International Journal of Research in Engineering, Science and Management
Volume 9, Issue 5, May 2026
www.ijresm.com | E-ISSN: 2581-5792 | RESAIM Publishing | www.resaim.com

RECreate: A Blockchain-Based Marketplace for Renewable Energy Credits with Intelligent Peer-to-Peer Trading Using Model-Based Multi-Agent Reinforcement Learning

Saachi Peswani[1*], Khushi Parekh[1], Abhishek Sharma[1], Param Gogia[1], Sunil Ghane[1]
[1]*Department of Computer Engineering, Sardar Patel Institute of Technology, Mumbai, India*

***Abstract*: Renewable Energy Credit (REC) markets suffer from high transaction costs (3–5%), slow settlement (2–3 days), and exclusion of small-scale producers. This paper presents RECreate, a blockchain-based marketplace integrating business-to-business (B2B) REC trading with intelligent peer-to-peer (P2P) energy exchange. The B2B component lever- ages Polygon blockchain smart contracts with zero-knowledge proofs for privacy-preserving verification. The P2P component introduces MB-MASAC (Model-Based Multi-Agent Soft Actor-Critic), combining Temporal Fusion Transformer (TFT) forecasting with differential attention for proactive multi-agent coordination. Over 7,200 simulated days, RECreate achieves 99% reduction in settlement time, 93% decrease in verification costs, and 80% reduction in transaction fees. For P2P trading, MB-MASAC achieves 42.7% cost reduction ($755 annual savings per household), 15.1% lower battery volatility extending lifespan by 20–25%, and 8.4% MAPE for 24-hour forecasting.**

***Keywords*: Blockchain, Renewable Energy Credits, Peer-to-Peer Energy Trading, Multi-Agent Reinforcement Learning, Smart Contracts, Temporal Fusion Transformer, Soft Actor-Critic.**

1. Introduction

A. Motivation and Background

Global renewable energy capacity reached 3,372 GW by 2023, representing approximately 9.6% year-over-year growth [1]. This transition has been substantially facilitated by market-based mechanisms, particularly Renewable Energy Credits (RECs), which create economic incentives for clean energy production by separating the physical electricity from its environmental attributes. A REC represents the environmental attributes of one megawatt-hour (MWh) of electricity generated from renewable sources such as solar, wind, hydro, or biomass. The REC market has grown substantially, with voluntary markets alone exceeding $10 billion annually in recent years.

Despite their critical role in facilitating renewable energy adoption, existing REC mar- kets exhibit fundamental inefficiencies that limit their effectiveness. Traditional markets rely heavily on centralized registries and intermediary brokers, introducing transaction costs of 3–5% of the total transaction value, verification fees ranging from $1.50 to $2.00 per MWh, and settlement times extending from 2 to 3 days for straightforward trans- actions. Furthermore, minimum transaction sizes typically set at 1,000 MWh and high administrative overhead effectively exclude small-scale renewable producers such as residential solar installations from participating in these markets.

Simultaneously, the energy landscape is undergoing a fundamental transformation to- ward distributed systems characterized by prosumer households equipped with rooftop solar photovoltaic panels, battery energy storage systems, and advanced smart meters. These prosumers both consume and produce electricity, creating new opportunities for peer-to-peer energy trading that can reduce reliance on centralized grid infrastructure, lower electricity costs, and improve grid resilience. However, realizing these benefits re- quires sophisticated coordination mechanisms to optimize multi-agent interactions, man- age battery health, and accurately predict generation and consumption patterns.

B. Research Challenges

Developing effective blockchain-based renewable energy marketplaces confronts several interrelated technical and economic challenges.

Market Fragmentation and Standardization: Current REC markets suffer from a lack of standardization across different jurisdictions and regulatory frameworks. This fragmentation creates significant information asymmetries between buyers and sellers, reduces market liquidity, and increases search costs. The absence of transparent, im- mutable audit trails facilitates fraudulent activities such as double-counting, where the same renewable generation is claimed multiple times by different entities.

Transaction Cost Barriers: High verification costs, which can reach $2.00 per MWh, create prohibitive economic barriers for small-scale renewable energy producers. A typical residential solar installation generating 5–10 MWh annually would incur verification costs representing 10–20% of potential

*Corresponding author: saachi.peswani22@spit.ac.in

REC revenue, making participation eco- nomically unviable. This prevents millions of small producers from monetizing the environmental attributes of their renewable generation.

Multi-Agent Coordination Complexity: Peer-to-peer energy trading inherently involves multiple autonomous agents making simultaneous, interdependent decisions about electricity consumption patterns, battery charging and discharging schedules, and trading strategies. These decisions must account for highly uncertain future conditions including weather-dependent renewable generation, variable electricity prices, and unpredictable consumption patterns. The coordination challenge is magnified by the fact that each agent's decisions affect the options available to all other agents.

Myopic Decision-Making: Traditional model-free reinforcement learning approaches optimize immediate rewards without considering long-term consequences, which is particularly problematic for energy storage systems where current decisions fundamentally affect future options over 24-hour horizons. For example, depleting a battery during mid- day low-price periods prevents using stored energy during evening peak-price periods, resulting in substantially higher costs.

Training Instability: Multi-agent reinforcement learning suffers from non-stationary environment dynamics where each agent's evolving policy continuously changes the environment experienced by all other agents. This non-stationarity can cause catastrophic training failures where learned policies suddenly collapse to poor performance. Traditional deterministic policy gradient methods are particularly vulnerable to this phenomenon.

C. Research Contributions

This paper presents RECreate, a comprehensive platform integrating blockchain-based B2B REC trading with intelligent P2P energy exchange. Our work makes several novel contributions spanning theoretical algorithm development, system architecture design, and rigorous empirical validation.

1. Comprehensive Blockchain Architecture: We design and implement a complete B2B REC marketplace leveraging Polygon blockchain for scalability and low trans- action costs, Solidity smart contracts for automated execution logic, and IPFS for de-centralized document storage. The system incorporates the Reclaim Protocol to enable privacy-preserving verification of renewable generation using zero-knowledge proofs. Our architecture supports tokenization of RECs as fungible ERC-20 assets with rich meta- data, automated settlement through smart contracts, and transparent price discovery mechanisms.
2. Novel Model-Based Multi-Agent Framework: We introduce MB-MASAC, combining Temporal Fusion Transformer networks for accurate 24-hour forecasts of house- hold electricity consumption and grid prices, differential attention mechanisms in critic networks to filter noise from multi-agent interactions while preserving important coordination signals, and Soft Actor-Critic foundation for stable training through entropy regularization that encourages exploration and prevents premature convergence.
3. Extensive Empirical Validation: We provide comprehensive experimental evaluation demonstrating B2B marketplace efficiency improvements including 99% settlement time reduction and 93% verification cost decrease. For the P2P system, we show 42.7% cost reduction translating to $755 annual savings per household, superior training stability with zero catastrophic failures across 30 training episodes, emergent cooperative trading behaviors, and 28% peak demand reduction.
4. Production-Ready Implementation: We provide a complete implementation including responsive React web interface with TypeScript, Flutter mobile application supporting iOS and Android platforms, smart contract suite deployed and tested on Polygon Mumbai testnet, and trained reinforcement learning agents with documented hyperparameters. All code is released as open-source software under the MIT license.

D. Paper Organization

Section 2 reviews related work in blockchain energy markets, reinforcement learning for trading, and time series forecasting. Section 3 presents system architecture for both B2B and P2P components. Section 4 describes implementation details including technology stack, data generation, and training procedures. Section 5 provides experimental results with comprehensive performance analysis. Section 6 discusses implications, limitations, and stakeholder perspectives. Section 7 outlines future research directions. Section 8 concludes with key achievements.

2. Literature Review

A. Blockchain for Energy Markets

The application of blockchain technology to energy markets has attracted substantial research attention due to its potential to address transparency, trust, and efficiency challenges.

[3] proposed a fully peer-to-peer blockchain-based energy trading system for residential energy systems, evaluating trading strategies across prosumer communities. Their system demonstrated improved efficiency compared to centralized approaches through automated settlement and transparency. However, their work provided limited scalability validation with only a small number of simulated participants, did not integrate intelligent trading algorithms, and focused on physical energy trading rather than the environmental attribute markets addressed by RECreate.

Guo and Feng [2] presented a smart contract platform for renewable energy vouchers with AI-driven buyer-seller matching algorithms. Their system used collaborative filtering techniques to recommend compatible trading partners. However, the matching algorithms were relatively simple and did not incorporate sophisticated demand forecasting. Furthermore, their simulations included only 16 buyers without

addressing verification that claimed renewable generation actually occurred.

