Journal of Computers Vol. 36 No. 3, June 2025, pp. 241-255
https://doi.org/10.63367/19911599202506360316

The Research on Temperature Control and Microstructure Optimization of Regenerative Heat Treatment Furnace Based on Deep Reinforcement Learning

Yan Lu, Da-Lei Zhang, Dong-Hua Lu, and Hao-Sheng Lu*

Qinhuangdao Vocational and Technical College,
Qinhuangdao City 066100, Hebei Province, China

jdgcx_ly@163.com, zhangdalei2025@163.com,
396726949@qq.com, haosheng89@126.com

Received 20 May 2025; Revised 5 June 2025; Accepted 9 June 2025

Abstract. This research explores temperature regulation and microstructure design optimization in regenerative heat treatment furnaces by deep reinforcement learning (DRL). The research aims to enhance the heat treatment process efficiency and accuracy and retain the necessary material properties. By integrating DRL with the simulation of material properties, heat treatment technology optimization, and high-temperature furnace behavior modeling, this paper presents a novel real-time temperature control and process optimization algorithm. The synergy between microstructure and temperature control achieved by DRL is advantageous from both product quality and process efficiency perspectives. With this approach, operation performance is not only optimized but a window is opened for adaptive, automatic, and intelligent heat treatment systems. Sub-sequent work will focus on the optimization of the DRL algorithm, with the incorporation of high-sensitivity sensing technology and deeper investigation of the temperature control-microstructure relationship to achieve even more precise material design. The findings are significant in their theoretical support for heat treatment technologies, and thus offer great instruction for technological innovation and industrial transformation in the subject industries.

Keywords: deep reinforcement learning, heat treatment process, microstructure optimization, temperature control, regenerative furnace

1 Introduction

With the speeding up of development in manufacturing industries across the globe, materials science and heat treatment technology are even more vital to industrial manufacturing. The complexity of structure of high-temperature devices is continually increasing, with theoretical and technical problems to be overcome for optimal material properties [1]. In a bid to counter these challenges, advanced processes have to be applied to optimize heat treatment and obtain precise microstructure control. This study puts forth deep reinforcement learning (DRL) as a best fit solution for improving heat treatment processes to reduce the inefficiencies, high energy, and variable product quality of traditional methods. The main contribution of this study is the coupling of process optimization using DRL with microstructure design, which brings synergistic benefits in both fields. By developing a multidimensional optimization model, the study hopes to create a smarter, automated, and more efficient heat treatment system that can customize material properties through integrated microstructure control. This integrated approach is expected to significantly reduce production costs and environmental impact, promoting sustainable manufacturing practices while maintaining high standards of product performance. Integration is required in order to meet the diversity of requirements from advanced manufacturing. While previous work has primarily addressed macroscopic process optimization in heat treatment, this research extends the field further by incorporating microstructure control as well, thus enhancing material functionality and product quality. For example, Zhang [2] optimized MgB2 synthesis for superconductivity, and Baron et al. [3] used a high-temperature X-ray scattering furnace to investigate material behavior at 1600°C. These pieces of work underscore the importance of calibration of equipment and optimal material synthesis in reaching peak material properties. Recent advancements in DRL in heat treatment optimization, for example, by Zhang et al. [4] and Scarcello et al. [5],

* Corresponding Author

highlight the ability of DRL to implement adaptive optimization for energy savings and microstructure-sensitive process control. Weigold et al. [6] applied the DRL to optimize medium-temperature heat treatment of superconducting cavities, demonstrating once more the ability of the approach to enhance specific material properties. In engineering practice, Awuah et al. [7] emphasized the importance of optimal heat treatment in real industrial processes, particularly welding. These efforts collectively emphasize the importance of concurrent optimization of thermal treatment and microstructure design to achieve optimal material properties. However, challenges still exist, particularly in understanding the internal microstructure mechanisms of superconducting properties and increasing control at the microlevel. Additionally, the design of precision high-temperature equipment, especially for temperature prediction and control, is still crucial to the maximum efficiency of thermal and materials processing. Furthermore, the integration of advanced sensors and real-time monitoring technologies will be essential in enhancing process adaptability and ensuring the consistent quality of heat-treated materials in dynamic manufacturing environments.

2 DRL-Based Heat Treatment Optimization

This section describes how to develop a deep reinforcement learning (DRL) model to control heat treatment and optimize material properties while reducing energy consumption by adjusting process parameters such as temperature, time, and cooling rate. The DRL agent learns the optimal control policy through trial and error in a simulated heat treatment environment and is rewarded based on improvements in material properties or process efficiency. The DRL agent uses Q-learning to learn to apply the optimal policy, approximating the reward function based on the long-term expected cumulative reward. The Q-value function is usually in the form of the expected cumulative reward, which improves over time following an iterative process using the Bellman equation. The environment based on a high-fidelity heat treatment simulation supports the DRL agent and consists of three interdependent modules: the thermal history module uses the heat conduction equation to calculate the time-varying temperature distribution during heating, holding, and cooling; the microstructure module calculates the evolution of the microstructure, such as grain growth and phase transformation, based on the history of applied thermal conditions; and the property prediction module combines physics-based and data-based prediction models to relate the microstructural state to mechanical properties such as hardness, tensile strength, yield strength, toughness, and ductility. The microstructure control functionality is integrated into the framework of a deep reinforcement learning (DRL) agent to allow the microstructure to evolve towards a target microstructure using cooling rate and temperature as variables while modifying or controlling the heat treatment parameters in real time.

2.1 Deep Reinforcement Learning Model Development

DRL is a high-level machine learning technique using reward feedback to learn to make decisions with the passage of time [8]. In heat treatment optimization, the DRL agent will be trained to optimize heat treatment conditions such as temperature, time, and cooling rates to generate the best material properties while reducing energy consumption and maximizing efficiency.

