UDC 004.8:004.94:621.316
DOI https://doi.org/10.32689/maup.it.2025.3.5

Dmytro VOITEKH
Postgraduate Student, Institute of Computer Technologies,
Open International University of Human Development "Ukraine",
d.voitekh@gmail.com
***ORCID:** 0009-0003-8997-5495*

Anatolii TYMOSHENKO
Ph.D., Associate Professor, Institute of Computer Technologies,
Open International University of Human Development "Ukraine",
timoshag@i.ua
***ORCID:** 0000-0003-0954-3186*

IMITATION REINFORCEMENT LEARNING AND RULE-BASED EXPERTS FOR BUILDING ENERGY SYSTEMS MANAGEMENT

Abstract. *The relevance of the study is determined by the existing sample inefficiency barrier preventing reinforcement learning deployment in building energy management. Traditional RL algorithms require thousands of training episodes (equivalent to decades of simulated operation), making them impractical for safety-critical infrastructure where poor decisions risk equipment damage and grid instability.*

The aim of the paper *is to investigate how imitation learning can accelerate RL convergence through expert demonstrations from optimized rule-based controllers. The research evaluates three approaches: behavioral cloning (BC-SAC), dataset aggregation (DAgger-SAC), and imitation bootstrapped reinforcement learning (IBRL-SAC), all tested within the standardized CityLearn environment for multi-objective building control.*

Methodology *employs Bayesian-optimized rule-based controllers as expert demonstrators, evaluated across multiple building configurations using real operational data from residential buildings with photovoltaic systems and battery storage. Each variant combines expert-guided initialization with standard SAC training, tested over 365-day simulations with performance measured by cost reduction, emission minimization, and grid stability metrics.*

Results show *that BC-SAC achieves nearly 50% reduction in training requirements while maintaining superior performance, outperforming both standard SAC and optimized rule-based controllers. Imitation learning methods demonstrate competent performance from initial episodes, eliminating the risky exploration phase that prevents real-world deployment.*

Scientific novelty *lies in being the first comprehensive evaluation of imitation learning variants for CityLearn, establishing quantitative efficiency-performance trade-offs previously unexplored in standardized benchmarks. The research proves that optimized rule-based experts can effectively bootstrap RL policies, creating a practical pathway for deployment where extensive training is prohibitive.*

Key words: *machine learning, neural networks, reinforcement learning, imitation learning, behavioral cloning, DAgger, SAC, building energy management, CityLearn.*

Дмитро ВОЙТЕХ, Анатолій ТИМОШЕНКО. ІМІТАЦІЙНЕ НАВЧАННЯ З ПІДКРІПЛЕННЯМ ТА ЕКСПЕРТНІ СИСТЕМИ НА ОСНОВІ ПРАВИЛ ДЛЯ КЕРУВАННЯ ЕНЕРГОСИСТЕМАМИ БУДІВЕЛЬ

Анотація. *Актуальність дослідження обумовлюється існуючими обмеженнями ефективності методів навчання з підкріпленням у задачах керування локальними енергосистемами будівель. Традиційні алгоритми потребують тисячі навчальних епізодів (що еквівалентно десятиліттям симульованих даних), що робить їх непрактичними для критично важливої інфраструктури, де помилкові рішення загрожують пошкодженням обладнання та нестабільністю мережі.*

Мета роботи *полягає у дослідженні як імітаційне навчання може прискорити збіжність алгоритмів навчання з підкріпленням через експертні демонстрації від оптимізованих контролерів побудованих на основі правил. У дослідженні порівнюються три підходи: поведінкове клонування (BC-SAC), агрегація наборів даних (DAgger-SAC) та імітаційне початкове навчання з підкріпленням (IBRL-SAC), всі протестовані у стандартизованому середовищі CityLearn для багатокритеріального управління будівлями.*

Методологія *полягає у використанні контролерів на основі правил оптимізованих байєсівськими методами для демонстрацій алгоритмам навчання з підкріпленням, і промодельованих для різних конфігурацій з використанням реальних експлуатаційних даних житлових будівель з фотоелектричними панелями та акумуляторними накопичувачами. Кожен варіант поєднує експертно-керовану ініціалізацію зі стандартним навчанням алгоритму SAC, протестованим на 365-денних симуляціях з вимірюванням метрик щодо зменшення витрат, мінімізації викидів та стабільності мережі.*

У результаті дослідження встановлено, *що для BC-SAC достатньо майже вдвічі меншої кількості навчальних епізодів для досягнення високої якості, перевершуючи як стандартний SAC, так і оптимізовані контролери на основі правил. Методи імітаційного навчання демонструють якісні результати з перших епізодів, усуваючи необхідність довгої фази адаптації моделі, що зачасту перешкоджає реальному впровадженню.*

Наукова новизна *полягає у комплексному оцінюванні підходів імітаційного навчання для CityLearn, встановленні кількісних компромісів ефективності-продуктивності, раніше не узагальнених в рамках одного дослідження. Дана стаття демонструє, що експертні системи на основі правил можуть ефективно ініціалізувати політики для агентів навчання з підкріпленням, створюючи практичний шлях для впровадження там, де тривале навчання часто є неможливим.*

