, non-zero oracle battery actions, tariff periods, and ordinary intervals are all represented. The rescue and finalist path trains on the sparse preference grid $w \in \{0.00, 0.30, 0.40, 0.80, 0.90\}$ and evaluates on the dense grid $w \in \{0.0, 0.1, \dots, 0.9\}$.

The current main evidence path did not perform in-training model selection. Each cell was trained for its fixed dose, exported to its final actor checkpoint, and replayed on the validation data. There was no validation-based early stopping, best-checkpoint search, train-loss rollback, or seed selection during training. This choice reduces the validation-driven selection bias in the parameter-performance comparison; late-training instability, budget reversals, and seed variance are reported rather than hidden by a checkpoint search.

The active validation evidence used five seeds during screening and ten seeds for the final full-train/-validation freeze; the locked-test replay used the validation-frozen 100-cell finalist manifest with ten seeds per retained family-budget cell.

Baselines and oracle

The no-storage policy provides a normalization floor; the perfect-foresight oracle provides a diagnostic opportunity ceiling and regret reference but is not a deployable policy. The rule-based peak-shaving baseline uses no future load forecast. It charges toward a reserve target during off-peak periods and discharges during peak or part-peak periods when grid imports exceed a validation-selected demand cap. The cap, reserve target, and charge/discharge thresholds were only selected for the validation split. The validation-tuned rule-based controller, evaluated under the same dense preference grid and bounded-objective regret as the learned policies, achieved a mean normalized regret of 0.953 on the validation split against a no-storage normalization regret of 1.000, i.e., it recovered approximately 5% of the oracle-relative opportunity on average. The corresponding validation-split learned MLP-Cap row at the matched 1.4k budget has a mean regret of 0.5863 (see Supplementary Table S17); therefore, the rule-based controller is far weaker than every retained learned family-budget row

within the validation split used to tune it. Therefore, the rule-based controller is reported as a sanity floor in Table 3 rather than as a competitive baseline, and the locked-test replay is restricted to the learned families and oracle/no-storage references so that the locked-test computation is not spent on an already-eliminated baseline. A stronger model-predictive-control comparator with a fitted load forecast was considered during protocol freeze and was deferred to a separate study.

Statistical analysis

The primary uncertainty intervals used a building-level paired bootstrap with 10,000 resamples and a bootstrap seed of 20,260,519. Buildings were resampled with replacement; all June–September months for the selected buildings were kept together. The same bootstrap sample was applied to every model to preserve paired differences. Seeds were averaged within each family-budget-building unit before the building-level resampling. Herein, the 95% CIs for model-family means, matched-budget paired differences, high-weight operational-burden metrics, and threshold-achievement metrics are reported.

A two-sided sign-flip permutation null with the building as the exchangeability unit was used for the hypothesis tests on matched-budget hybrid-versus-MLP paired differences. For each of the six pre-specified hybrid-versus-MLP paired contrasts (VQC-Cap and VQC-RegimeCap each at 0.5k, 1.4k, and 4k), the test statistic was the mean of the per-building paired difference in regime-balanced normalized regret after averaging the seeds within each family-budget-building unit; the null distribution was obtained by independently flipping the sign of each building’s paired difference with probability 0.5 (10,000 permutations, permutation seed 20,260,519). The two-sided p -value is the fraction of permutations with absolute test statistics at or above the observed value. The family-wise error rate across the six pre-specified contrasts was controlled via the Holm–Bonferroni step-down procedure applied to the six raw permutation p -values, i.e., the Holm-adjusted p -values reported in Table 5. Cohen’s d_z for each contrast is defined as the mean paired difference divided by the standard deviation of the per-building paired differences. Bootstrap and permutation seeds were registered in the locked-test scope document before the S8 replay was executed.

Seed variability is reported separately from the building-level uncertainty. Herein, the validation-selected checkpoints are reported, whereas the full seed ledger, failed runs, validation metrics, and training computation are reported in the Supplementary Information.

Resource accounting

The deployable actor parameters are counted as all stored parameters required for policy inference. Train-only critic parameters, optimizer states, and replay buffers are excluded from the primary count and are reported separately. For VQC policies, the reported resources include qubits, circuit depth, trainable quantum parameters, observables, shots or analytic expectation modes, gradient methods, circuit evaluations per action, and latency. For structured classical feature policies, the stored feature matrices, prototype centers, and fixed random features are counted or reported separately according to the deployment definition.