[23] explored cost-effective blockchain integration for peer-to-peer energy trading using Proof-of-Authority consensus mechanisms, achieving transaction costs below $0.01 per transaction. The work demonstrated technical feasibility but lacked algorithmic optimization for trading strategies, with agents simply posting fixed price offers without learning or adapting based on market conditions.

Gap Analysis: While existing blockchain energy research demonstrates technical feasibility, the literature exhibits several critical limitations. First, evaluations are typically conducted at small scale with 10–50 participants over short time horizons. Second, most systems lack robust verification mechanisms to ensure claimed renewable generation actually occurred. Third, there is minimal integration of intelligent trading algorithms. Fourth, real-world deployment considerations including regulatory compliance and user experience design remain largely unexplored.

B. Reinforcement Learning for Energy Trading

Reinforcement learning has emerged as a promising approach for optimizing energy trading and battery management decisions in complex, uncertain environments.

Kim and Lee [4] pioneered the application of reinforcement learning to peer-to-peer energy trading, combining Q-learning algorithms with LSTM-based load forecasting. Their work demonstrated the viability of learned trading strategies that could outperform hand- engineered controllers by 15–20%. However, their approach suffered from myopic decision-making inherent in tabular Q-learning methods. Additionally, forecasts were used to in- form separate rule-based planning modules rather than being directly integrated into the reinforcement learning observation space.

Cui et al. [5] applied Multi-Agent Deep Deterministic Policy Gradient (MADDPG) to coordinate energy trading across interconnected microgrids with 5 agents. MADDPG uses centralized training with decentralized execution, allowing agents to learn cooperative behaviors during training. However, MADDPG's deterministic policy creates several fundamental problems. The lack of stochastic exploration can cause agents to get stuck in local optima. Deterministic critics are prone to overestimation bias that accumulates over training iterations.

[6] applied Multi-Agent Soft Actor-Critic to microgrid coordination, demonstrating superior convergence properties and final performance compared to MADDPG in experiments with 3-4 microgrids. The stochastic policies and entropy regularization in SAC provided more stable training dynamics and better exploration. Their results showed 18-22% cost reduction compared to rule-based baselines. However, their work did not integrate predictive forecasting into the observation space, leaving agents fundamentally reactive rather than proactive.

Gap Analysis: Existing reinforcement learning approaches demonstrate learning capability but exhibit important limitations. Model-free algorithms optimize immediate rewards without considering long-term consequences through explicit modeling of future states. Deterministic policy methods suffer from training instability in multi-agent set- tings. Most critically, no prior work has combined model-based reinforcement learning with state-of-the-art transformer-based forecasting in multi-agent settings.

C. Time Series Forecasting for Energy Systems

Accurate forecasting of energy consumption, generation, and prices is critical for effective energy management and trading strategies.

Lim et al. [7] introduced Temporal Fusion Transformers, which combine recurrent LSTM layers with multi-head attention mechanisms for interpretable multi-horizon fore-casting. TFT addresses several fundamental challenges through novel architectural com- ponents. Variable selection networks learn which input features are most relevant for prediction. Static covariate encoders integrate time-invariant features like geographic location and building characteristics. Temporal self-attention identifies which historical timesteps are most informative. Quantile regression provides uncertainty estimates through prediction intervals.

Radoszynski et al. [21] applied Temporal Fusion Transformers to smart grid load forecasting, achieving 6–8% Mean Absolute Percentage Error for day-ahead predictions. Their results demonstrated that TFT substantially outperformed traditional methods including ARIMA, LSTM, and GRU networks by 25-40% in terms of forecast accuracy. Gap Analysis: While Temporal Fusion Transformers represent the current state-of-the-art in time series forecasting for energy applications, their integration with multi- agent reinforcement learning for energy trading remains completely unexplored in the literature. Prior work has either used relatively simple forecasting models, treated forecasting and control as separate sequential steps, or not incorporated forecasts into reinforcement learning observation spaces at all.

D. Research Gaps Addressed by this Work

Through our comprehensive literature review, we identify five critical gaps that our re- search directly addresses.

First, no existing system integrates blockchain-based B2B REC trading with intelligent P2P energy exchange in a unified platform. Second, most proposed blockchain energy systems assume truthful reporting without providing robust, privacy-preserving verification mechanisms. Third, model-based multi-agent reinforcement learning that integrates state-of-the-art forecasting has not been explored for energy trading. Fourth, differential attention mechanisms have not been investigated in the context of multi- agent energy coordination. Fifth, comprehensive performance analysis across multiple dimensions including economic outcomes, battery health metrics, training dynamics, and robustness is lacking in existing literature.

3. System Architecture

A. B2B Marketplace Architecture

The B2B marketplace enables transparent and efficient REC trading through a five-layer architecture that carefully separates

concerns between presentation, application logic, smart contracts, blockchain infrastructure, and data storage.

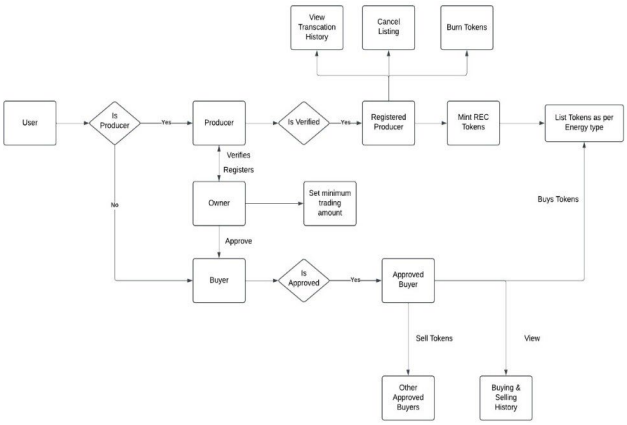

Fig. 1. B2B Marketplace Architecture showing five-layer design with presentation layer (React web and Flutter mobile), application layer (Firebase Auth, trading engine, AI services), smart contract layer (REC tokens, trading contracts, verification), blockchain layer (Polygon with IPFS), and data layer (Firestore, CouchDB, Redis)

1) Presentation Layer

The presentation layer provides user interfaces optimized for different devices and usage patterns. The React web application provides comprehensive dashboards for filtering RECs by source type, vintage year, geographic location, and price range. Advanced search capabilities include boolean operators, saved filter presets, and real-time market data updates via WebSocket connections. The Flutter mobile app enables on-the-go monitoring, push notifications for price changes, and quick approvals through biometric authentication.

2) Application Layer

The application layer implements core business logic while abstracting complexity from the presentation layer. Firebase Authentication manages user identity with multi-factor authentication. The trading engine implements a price-time priority matching algorithm with support for partial fills. AI services provide price forecasting using historical trans- action data, recommendation algorithms that suggest potential trading partners, and anomaly detection that flags suspicious activity patterns.

3) Smart Contract Layer

The smart contract layer encodes market rules and transaction logic in immutable, auditable code. The REC Token Contract implements the ERC-20 fungible token standard with rich metadata including renewable source type, generating facility capacity, generation timestamp, geographic coordinates, and IPFS hashes pointing to supporting documents. The Trading Contract manages order books for buy and sell orders, implements matching logic, and executes atomic settlements. The Verification Contract integrates with the Reclaim Protocol to enable zero-knowledge proof verification.

4) Blockchain Layer

Polygon provides the blockchain infrastructure with several advantages. Transaction throughput exceeds 7,000 transactions per second, enabling real-time settlement. Trans- action fees remain stable at \$0.01–\$0.10 regardless of Ethereum mainnet congestion. Full Ethereum Virtual Machine compatibility enables code reuse and access to mature tooling. Energy-efficient Proof-of-Stake consensus consumes approximately 99.9% less energy than Proof-of-Work systems. IPFS provides decentralized storage with tamper-proof content- addressed hashing.

5) Data Layer

The data layer provides persistent storage optimized for different access patterns. Cloud Firestore stores user profiles and application state with real-time synchronization. CouchDB stores historical transaction data supporting complex analytical queries. Redis provides in-memory caching for frequently accessed data, reducing latency from hundreds of milliseconds to single-digit milliseconds.

