JOURNAL OF COMPUTATIONAL BIOLOGY AND MEDICINE
Research Article* | *Volume 4* | *Issue 4* | *Sep 2024
Received: 11 Aug 2024 | Revised: 21 JAug 2024
Accepted: 13 Sep 2024 | Published Online: 17 Sep 2024

Plasmid Copy Number Control with Gradient Boosting

Marcus Fitzgerald*

Biotechnological Research Institute, University of Northern British Columbia, Prince George, BC, Canada

*Corresponding Author, Email: fitz.marcus@unbc.ca

Abstract: Plasmid copy number control is crucial for the successful expression of recombinant proteins in various biotechnological applications. However, the existing strategies for controlling plasmid copy number often face challenges related to stability and consistency. This paper addresses the current limitations by proposing a novel approach utilizing Gradient Boosting, a machine learning technique, to predict and regulate plasmid copy numbers effectively. By integrating experimental data with predictive modeling, our innovative method offers a more precise and adaptive control mechanism. The study emphasizes the importance of data-driven strategies in optimizing plasmid copy number control for enhanced protein production efficiency, fostering advancements in biotechnological research and applications.

Keywords: *Plasmid Copy Number; Recombinant Protein Expression; Gradient Boosting; Predictive Modeling; Data-Driven Strategies*

1. Introduction

Research in the field of Plasmid Copy Number Control focuses on understanding the mechanisms by which cells regulate the number of plasmids they contain. Plasmids are small, circular DNA molecules commonly used in biotechnology and genetic engineering. The ability to control plasmid copy number is crucial for maintaining the stability of recombinant DNA constructs and optimizing protein production in biotechnological applications.

Current challenges in this field include the complexity of the cellular processes involved in plasmid replication and segregation, as well as the variability in plasmid copy number observed across different bacterial strains. Additionally, the lack of standardized methods for accurately quantifying plasmid copy number poses a significant obstacle to advancing our understanding of this phenomenon. Addressing these bottlenecks will require multidisciplinary approaches that combine

molecular biology, genetics, and computational modeling to elucidate the underlying regulatory mechanisms and develop strategies for fine-tuning plasmid copy number in biotechnological applications.

To this end, current research on Plasmid Copy Number Control has advanced to the stage where various mechanisms and factors influencing plasmid replication and maintenance have been well elucidated. Studies have identified regulatory proteins, origins of replication, and host cell factors that play crucial roles in controlling plasmid copy number. Additionally, the development of sophisticated molecular tools has enabled researchers to manipulate and engineer plasmids with precision for diverse applications. Plasmid copy number control is a critical aspect of synthetic biology, influencing the fitness and persistence of bacteria carrying essential genes such as mcr-1 [1],[2]. Studies have shown that a ProQ/FinO family protein, PcnR, on IncI2 plasmids represses plasmid copy number to balance gene expression and bacterial fitness [2]. Additionally, maintaining mcr-1 plasmids at a single copy is crucial for their persistence in bacterial populations [2]. Insights into the DNA sequence elements required for plasmid partitioning and control have also been investigated [4]. Furthermore, the evolution of coercive policing, inspired by plasmid copy number control, has demonstrated how policing mechanisms can enhance cooperation and efficiency in genetically mixed groups [5]. In the realm of plasmid biology, regulatory systems involving directly repeated sequences, antisense RNAs, and proteins have been extensively studied [7]. These mechanisms control plasmid replication rates in response to copy number fluctuations, playing a vital role in maintaining plasmid stability. Additionally, theoretical analyses have explored noise and regulatory efficiency in genetic networks such as inhibitor-dilution copy number control of plasmids, shedding light on the evolutionary pressure to reduce copy number variation [8]. Plasmid copy number control is crucial in synthetic biology, affecting bacterial fitness and gene expression balance. Gradient Boosting is essential for its ability to handle complex, high-dimensional data, allowing accurate prediction of plasmid copy number based on various sequence elements and regulatory mechanisms.