Data availability

The ComStock 2025 Release 2 (AMY2018) dataset used in this study is publicly available through the Open Energy Data Initiative (OEDI) data lake (<https://data.openei.org/submissions/4520>; DOI: 10.25984/1876417) and the ComStock data viewer (<https://comstock.nlr.gov/page/datasets>), maintained by the National Laboratory of the Rockies (formerly the National Renewable Energy Laboratory). Raw building time-series files were not redistributed. The source code, frozen experimental manifests, evaluation framework, and reproducibility artifacts are publicly available at <https://github.com/meritlab/ess-demand-charge-vqc>. Additional processed data, oracle labels, and replay outputs supporting the findings of this study are available from the corresponding author upon reasonable request.

Code availability

The code for environment construction, oracle solving, behavior-cloning training, replay evaluation, bootstrap analysis, and figure generation is organized under `src/pcqmorl/` (the core package) and `scripts/` (curated run, analysis, and figure-generation scripts). The frozen family-budget and finalist manifests are stored under `artifacts/configs`, and the selected checkpoint paths and per-checkpoint SHA256 hashes are recorded in `artifacts/configs/s8_locked_test_finalist_manifest_v1.csv`. The core package, curated scripts, frozen manifests, and results summaries are released under the MIT license at the public GitHub repository (<https://github.com/meritlab/ess-demand-charge-vqc>).

References

1. Mary, N. & Dessaint, L.-A. Robust model predictive control of battery energy storage with neural network forecasting for peak shaving in university campus. *Journal of Building Engineering* **107**, 112445 (2025).
2. Parray, M. T. & Martinsen, T. Multi-year analysis for optimal behind-the-meter battery storage sizing and scheduling: A Norwegian case study. *Journal of Energy Storage* **110**, 115304 (2025).
3. Liu, Y., Feng, G., Chang, S., Cheng, Y. & Zhang, L. ToU-driven energy storage scheduling for enhancing energy flexibility in a nearly zero-energy building. *Energy* **340**, 139314 (2025).
4. Oh, E. Fair virtual energy storage system operation for smart energy communities. *Sustainability* **14**, 9413 (2022).
5. Weber, L., Bušić, A. & Zhu, J. Reinforcement learning based demand charge minimization using energy storage. in *Proceedings of the 2023 62nd IEEE conference on decision and control (CDC)* 4351–4357 (2023).
6. Cardo-Miota, J., Khadem, S. & Bahloul, M. Deep reinforcement learning based electricity bill minimization strategy for residential prosumer. *Mathematics and Computers in Simulation* **238**, 296–305 (2025).
7. Im, J., Seo, H., Kim, J. & Choi, J. K. Reinforcement learning-based energy management system in the complex electric tariff environment. *International Journal of Electrical Power & Energy Systems* **172**, 111038 (2025).
8. Mussi, M., Pellegrino, L., Pindaro, O. F., Restelli, M. & Trovò, F. A reinforcement learning controller optimizing costs and battery state of health in smart grids. *Journal of Energy Storage* **82**, 110572 (2024).
9. Lee, J.-O. & Kim, Y.-S. Novel battery degradation cost formulation for optimal scheduling of battery energy storage systems. *International Journal of Electrical Power & Energy Systems* **137**, 107795 (2022).
10. Preskill, J. Quantum computing in the NISQ era and beyond. *Quantum* **2**, 79 (2018).
11. Pérez-Salinas, A., Cervera-Lierta, A., Gil-Fuster, E. & Latorre, J. I. Data re-uploading for a universal quantum classifier. *Quantum* **4**, 226 (2020).
12. Sim, S., Johnson, P. D. & Aspuru-Guzik, A. Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms. *Advanced Quantum Technologies* **2**, 1900070 (2019).
13. Lockwood, O. & Si, M. Reinforcement learning with quantum variational circuit. in *Proceedings of the AAAI conference on artificial intelligence and interactive digital entertainment* vol. 16 245–251 (2020).
14. Skolik, A., Jerbi, S. & Dunjko, V. Quantum agents in the gym: A variational quantum algorithm for deep q-learning. *Quantum* **6**, 720 (2022).
15. Sequeira, A., Santos, L. P. & Barbosa, L. S. Policy gradients using variational quantum circuits. *Quantum Machine Intelligence* **5**, 18 (2023).
16. Meyer, N., Scherer, D. D., Plinge, A., Mutschler, C. & Hartmann, M. J. Quantum policy gradient algorithm with optimized action decoding. in *Proceedings of the 40th international conference on machine learning* vol. 202 24592–24613 (2023).
17. Wu, S., Jin, S., Wen, D., Han, D. & Wang, X. Quantum reinforcement learning in continuous action space. *Quantum* **9**, 1660 (2025).
18. Meyer, N. *et al.* Benchmarking quantum reinforcement learning. in *Proceedings of the 42nd international conference on machine learning* vol. 267 43934–43964 (2025).
19. Ajagekar, A. & You, F. Variational quantum circuit based demand response in buildings leveraging a hybrid quantum-classical strategy. *Applied Energy* **364**, 123244 (2024).
20. Ajagekar, A. & You, F. Variational quantum circuit learning-enabled robust optimization for AI data center energy control and decarbonization. *Advances in Applied Energy* **14**, 100179 (2024).
21. Wang, S., Fontana, E., Cerezo, M., Sharma, K. & Cincio, L. Noise-induced barren plateaus in variational quantum algorithms. *Nature Communications* **12**, 6961 (2021).
22. Franz, M. *et al.* Uncovering instabilities in variational-quantum deep Q-networks. *Journal of the Franklin Institute* **360**, 13822–13844 (2023).
23. National Renewable Energy Laboratory. ComStock 2025 release 2, AMY2018 public dataset. (2025).
24. Pacific Gas and Electric Company. Electric schedule b-10: Medium general demand-metered service. (2026).
25. Roijers, D. M., Vamplew, P., Whiteson, S. & Dazeley, R. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research* **48**, 67–113 (2013).
26. Schaul, T., Horgan, D., Gregor, K. & Silver, D. Universal value function approximators. in *Proceedings of the 32nd international conference on machine learning* vol. 37 1312–1320 (2015).
27. Reymond, M., Bargiacchi, E. & Nowé, A. Pareto conditioned networks. in *Proceedings of the 21st international conference on autonomous agents and multiagent systems (AAMAS)* 1110–1118 (2022).
28. Felten, F., Talbi, E.-G. & Danoy, G. Multi-objective reinforcement learning based on decomposition: A taxonomy and framework. *Journal of Artificial Intelligence Research* **79**, 679–723 (2024).
29. Basaklar, T., Gumussoy, S. & Ogras, U. Y. PD-MORL: Preference-driven multi-objective reinforcement learning algorithm. in *Proceedings of the eleventh international conference on learning representations (ICLR)* (2023).
30. Yang, Y., Zhou, T., Pechenizkiy, M. & Fang, M. Preference controllable reinforcement learning with advanced multi-objective optimization. in *ICML 2025 poster* (2025).