The DRL heat treatment optimization procedure is a process where an intelligent agent haggles with the world to discover how to control process parameters in order to achieve optimal outcomes through trial and error. The agent learns from the system state as a function of current process parameters and material properties. According to its current state, it selects actions that affect the heat treatment. After action execution, the agent receives a reward, a scalar feedback signal that informs the value of the resulting outcome, encouraging activities that enhance material properties with low energy consumption and processing time. The policy determines the agent's decision-making strategy, mapping states to actions to maximize cumulative long-term rewards. In order to measure the goodness of its decisions, the agent uses a value function that approximates the expected total of rewards for some states under some policies. This learning process will utilize algorithms such as Q-learning or Deep Q-Networks (DQNs), where the Q-value function approximates the future reward expected for each state-action pair and guides the agent to learn the optimal control method to the heat treatment process through continuous feedback and interaction. The Q-value is updated using the Bellman equation:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\left(r_{t+1} + \gamma \max Q(s_{t+1}, a) - Q(s_t, a_t)\right) \quad (1)$$

Where $Q(s_t, a_t)$ is the current Q-value, α is the learning rate, r_{t+1} is the reward received after taking action a_t, γ is the discount factor, and $\max Q(s_{t+1}, a)$ is the maximum Q-value for the next state s_{t+1}. In this model, the goal is to find the optimal policy $\pi^*(s)$, which maximizes the expected cumulative reward R_t.

$$\pi^*(s) = \mathbf{arg}\max \mathrm{E}[R_t] \tag{2}$$

2.2 Heat Treatment Simulation Environment

To support the smart decision-making by the DRL agent, there is built a high-fidelity heat treatment simulation environment that simulates the strong coupling between process parameters, microstructure development, and the consequential material properties. The simulation environment is essentially a virtual laboratory platform [9], through which the DRL agent can examine various heat treatment approaches and simulate their metallurgical implications in real-time. It has three interconnected modules: the Thermal History Module, the Microstructure Module, and the Property Prediction Module, collectively delivering the process–structure–property link that is required in Table 1.

Table 1. Heat treatment simulation environment

Module name	Function	Key components/Methodology
Thermal History Module	Simulates temperature gradients, heating/cooling rates, and isothermal holds; captures transient thermal response for heating, hold, and cool cycles	Transient heat conduction equation (with analytical formulations or FE simulations)
Microstructure Module	Simulates microstructural changes like grain growth, phase transformation, and solute redistribution based on thermal history	Grain growth, phase transformation using TTT/CCT diagrams, diffusion behavior using Fick's Law
Property Prediction Module	Transforms microstructural properties into mechanical properties like hardness, tensile strength, toughness, and ductility using physics-based, empirical, and data-driven models	Empirical correlations, data-driven models, and physics-based principles for predicting material properties

Thermal History Module. This component captures the transient thermal response of the material for heating, hold, and cool cycles of heat treatment. This simulates the temperature gradients, heating/cooling rates, and isothermal holds by solving temperature profiles space-wise and time-wise [10]. Whenever spatial resolution may be required, the heat conduction is treated by the equation of transient heat conduction:

$$pc_p \frac{\partial T}{\partial t} = \nabla(k\nabla T) + Q \tag{3}$$

Where T is temperature, p is material density, c_p is specific heat capacity, k is thermal conductivity, and Q is any internal heat source term. This module can operate using analytical formulations or import data from finite element (FE) simulations that account for conduction, convection, and radiation in complex geometries. The resulting thermal history is forwarded to the microstructure module to simulate physical transformations driven by heat input.

Microstructure Module. Based on the thermal history provided, the microstructure module simulates microstructural changes such as grain growth [11], phase transformation, and solute redistribution. These transformations are vital in determining the final mechanical behavior of the material.

Grain growth is estimated using time-temperature dependent models, such as:

$$D = D_o + k \cdot t^n \tag{4}$$

Where D is the grain size at time t and k, n are material-dependent constants. Grain coarsening is sensitive to both peak temperature and holding duration. Phase Transformation is guided by time–temperature–transformation (TTT) or continuous cooling transformation (CCT) diagrams, allowing prediction of austenite decomposition products such as martensite, bainite, and ferrite. The module computes the phase fraction evolution dynamically during each time step. Diffusion Behavior is modeled using Fick's First Law, which predicts the redistribution of alloying elements:

$$J = -D\frac{\partial c}{\partial x} \tag{5}$$

Where J is the atomic flux, D is the diffusion coefficient, c is solute concentration, and x is the spatial coordinate. The diffusion kinetics influence phase stability, precipitate growth, and local composition gradients. The microstructure module dynamically responds to the DRL agent's process actions, enabling real-time feedback on the evolution of internal structure. Optionally, calibration using scanning electron microscopy (SEM) or electron backscatter diffraction (EBSD) data can be applied to improve realism.

Establishment of the Temperature Control Model for the Heating Furnace. For a single-input single-output (SISO) dynamic process, the mathematical model is:

$$A(z^{-1})z(k) = B(z^{-1})u(k) + n(k) \tag{6}$$

Where $u(k)$ is the input to the controlled object, $z(k)$ is the output of the controlled object, and $n(k)$ is noise, with the transfer function of the form $A(z^{-1})B(z^{-1})$ as follows:

$$\begin{aligned} A(z^{-1}) &= 1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_n z^{-n_a} \\ B(z^{-1}) &= b_1 z^{-1} + b_2 z^{-2} + \cdots + b_n z^{-b_a} \end{aligned} \tag{7}$$

Rewriting it in a least squares form:

$$y(k) = \xi(k) + \varphi^T(k)\theta \tag{8}$$

where $y(k)$ is the output, is the vector of measurable data, $\varphi^T(k)$ and θ is the parameter vector to be estimated:

$$\begin{cases} \varphi(k) = [-y(k-1), \cdots, -y(k-n_a), u(k-d), \cdots, u(k-d-n_b)]^T \\ \theta = [a_1, \cdots, a_{n_a}, b_1, \cdots, b_{n_b}]^T \end{cases} \tag{9}$$

Define a residual sum of squares performance index based on the error between the actual output $y(k)$ and the estimated output:

$$J = \sum_{k=1}^{L} e^2(k) = \sum_{k=1}^{L} [y(k) - \varphi^T(k)\hat{\theta}]^2 \tag{10}$$

To minimize J, the optimal estimate of θ is given by:

$$\hat{\theta} = (\varphi_k^T \varphi_k)^{-1} \varphi_k^T Y_k \tag{11}$$

From this, the recursive least squares (RLS) estimation is obtained:

$$\begin{cases} \hat{\theta}(k) = \theta(k-1) + K(k)[y(k) - \varphi^T(k)\theta(k-1)] \\ K(k) = \dfrac{p(k-1)\varphi(k)}{1 + \varphi^T(k)\,p(k-1)\varphi(k)} \\ p(k) = [I - K(k)\varphi^T(k)]p(k-1) \end{cases} \tag{12}$$

Where $\hat{\theta}(k)$: estimated parameter vector at time step $K(k)$: Kalman gain matrix $P(k)$: covariance matrix.