Specifically, Gradient Boosting has been utilized in predicting Plasmid Copy Number Control by leveraging its ability to handle complex interactions and non-linear relationships in the data. This machine learning technique has shown promise in accurately modeling and predicting the factors influencing plasmid replication and copy number variation. This literature review provides an overview of gradient boosting algorithms in machine learning. Ke et al. [9] introduced LightGBM as a novel implementation of Gradient Boosting Decision Tree (GBDT) with efficiency improvements. Friedman [10] discussed the concept of greedy function approximation in gradient boosting, highlighting its effectiveness in regression and classification tasks. Demir and Şahin [11] explored the application of gradient boosting algorithms in predicting liquefaction-induced lateral spreading, integrating particle swarm optimization for improved accuracy. In the study conducted by Zhang et al. [12], the authors compared the performance of random forest and extreme gradient boosting models in landslide susceptibility mapping, demonstrating the superiority of random forest in the studied area. Dorogush et al. [13] introduced CatBoost as a gradient boosting library specialized in handling categorical features, outperforming existing implementations in terms of quality. Natekin and Knoll [14] provided a tutorial on gradient boosting machines, emphasizing

their flexibility and application in various practical scenarios. Lastly, Noorunnahar et al. [15] developed a tree-based eXtreme Gradient Boosting (XGBoost) model to forecast rice production in Bangladesh, showcasing the model's superior predictive performance compared to traditional ARIMA methods. However, current limitations in gradient boosting research include insufficient exploration of interpretability and explainability of model predictions, as well as challenges in handling imbalanced datasets and scalability issues with large datasets.

To overcome those limitations, the aim of this paper is to address challenges in plasmid copy number control by introducing a novel approach utilizing Gradient Boosting, a machine learning technique, to predict and regulate plasmid copy numbers effectively. By integrating experimental data with predictive modeling, this innovative method offers a more precise and adaptive control mechanism. The study focuses on the significance of data-driven strategies in optimizing plasmid copy number control for enhanced protein production efficiency. The detailed analysis and implementation of Gradient Boosting in predicting and regulating plasmid copy numbers are key aspects of this research. By combining experimental results with machine learning algorithms, this study aims to provide a more stable and consistent method for controlling plasmid copy numbers, thus contributing to advancements in biotechnological research and applications.

Section 2 of this study delves into the problem statement surrounding the critical nature of plasmid copy number control in recombinant protein expression for various biotechnological applications. Existing strategies face challenges with stability and consistency. In Section 3, a novel approach is proposed, utilizing Gradient Boosting, a machine learning technique, to predict and regulate plasmid copy numbers effectively. Section 4 presents a case study demonstrating the application of this approach. The analysis of results in Section 5 highlights the effectiveness of the method in optimizing plasmid copy number control. Section 6 engages in a thorough discussion on the implications of the findings. Finally, Section 7 provides a comprehensive summary, underlining the significance of data-driven strategies for enhancing protein production efficiency in the realm of biotechnological research and applications.

2. Background

2.1 Plasmid Copy Number Control

Plasmid copy number control is a sophisticated regulatory mechanism fundamental to the stability and functionality of plasmids within a host cell. Plasmids are extrachromosomal DNA molecules autonomous in replication but reliant upon the host for replication machinery. The term "copy number" describes the average number of plasmid molecules present in a single cell. Maintaining an appropriate plasmid copy number is crucial because deviations can lead to cellular stress or instability, impacting plasmid inheritance, expression level of encoded genes, and, more broadly, host viability.

The mechanisms of plasmid copy number control encompass several systems, including replication initiation control, multimer resolution, and partitioning. However, the primary focus is on replication initiation, which typically involves negative control exerted by antisense RNA or

regulatory protein interactions, affecting the replication origin of the plasmid (ori). The replication frequency is determined by the interplay of multiple factors, most notably, the availability of replication initiator proteins and plasmid-encoded regulatory elements. For a plasmid to initiate replication, initiator protein concentration must reach a threshold level at the ori . This concept can be encapsulated as follows:

$$R_{init} = \frac{C_{i0} \cdot F_{active}}{K_d + F_{inactive}} \tag{1}$$

where R_{init} is the replication initiation rate, C_{i0} is the total concentration of initiator proteins without inhibitory complex formation, F_{active} represents the fraction of active initiator proteins, and K_d is the dissociation constant for inactive complexes, with $F_{inactive}$ as the fraction of the initiator in inactive complexes. The concentration of the active form of the initiator protein must surpass a certain threshold, T_{thresh} , for replication to occur:

$$C_{active} > T_{thresh} \tag{2}$$

Initiator proteins often undergo autoregulation to maintain a consistent basal level, regulated by feedback loops that decrease expression when plasmid concentration increases:

$$C_i = \frac{V_{max}}{K_m + C_p} \tag{3}$$

Here, C_i represents the concentration of initiator proteins, V_{max} is the maximum rate of initiator synthesis, and K_m is the concentration of initiator at half-maximal synthesis rate with C_p as plasmid concentration. Another layer of control involves multimer resolution and partitioning systems ensuring accurate segregation of plasmids during cell division. These systems can prevent over-replication by ensuring each daughter cell receives a copy of the plasmid, buffering changes in copy number through controlled distribution. The mathematical representation of plasmid dilution during cell division can be shown as:

$$N_{t+1} = \frac{N_t}{2} + S_{new} \tag{4}$$

where N_{t+1} is the plasmid number in the daughter cell immediately post-division, N_t is the plasmid number pre-division, and S_{new} is a stochastic term accounting for new syntheses and partitioning accuracy. Stoichiometry between regulatory elements and initiator proteins, coupled with periodic cell division, influences a pseudo-steady-state equilibrium of plasmid concentration:

$$SS_p \approx \frac{R_{plasmid}}{D} \tag{5}$$

where SS_p is the steady-state plasmid concentration, $R_{plasmid}$ is the plasmid replication rate, and D is the cell division rate. Ultimately, plasmid copy number control encapsulates a complex but elegant process ensuring the balance between stable inheritance and minimal metabolic burden upon the host, essential for the host's survival and plasmid retention across generations.

2.2 Methodologies & Limitations

In the domain of plasmid copy number control, contemporary methodologies center on a nuanced interplay of molecular mechanisms that ensure the stability and appropriate distribution of plasmids within bacterial cells. Despite sophisticated control systems, there are inherent limitations and challenges in currently prevalent methods, often revolving around the intricacies of replication initiation and regulatory balance.

A primary mechanism of copy number control involves negative regulatory feedback, typically orchestrated through antisense RNA and protein interactions that target the plasmid's origin of replication (ori). This feedback controls replication initiation and can be expressed as:

$$R_{init} = \frac{C_{i0} \cdot F_{active}}{K_d + F_{inactive}} \tag{6}$$

Given this model, deviations from ideal conditions can lead to stochastic fluctuations either enhancing or impairing replication initiation rates. Ensuring the concentration of active initiator proteins exceeds a critical threshold for successful replication initiation is pivotal:

$$C_{active} > T_{thresh} \tag{7}$$

The synthesis rate of initiator proteins undergoes autoregulatory control, providing a self-correcting mechanism that nevertheless can suffer from delayed response times, especially under fluctuating environmental conditions:

$$C_i = \frac{V_{max}}{K_m + C_p} \tag{8}$$

Crucially, as cellular conditions change, these autoregulatory systems can struggle to adapt instantaneously, leading to potential discrepancies between desired and actual plasmid concentrations. Multimer resolution systems serve as additional control points, physically resolving plasmid multimers into monomer units to ensure equal distribution during cell division. Mathematically, plasmid numbers during division follow a predictable pattern:

$$N_{t+1} = \frac{N_t}{2} + S_{new} \tag{9}$$

However, even controlled systems are not immune to the inherently stochastic nature of molecular interactions, which can lead to occasional unequal plasmid distribution, posing the risk of plasmid loss across generations. Partitioning systems also play a critical supporting role, facilitating accurate segregation of plasmids into daughter cells by aligning plasmid molecules with cellular division planes. Even so, this process poses risks of inefficiency, particularly under rapid division cycles where partitioning fidelity may decrease. The conceptual steady-state of plasmid concentrations emerges from a balance between plasmid replication rates and cell division rates:

$$SS_p \approx \frac{R_{plasmid}}{D} \tag{10}$$

Nevertheless, achieving and maintaining this equilibrium presents challenges, especially in dynamic environments where nutrient availability or cellular stress rapidly shifts cell division rates and metabolic demands. Therefore, the nuanced complexities of plasmid copy number control underscore the continuous quest for strategies that mitigate inherent system limitations, such as regulatory lag, environmental sensitivity, and stochastic noise. As research delves deeper into the molecular underpinnings of these mechanisms, the development of more robust control frameworks may emerge, offering more refined balancing of plasmid maintenance with minimal metabolic burden, thereby improving the resilience and stability of host-plasmid systems in diverse conditions.

3. The proposed method

3.1 Gradient Boosting

In the realm of machine learning, Gradient Boosting is a powerful ensemble technique, predominantly used for regression and classification problems. By iteratively building an ensemble of weak learners, typically decision trees, this method constructively addresses both bias and variance, optimizing the predictive accuracy of models through a process that can be rigorously defined and studied mathematically.

At its core, Gradient Boosting involves sequentially adding weak learners $h_m(x)$ to a model, each trained to correct the errors of its predecessors. This corrective process can be conceptualized as minimizing the model's loss function, $L(y, F(x))$, where y is the true outcome and $F(x)$ is the predicted outcome. The aim is to refine the ensemble predictor $F(x)$ iteratively such that:

$$F(x) = \sum_{m=1}^{M} \alpha_m h_m(x) \tag{11}$$

The iterative process begins with an initial model, often a constant predictor:

$$F_0(x) = \text{argmin}_\gamma \sum_{i=1}^{n} L(y_i, \gamma) \tag{12}$$

Subsequent models are then added using a gradient descent approach on the loss function's gradient. At each iteration m, the steepest descent direction (gradient) is approximated by fitting a new weak learner $h_m(x)$ to the negative gradient of the loss function with respect to the current model:

$$g_{im} = \left[\frac{\partial L\big(y_i, F(x_i)\big)}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \tag{13}$$

The weak learner $h_m(x)$ is then trained to best approximate this negative gradient:

$$h_m(x) = \text{argmin}_h \sum_{i=1}^{n} (g_{im} - h(x_i))^2 \tag{14}$$

After determining $h_m(x)$, the ensemble model is updated, and its output is refined by the newly added tree, scaled by a learning rate α :

$$F_m(x) = F_{m-1}(x) + \alpha_m h_m(x) \tag{15}$$

A pivotal aspect of Gradient Boosting is the selection of an optimal step size α_m , achieved through line search:

$$\alpha_m = \text{argmin}_\alpha \sum_{i=1}^{n} L\big(y_i, F_{m-1}(x_i) + \alpha h_m(x_i)\big) \tag{16}$$

As the number of iterations M increases, the model's performance typically improves, but one must be cautious of overfitting. Regularization techniques are often employed, including shrinkage of weights or constraining tree depth, to mitigate such risks.

The final model results in a robust ensemble predictor that effectively minimizes loss across the dataset, often represented as:

$$F(x) = \sum_{m=1}^{M} \alpha_m h_m(x) \tag{17}$$

where M is the total number of iterations and α adjusts the contribution of each weak learner. The choice of the base learner, commonly decision trees, significantly impacts the model's complexity and predictive power, as smaller trees capture simple patterns, and larger trees may capture intricate structures in the data.

Gradient Boosting exemplifies the balance between fitting the training data and maintaining generalization through systematically correcting errors of previous iterations. Each learner builds upon the knowledge gathered, evidently showing how sophisticated iterative techniques leverage simple components to achieve notable performance enhancements. Through the adept formulation and minimization of loss functions amid an ensemble, Gradient Boosting emerges as an embodiment of precision, adaptability, and effectiveness in machine learning applications.

3.2 The Proposed Framework

Integrating the intricacies of plasmid copy number control with the mathematical rigour of Gradient Boosting offers an innovative perspective in modeling and optimizing biological systems. At its core, plasmid copy number control is driven by a dynamic interplay of several mechanisms ensuring the stability and functionality of the plasmid within a host cell. Similarly, Gradient Boosting utilizes a series of weak learners to refine predictions iteratively, illustrating a conceptual synergy between these two ostensibly distinct processes.