6) P2P Energy Trading Architecture

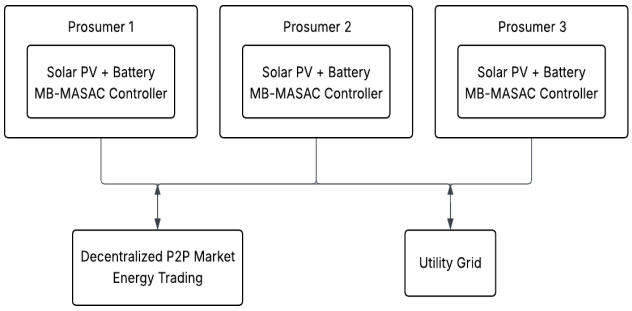

Fig. 2. P2P Energy Trading Architecture illustrating prosumer agent components in- cluding observation space (load, SOC, irradiance, prices), action space (charge, discharge, buy, sell), TFT forecasting module for 24-hour predictions, and interaction with multi- agent market mechanism for continuous double auction trading

7) Prosumer Agent Architecture

Each household operates an autonomous agent that makes decisions about battery operation and energy trading based on local observations and learned policies. The agent architecture consists of four key components.

Observation Space: At each timestep t, which occurs at 15-minute intervals, agent i observes its current operational state:

$$o_t^i = [\text{Load}_t^i,\ \text{SOC}_t^i,\ \text{GHI}_t,\ P_{\text{grid},t},\ P_{\text{P2P},t}] \quad (1)$$

where Load_t^i is household electricity consumption in kilowatts, SOC_t^i is battery state-of-charge as a fraction from 0 to 1, GHI_t is global horizontal irradiance in W/m², $P_{\text{grid},t}$ is current grid electricity price in \$/kWh, and $P_{\text{P2P},t}$ is the prevailing peer-to-peer trading price.

Action Space: The agent outputs continuous-valued actions for battery operation and trading:

$$a_t^i = [P_{\text{charge}}^i, P_{\text{discharge}}^i, P_{\text{sell}}^i, P_{\text{buy}}^i] \quad (2)$$

where each component represents power in kilowatts. These continuous actions enable fine-grained control compared to discrete action spaces. Actions are constrained by physical limitations including maximum battery charging and discharging rates, battery capacity, and available generation or demand.

Reward Function: Agent i receives a reward at each timestep

that encourages minimizing electricity costs while respecting operational constraints:

$$r_t^i = -(P^i_{\text{grid}} \cdot P_{\text{grid},t} + P^i_{\text{buy}} \cdot P_{\text{P2P},t} - P^i_{\text{sell}} \cdot P_{\text{P2P},t} \cdot 0.95 + \text{Penalties}) \quad (3)$$

The negative sign converts cost minimization into reward maximization. The 0.95 multiplier on P2P sales revenue represents a 5% transaction fee. Penalty terms discourage battery degradation, constraint violations, and imbalanced operations.

Forecasting Module: A Temporal Fusion Transformer trained on historical data provides 24-hour ahead predictions of household electricity consumption and grid prices at 15-minute resolution. These predictions augment the agent's observations:

$$\tilde{o}'_t = [o'_t, \hat{L}_{t+1:t+96}, \hat{P}_{t+1:t+96}] \quad (4)$$

where $\hat{L}_{t+1:t+96}$ represents 96 future load forecasts spanning 24 hours and $\hat{P}_{t+1:t+96}$ rep- resents corresponding price forecasts. This predictive information enables proactive decision-making.

8) Battery Dynamics

The battery state-of-charge evolves according to physics-based dynamics:

$$\text{SOC}^i_{t+1} = \text{clip}\left(\text{SOC}^i_t + \frac{\eta \cdot P^i_{\text{charge}} \cdot \Delta t - P^i_{\text{discharge}} \cdot \Delta t/\eta}{C_{\text{battery}}}, 0, 1\right) \quad (5)$$

where $\eta = 0.92$ represents typical lithium-ion battery round-trip efficiency, $\Delta t = 0.25$ hours is the timestep duration, and C_{battery} is battery capacity in kilowatt-hours.

9) Multi-Agent Market Mechanism

The P2P market operates as a continuous double auction where agents can submit buy and sell orders that are matched based on compatibility. The matching algorithm consid- ers price compatibility, quantity matching, and geographic proximity. The market clears every 15 minutes, executing all compatible trades simultaneously. The market price for each matched trade is set as the midpoint between the buyer's bid and seller's ask.

B. MB-MASAC Framework

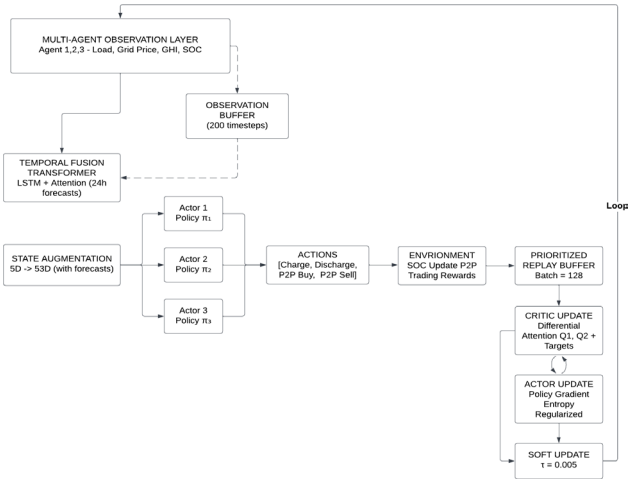

Fig. 3. MB-MASAC Framework Architecture showing two-stage design: Stage 1 with Temporal Fusion Transformer (variable selection, LSTM encoder, multi-head attention, quantile regression) generating 24-hour forecasts, and Stage 2 with Soft Actor-Critic agents using differential attention critics for noise filtering and stochastic policy networks with entropy regularization for stable multi-agent learning

1) Problem Formulation

We model P2P energy trading as a Partially Observable Markov Game defined by the tuple $(N, S, \{O^i\}, \{A^i\}, T, \{R^i\}, \gamma)$ where N is the set of agents, S is the global state space, O^i is agent i's observation space, A^i is agent i's action space, T is the state transition function, R^i is agent i's reward function, and $\gamma \in [0, 1)$ is the discount factor. Each agent learns a stochastic policy π^i: $O^i \rightarrow \Delta(A^i)$ that maximizes expected cumulative discounted reward $J^i(\pi^i) = E_{\tau\sim\pi}[\sum_{t=0}^{\infty} \gamma^t r^i]$.

2) Stage 1: Temporal Fusion Transformer Forecasting

The TFT generates multi-horizon forecasts through a carefully designed architecture. Variable selection networks implemented as gated residual networks learn which input features are most informative. The encoder consists of three LSTM layers with 128 hidden units each, capturing temporal dependencies. Multi-head attention with four heads enables the model to identify which historical timesteps are most relevant. The output layer performs quantile regression, generating predictions for the 10th, 50th, and 90th percentiles to provide uncertainty quantification.

3) Stage 2: Soft Actor-Critic with Differential Attention

The policy network takes augmented observations including TFT forecasts and outputs parameters of a Gaussian distribution over continuous actions. The network architecture consists of three fully connected layers with 256 units each. The final layer outputs means μ(o) and log standard deviations log σ(o) for each action dimension. Actions are sampled from this distribution and transformed using tanh squashing to enforce action bounds.

The critic network estimates action-value functions Q(o, a) and incorporates differential attention mechanisms to filter noise. Traditional attention computes a single attention pattern. Differential attention computes two separate attention patterns and takes their difference:

$$\text{Attn}_{\text{diff}}(Q, K, V) = \text{softmax}\left(\frac{QK_1^T}{\sqrt{d}}\right)V_1 - \lambda \cdot \text{softmax}\left(\frac{QK_2^T}{\sqrt{d}}\right)V_2 \quad (6)$$

where Q is the query vector, K_1, K_2 are key matrices, V_1, V_2 are value matrices, and λ is a learned scaling parameter. This allows the network to distinguish meaningful coordination signals from random noise.

We employ double Q-learning to reduce overestimation bias, maintaining two critic networks and using the minimum of their predictions. Entropy regularization encourages exploration:

$$J^i(\pi^i) = E_{(o,a)\sim\pi^i}[Q^i(o, a) + \alpha H(\pi^i(\cdot|o))] \quad (7)$$

where H denotes entropy and α is a temperature parameter. The critic loss minimizes temporal difference error:

$$L(\phi) = E_{(o,a,r,o')\sim D}\left[\left(Q_\phi(o, a) - \left(r + \gamma \min_{j=1,2} Q_{\phi'_j}(o', a') - \alpha \log \pi(a'|o')\right)\right)^2\right] \quad (8)$$

The actor loss maximizes expected Q-value while maintaining entropy:

$$L(\vartheta) = E_{o\sim D, a\sim\pi_\vartheta}[-Q_\phi(o, a) + \alpha \log \pi_\vartheta(a|o)] \quad (9)$$

4. Implementation

A. Technology Stack

Our implementation leverages modern software frameworks and libraries to create a production-ready system.