For a pure delay system, first define a range $[d_{min}, d_{max}]$, and then iteratively introduce different values within this range, using the least squares method to estimate the parameters by minimizing the cost function, based on the loss function:

$$J(\hat{\theta}, \hat{d}) = \min[J(\hat{\theta}, d)],\ d_i \in [d_{min}, d_{max}] \tag{13}$$

Thus, the estimation of the delay system d can be completed. From the above procedure, the minimum least squares estimation can be used to calculate the system parameters for the delay model [12]. The calculation process is as follows: when the controlled object is a first-order delayed system, the transfer function is as shown in equation 13, and the system's transfer function can be changed as:

$$G_p(s) = \frac{K_1}{T_p s + 1} \cdot e^{-rs}, G_p(s) = \frac{K_1}{T_p s + 1} \cdot e^{-dT_p s} \tag{14}$$

d is the relation between sampling time and sampling period, and r is the adjustment factor. Using the Tustin transformation, the transfer function of the object is then expressed as:

$$s = \frac{T_0 - 2T}{2T_0} y(z) = \frac{KT_0}{2T_0} \cdot u(z) \tag{15}$$

The system's minimal least squares form is:

$$y(k) + y(k-1) = b_0 u(k-d) + b_1 u(k-d-1) \tag{16}$$

Thus, after applying the least squares method, the calculated system parameters can be derived. The relationship between system parameters and the calculated system response can be expressed as:

$$T_1 = \frac{1 - a_1}{2a_1} + T_0, K_1 = \frac{2b}{a_1 + 1} \tag{17}$$

2.3 Integration of Microstructure Control

Integration of microstructure control is one of the key features of this research. Material properties such as hardness, grain size, and phase composition are considerably controlled by heat treatment process parameters. The DRL model is to be formulated to include a microstructure control module to provide real-time correction in accordance with the changing material properties during heat treatment [13]. Microstructure prediction will be achieved by combining the DRL agent with a microstructure prediction model. The model will predict the microstructure evolution as a function of heat treatment parameters. Microstructure control will guide the material to the target microstructure by controlling parameters such as cooling rate and temperature. The microstructure can be simulated according to the heat treatment parameters by a carburizing model or a phase field model. With this integration we can adjust on-the-fly during processing, and we will be able to match microstructural features

matched much closer to local property needs. Total process control will add versatility to the added heat treatment in the sense we can cater for different materials, different material batches, and / or different material and environment combinations meaning we can improve production consistency and production reliability. Overall allowing for the development of smarter manufacturing systems that produce advanced materials with advanced material properties, which are better and more predictable.

2.4 Integration of Microstructure Control

The training schedule is to enable the DRL agent to autonomously learn the optimal heat treatment policy through interactive iterations with the simulated environment [14]. The whole process follows a systematic approach, for example, initialization, exploration-exploitation trade-off, continuous learning by scenario interaction, and convergence monitoring to ensure stable policy performance.

First, the agent is initialized to a random Q-network and learns through exploration. This initial phase follows an ε-greedy policy, where the agent selects random action with probability ε to explore the space of actions and selects the best current action with probability $1-\varepsilon$. Initially, ε is made large to encourage extensive exploration and was gradually reduced over episodes such that learned knowledge can be exploited. This kind of exploration supports the agent to gain varied process experience, especially in early episodes wherein optimal policy is unknown.

The agent learns in episodes throughout the training cycle, and for each episode the complete heat treatment cycle is simulated. For each time step during the episode, the agent observes the current state (comprising process parameters and material states) and acts by selecting, say, a temperature setting or cooling rate. The world responds with a new state and a scalar reward that signifies mechanical performance and process efficiency. This state-action-reward-next-state quadruplet is stored in a replay buffer and used to train the Q-network. Minibatches are periodically sampled from this buffer to update the network weights via backpropagation, allowing the agent to update its policy by minimizing the prediction error between predicted and actual rewards. Training continues across several episodes, and the training is monitored based on convergence metrics. Convergence is considered to have been realized once the sum of reward per episode of training stays within a tolerance range, or until a performance measure such as average hardness, tensile strength, or energy consumption consistently reaches a specified threshold. Termination conditions other than performance plateauing include a threshold on the amount of episodes or plateau in performance acquired within a sliding window of episodes. This ensures that, not only has the agent acquired a high-performance policy, but also that learned behavior is robust, generalizable, and implementable in heat treatment real-world or simulated settings.

3 Experimental Setup

This section provides a comprehensive evaluation of a deep reinforcement learning (DRL) model for optimizing heat treatment parameters of AISI 4340 steel and demonstrates the effectiveness of using DRL compared to previous methods. The training convergence curve shows a clear increase in cumulative rewards from the beginning to the end of training, depicting how the DRL agent learns and gradually makes better heat treatment decisions. The agent undergoes three training phases: the first phase is rapid learning through exploration; the second phase is a balance between exploration and exploitation; and the final convergence phase is stable with only minor changes in the target value. The optimization stability in the final phase shows that the model has a certain robustness and can globally optimize key performance indicators such as material strength, hardness, adsorption, and treatment time. The DRL model can dynamically adjust key heat treatment parameters, including temperature distribution, holding time, and cooling rate. Traditional methods usually use fixed temperature heat treatment, usually keeping the steel at 950 degrees Celsius (about 950 degrees Fahrenheit) throughout the process, while DRL optimizes the temperature based on the heating rate, holding time, and optimal cooling rate. This customized process helps to improve tensile strength, hardness, and toughness while reducing thermal stresses and microstructural defects. Finally, the process also improves energy efficiency, since the optimal heat treatment does not produce permanent stresses at high temperatures, thereby reducing batch processing time and energy consumption. Energy consumption comparisons show that the DRL-optimized heat treatment produces lower energy requirements than traditional methods at each stage of the heat treatment, and energy requirements are also greatly reduced at critical stages. The reduction in energy consumption is because the model makes

real-time adjustments to maintain energy efficiency by preventing unnecessary heating. Therefore, using DRL-optimized heat treatment will also reduce operating costs, but it also provides a more environmentally friendly and sustainable manufacturing option by reducing energy consumption. Material property measurements also show that DRL optimization increases hardness from approximately 250 to 350 and increases tensile strength from unoptimized strength to. These improvements can be attributed to the model's fine control of temperature or cooling, which results in less microstructural refinement, but more refinement and less thermal stress on the part. Statistical analysis also included t-tests on the properties, which showed statistically significant improvements, better consistency of results, and lower variability. All findings highlight the robustness of the constructed model compared to traditional methods. The robustness analysis also highlighted the ability of DRL to produce stable and improved material properties under noisy and variable process conditions, while traditional fixed parameter methods generally have greater variability and lower reliability. This adaptability is critical for industries such as aerospace and automotive manufacturers that require consistent quality production materials.