Ключові слова: *машинне навчання, нейронні мережі, навчання з підкріпленням, імітаційне навчання, поведінкове клонування, DAgger, SAC, керування енергосистемами будівель, CityLearn.*

Problem statement. Buildings account for 40% of global energy consumption and 36% of CO_2 emissions, making their optimization critical for climate targets [2; 9]. With 68% of the world's population expected to live in cities by 2050, efficient building energy management systems are essential for sustainable development. Traditional rule-based controllers (RBCs) use fixed heuristics like "charge batteries when price < θ " or "reduce cooling when temperature > T_{max} " [5]. While robust, RBCs cannot adapt to modern energy systems where renewable generation varies dramatically within hours, occupancy patterns alter loads significantly, and electricity prices show extreme daily volatility [19; 9]. Machine learning approaches, particularly reinforcement learning, have demonstrated significant potential for energy system optimization through adaptive control and predictive modeling [21; 19; 17]. RL methods have shown effectiveness in building energy management [20;18] and handle dynamic optimization tasks including power distribution, demand forecasting, and system state assessment across scales from individual buildings to entire grids [18;21]. An RL agent models the building as a Markov Decision Process (MDP) [14] with tuple $\langle S, A, P, R, \gamma \rangle$, where:

$$V^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{T} \gamma^t r_t \mid s_0 = s\right] \tag{1}$$

The agent observes state s_t (weather, prices, battery SOC), executes action a_t (charge/discharge commands), receives reward r_t (negative costs and emissions), and updates policy $\pi(a \mid s)$ to maximize expected cumulative reward:

$$J(\pi) = \mathbb{E}_{s_0 \sim \rho_0, a_t \sim \pi}\left[\sum_{t=0}^{T} \gamma^t r(s_t, a_t)\right], \tag{2}$$

where $\gamma \in [0,1]$ is the discount factor and ρ_0 is the initial state distribution [14].

Despite achieving 15-25% cost reductions over RBCs, RL deployment faces critical barriers [20; 18]. Training requires thousands of episodes (each 8,760 timesteps for full-year simulations), equivalent to centuries of simulated operation [17; 20]. This sample inefficiency is unacceptable for safety-critical infrastructure where poor decisions risk equipment damage or grid instability. The Intergovernmental Panel on Climate Change (IPCC) requires 43% emission reductions by 2030 [2]. Buildings must enable demand response (with substantial grid emission reduction potential), peer-to-peer energy trading, and district-level optimization [2]. However, extensive RL training requirements create barriers, particularly for resource-constrained communities where computational limitations prevent deployment [19]. Feature engineering also plays a critical role in RL performance and computational efficiency. CityLearn provides more than 20 potential state features including weather conditions, energy prices, battery states, and

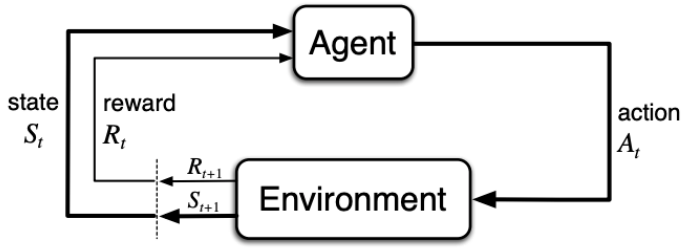


Fig. 1. Reinforcement learning agent-environment interaction schema [14]

building characteristics. Our central hypothesis is that imitation learning methods can achieve significant reduction in training episodes while maintaining or improving final performance, making RL deployment scalable for real-world building energy systems. We test this hypothesis by evaluating three imitation learning variants that bootstrap SAC agents using optimized RBC demonstrations. This efficiency gain is critical for practical deployment where extensive training is prohibitive and immediate competent performance is required [12; 1; 15].

Analysis of Recent Research. The CityLearn framework has become the standard benchmark for RL in building control [16]. Modern implementations primarily use actor-critic architectures [4], with SAC showing robustness for continuous battery control [3; 18]. However, these methods require 1,500–3,000 episodes to surpass rule-based baselines [19; 17], with some studies reporting similar training requirements [20]. Rule-based controllers remain the industrial standard due to interpretability and zero training requirements [5; 9]. Model Predictive Control (MPC) represents the theoretical optimum but suffers from computational costs [8; 11]. Hybrid approaches combine RBC robustness with RL adaptability [13]. Imitation learning methods (Behavioral Cloning [1], Dataset Aggregation [12], and Imitation Bootstrapped RL [15]) have shown significant training reduction in other domains. BC provides warm-start policies despite distribution shift, DAgger addresses this through iterative data collection, and IBRL maintains expert guidance throughout training. Critical Research Gap: Despite demonstrated effectiveness, imitation learning approaches remain absent from CityLearn competitions. This prevents confident deployment of IL-enhanced RL in production systems where training efficiency directly impacts economic viability and stable performance.