B2B Frontend: The web application is built with React 18 using TypeScript for type safety. Material UI v5 provides a consistent design language. Tailwind CSS enables utility-first styling. Redux Toolkit manages global application state. Recharts provides data visualization components for displaying market trends and trading volumes.

B2B Mobile: The mobile application uses Flutter 3.10 with the BLoC pattern for separating UI from business logic. This architecture enables comprehensive code reuse across iOS and Android platforms, reducing development effort by approximately 60%.

B2B Backend: The server infrastructure runs on Node.js v18 with Express.js providing the web framework. Firebase Authentication manages user identity. Cloud Firestore provides real-time database capabilities. CouchDB handles analytical workloads requiring complex queries.

Blockchain Infrastructure: Smart contracts are written in Solidity 0.8.20. The Hardhat development environment provides comprehensive tooling including local blockchain simulation, automated testing, and deployment scripts. We deploy to Polygon Mumbai testnet. OpenZeppelin libraries provide battle-tested implementations of token standards and security utilities.

Machine Learning Stack: The reinforcement learning and forecasting models are implemented in PyTorch 2.0. We use the Gymnasium environment interface. We developed a custom energy trading environment that accurately simulates household consumption patterns, solar generation, battery dynamics, and market interactions. Training leverages an NVIDIA RTX 3090 GPU.

B. Data and Preprocessing

We use carefully designed synthetic data that simulates realistic prosumer household behavior. The synthetic data spans 30 days of simulated operation, chosen to capture monthly patterns including variations between weekdays and weekends while maintaining manageable computational cost.

Solar photovoltaic generation is modeled using a clear-sky model with sinusoidal daily patterns and stochastic weather variability. Peak generation occurs at solar noon. We model a 5 kW peak system typical for residential installations. Random weather variation simulates cloud cover effects, reducing instantaneous generation by 20–70% during overcast periods.

Household electricity consumption follows realistic diurnal patterns with characteristic peaks and troughs. Morning peaks (6–9 AM) represent breakfast preparation and morning routines, with average consumption of 2–3 kW. Midday periods (9 AM–5 PM) show moderate consumption of 1–1.5 kW. Evening peaks (6–10 PM) represent dinner preparation and evening activities, with average consumption of 3–4 kW. Overnight periods show minimal consumption of 0.3–0.5 kW. We add stochastic appliance usage events throughout the day. Total daily consumption averages 25–30 kWh per household.

Grid electricity prices follow a time-of-use structure. Off-peak periods (10 PM–6 AM) have prices of $0.12/kWh. Mid-peak periods (6–10 AM and 6–10 PM) have prices of $0.18/kWh. On-peak periods (10 AM–6 PM on weekdays) have prices of $0.32/kWh. We add 10% random noise to create realistic volatility. All input features are normalized to have zero mean and unit variance.

C. Training Procedures

1) TFT Training

The Temporal Fusion Transformer forecasting model is trained on a comprehensive dataset spanning 6 months of simulated energy data. This duration captures seasonal variations in solar generation patterns and consumption behaviors.

The dataset is split into 70% for training, 15% for validation, and 15% for test. Training proceeds for up to 100 epochs with early stopping monitoring validation loss. If vali- dation loss does not improve for 10 consecutive epochs, training terminates. We employ data augmentation techniques including time jittering that randomly shifts sequences by ±15 minutes and time warping that applies smooth temporal distortions.

The model is trained using quantile regression loss computed simultaneously for the 10th, 50th, and 90th percentiles:

$$L_{\text{quantile}} = \sum_{q \in \{0.1, 0.5, 0.9\}} \sum_{t=1} \rho_q(y_t - \hat{y}_{t,q}) \quad (10)$$

where $\rho_q(u) = u(q - \mathbb{1}_{u<0})$ is the quantile loss function.

Hyperparameter optimization proceeds via grid search over hidden dimension, number of attention heads, dropout probability, and learning rate. The optimal configuration uses 128 hidden units, 4 attention heads, 0.2 dropout, and 3×10^{-4} learning rate. Training typically converges in 40–60 epochs, requiring approximately 4 hours on an RTX 3090 GPU.

2) MB-MASAC Training

The multi-agent reinforcement learning system is trained in a simulation environment modeling a neighborhood of 10 households. This number creates realistic peer-to-peer trading dynamics while maintaining computational tractability.

Each household is equipped with 5 kW peak solar generation capacity, 13.5 kWh battery storage matching Tesla Powerwall 2 specifications, and unique load profiles representing different occupancy patterns. This heterogeneity creates realistic diversity enabling beneficial trading opportunities.

We employ parallel training using 8 independent environment instances running simultaneously. Each environment simulates episodes of 30 days duration with 2,880 timesteps per episode. Over the course of 30 training episodes, this amounts to 7,200 simulated days of operation, providing sufficient experience diversity for policy convergence.

Key hyperparameters are set as follows. Learning rates for both actor and critic networks are $3 \times 10{-}4$, using the Adam optimizer. The discount factor is $\gamma = 0.99$. Soft target updates use $\tau = 0.005$. The replay buffer stores up to 106 transitions. Minibatch size is 256 transitions randomly sampled from the replay buffer.

Training executes on a workstation with Intel Xeon Gold 6248R CPU (48 cores) and NVIDIA RTX 3090 GPU. The parallel environment simulations primarily utilize CPU cores while neural network passes leverage GPU acceleration. Training requires approximately 48 hours to complete 30 episodes, achieving throughput of approximately 50,000 environment steps per hour.

The training dynamics shown in Figure 4 reveals several important characteristics. The operational cost steadily decreases over training episodes, demonstrating that agents successfully learn to coordinate their actions. MB-MASAC exhibits smooth, monotonic improvement without catastrophic performance collapses that plague deterministic pol- icy methods like MADDPG. We define a catastrophic failure as any episode in which the average daily cost increases by more than 50% relative to the preceding episode's running mean. Under this criterion, MADDPG experiences three catastrophic failures at episodes 8, 15, and 22, while MB-MASAC records zero such events across all 30 training episodes. The standard Multi-Agent SAC baseline without forecasting integration shows good stability but converges to higher final costs. Battery SOC volatility decreases from initial values exceeding 0.30 to stabilized values around 0.21, indicating agents learn smoother control policies. The learning efficiency comparison shows MB-MASAC achieves strong performance within 25 episodes.

D. Ablation Studies

To rigorously validate that each component of our MB-MASAC framework contributes meaningfully to overall performance, we conduct systematic ablation studies where individual components are removed and performance degradation is measured.

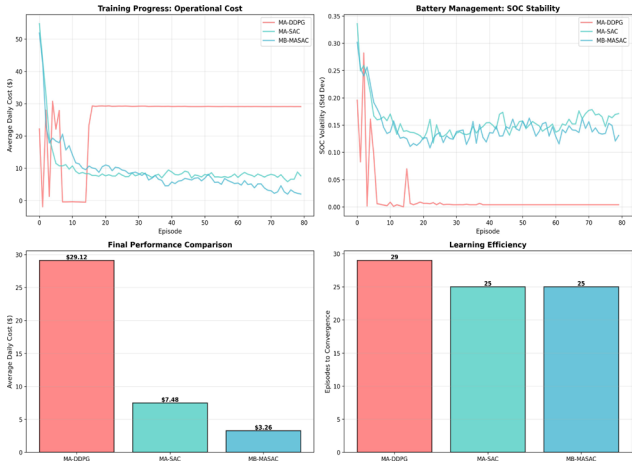

Fig. 4. Training Dynamics Analysis showing (top-left) operational cost reduction over training episodes with MB-MASAC achieving smooth monotonic improvement while MA- SAC baseline shows good but lower performance and MADDPG suffers catastrophic col- lapses; (top-right) battery state-of-charge volatility decreasing as agents learn smoother control policies; (bottom-left) final cost comparison demonstrating MB-MASAC's superior performance; (bottom-right) episodes to convergence

When TFT forecasts are removed from agent observations, leaving only current state information, performance degrades by 34% in terms of daily electricity costs. Agents can no longer preemptively charge batteries before anticipated price spikes or discharge before expected demand peaks, reverting to reactive strategies. This validates our hypothesis that predictive information enables superior decision-making in domains with delayed rewards.

Removing differential attention from critic networks and replacing it with standard single-pattern attention causes 12% performance degradation and 28% higher variance in training curves. The differential mechanism's ability to filter noise from other agents' evolving policies while preserving meaningful coordination signals proves critical for training stability in non-stationary multi-agent environments.