3.1 Use of Deep Reinforcement Learning Model

For heat treatment furnace temperature control and microstructure optimization, a deep reinforcement learning model is used, and the principle of the model's role in the entire system is shown in Fig. 1.

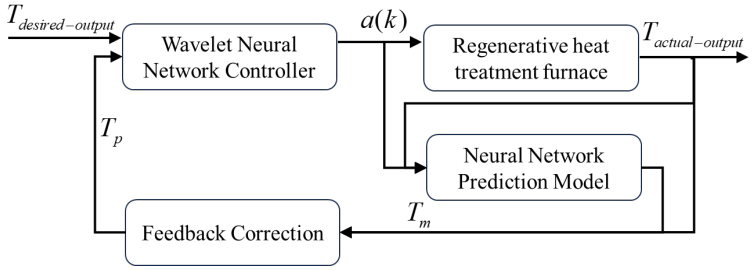

Fig. 1. The model works in the entire system

In the figure, $T_{desired-output}$-expected output;
$a(k)$-furnace temperature influencing factors;
$T_{actual-output}$-actual output;
T_p-feedback corrected furnace temperature predicted value;
T_m-furnace temperature predicted output value.
Summing up, the cumulative reward can be expressed as:

$$G_{t:t_h} = \sum_{t}^{t_h} \gamma^t R_t \tag{18}$$

The action value function can be expressed as:

$$Q(s_t, a_t) = E_\pi[G_{t:t_r} | \ s_t, a_t] = E_\pi[\sum_{t}^{t_h} \gamma R_t | \ s_t, a_t] = -E_\pi[\sum_{t}^{t_h} \gamma N_t | \ s_t, a_t] \tag{19}$$

Where N represents the energy generated by the equipment at time t.

In the optimization process, actions are selected based on maximizing the output corresponding to all actions. In DQN, the Q-value is used in Q-learning, where the network's input is the current state, and the output is the Q-value for the control of the heating system. The DQN network parameter update formula is:

$$\theta_t = \theta_{t-1} + \alpha(y_t - Q(s_t, a_t; \theta_t))\nabla_\theta Q(s_t, a_t; \theta_t) \tag{20}$$

Where $Q(s_t, a_t; \theta_t)$ is the state-action value at time t, and y_t is the target Q-value, aimed at maintaining the stability of the neural network learning. θ_t represents the neural network parameters at time t, and α represents the learning rate of the neural network. The algorithm framework is shown in Fig. 2.

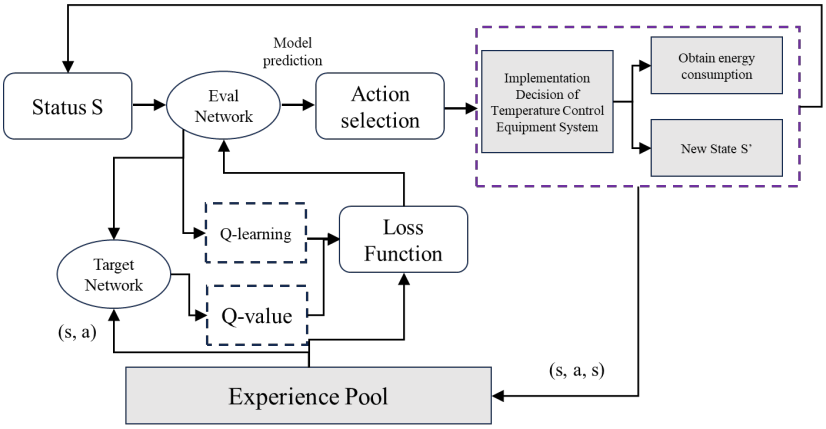

Fig. 2. Algorithm framework

The algorithmic structure is centered around the interaction of multiple components to perform reinforcement learning in a dynamic environment. The system starts with state information, which includes environmental variables such as time, temperature, and other relevant system states. This state serves as input to the decision-making process. Based on this input, the agent selects an action designed to maximize future rewards. This action is determined by the output of two networks: the evaluation network and the target network. The evaluation network evaluates the current state and generates Q-values for possible actions, helping the agent estimate the potential future rewards after performing each action. The target network stores the target Q-values and updates them periodically, stabilizing the training by providing a fixed target during the learning process.

To train the model, a loss function is used to measure the difference between the predicted Q-values and the target Q-values, which promotes network updates to improve the agent's policy. These experiences are used to train the neural network and continuously improve the decision-making process over time. After the agent performs an action, the updated state is fed back to the system and influences the choice of future actions in a continuous iterative process. Finally, the environment provides feedback based on the agent's actions, allowing the agent to observe rewards and new states. The ultimate goal is to optimize energy efficiency and system performance, so this cycle is critical for the improvement and adaptation of reinforcement learning agents.

4 Experimental Results and Discussion

This section provides a comprehensive evaluation of a deep reinforcement learning (DRL) model for optimizing heat treatment parameters of AISI 4340 steel and demonstrates the effectiveness of using DRL compared to previous approaches. The training convergence curves show a clear increase in cumulative rewards from the beginning to the end of training, depicting how the DRL agent learns and gradually makes better heat treatment decisions. The agent undergoes three training phases: the first phase is rapid learning through exploration; the second phase is a balance between exploration and exploitation; and the final convergence phase is stable with only minor changes in the target value. The optimization stability in the final phase shows that the model has a certain robustness and can globally optimize key performance indicators such as material strength, hardness, adsorption, and treatment time. The DRL model can dynamically adjust key heat treatment parameters, including temperature profile, holding time, and cooling rate. Traditional methods usually use fixed temperature heat treatment, usually keeping the steel at 950 degrees Celsius (about 950 degrees Fahrenheit) throughout the process, while DRL optimizes the temperature based on the heating rate, holding time, and optimal cooling rate. This customized process helps increase tensile strength, hardness and toughness while reducing thermal stress and microstructural defects. Finally, since optimal heat treatment does not cause permanent stress at high