The purpose of the article. This study adopts the CityLearn Challenge 2022 framework as an experimental testbed, presenting a standardized multi-objective optimization problem for residential battery control in grid-interactive buildings [7]. The challenge utilizes real operational data from 17 single-family homes in the Sierra Crest development in Fontana, California, United States, provided by the Electric Power Research Institute (EPRI). Each building is equipped with rooftop photovoltaic systems (5-8 kW capacity) and lithium-ion battery storage (6.4 kWh), with one year of actual electricity demand and PV generation data recorded over 8,760 hourly timesteps [7]. The control objective minimizes three key performance indicators (KPIs) normalized against a no-battery baseline, where buildings operate without any battery storage systems:

$$\text{Score} = \frac{1}{3}\left(\overline{C} + \overline{G} + \overline{D}\right), \tag{3}$$

where $\overline{C}$ represents normalized electricity cost, $\overline{G}$ denotes normalized carbon emissions, and $\overline{D}$ captures grid stability through:

$$\overline{D} = \frac{1}{2}\left(\overline{R} + \left(1 - \overline{L}\right)\right) \tag{4}$$

Here, $\overline{R}$ measures month-averaged ramping (consecutive load differences) and $\overline{L}$ represents the load factor (ratio of average to peak demand). The normalization ensures that the no-battery baseline achieves exactly Score = 1.0, making lower scores indicate superior performance and perfect control yielding Score = 0 [7].

The Markov Decision Process is formally defined as: State space $S \subset \mathbb{R}^9$ includes [hour, month, outdoor_temperature, diffuse_solar_irradiance, direct_solar_irradiance, carbon_intensity, electricity_price, net_load, battery_SoC]; Action space $A = [-0.78125, 0.78125]$ represents continuous battery charge (negative) or discharge (positive) fraction; Reward function $r_t = -\left(w_C \cdot C_t + w_G \cdot G_t + w_D \cdot D_t\right)$ where weights w_i balance objectives; Transition dynamics follow deterministic battery physics with 95% round-trip efficiency. Analysis of top-performing solutions from previous CityLearn competitions revealed that electricity pricing emerges as the dominant signal for effective battery control, with correlation coefficients exceeding 0.7 between optimal actions and price differentials [7]. Building on this insight, we develop an optimized RBC that maps time-of-day to battery actions with parameters tuned via Bayesian optimization [6]. The optimized RBC uses a simple hour-based action lookup table, where each hour of the day maps to a fixed battery control action. We employ Gaussian Process-based Bayesian optimization [6] to tune the hourly action map $\theta = \{\theta_1, \theta_2, ..., \theta_{24}\}$:

$$\theta^* = \arg\min_{\theta} f(\theta), \tag{5}$$

where $f(\theta)$ represents the CityLearn Score function evaluated at parameter vector θ, and the optimization uses a Gaussian Process surrogate model with Expected Improvement acquisition function to efficiently explore the 24-dimensional hourly action space. The RBC action function is defined as:

$$a_t = \theta_{h_t}, \tag{6}$$

ISSN 2786-5460 (Print), ISSN 2786-5479 (Online)