31. Kandala, A. *et al.* Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets. *Nature* **549**, 242–246 (2017).
32. Henderson, P. *et al.* Deep reinforcement learning that matters. in *Proceedings of the thirty-second AAAI conference on artificial intelligence (AAAI-18)* 3207–3214 (2018).
33. Jordan, S., Chandak, Y., Cohen, D., Zhang, M. & Thomas, P. Evaluating the performance of reinforcement learning algorithms. in *Proceedings of the 37th international conference on machine learning* vol. 119 4962–4973 (PMLR, 2020).
34. Shi, Y., Xu, B., Tan, Y. & Zhang, B. A convex cycle-based degradation model for battery energy storage planning and operation. *arXiv preprint arXiv:1703.07968* (2017).
35. Schuld, M., Bergholm, V., Gogolin, C., Izaac, J. & Killoran, N. Evaluating analytic gradients on quantum hardware. *Physical Review A* **99**, 032331 (2019).
36. McClean, J. R., Boixo, S., Smelyanskiy, V. N., Babbush, R. & Neven, H. Barren plateaus in quantum neural network training landscapes. *Nature Communications* **9**, 4812 (2018).
37. Cerezo, M., Sone, A., Volkoff, T., Cincio, L. & Coles, P. J. Cost function dependent barren plateaus in shallow parametrized quantum circuits. *Nature Communications* **12**, 1791 (2021).
38. Bergholm, V. *et al.* PennyLane: Automatic differentiation of hybrid quantum-classical computations. *arXiv preprint arXiv:1811.04968*, (2018).

Acknowledgements

The author acknowledges the public ComStock dataset and documentation maintained by the National Laboratory of the Rockies (formerly NREL) and the Open Energy Data Initiative.

Author contributions

E.O. conceived the study, designed the benchmark, implemented the experiments, analyzed the results and wrote the manuscript.

Competing interests

The author declares no competing interests.

Funding declaration

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (Ministry of Science and ICT) (Grant No. RS-2025-16067790).

Additional information

Correspondence and requests for materials should be addressed to E.O.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementaryinformation.pdf](#)