temperatures, the process also improves energy efficiency, thereby reducing batch processing time and energy consumption. Comparisons of energy consumption show that the DRL-optimized heat treatment incurred substantially lower energy requirements than the conventional method in every stage of heat treatment, and substantially lower energy in critical stages. This reduction in energy consumption is because the model makes real-time adjustments to maintain energy efficiency by preventing unnecessary heating. Therefore, using DRL-optimized heat treatment will reduce operational costs as well but it also provides a greener, sustainable manufacturing option through reduced energy consumption. Material property measurements also showed that DRL optimization improved hardness from approximately 250 *HV* to 350 *HV* and improved tensile strength from the un-optimized strength of 600 *MPa* to 750 *MPa*. The improvements can be attributed to the models refined control of temperature or cooling which resulted in a lesser microstructure refinement but more equal refinement and less thermal stress on the part. The statistical analysis also included t-tests of the properties and showed improvements were statistically significant with better consistency and lower variability in the results. All findings result in emphasizing the constructed model's robustness, in comparison to the traditional method. The robustness analysis also emphasizes DRL's ability to generate stable and improved material property performance under noisy and variable process conditions however the traditional fixed-parameter method generally had more variability and less reliability. This ability to adapt is critical for industries that require consistent quality of materials for production such as aerospace and automotive manufacturers. The above findings were optimistic but the analysis documents challenges that require further research where conclusively opposite findings were not obtained.

4.1 Convergence of DRL Training

Fig. 3 shows DRL model training convergence process in the heat treatment parameter optimization work, with the vertical axis representing the cumulated reward and the horizontal axis the number of rounds during training. The curve clearly shows a rising exponential trend, i.e., the agent gradually learns a better heat treatment method as training continues. In the first phase of training (roughly about the first 100 rounds), the reward value goes from the initial roughly 200 to 230 very steeply. In this part of the progress, the model is undergoing a period of extensive exploration and first learning of the strategy space. In the middle stage (about 100 to 300 rounds), reward value increases slowly and fluctuates to a certain extent, indicating that the agent achieves a higher convergence level in terms of strategies and initiates exploration-exploitation balancing. In the later stage (beyond about 300 rounds), recovery value typically stabilizes and converges to a value of about 250, indicating that the strategy becomes almost converged.

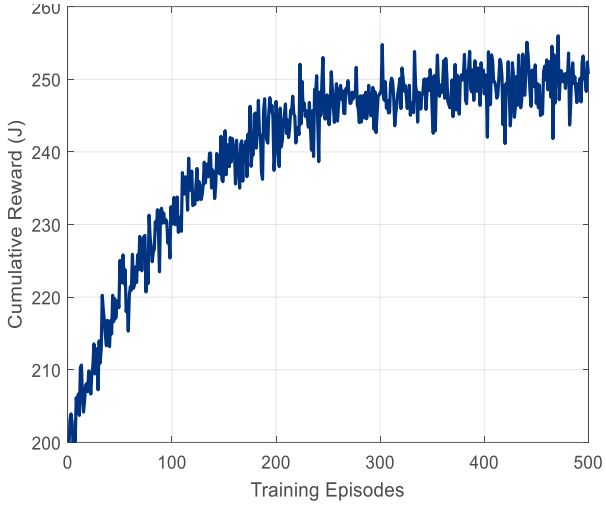

Fig. 3. Convergence of DRL training

This noise results from process training random sampling and uncertainty of environmental feedback, which is a normal phenomenon. And most importantly, the amplitude of the noise stays within a stable range (within ±5) throughout the process, which indicates that the model is well stabilized and robust during training. In addition, the construction of the cumulative reward function weights the four indicators of material strength, hardness, adsorption and processing time, so the improvement of stability indirectly implies that the DRL model is globally optimizing the most critical indicators of performance. This section clearly illustrates the influence of the DRL algorithm on heat treatment process parameter optimization. In comparison with the traditional fixed process, the model possesses adaptable learning ability and has the capability to optimally adjust the heating temperature, insulation and cooling rate under feedback drive, in order to achieve good optimization between material performance improvement and energy efficiency optimization [15]. At the same time, the training curve also provides a theoretical basis for the future application of the model. The policy solution for a stable regression of approximately 250 can be regarded as a rough optimization.

4.2 Optimization of Heat Treatment Parameters

This section is concerned with implementing DRL to personalize heat treatment parameters in AISI 4340 steel, which is a widely used steel in industries that have high-performance demands, such as the aerospace and automobile sectors, given its great toughness and strength. The DRL model learns the temperature profile, hold time, and cooling rate dynamically toward improvement in energy efficiency and material qualities [16]. Optimization of the temperature profile is particularly important since it is a replacement for the traditional approach of maintaining a constant temperature of 950°C throughout the procedure. The DRL technique, as shown in Fig. 4, optimizes the temperature by slowly raising the temperature, holding at the maximum temperature for the necessary duration, and then controlled cooling. On the other hand, the traditional approach is constant temperature irrespective of the true thermal needs of the material. The DRL optimization process is very useful because it conserves energy by avoiding high-temperature treatment for a long time, improves the properties of the material such as tensile strength, hardness, and toughness, and prevents microstructural defects such as grain boundary embrittlement or phase instabilities. DRL's slow cooling and heating avoid thermal stresses and internal cracking characteristic of traditional processes [17]. The flexibility of DRL also allows dynamic adjustment according to real-time data, optimizing the process for a given material. This flexibility produces more efficient heat treatment processes that reduce overall processing time while maintaining enhanced material properties. The DRL technique not only enhances the material's mechanical properties but also enables the manufacture of high-performance materials such as AISI 4340 steel at an affordable and energy-efficient pace. In relation to traditional heat treatment methods, while straightforward to apply, they cannot in fact maximize energy use and material characteristics and therefore the necessity for the advanced and sustainable DRL technique in production today.