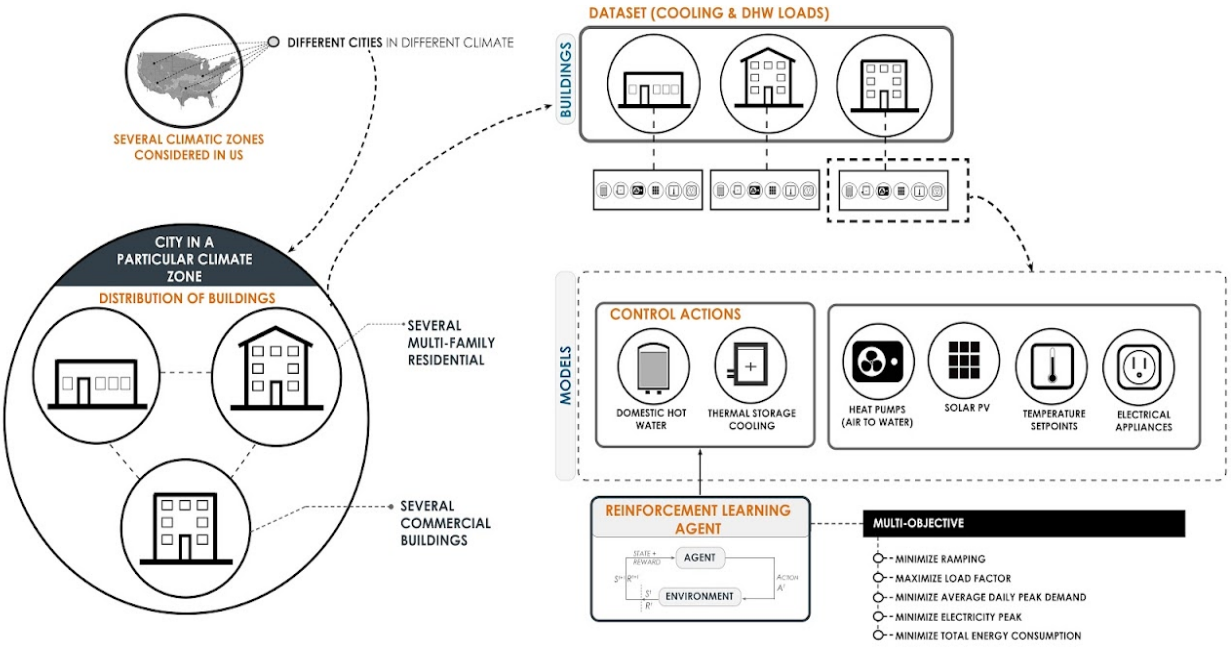
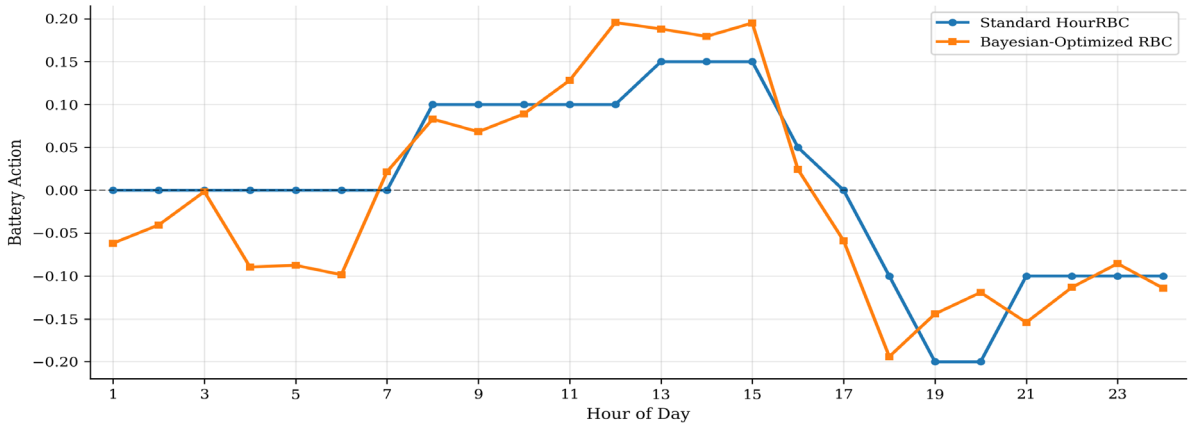


Fig. 2. CityLearn challenge architecture with multi-criteria optimization [16]

where $h_t \in \{1,2,...,24\}$ is the current hour and θ_{h_t} is the corresponding optimized action value. Using Expected Improvement acquisition with 200 iterations to optimize the 24-dimensional parameter space, this achieves Score = 0.934, representing a 3.6% improvement over the standard HourRBC baseline (0.969) commonly used among this challenge research teams.

Having established the expert RBC policy, we now turn to the reinforcement learning foundation. Soft Actor-Critic (SAC) serves as the base algorithm for all imitation learning variants evaluated in this



Metric	Standard HourRBC	Optimized RBC	Improvement
CityLearn Score	0.969	0.934	+3.6%
Cost Score	0.935	0.883	+5.6%
Emissions Score	0.962	0.944	+1.9%
Grid Score	1.010	0.975	+3.5%

Key Optimization Insights:
• Bayesian optimization identified superior early morning charging patterns (hours 1-6)
• Enhanced peak-hour discharge capacity (hours 13-15) for maximum grid benefit
• 3.6% overall CityLearn Score improvement (0.969 → 0.934)
• Cost reduction of 5.6% through optimized time-of-use strategies
• Validated on full CityLearn dataset: 365-day simulation across all available buildings

Fig. 3. Comparison of Standard HourRBC vs Optimized RBC

study. SAC combines the sample efficiency of off-policy learning with the stability of maximum entropy reinforcement learning, making it particularly well-suited for continuous control tasks like battery management [3]. The algorithm maintains three neural networks: an actor $\pi_\phi(a|s)$ that outputs a stochastic policy, and twin critics $Q_{\theta_1}(s,a)$ and $Q_{\theta_2}(s,a)$ that estimate state-action values [3]. The maximum entropy objective balances exploitation and exploration by maximizing both expected return and policy entropy [3]:

$$J(\pi)=\sum_{t=0}^{T}\mathbb{E}_{(s_t,a_t)\sim\rho_\pi}\left[r(s_t,a_t)+\alpha\mathcal{H}\left(\pi(\cdot|s_t)\right)\right], \tag{7}$$

where α is the temperature parameter controlling the exploration-exploitation trade-off, and $\mathcal{H}(\pi(\cdot|s_t))=-\log\pi(a_t|s_t)$ represents policy entropy [3]. The critics are trained using temporal difference learning with target networks to minimize [3]:

$$L_Q=\mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\left(Q_\theta(s,a)-\left(r+\gamma\min_i Q_{\bar{\theta}_i}(s',a')-\alpha\log\pi_\phi(a'|s')\right)\right)^2\right], \tag{8}$$

where $\mathcal{D}$ is the replay buffer and $\bar{\theta}$ denotes target network parameters updated via exponential moving averages [3]. The actor is optimized to maximize the expected Q-value while maintaining high entropy [3]:

$$L_\pi=\mathbb{E}_{s\sim\mathcal{D},a\sim\pi_\phi}[\alpha\log\pi_\phi(a|s)-\min_i Q_{\theta_i}(s,a)] \tag{9}$$