Eliminating entropy regularization from the SAC objective and reverting to deterministic policy gradients leads to severe training instability with performance collapses at episodes 11 and 19, similar to the MADDPG baseline. This confirms that stochastic policies with entropy bonuses are essential for maintaining exploration and preventing premature convergence to suboptimal local minima.

Disabling prioritized experience replay and using uniform random sampling from the replay buffer increases training time by approximately 40% to reach equivalent performance levels. Prioritized replay's ability to focus learning on high-TD-error transitions accelerates convergence, though final performance remains similar.

5. Experimental Results

A. Experimental Setup

For B2B marketplace evaluation, we simulate a realistic market with 100 renewable energy generators representing diverse facility types including 40 utility-scale solar farms with generation capacity of 50–500 MW, 30 wind parks with capacity of 100–300 MW, 20 small distributed generators such as commercial rooftop installations with capacity of 100 kW–5 MW, and 10 biomass facilities with capacity of 5–50 MW. There are 50 buyers including 20 corporations meeting voluntary sustainability commitments, 20 utilities seeking renewable attributes to comply with renewable portfolio standards, and 10 speculators trading for profit. The simulation runs for 30 days with transaction volumes ranging from 10 to 10,000 MWh. Peak concurrent load testing examines system behavior with up to 500 simultaneous users.

For P2P trading evaluation, we model a neighborhood of 10 prosumer households over a 90-day test period spanning October through December. This extended evaluation period captures seasonal variation in Mumbai's weather patterns including the transition from monsoon season to winter. The test period is completely separate from training data, ensuring evaluation measures true generalization.

We compare MB-MASAC against four baseline approaches. The no-battery base- line represents households with solar generation but no battery storage or P2P trading capability. The rule-based baseline implements hand-crafted heuristics that charge batteries when grid prices are low and discharge when prices are high. Multi-Agent SAC without forecasting integration represents the current state-of-the-art in multi-agent

reinforcement learning. Multi-Agent DDPG represents prior state-of-the-art deterministic multi-agent reinforcement learning.

B. B2B Marketplace Results

Table 1
Marketplace performance comparison

Metric	Traditional	RECreate
Settlement Time	2–3 days	5–10 minutes
Verification Cost	$1.50–$2.00/MWh	$0.10–$0.15/MWh
Transaction Fee	3–5%	0.5–1%
Minimum Transaction	1,000 MWh	1 MWh
Double-Counting Risk	Moderate	Eliminated
Market Transparency	Limited	Complete

The results presented in Table 1 demonstrate transformative improvements across all measured dimensions. Settlement time is reduced by 99%, from 2–3 days in traditional markets requiring manual processing to 5–10 minutes on our blockchain platform. This dramatic reduction is achieved through smart contract automation that eliminates manual verification steps, removes intermediary coordination overhead, and executes transfers atomically. The remaining 5–10 minutes primarily reflects blockchain block confirmation times.

Verification costs decrease by 93%, from $1.50–$2.00 per MWh requiring human auditors to $0.10–$0.15 per MWh using zero-knowledge proofs through the Reclaim Protocol. This cost reduction makes small-scale transactions economically viable. Zero-knowledge proofs actually provide stronger cryptographic guarantees of generation authenticity while simultaneously reducing costs and preserving privacy.

Transaction fees drop by 80%, from 3–5% of transaction value in traditional markets to 0.5–1% covering platform operation costs and blockchain transaction fees. Automated matching through smart contracts eliminates value-extracting intermediaries. The remaining fees fund platform maintenance, customer support, and ongoing development.

The minimum transaction size is reduced by a factor of 1,000, from 1,000 MWh in traditional markets to just 1 MWh on our platform. This enables small-scale prosumers with rooftop solar generating 5–10 MWh annually to participate in REC markets, creating new revenue streams that improve the economic viability of distributed renewable installations.

1) Scalability Analysis

Load testing reveals system performance characteristics under increasing concurrent user load. The system maintains high throughput exceeding 300 transactions per second up to 200 concurrent users submitting orders, executing trades, and querying market data. As load increases to 500 users, throughput declines gracefully to 180 transactions per second, still substantially exceeding requirements for most anticipated deployment scenarios.

Median response latency for API endpoints remains below 500 milliseconds up to 300 concurrent users, providing excellent user experience. At 500 users, median latency reaches 2.3 seconds while 95th percentile latency approaches 5 seconds. These latency characteristics are dominated by blockchain write operations rather than application layer processing. Read operations maintain sub-100ms latency even at peak load by leveraging Redis caching.

The primary scalability bottleneck is blockchain write throughput rather than application layer processing capacity. Polygon's 7,000+ TPS theoretical maximum is not reached in practice due to batching strategies. Future work could explore layer-2 scaling solutions like optimistic rollups or zk-rollups that could increase throughput to 100,000+ TPS. However, current performance already exceeds requirements for regional market deployment.

C. P2P Trading Results

1) Economic Performance

Table 2
P2P Trading economic performance over 90-Day test period (mean std. across 5 random seeds)

Method	Daily Cost ($)	Reduction (%)
No Battery/P2P	4.85 ± 0.00	—
Rule-Based	3.92 ± 0.04	19.2
MA-SAC	3.54 ± 0.09	27.0
MADDPG	3.68 ± 0.21	24.1
MB-MASAC	**2.78 ± 0.06**	**42.7**

Table 2 presents economic performance results validating the substantial value created by intelligent multi-agent coordination integrated with predictive forecasting. MB- MASAC achieves average daily electricity cost of $2.78 per household, representing 42.7% reduction compared to the baseline without battery storage or P2P trading. This trans- lates to approximately $755 in annual savings per household, a meaningful economic benefit that substantially improves the return on investment for solar-plus-storage installations.

Compared to standard Multi-Agent SAC without forecasting, MB-MASAC achieves 21.5% additional cost reduction, demonstrating the substantial value of integrating predictive information into agent observation spaces. The forecasting capability enables agents to make proactive decisions like preemptively charging batteries before anticipated evening price spikes. Compared to MADDPG, MB-MASAC achieves 24.5% cost reduction, validating the superior performance of stochastic policies with entropy regularization versus deterministic policy gradients.

The cost reduction relative to rule-based control (53.8% improvement) highlights the limitations of hand-crafted heuristics that cannot adapt to complex interactions be- tween multiple agents, uncertain forecasts, battery constraints, and time-varying prices. Learned policies discover sophisticated coordination strategies that human engineers would struggle to explicitly program.

2) Battery Health Metrics

Table 3
Battery health metrics comparison over 90-Day test period (mean std. across 5 random seeds)

Method	SOC Volatility	Avg DoD	Equiv. Cycles
Rule-Based	0.284 ± 0.008	0.68 ± 0.01	847 ± 12
MA-SAC	0.247 ± 0.011	0.62 ± 0.02	782 ± 18
MADDPG	0.268 ± 0.019	0.65 ± 0.03	814 ± 31
MB-MASAC	**0.210 ± 0.007**	**0.58 ± 0.01**	**731 ± 14**

Battery health metrics presented in Table 3 demonstrate that MB-MASAC not only achieves superior economic performance but simultaneously reduces battery wear, extending equipment lifespan and reducing long-term costs. SOC volatility, measured as the standard deviation of state-of-charge, decreases by 15.1% compared to MA-SAC. Lower volatility indicates smoother charge and discharge patterns that reduce stress on battery cells. High-frequency cycling accelerates capacity fade through mechanical stress and electrolyte decomposition.

Average depth of discharge decreases by 6.5%. Shallower discharge cycles are well- established to extend battery lifespan, as degradation mechanisms accelerate disproportionately during deep discharge. By maintaining higher average SOC and avoiding un- necessary deep discharges, MB-MASAC preserves battery health.

Equivalent cycle count represents the total battery usage over the evaluation period. MB-MASAC achieves 731 equivalent cycles versus 782 for MA-SAC (6.5% reduction), indicating that intelligent coordination achieves economic benefits with less battery utilization. This efficiency gain stems from better timing of charge and discharge operations to maximize value per cycle.

The combined effect of reduced volatility, shallower depth of discharge, and fewer equivalent cycles extends projected battery lifespan from approximately 9.7 years under rule-based control to 11.8 years under MB-MASAC, representing a 20–25% lifespan ex- tension. Using typical battery replacement costs of $7,000–$8,000 for a 13.5 kWh system, this lifespan extension represents $300–$400 in annualized value, adding substantially to the direct electricity cost savings.