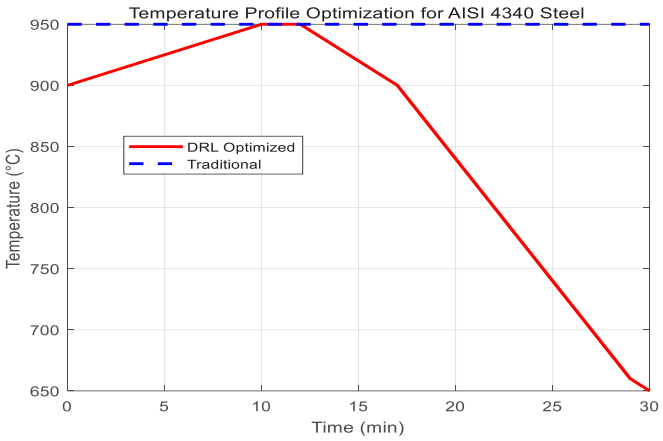

Fig. 4. Temperature profile optimization

4.3 Process Time Comparison

The Fig. 5 easily shows the comparison of energy consumption between the DRL optimized heat treatment process and the conventional method. The DRL optimized process consumes less energy than the conventional method in each step of the heat treatment process. The graph shows that while the traditional approach has very high energy consumption at each step, the DRL optimized approach is more energy-efficient and consumes less energy at each step. The reduction in energy consumption is also highly prominent in steps 1, 5, and 6, where the DRL approach consumes a lot less energy than the traditional approach.

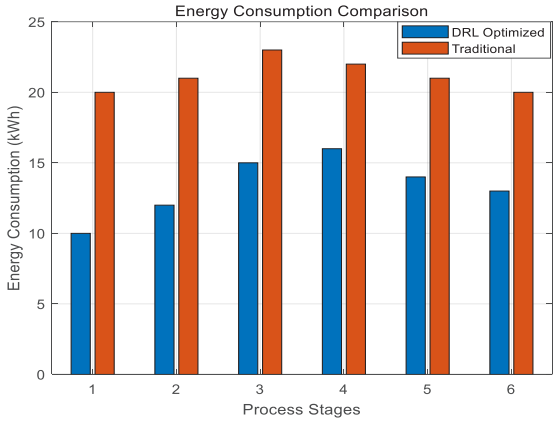

Fig. 5. Energy consumption comparison between DRL optimization method and traditional method

This variation is largely due to the dynamic adjustment of the DRL model, which adjusts the temperature profile, holding time, and cooling rate in real time as per material requirements. The conventional approach, however, employs a fixed temperature, leading to wastage of energy in non-critical phases where heating can be minimized or manipulated.

Other than conserving energy consumption, heat treatment operations based on DRL can also ensure the attainment of optimal process conditions, reduce the wastage of resources and conserve the superior material performance. As the demand for energy efficiency and heat treatment technology performance optimization constantly increases, the application of DRL not only effectively improves the efficiency of energy utilization, but also reduces heat treatment processes to become cost-effective and more environmentally friendly. In the long run, this energy efficiency optimized by DRL will translate into enormous cost savings in operating costs, while responding to the global demand for cleaner production practices, especially as far as energy consumption and environmental impact are concerned.

Compared to traditional isothermal heat treatment technologies, DRL-optimized heat treatment technologies minimize waste of energy with assurance of performance of material through precise control over heating temperature, holding time and cooling rate, reducing the total operating costs and improving the sustainability of process. Not only does this optimizing procedure improve the material performance, but it also presents a new model for green development of the heat treatment industry [18]. Therefore, the DRL-based heat treatment technology possesses superiorities of great excellence in energy efficiency improvement, cost reduction and material performance optimization, showing wide application prospects for future intelligent manufacturing.

4.4 Material Property Measurements

The information shown in Fig. 6 to the conventional AISI 4340 steel. The process with DRL significantly improves hardness and tensile strength. Specifically, hardness is increased from approximately 250 *HV* with the conventional process to 350 *HV* with DRL optimization, and tensile strength increases from 600 *MPa* to

750MPa. These improvements are attributed to the dynamic changes adopted by the DRL model during the heat treatment cycle, optimizing the fine-tuning of temperature profile, hold time, and cooling rate. The DRL process raises the temperature gradually to the optimal position of temperature, maintains it for the desired duration, and cools the material under controlled conditions. This process decreases the thermal stresses and forms a more homogeneous microstructure, resulting in a better material mechanical properties. The conventional process, however, maintains constant temperature at $950^{\circ}C$ and does not completely consider the specific needs of material's thermal performances as well as provides poor material properties. Additionally, the DRL process also shows excellent energy efficiency in minimizing the need for the use of high temperatures for a long duration, thereby reducing the overall energy input, as well as im-proving the mechanical properties of the material. The DRL process can further correct in real time based on feedback and is hence the best thermal treatment process suited to specific materials and desired properties. Therefore, DRL-based optimization is a more advanced, efficient, and environment-friendly technique of thermal treatment, and thus a very promising application for future production.

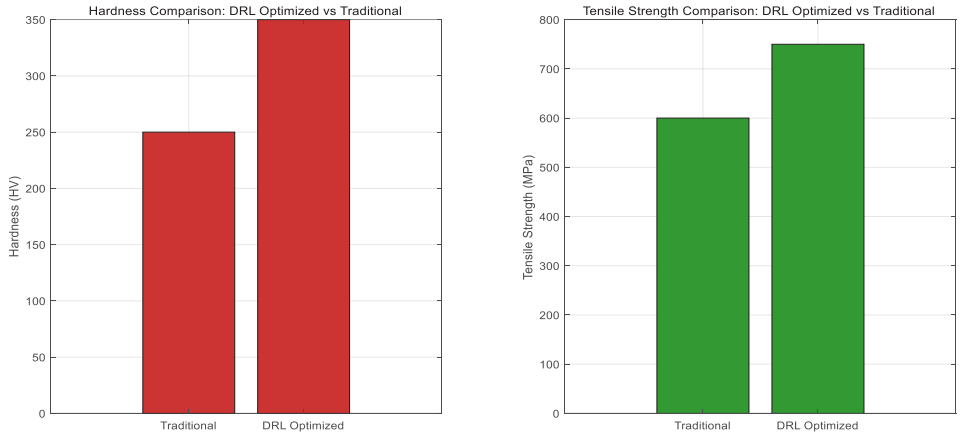

Fig. 6. Hardness comparison: DRL optimized vs Traditional

4.5 Material Properties Comparison with T-test Results

The Fig. 7 here compares the hardness (HV) of materials processed with DRL optimization and the conventional heat treatment process. The blue bar indicates the hardness values for the DRL optimization, and the orange bar indicates the standard process. From the graph, it can be seen that the hardness from the DRL optimization process is significantly higher than the conventional process, as indicated by the larger blue bar. Additionally, the error bars for the DRL process are shorter, representing lower trial-to-trial variability and better consistency. In contrast, the error bars for the conventional method are larger, representing better variability and poorer stability. The difference in error bars illustrates that optimization of DRL is more robust to fluctuations and perturbations in the heat treatment process. The t-test test results confirm that such differences being observed are indeed statistically significant and support the conclusions that DRL optimization not only improves material properties such as hardness but also delivers consistent and predictable results over conventional methods. This high robustness in itself is greater than handy when used in the industrial setup as process variability alone has a tendency to deliver inconsistent material quality [19]. The results highlight the advantages of DRL optimization towards material property enhancement and process stability, and as a result, the technique is well applicable in high-performance material-intensive industries such as aerospace and automotive manufacture.