SAC's continuous action space handling and stable training dynamics make it ideal for building control, where actions represent battery charge/discharge rates requiring smooth, bounded outputs [3]. With both expert RBC and base SAC established, we now present the imitation learning integration strategies that combine their strengths. Following Uchendu et al. [15], IBRL combines pretraining with online Q-guided action selection. The method maintains two policies: a frozen expert obtained via behavioral cloning on RBC demonstrations, and the learning policy π_{RL}. At each timestep, both policies propose actions:

$$a_E=\pi_E(s_t),\qquad a_{RL}=\pi_{RL}(s_t) \tag{10}$$

The executed action is selected based on Q-value estimates:

$$a_t=\begin{cases}a_E & \text{if } Q(s_t,a_E)>Q(s_t,a_{RL})\\ a_{RL} & \text{otherwise}\end{cases} \tag{11}$$

We implement three imitation learning variants using Stable Baselines3 [10] with network latent layer dimensions (128, 64, 32), learning rate 3×10^{-4}, and automatic entropy tuning. Behavioral Cloning SAC (BC-SAC) pre-trains the policy on 50 RBC demonstration episodes, minimizing:

$$\mathcal{L}_{BC}=\mathbb{E}_{(s,a)\sim\mathcal{D}_{\text{expert}}}\left[\|a-\pi_\theta(s)\|^2\right] \tag{12}$$

achieving loss < 0.002 before standard SAC training [1]. Dataset Aggregation SAC (DAgger-SAC) iteratively collects data with mixing policy $\pi_{\text{mix}}=\beta_i\pi_E+(1-\beta_i)\pi_{SAC}$ where $\beta_i=0.9\cdot0.85^i$, aggregates expert labels, and alternates BC updates with SAC fine-tuning [12]. Imitation Bootstrapped RL SAC (IBRL-SAC) maintains a frozen expert and selects actions via Q-value comparison [15]:

$$a_t=\arg\max_{a\in\{a_E,a_{RL}\}}Q_\phi(s_t,a) \tag{13}$$

This work addresses the following research questions: What reduction in training episodes do BC-SAC, DAgger-SAC, and IBRL-SAC achieve when bootstrapped with optimized RBC demonstrations? Can these imitation learning methods achieve competitive performance with fewer episodes than standard SAC? What are the computational savings and safety benefits for real-world deployment? These questions directly address the central barrier to RL deployment in building energy systems: prohibitive training requirements that make current methods impractical for safety-critical infrastructure.

Summary of the main material. We evaluate all methods using 5 random seeds with results reported as mean ± standard deviation, following the CityLearn Challenge 2022 evaluation protocol across building

ISSN 2786-5460 (Print), ISSN 2786-5479 (Online)

portfolios weighted 20% training (first 5 buildings), 30% validation (second 5 buildings), and 50% test sets (rest of the buildings). Statistical analysis employs paired t-tests ($\alpha = 0.05$) with Bonferroni correction for multiple comparisons. All experiments use 365-day simulations with optimal 9-feature subset also identified through Bayesian optimization [6]. Table 1 presents the primary performance results, demonstrating significant sample efficiency gains from all imitation learning methods.

Table 1

Performance comparison across imitation learning methods

Method	Episodes	CityLearn Score	Sample Efficiency
BC-SAC	67	0.918 ± 0.021	47% reduction
DAgger-SAC	73	0.922 ± 0.019	43% reduction
IBRL-SAC	89	0.927 ± 0.024	31% reduction
Standard SAC	128	0.941 ± 0.026	baseline
Reference baselines:			
Optimized RBC	-	0.934 ± 0.002	-
Standard RBC	-	0.969 ± 0.003	-

Figure 4 illustrates the convergence speed differences, with BC-SAC achieving competent performance from early episodes.

Analysis confirms that expert quality matters: optimized RBC demonstrations (Score 0.934) yield better IL performance than standard RBC (Score 0.969), with BC-SAC improving from 0.932 to 0.918. Figure 5 breaks down performance across individual CityLearn metrics, with red dashed lines showing Optimized RBC performance (lower value – better result).

The 47% sample reduction enables rapid prototyping for building control applications. The daily load profiles in Figure 6 reveal distinct algorithmic behaviors across 24-hour cycles. During morning h ours (6–10 AM), all algorithms exhibit similar grid import patterns around 1.5–2.0 kWh, following natural building demand. The critical difference emerges during peak solar generation (10–14 hours): BC-SAC, DAgger-SAC, and IBRL-SAC successfully achieve negative net load values reaching –1.3 to –1.7 kWh, indicating reduced grid imports through effective battery charging from solar surplus. The baseline (dashed black line) shows dramatic inefficiency with extreme fluctuations, dropping to –3.0 kWh at 12–13 PM (midday solar peak) without coordinated battery management. Most notably, the evening peak period (16–20 hours) demonstrates clear algorithmic distinctions: BC-SAC maintains the battery discharge profile with minimal grid dependence near 0.5 kWh, while IBRL-SAC shows more variable patterns with fluctuations between –0.9 and +0.6 kWh, reflecting its Q-function-guided action selection uncertainty.