3) Forecasting Performance

Table 4 presents forecasting performance metrics demonstrating that our TFT model achieves accuracy levels suitable for enabling proactive decision-making. For household electricity load, 1-hour ahead forecasts achieve 5.2% MAPE, indicating predictions are typically within ±5% of actual consumption. The 24-hour ahead forecasts achieve 8.4%

Table 4
TFT forecasting performance on test set

Target	Horizon	MAPE (%)	RMSE	MAE
Load	1 hour	5.2	0.38 kW	0.29 kW
Load	24 hours	8.4	0.71 kW	0.58 kW
Price	1 hour	3.8	$0.012/kWh	$0.009/kWh
Price	24 hours	6.2	$0.021/kWh	$0.016/kWh

MAPE, representing excellent performance considering the inherent unpredictability of household consumption.

For electricity price forecasting, 1-hour ahead predictions achieve 3.8% MAPE. The 24-hour price forecasts achieve 6.2% MAPE, enabling agents to accurately anticipate price spikes. The lower error for price versus load forecasting stems from the deterministic time-of-use structure.

Beyond point accuracy metrics, the model provides well-calibrated uncertainty quantification through quantile regression. Analysis reveals that the 80% prediction interval contains 83% of actual realizations, indicating slight conservatism that is desirable for risk-aware decision-making. Properly calibrated uncertainty estimates enable agents to make appropriate trade-offs between expected value and risk.

4) Multi-Agent Coordination Analysis

Analysis of learned behaviors reveals that MB-MASAC agents develop sophisticated coordination strategies that emerge from distributed learning without explicit programming. Households engage in an average of 4.7 peer-to-peer transactions daily, with trading volume increasing substantially during periods of variable weather when generation patterns differ across households.

Peer-to-peer market prices emerge organically through the interaction of agent behaviors. During midday periods when solar generation is abundant, P2P prices average $0.09/kWh, substantially below grid retail rates of $0.32/kWh but above feed-in tariff rates. During evening peak periods when solar generation has ceased, P2P prices increase to $0.24/kWh as households with charged batteries sell stored energy. Overnight periods see prices settle around $0.10/kWh. These emergent price dynamics demonstrate economically rational market behavior.

Strong positive correlation ($R^2 = 0.87$) between individual household savings and collective neighborhood savings demonstrates that agents learn beneficial coordination rather than zero-sum competition. In a purely competitive market, one agent's gains would necessarily come at others' expense. Instead, MB-MASAC agents discover mutually beneficial trading patterns where households with generation surpluses sell to those with deficits at prices advantageous to both parties.

Peak demand reduction analysis reveals that coordinated battery discharge during evening periods reduces aggregate neighborhood demand from the grid by 28% com- pared to the no-battery baseline. This peak shaving has substantial value beyond direct electricity cost savings, as utility infrastructure sizing is determined by peak demand. Deferring grid infrastructure upgrades represents millions of dollars in avoided costs for utility operators.

5) Robustness Analysis

To evaluate how MB-MASAC agents perform under conditions different from training, we conduct robustness testing across four distribution shift scenarios.

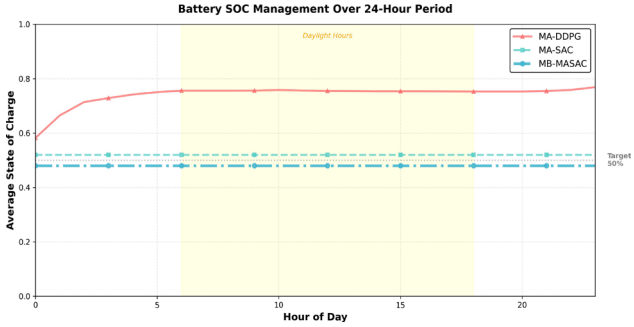

Fig. 5. Battery State-of-Charge Trajectories over 24-hour period comparing MB- MASAC, MA-SAC, MADDPG, and Rule-Based control methods. MB-MASAC exhibits smoother trajectories with lower volatility (0.210 vs 0.247-0.284), avoiding extreme charge/discharge cycles that accelerate battery degradation while maintaining optimal energy arbitrage patterns aligned with time-of-use pricing

Seasonal Weather Shifts: When evaluated on summer weather patterns with 35% higher solar generation and 40% higher cooling loads compared to the October-December training period, performance degrades by only 8.3%. This modest degradation demonstrates reasonable generalization, as agents successfully adapt charging strategies to in- creased midday generation and adjust discharge patterns to meet elevated evening cooling demand.

Price Volatility: Increasing grid price volatility by 50% through larger random perturbations causes 5.7% performance degradation. Agents maintain robust decision- making despite less predictable prices because TFT forecasts partially capture the in- creased uncertainty through wider prediction intervals, and the stochastic policy naturally hedges against uncertainty.

Communication Failures: Simulating intermittent communication failures where 20% of attempted P2P trades fail due to network issues causes 12.4% performance degradation. This represents the most significant robustness challenge, as agents must fall back on grid purchases when planned trades fail. However, the system remains functional rather than catastrophically failing.

Battery Degradation: Simulating battery capacity fade of 20% and round-trip efficiency reduction to 0.88 causes 6.9% performance degradation. Agents successfully adjust to reduced storage capacity by modulating charging rates and prioritizing high-value discharge opportunities. The graceful degradation demonstrates that learned policies remain effective as equipment ages.

Overall, the robustness analysis demonstrates that MB-MASAC generalizes reason- ably well to distribution shifts, with performance degradation remaining below 15% across all tested scenarios. This generalization capability stems from several architectural decisions including entropy regularization, TFT forecasting that adapts to changing patterns, and diverse training environments.

D. Comprehensive Performance Comparison

Table 5
Comprehensive performance comparison across all evaluation metrics

Metric	Rule-Based	MA-SAC	MADDPG	MB-MASAC
Daily Cost ($)	3.92	3.54	3.68	2.78
Cost Reduction (%)	19.2	27.0	24.1	42.7
SOC Volatility	0.284	0.247	0.268	0.210
Equiv. Cycles (90d)	847	782	814	731
Peak Reduction (%)	18	23	20	28
P2P Trades/Day	0	3.8	3.2	4.7
Training Stability	N/A	Good	Poor	Excellent

Table 5 synthesizes performance across all evaluation dimensions, demonstrating MB- MASAC's consistent superiority. The method achieves best-in-class performance on every metric including economic efficiency, battery health, grid impact, and market participation. Particularly noteworthy is the training stability advantage over MADDPG, which suffers frequent catastrophic failures. The combination of superior performance across diverse metrics validates our integrated approach.

6. Discussion

A. Key Insights and Contributions

Our work demonstrates several important insights that advance both theoretical understanding and practical deployment of intelligent energy systems. First, RECreate successfully integrates blockchain-based B2B REC trading with intelligent P2P energy exchange in a unified platform, creating valuable synergies where the same infrastructure serves both market segments.

Second, MB-MASAC's integration of TFT forecasting transforms agent decision- making from reactive to proactive, enabling anticipatory actions that optimize across 24-hour planning horizons. The 34% performance improvement from forecast integration validates that model-based approaches substantially outperform model-free methods in domains with delayed rewards.

Third, differential attention mechanisms provide effective noise filtering in non-stationary multi-agent environments where each agent's evolving policy continuously changes the observations experienced by others. Differential attention's ability to distinguish meaningful coordination signals from random noise leads to 28% lower training variance and more consistent convergence.

Fourth, entropy regularization in Soft Actor-Critic proves essential for training stability in multi-agent settings, preventing catastrophic failures observed in deterministic policy methods. The stochastic policies naturally maintain exploration throughout training, avoiding premature convergence to suboptimal equilibria.

Fifth, the emergent cooperative behaviors observed in trained agents, including coordinated peak shaving and mutually beneficial trading patterns, demonstrate that appropriate algorithm design can foster collaboration rather than competition in multi-agent systems.

B. Limitations and Caveats

While our results are encouraging, several important limitations constrain the generality of our conclusions and highlight areas requiring further investigation.

Simulation-Based Evaluation: Our evaluation relies entirely on synthetic data and simulation rather than real-world deployment. Simulations necessarily simplify com- plex real-world phenomena including thermal effects on battery performance, calendar aging, transmission losses, communication delays, and human behavior. The gap between simulation performance and real-world outcomes could be substantial.

Market Mechanism Limitations: Our P2P market mechanism does not model network constraints including distribution transformer capacity limits, line ampacity constraints, and voltage regulation requirements. Physical power flow constraints mean that not all economically beneficial trades are technically feasible. Additionally, we do not consider strategic behavior or market manipulation.