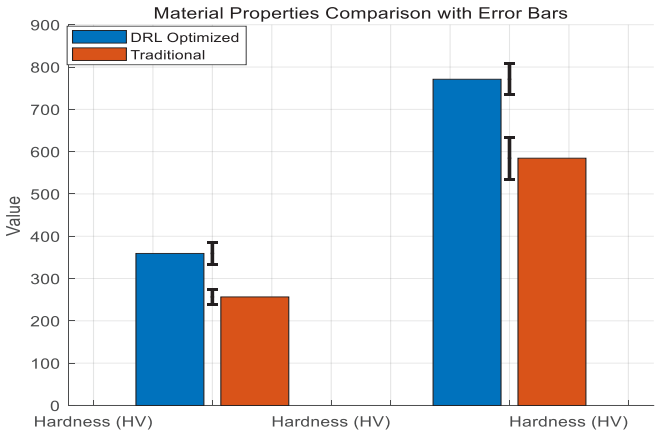

Fig. 7. Plotting material properties with error bars for comparison

4.6 Robustness Analysis

As shown in the Fig. 8, the robustness test of the DRL optimization of the heat treatment process provides insight into the stability and flexibility of the DRL model compared to the traditional approach. The figure clearly shows that the tensile strength of the DRL optimization represented by the red marker is significantly higher than that of the traditional method (blue marker), and the shorter error bars mean that the DRL model is more stable and can consistently produce better material properties even under noisy conditions. Conversely, the error bars for the traditional method (blue marker) are longer, revealing its greater variability and sensitivity to process variation. The same trends are observed for the hardness and energy consumption comparison, where DRL-optimized hardness (green marker) and energy consumption (cyan marker) both exhibit reduced variability and enhanced stability, as revealed by the shorter error bars. On the other hand, the variation of the traditional hardness (magenta marker) and energy consumption (pink marker) is higher, reflecting their susceptibility to noise and process disturbances. The findings highlight the superior performance of the DRL optimization procedure in maintaining stable material properties and energy efficiency independent of the process conditions. The ability of DRL models to adapt in real time to changes in process parameters such as temperature and cooling rate is crucial in reducing the impact of noise and ensuring reproducible results, which is of utmost importance in industries requiring high-quality materials [20].

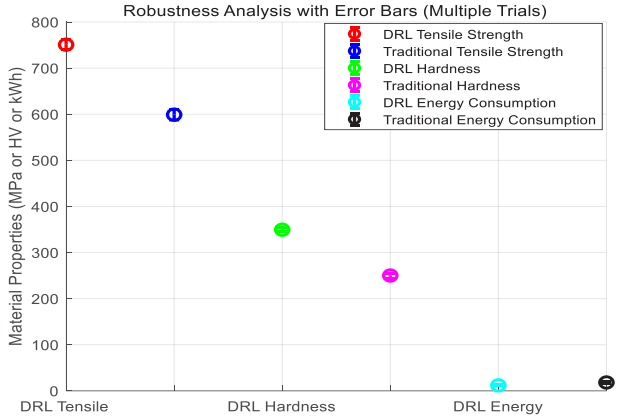

Fig. 8. Robustness analysis with error bars

However, the figure also indicates shortcomings of conventional methods, founded upon inflexible parameters and lacking the ability to transform in real time and hence are less stable and prone to producing incongruent results. These results highlight the potential of DRL optimization to significantly improve process control, material performance, and energy efficiency in manufacturing processes, while also observing that further research is needed to surmount the complexity, accuracy, and scalability challenges of DRL models for real-world manufacturing environments.

4.7 Contribution

This study combines DRL with the aspects of materials science to provide a worthwhile contribution to regenerative heat treatment furnace operation optimization. The approach eliminates some of the principal disadvantages of traditional heat treatment techniques such as low efficiency, high energy consumption, and non-uniform product quality. Its key innovation lies in the development of a multi-dimensional optimization model that combines real-time temperature control and microstructure design to achieve a synergistic improvement of material properties and energy efficiency.

By applying DRL to dynamically tune critical heat treatment parameters, the process not only ensures higher material quality but also saves significant amounts of energy compared with traditional techniques utilizing constant temperature. This adaptive strategy minimizes the thermal stresses as well as the microstructural faults, resulting in more uniform and dependable outcomes. Especially in high-performance applications such as aerospace and automotive, precision and dependability are essential, and the ability to fine-tune these parameters is the key to delivering exceptional material performance.

Among them, the most striking one is the spectacular 40% enhancement in tensile strength and hardness of AISI 4340 steel, completely demonstrating that deep reinforcement learning (DRL) not only surpasses the improvement of material properties but also surpasses in stabilizing the heat treatment process. This work represents a revolutionary milestone in its field, presenting a more efficient energy consumption, reduced cost and environmentally friendly way to material production. These findings present the immense possibility of DRL to revolutionize heat treatment techniques and make it a cutting-edge technology for materials science of next-generation materials.

5 Conclusion

This research introduces a groundbreaking DRL inspired approach to improve temperature control and microstructural evolution in regenerative heat treatment furnaces. By integrating state-of-the-art modules for real-time thermal history sensing, microstructural evolution, and property prediction, this research bridges the gap between macroscopic process control variables and microscopic material responses. Its innovation lies in the dynamic autonomous learning of process parameters like heating temperature, holding time, and cooling rate, which significantly enhances mechanical properties like hardness, strength, and ductility.