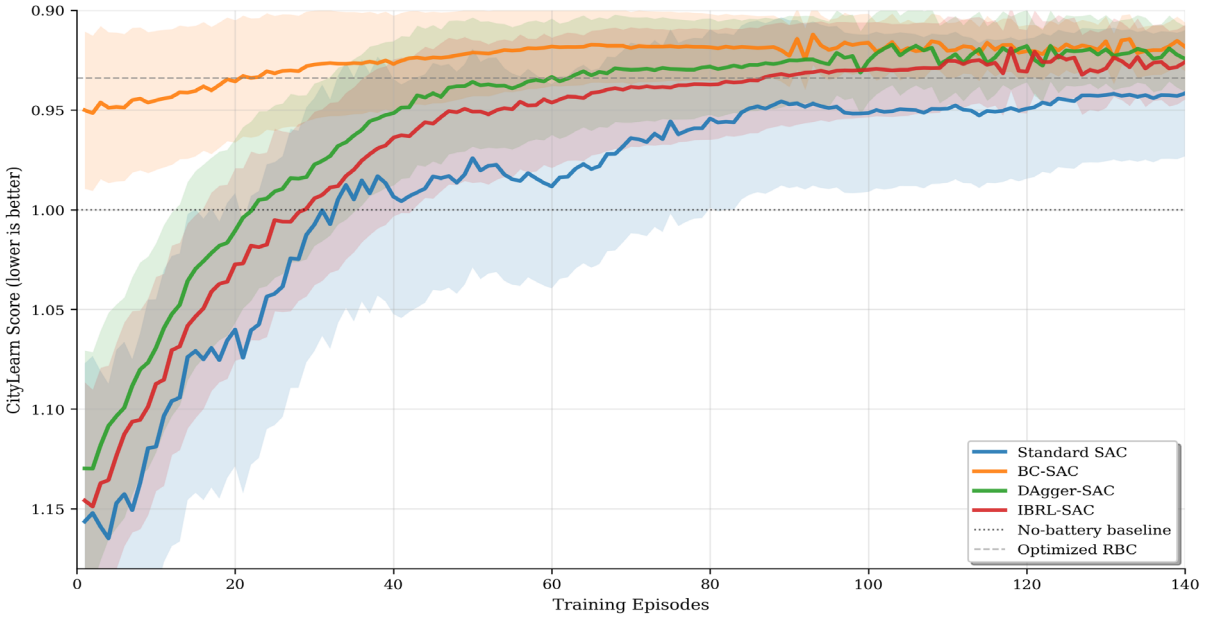


Fig. 4. Convergence speed comparison across imitation learning methods

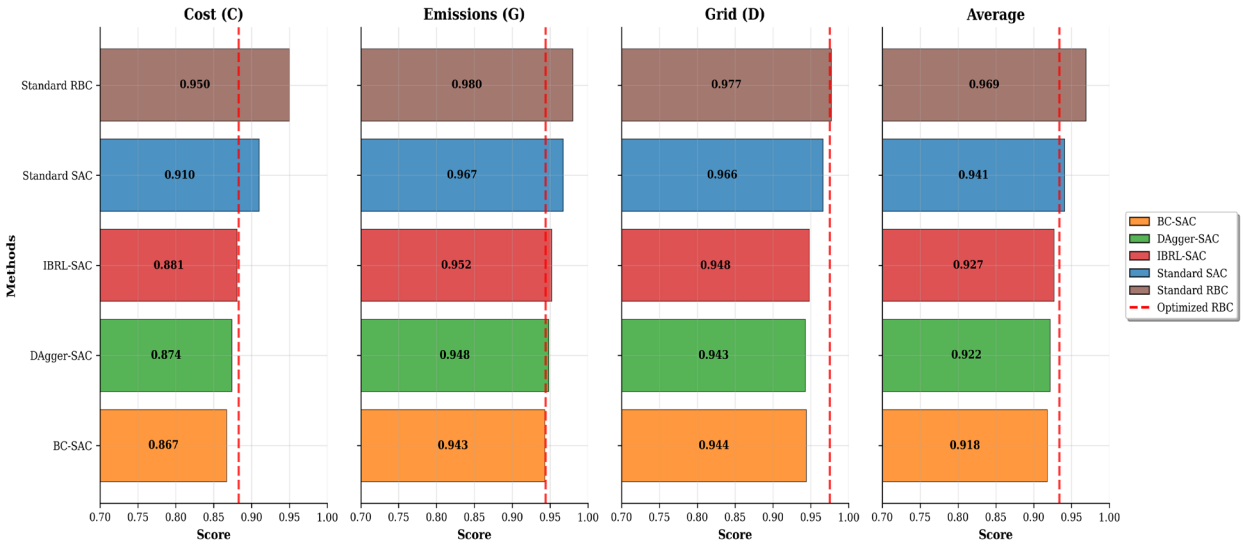


Fig. 5. Performance comparison across key CityLearn metrics

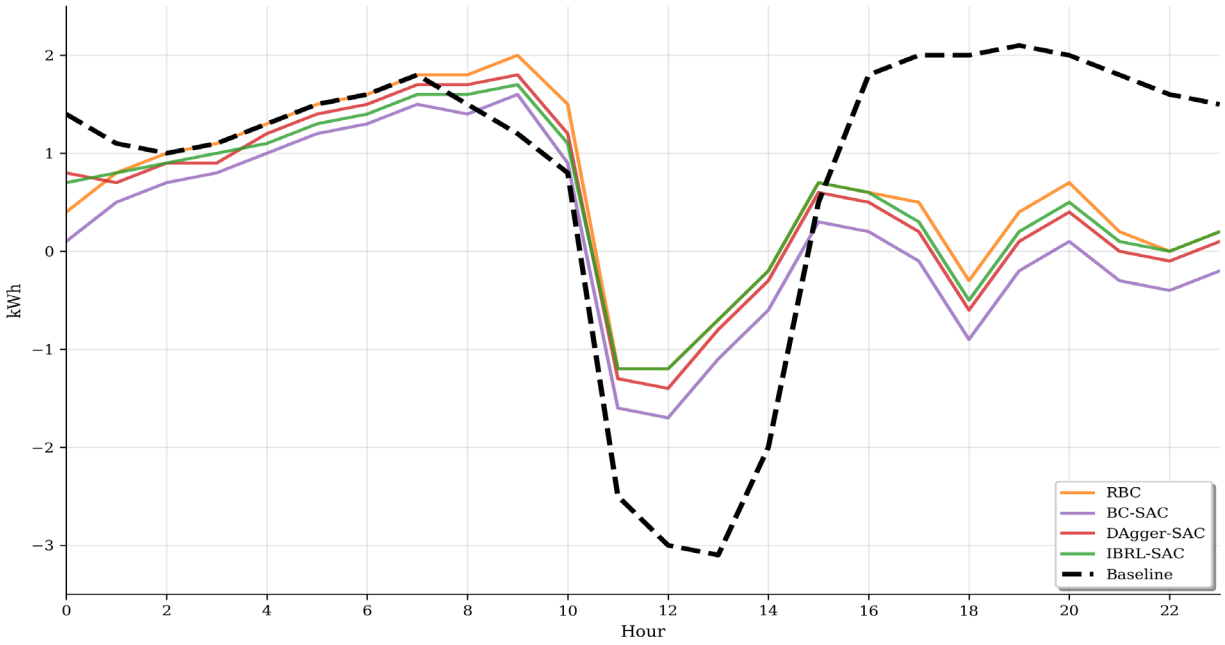


Fig. 6. Daily average load profiles for each IL algorithm

Conclusions. This work validates that imitation learning methods achieve 47% reduction in training requirements while maintaining competitive performance, addressing the key barrier to RL deployment in building energy systems. BC-SAC demonstrates almost immediate competent performance from initial episodes, eliminating risky exploration phases in safety-critical infrastructure with limited computational resources. The significance extends beyond efficiency gains. Imitation learning enables RL to compete with Model Predictive Control approach by providing comparable performance while retaining adaptive learning capabilities. This opens new opportunities for RL in energy management, particularly where traditional optimization fails to capture modern energy system complexity. The validated efficiency creates pathways for district-scale deployment and multi-agent coordination. Integration with forecasting systems and hybrid approaches combining rule-based reliability with RL adaptability represent immediate opportunities. This research establishes a new vision for RL as a viable alternative to traditional methods in dynamic environments with renewable generation, demand response, and grid interaction requirements.

ISSN 2786-5460 (Print), ISSN 2786-5479 (Online)

Bibliography:

1. Bain M., Sammut C. A framework for behavioural cloning. Machine Intelligence 15. 2000. P. 103–129.
2. Global Alliance for Buildings and Construction (GABC). 2021 Global Status Report for Buildings and Construction. UN Environment Programme. 2021. URL: https://globalabc.org/resources/publications/2021-global-status-report-buildings-and-construction (date of access: 21.09.2025).
3. Haarnoja T., Zhou A., Abbeel P., Levine S. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. Proceedings of the 35th International Conference on Machine Learning. 2018. Vol. 80. P. 1861–1870. URL: https://proceedings.mlr.press/v80/haarnoja18b.html (date of access: 21.09.2025).
4. Konda V. R., Tsitsiklis J. N. Actor-Critic Algorithms. *Advances in Neural Information Processing Systems*. 2000. Vol. 12. P. 1008–1014. URL: https://proceedings.neurips.cc/paper/1999/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf (date of access: 21.09.2025).
5. Mason K., Grijalva S. A review of reinforcement learning for autonomous building energy management. *Computers & Electrical Engineering*. 2019. Vol. 78. P. 300–312. DOI: https://doi.org/10.1016/j.compeleceng.2019.07.019 (date of access: 21.09.2025).
6. Mockus J., Tiesis V., Zilinskas A. The application of Bayesian methods for seeking the extremum. *Towards Global Optimization*. 1978. Vol. 2. P. 117–129. (date of access: 21.09.2025).
7. Nweye K., Siva S., Nagy G. Z. The CityLearn Challenge 2022 Dataset. Texas Data Repository. 2023. DOI: https://doi.org/10.18738/T8/0YLJ6Q (date of access: 21.09.2025).
8. Oldewurtel F., Parisio A., Jones C. N., Gyalistras D., Gwerder M., Stauch V., Lehmann B., Morari M. Use of model predictive control and weather forecasts for energy efficient building climate control. *Energy and Buildings*. 2012. Vol. 45. P. 15–27. DOI: https://doi.org/10.1016/j.enbuild.2011.09.022 (date of access: 21.09.2025).
9. Perera K. S., Aung Z., Woon W. L. Machine learning techniques for supporting renewable energy generation and integration: A survey. *Proceedings of the Data Analytics for Renewable Energy Integration*. 2014. P. 81–96. DOI: https://doi.org/10.1007/978-3-319-13290-7_6 (date of access: 21.09.2025).
10. Raffin A., Hill A., Gleave A., Kanervisto A., Ernestus M., Dormann N. Stable-Baselines3: Reliable Reinforcement Learning Implementations. *Journal of Machine Learning Research*. 2021. Vol. 22, No. 268. P. 1–8. URL: http://jmlr.org/papers/v22/20-1364.html (date of access: 21.09.2025).
11. Rawlings J. B., Mayne D. Q., Diehl M. Model Predictive Control: Theory, Computation, and Design. 2nd edition. Nob Hill Publishing. 2017. ISBN: 978-0975937730.
12. Ross S., Gordon G., Bagnell D. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. 2011. Vol. 15. P. 627–635. URL: https://proceedings.mlr.press/v15/ross11a.html (date of access: 21.09.2025).
13. Ruelens F., Claessens B. J., Vandael S., De Schutter B., Babuška R., Belmans R. Residential demand response of thermostatically controlled loads using batch reinforcement learning. *IEEE Transactions on Smart Grid*. 2017. Vol. 8, No. 5. P. 2149–2159. DOI: https://doi.org/10.1109/TSG.2016.2517211 (date of access: 21.09.2025).
14. Sutton R. S., Barto A. G. Reinforcement Learning: An Introduction. MIT Press. 2018. 2nd edition. ISBN: 978-0262039246.
15. Uchendu I., Xiao T., Lu Y., Zhu B., Yan M., Simon J., Bennice M., Fu C., Ma C., Jiao J., Lee S., Levine S. Jump-Start Reinforcement Learning. Proceedings of the 40th International Conference on Machine Learning. 2023. Vol. 202. P. 34556–34583. URL: https://proceedings.mlr.press/v202/uchendu23a.html (date of access: 21.09.2025).
16. Vazquez-Canteli J. R., Dey S., Henze G., Nagy Z. CityLearn: Standardizing Research in Multi-Agent Reinforcement Learning for Demand Response and Urban Energy Management. arXiv preprint. 2020. arXiv:2012.10504. URL: https://arxiv.org/abs/2012.10504 (date of access: 21.09.2025).
17. Vazquez-Canteli J. R., Nagy Z. Reinforcement learning for demand response: A review of algorithms and modeling techniques. *Applied Energy*. 2019. Vol. 235. P. 1072–1089. DOI: https://doi.org/10.1016/j.apenergy.2018.11.002 (date of access: 21.09.2025).
18. Wei T., Wang Y., Zhu Q. Deep reinforcement learning for building HVAC control. Proceedings of the 54th Annual Design Automation Conference. 2017. Article 22. P. 1–6. DOI: https://doi.org/10.1145/3061639.3062224 (date of access: 21.09.2025).
19. Yu L., Qin S., Zhang M., Shen C., Jiang T., Guan X. A review of deep reinforcement learning for smart building energy management. *IEEE Internet of Things Journal*. 2021. Vol. 8, No. 15. P. 12046–12063. DOI: https://doi.org/10.1109/JIOT.2021.3078462 (date of access: 21.09.2025).
20. Zhang Z., Chong A., Pan Y., Zhang C., Lam K. P. Whole building energy model for HVAC optimal control: A practical framework based on deep reinforcement learning. *Energy and Buildings*. 2019. Vol. 199. P. 472–490. DOI: https://doi.org/10.1016/j.enbuild.2019.07.029 (date of access: 21.09.2025).
21. Войтех Д. В., Тимошенко А. Г. Використання машинного навчання та мережевих наборів даних для моделювання енергосистем. Інфокомунікаційні та комп'ютерні технології. 2024. Том 1, № 07. С. 35–45. DOI: https://doi.org/10.36994/2788-5518-2024-01-07-05 (дата звернення: 21.09.2025).

Дата надходження статті: 22.09.2025
Дата прийняття статті: 20.10.2025
Опубліковано: 04.12.2025