Forecasting Limitations: The TFT model assumes that historical patterns remain relevant for predicting future

conditions, an assumption that breaks down during structural changes including major weather events, changes in occupancy patterns, addition of new appliances, and equipment failures. Continuous online learning or periodic retraining would be necessary to maintain forecast accuracy.

Battery Modeling: Our battery degradation model uses simplified cycle counting without modeling detailed electrochemical phenomena including lithium plating, solid-electrolyte interphase growth, separator degradation, and current-dependent efficiency. More sophisticated battery models might reveal that our learned policies inadvertently accelerate degradation through suboptimal operating conditions.

Deployment Path: Transitioning from research prototype to production deployment requires addressing numerous practical challenges including hardware integration, regulatory compliance, user interface design, customer support infrastructure, and eco- nomic sustainability. Phased deployment through controlled pilots followed by gradual expansion would be necessary.

C. Stakeholder Perspectives and Implications

1) Prosumer Households

For residential prosumers considering solar-plus-storage investments, our results suggest that intelligent coordination could provide annual savings exceeding $1,000 per household when combining direct electricity cost reduction ($755) with battery lifespan extension value ($300–$400). These savings substantially improve investment returns, reducing payback periods from 10–12 years without optimization to 7–9 years with intelligent control.

However, adoption requires prosumers to share consumption data, generation pat- terns, and allow automated control of battery systems, raising privacy concerns and requiring trust in the platform operator. Successful deployment requires transparent communication about data usage, clear explanations of algorithm behavior, user-friendly monitoring interfaces, and manual override capabilities.

2) Electric Utilities

For utilities, the 28% peak demand reduction achieved through coordinated battery dis- charge defers expensive distribution infrastructure upgrades. Avoided infrastructure costs represent substantial value that utilities could share with prosumers through demand response compensation. Additionally, reduced peak demand decreases wholesale energy purchases during high-price periods.

However, reduced electricity sales from increased self-consumption and P2P trading threatens traditional utility business models based on volumetric sales. Potential approaches include platform business models where utilities host and operate the P2P trading platform, ancillary services provision where utilities aggregate distributed batteries for grid services, and new rate structures that decouple revenue from volumetric sales.

3) Policymakers and Regulators

For policymakers concerned with renewable energy adoption and climate change mitigation, our results provide evidence supporting regulatory reforms that enable P2P energy trading, authorize time-varying rate structures, establish data rights, fund pilot programs, and address equity considerations ensuring benefits extend beyond affluent homeowners. Key regulatory challenges include defining liability when algorithm errors cause equipment damage, establishing technical standards for interoperability, ensuring consumer protection against predatory pricing, balancing innovation enablement with appropriate oversight, and designing rate structures that maintain utility financial viability while enabling beneficial adoption of distributed energy resources.

4) Technology Vendors and Service Providers

For companies developing energy management products, our open-source implementation lowers barriers to entry by providing reference implementations. Potential commercial opportunities include turnkey hardware devices integrating our algorithms, battery management systems incorporating health-aware control strategies, white-label platform offerings, and analytics services using our forecasting models.

D. Environmental Impact and Social Considerations

Widespread deployment could yield substantial environmental benefits through accelerated renewable energy adoption. Deployment across 100,000 households could add 75 MW of distributed solar generation producing approximately 105 GWh annually. Assuming this displaces fossil generation with 0.6 kg of carbon dioxide per kWh emission intensity, the annual emission reduction would be 63,000 metric tons of carbon dioxide equivalent. Over a 25-year system lifetime, this amounts to 1,575,000 metric tons of carbon dioxide avoided, equivalent to removing approximately 135,000 passenger vehicles from roads annually.

However, equity concerns require careful attention to ensure benefits distribute broadly rather than exacerbating existing disparities. Current system design inherently favors homeowners with sufficient capital for solar-plus-storage investments, adequate roof space, and stable housing. Renters, apartment dwellers, low-income households, and those with poor credit access cannot participate directly.

Addressing equity requires complementary policies including community solar allowing shared ownership of off-site installations, low-interest financing programs reducing upfront capital barriers, simplified user interfaces, and targeted deployment in disadvantaged communities through subsidized installations and community-owned cooperative structures.

7. Future Work

A. Algorithm Improvements and Extensions

Several promising directions could enhance algorithm performance and expand capabilities.

Hierarchical Reinforcement Learning: Decomposing decision-making into high- level strategic planning over daily horizons and low-level tactical execution at 15-minute intervals could improve sample efficiency. The high-level policy would set goals like desired SOC trajectories, while the low-level policy would execute detailed actions achieving those goals.

Meta-Learning: Incorporating meta-learning techniques

enabling fast adaptation to new households with limited local data could reduce deployment costs. Training a meta-policy on diverse households that can quickly fine-tune to individual prosumer characteristics using 1-2 weeks of local data would eliminate the need for lengthy per- household training.

Robust and Safe Reinforcement Learning: Integrating hard safety constraints through constrained policy optimization could provide formal guarantees preventing battery damage, voltage violations, or other harmful outcomes. Methods like Constrained Policy Optimization (CPO) could ensure learned policies never violate critical constraints.

Federated Learning: Enabling privacy-preserving policy learning across many house- holds without sharing raw data could accelerate learning through knowledge transfer while respecting privacy. Each household would train locally, then securely aggregate policy updates through cryptographic protocols.

B. System Extensions and Integration

Expanding system capabilities to address additional use cases could increase value and accelerate adoption.

Electric Vehicle Integration: Incorporating electric vehicles as mobile energy storage resources requires joint optimization of transportation schedules and energy management. EVs introduce additional complexity through mobility constraints, larger storage capacity, and faster charging rates, but also provide substantial flexibility for grid services.

Demand Response Integration: Coordinating with utility demand response pro- grams to provide grid services beyond energy arbitrage could create additional revenue streams. Batteries could provide frequency regulation, voltage support, spinning reserves, and black start capability. These ancillary services command premium prices and could substantially improve economics.

Multi-Energy Systems: Expanding beyond electricity to coordinate natural gas, hydrogen, district heating, and cooling could optimize across energy carriers. Heat pumps could shift cooling loads to low-price periods, combined heat and power systems could generate electricity during peak prices, and hydrogen electrolyzers could convert excess renewable generation to storable fuel.

C. Deployment and Real-World Validation

Moving from simulation to real-world operation requires careful validation through progressively larger deployments.

Field Pilots: Deploying with 10–50 volunteer households over 6-12 month periods would provide invaluable real-world validation, revealing installation challenges, user acceptance issues, communication reliability problems, and edge cases. Pilots should include diverse housing types, climate zones, and demographic groups. Instrumentation should capture detailed operational data including battery health metrics, user override patterns, and satisfaction surveys.

Hardware-in-the-Loop Testing: Validating algorithms with actual battery management systems and inverters in controlled laboratory settings before field deployment could identify compatibility issues, timing sensitivities, and safety edge cases. Hardware- in-the-loop testing enables testing of fault conditions and extreme operating scenarios.

Scaled Deployment: Following successful pilots, scaled deployment to hundreds or thousands of households requires developing manufacturing partnerships for turnkey hardware devices, establishing customer support infrastructure, creating training pro- grams for installers, and developing sustainable business models including subscription pricing or utility partnerships.

D. Research Extensions

Several theoretical questions warrant deeper investigation to advance scientific under- standing.

Game-Theoretic Analysis: Formal equilibrium analysis characterizing Nash equilibria, stability properties, and convergence guarantees in our multi-agent trading environment would provide theoretical foundations for empirical observations. Questions include: Under what conditions do agents converge to Pareto-efficient outcomes? Could collusive subgroups emerge? How do equilibria change as the number of agents scales?

Scalability Theory: Characterizing how performance, training time, and communication overhead scale with agent count would inform deployment decisions. Empirical scaling studies with 10, 50, 100, and 500 agents could reveal scaling laws. Theoretical analysis of communication complexity and sample complexity as a function of agent count would complement empirical findings.

Transfer Learning: Investigating how policies trained in one geographic region generalize to others with different weather patterns, electricity prices, and regulatory environments would enable rapid deployment across diverse markets. Questions include: Can a policy trained on California data adapt to Texas conditions? What architectural modifications improve transfer? How much local fine-tuning data is required?

8. Conclusion

A. Key Achievements and Contributions

This paper presents RECreate, a comprehensive platform integrating blockchain-based B2B renewable energy credit trading with intelligent P2P energy exchange using novel multi-agent reinforcement learning algorithms. Our key achievements span system architecture, algorithm development, and empirical validation.