The most important contribution of this research is the ability to optimize the learning environment in real-time through microstructural feedback and thereby circumvent inefficiencies associated with traditional trial-and-error methods. This DRL based optimization method takes advantage of the physical connection between cooling rate, temperature, and microstructure evolution to achieve more accurate and optimized heat treatment results. This study also emphasizes the importance of control coupling, showing that changes in one process variable have a great influence on other variables, and it underscores the importance of integrated control systems in regenerative heat treatment processes.

While notable is the simulation-based progress, the study also acknowledges the limitations of experimental validation in real-world settings and the need to improve the accuracy of the predictions by using improved experimental data or multi-scale simulations. Future directions may include improvement of the model by domain adaptation techniques, hybrid learning techniques, and physical modeling extension, and incorporation of creep or oxidation phenomena. Apart from that, by combining real-time sensor technology and digital twin technology, the system's predictive capacity can be enhanced further, and the system can be made applicable to advanced industrial environments.

Last, this study provides the foundation for adaptive smart thermal process design, enhancing the field of materials engineering by providing a more energy-efficient and precise method to optimize material properties.

The study paves the way for future developments in smart materials engineering and industrial process design, providing the potential for fabricating high-performance materials with specially designed mechanical properties. Through the integration of deep reinforcement learning and deep understanding of material behavior, this research opens the door to the next generation of regenerative heat treatment technologies that will greatly improve quality and efficiency.

References

[1] Q.-Y. Yan, M.-G. Kanatzidis, High-performance thermoelectrics and challenges for practical devices, Nature Materials 21(5)(2022) 503-513.

[2] X.-Y. Zhang, Y.-L. Li, X.-D. Kong, L. Han, Preparation of MgB_2 Superconducting Microbridge Based on FIB/HPCVD, Micronanoelectronie Technology 58(12)(2021) 1100-1107.

[3] N. Yasuda, S.-I. Takagi, A. Toriumi, Spectral shape analysis of infrared absorption of thermally grown silicon dioxide films, Applied Surface Science 117-118(1997) 216-220.

[4] Y.-Z. Zhang, Q.-Z. Xin, Y.-L. Wang, X.-M. Kong, F. Liu, Z.-S. Yang, Research on Classification and Recognition Method of Steel Metallographic Structure Based on Transfer Learning, Materials Reports 35(24)(2021) 24152-24157.

[5] L. Scarcello, F. Cicirelli, A. Guerrieri, C. Mastroianni, G. Spezzano, A. Vinci, Pursuing Energy Saving and Thermal Comfort with a Human-Driven DRL Approach, IEEE Transactions on Human-Machine Systems 53(4)(2023) 707-719.

[6] M. Weigold, H. Ranzau, S. Schaumann, T. Kohne, N. Panten, E. Abele, Method for the application of deep reinforcement learning for optimised control of industrial energy supply systems by the example of a central cooling system, CIRP Annals 70(1)(2021) 17-20.

[7] G.-B. Awuah, H.-S. Ramaswamy, A. Economides, Thermal processing and quality: Principles and overview, Chemical Engineering and Processing: Process Intensification 46(6)(2007) 584-602.

[8] X. Wang, S. Wang, X.-X. Liang, D.-W. Zhao, J.-C. Huang, X. Xu, B. Dai, Q.-G. Miao, Deep Reinforcement Learning: A Survey, IEEE Transactions on Neural Networks and Learning Systems 35(4)(2024) 5064-5078.

[9] S.-X. Wang, J.-L. Jiao, Z.-W. Huang, The Balance Scheduling of Cloud Platform Resources in Virtual Laboratory, Computer Knowledge and Technology 13(29)(2017) 172-173+190.

[10] M.-A. Ezan, M. Kalfa, Natural Convection of Water Near 4 °C Inside Partially Heated and Cooled Vertical Walls, Isi Bilimi ve Teknigi Dergisi 37(1)(2017) 1-12.

[11] T. Pinomaa, M. Lindroos, M. Walbrühl, N. Provatas, A. Laukkanen, The significance of spatial length scales and solute segregation in strengthening rapid solidification microstructures of 316L stainless steel, Acta Materialia 184(2020) 1-16.

[12] S. Sodjit, V. Uthaisangsuk, Microstructure based prediction of strain hardening behavior of dual phase steels, Materials & Design 41(2012) 370-379.

[13] C.-E. Mazgualdi, T. Masrour, I.-E. Hassani, A. Khdoudi, A Deep Reinforcement Learning (DRL) Decision Model for Heating Process Parameters Identification in Automotive Glass Manufacturing, Advances in Intelligent Systems and Computing (2020) 77-87.

[14] S. Liu, X.-H. Liu, T. Zhang, C.-L. Wang, W. Liu, Joint optimization for temperature and humidity independent control system based on multi-agent reinforcement learning with cooperative mechanisms, Applied Energy 375(C)(2024) 123968.1-123968.16.

[15] Y.-J. Wang, J.-W. Huang, C. Su, H.-G. Li, Furnace thermal efficiency modeling using an improved convolution neural network based on parameter-adaptive mnemonic enhancement optimization, Applied Thermal Engineering 149(2019) 332-343.

[16] A. Silvestri, D. Coraci, S. Brandi, A. Capozzoli, E. Borkowski, J. Köhler, D. Wu, M.-N. Zeilinger, A. Schlueter, Real building implementation of a deep reinforcement learning controller to enhance energy efficiency and indoor temperature control, Applied Energy 368(2024) 123447.1-123447.19.

[17] E. Hachem, A. Vishwasrao, M. Renault, J. Viquerat, P. Meliga, Reinforcement learning for cooling rate control during quenching, International Journal of Numerical Methods for Heat & Fluid Flow 34(8)(2024) 3223-3252.

[18] P. Michailidis, I. Michailidis, E. Kosmatopoulos, Reinforcement Learning for Optimizing Renewable Energy Utilization in Buildings: A Review on Applications and Innovations, Energies 18(7)(2025) 1724.

[19] M. Colledani, T. Tolio, A. Fischer, B. Iung, G. Lanza, R. Schmitt, J. Váncza, Design and management of manufacturing systems for production quality, CIRP Annals 63(2)(2014) 773-796.

[20] Y.-X. Guo, S.-L. Qu, C. Wang, Z.-W. Xing, K.-W. Duan, Optimal dynamic thermal management for data center via soft actor-critic algorithm with dynamic control interval and combined-value state space, Applied Energy 373(2024) 123815.1-123815.25.