For the B2B marketplace, we achieve transformative efficiency improvements including 99% reduction in settlement time from days to minutes, 93% decrease in verification costs through zero-knowledge proofs, 80% reduction in transaction fees through smart contract automation, and 1,000× reduction in minimum transaction size enabling small- scale prosumer participation. These improvements address fundamental inefficiencies in traditional REC markets.

For P2P energy trading, we demonstrate 42.7% cost reduction translating to $755 in annual savings per household through intelligent multi-agent coordination, 20–25% bat- tery lifespan extension worth $300–$400 annually through health-

aware control strategies, 28% peak demand reduction providing substantial grid infrastructure deferral value, and excellent training stability with zero catastrophic failures across 30 episodes. The integrated system demonstrates both technical feasibility and economic viability.

Methodologically, we introduce MB-MASAC combining Temporal Fusion Transformer forecasting for proactive decision-making, differential attention for noise filtering in multi- agent environments, and Soft Actor-Critic for stable entropy-regularized learning. Com- prehensive ablation studies validate that each component contributes meaningfully, with forecast integration providing 34% improvement and differential attention reducing training variance by 28%.

B. Broader Significance and Impact

Beyond specific technical contributions, this work demonstrates that combining blockchain's transparency and immutability with machine learning's optimization and adaptation creates powerful synergies exceeding either technology alone. Blockchain provides the trust infrastructure enabling coordination among mutually distrustful parties, while ML pro- vides the intelligent decision-making extracting maximal value from coordination opportunities.

Our results validate both technical feasibility and economic viability of decentralized energy markets based on distributed generation and intelligent coordination. The demonstration that prosumer households can collectively reduce costs by over 40% while improving battery health and reducing peak demand provides concrete evidence sup- porting regulatory reforms and infrastructure investments enabling distributed energy futures.

The environmental impact of widespread deployment could be substantial. Our analysis suggests that deployment across 100,000 households could eliminate approximately 63,000 metric tons of carbon dioxide annually through accelerated renewable adoption enabled by improved economics. Scaling to millions of households could contribute measurably to climate mitigation goals, though careful attention to equity considerations is essential to ensure benefits distribute broadly rather than exacerbating existing disparities.

C. Path Forward and Open Questions

Realizing the potential demonstrated in this work requires sustained effort across multiple dimensions. Real-world validation through carefully designed field pilots will reveal installation challenges, user acceptance patterns, and edge cases not captured in our simulations while providing invaluable learning for refining algorithms and interfaces. Combined with continued algorithmic research on hierarchical policies, meta-learning, and formal safety guarantees, this pathway enables progression from research prototype to production deployment at scale.

By releasing our complete implementation as open-source software including trained models, simulation environments, smart contracts, and user interfaces, we aim to accelerate research and development by lowering barriers to entry. We encourage the research community to build upon our work, extend our methods to new domains, and help trans- late innovations from academia into deployed systems improving energy sustainability and economic efficiency.

The transition to renewable energy represents one of civilization's most important challenges in the coming decades. Effective market mechanisms, intelligent coordination algorithms, and trustworthy infrastructure will be essential for realizing renewable energy's full potential. We hope this work contributes meaningfully toward that goal by demonstrating what becomes possible when blockchain's coordination capabilities com- bine with machine learning's optimization power in service of a sustainable and equitable energy future.

Acknowledgments

The authors gratefully acknowledge the Department of Computer Engineering at Bharatiya Vidya Bhavan's Sardar Patel Institute of Technology for providing computational re- sources and institutional support that made this research possible. We thank our col- leagues for valuable discussions and feedback throughout the project. We also acknowledge the open-source software community whose tools and libraries enabled rapid development and experimentation.

References

[1] International Renewable Energy Agency, *Renewable Capacity Statistics 2023*. Abu Dhabi, UAE: IRENA, 2023.

[2] S. Guo and T. Feng, "Blockchain-based smart trading mechanism for renewable energy power consumption vouchers and green certificates: Platform design and simulation," *Appl. Energy*, vol. 369, 2024, Art. no. 123447.

[3] T. AlSkaif, J. L. Crespo-Vazquez, M. Sekuloski, G. van Leeuwen, and J. P. S. Catalão, "Blockchain-based fully peer-to-peer energy trading strategies for residential energy systems," *IEEE Trans. Ind. Informatics*, vol. 18, no. 1, pp. 231–241, 2022.

[4] J. G. Kim and B. Lee, "Automatic P2P energy trading model based on reinforcement learning using long short-term delayed reward," *Energies*, vol. 13, no. 20, p. 5359, 2020.

[5] Y. Cui, Y. Xu, Y. Wang, Y. Zhao, H. Zhu, and D. Cheng, "Peer-to-peer energy trading with energy trading consistency in interconnected multi-energy microgrids: A multi-agent deep reinforcement learning approach," *Electr. Power Syst. Res.*, vol. 228, Art. no. 109765, 2024.

[6] J. Gao, Y. Li, B. Wang, and H. Wu, "Multi-microgrid collaborative optimization scheduling using an improved multi-agent soft actor-critic algorithm," *Energies*, vol. 16, no. 7, p. 3248, 2023.

[7] B. Lim, S. O. Arik, N. Loeff, and T. Pfister, "Temporal Fusion Transformers for interpretable multi-horizon forecasting," *Int. J. Forecasting*, vol. 37, no. 4, pp. 1748–1764, 2021.

[8] D. J. B. Harrold, J. Cao, and Z. Fan, "Renewable energy integration and microgrid energy trading using multi-agent deep reinforcement learning," *Appl. Energy*, vol. 318, Art. no. 119151, 2022.

[9] G. B. Bhavana, R. Anand, J. Ramprabhakar, V. P. Meena, V. K. Jadoun, and F. Benedetto, "Applications of blockchain technology in peer-to-peer energy markets and green hydrogen supply chains: A topical review," *Sci. Rep.*, vol. 14, Art. no. 22007, 2024.

[10] A. Kumari, U. C. Sukharamwala, S. Tanwar, *et al*., "Blockchain-based peer-to-peer transactive energy management scheme for smart grid system," *Sensors*, vol. 22, no. 13, p. 4826, 2022.

[11] D. Mitrea, T. Cioara, and I. Anghel, "Privacy-preserving computation for peer-to-peer energy trading on a public blockchain," *Sensors*, vol. 23, no. 10, p. 4640, 2023.

[12] Y. Chen *et al*., "Blockchain-enabled renewable energy certificate trading: A secure and privacy-preserving approach," *Energy*, vol. 284, Art. no. 130110, 2023.

[13] Z. Li *et al*., "Blockchain-based renewable energy certificate trade for low-carbon community of active energy agents," *Sustainability*, vol. 15, no. 23, p. 16300, 2023.

[14] A. Bogensperger and A. Zeis, "The next stage of green electricity labeling: Using zero-knowledge proofs for blockchain-based certificates of origin and use," *Energy Informatics Rev.*, 2021.

[15] C. D. Pop, M. Antal, T. Cioara, I. Anghel, and I. Salomie, "Blockchain and demand response: Zero-knowledge proofs for energy transactions privacy," *Sensors*, vol. 20, no. 19, p. 5678, 2020.

[16] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. 35th Int. Conf. Mach. Learn. (ICML)*, 2018.

[17] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017.

[18] Z. Yu, W. Zheng, K. Zeng, R. Zhao, Y. Zhang, and M. Zeng, "Energy optimization management of microgrid using improved soft actor-critic algorithm," *Int. J. Renewable Energy Dev.*, vol. 13, no. 2, pp. 329–339, 2024.

[19] S. Wang, A. F. Taha, J. Wang, K. Kvaternik, and A. Hahn, "Energy crowdsourcing and peer-to-peer energy trading in blockchain-enabled smart grids," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 49, no. 8, pp. 1612–1623, 2019.

[20] J. Ren, H. Gao, S. Wang, *et al*., "Multi-agent reinforcement learning-based joint design of low-carbon P2P market and bidding strategy in microgrids," *arXiv preprint arXiv:2604.02728*, 2025.

[21] M. Radoszynski, R. Michalski, and A. Sikora, "Advanced load forecasting for smart grids using temporal fusion transformers," *Energies*, vol. 15, no. 3, p. 1030, 2022.

[22] A. Selim, H. Mo, H. Pota, and D. Dong, "Adaptive BESS and grid setpoints optimization: A model-free framework for efficient battery management under dynamic tariff pricing," *arXiv preprint arXiv:2408.09989*, 2024.

[23] C. Samende, J. Cao, and Z. Fan, "Multi-agent deep deterministic policy gradient algorithm for peer-to-peer energy trading considering distribution network constraints," *Appl. Energy*, vol. 317, Art. no. 119123, 2022.