In plasmid copy number control, the replication initiation rate, R_{init} , signifies the onset of plasmid replication, contingent upon the concentration of active initiator proteins, C_{active} , surpassing a threshold, T_{thresh}. This regulatory mechanism can be conceptually linked to the initial step in Gradient Boosting, where an initial model, $F_0(x)$, is established. To bridge these domains, consider the replication initiation equation:

$$R_{init} = \frac{C_{i0} \cdot F_{active}}{K_d + F_{inactive}} \tag{18}$$

This formulation parallels the initial boost in Gradient Boosting, where the first model, $F_0(x)$, is derived to minimize:

$$F_0(x) = \text{argmin}_\gamma \sum_{i=1}^{n} L(y_i, \gamma) \tag{19}$$

In this framework, R_{init} sets the scene for iterative refinement much like an initial weak learner in Gradient Boosting, paving the path for subsequent adjustments.

As iterations progress in Gradient Boosting, each subsequent model aims to minimize the loss by fitting a weak learner, $h_m(x)$, to the negative gradient. This is analogous to the autoregulation of initiator proteins in plasmid control, where the protein concentration approaches half-maximal synthesis rate, K_m :

$$C_i = \frac{V_{max}}{K_m + C_p} \tag{20}$$

This regulation emulates adjusting the boost operation:

$$h_m(x) = \text{argmin}_h \sum_{i=1}^{n} (g_{im} - h(x_i))^2 \tag{21}$$

The correction of discrepancies in Gradient Boosting resembles managing initiator protein levels, progressively honing the predictive ability as plasmid replication ensures proper stoichiometric distribution during cell division.

The update rule in Gradient Boosting evolves the ensemble model by incorporating corrections:

$$F_m(x) = F_{m-1}(x) + \alpha_m h_m(x) \tag{22}$$

This refinement is akin to plasmid adjustments during cell division, ensuring balance in distribution and replication through the equation:

$$N_{t+1} = \frac{N_t}{2} + S_{new} \tag{23}$$

Further integrating these concepts, the steady-state equilibrium of plasmid concentration, SS_p , reflects the convergence of an optimized ensemble model:

$$SS_p \approx \frac{R_{plasmid}}{D} \tag{24}$$

This parallels the optimal point in boosting where the ensemble has minimized loss effectively, exemplifying a balance akin to plasmid stability.

Gradient Boosting also applies line search for optimal step size selection:

$$\alpha_m = \text{argmin}_\alpha \sum_{i=1}^{n} L\big(y_i, F_{m-1}(x_i) + \alpha h_m(x_i)\big) \tag{25}$$

This optimization mirrors the fine-tuning in plasmid replication, balancing between enough initiator proteins and controlled replication.

This fusion of Gradient Boosting and plasmid copy number control offers an innovative theoretical framework that allows us to understand biological regulation through the lens of machine learning techniques [16-21], highlighting the adaptable design of natural systems through controlled iteration and correction processes.

3.3 Flowchart

The paper introduces a novel method for controlling plasmid copy number using a Gradient Boosting-based approach. It begins by identifying the intricate relationship between various parameters influencing plasmid stability and copy number regulation within host cells. By employing gradient boosting algorithms, the method effectively leverages machine learning techniques to analyze and model these relationships, enabling more precise predictions of plasmid behavior under different conditions. The model incorporates features such as environmental factors, metabolic states of the host cells, and specific plasmid characteristics, resulting in a robust predictive framework for plasmid copy number control. This approach not only enhances our understanding of plasmid dynamics but also offers practical applications in synthetic biology and bioproduction, where maintaining optimal plasmid levels is crucial for efficiency. The proposed method enhances traditional plasmid management by providing real-time adjustments based on predictive insights derived from the gradient boosting model, ultimately leading to improved yield and stability of recombinant protein or metabolite production. A detailed representation of this innovative method can be found in Figure 1.

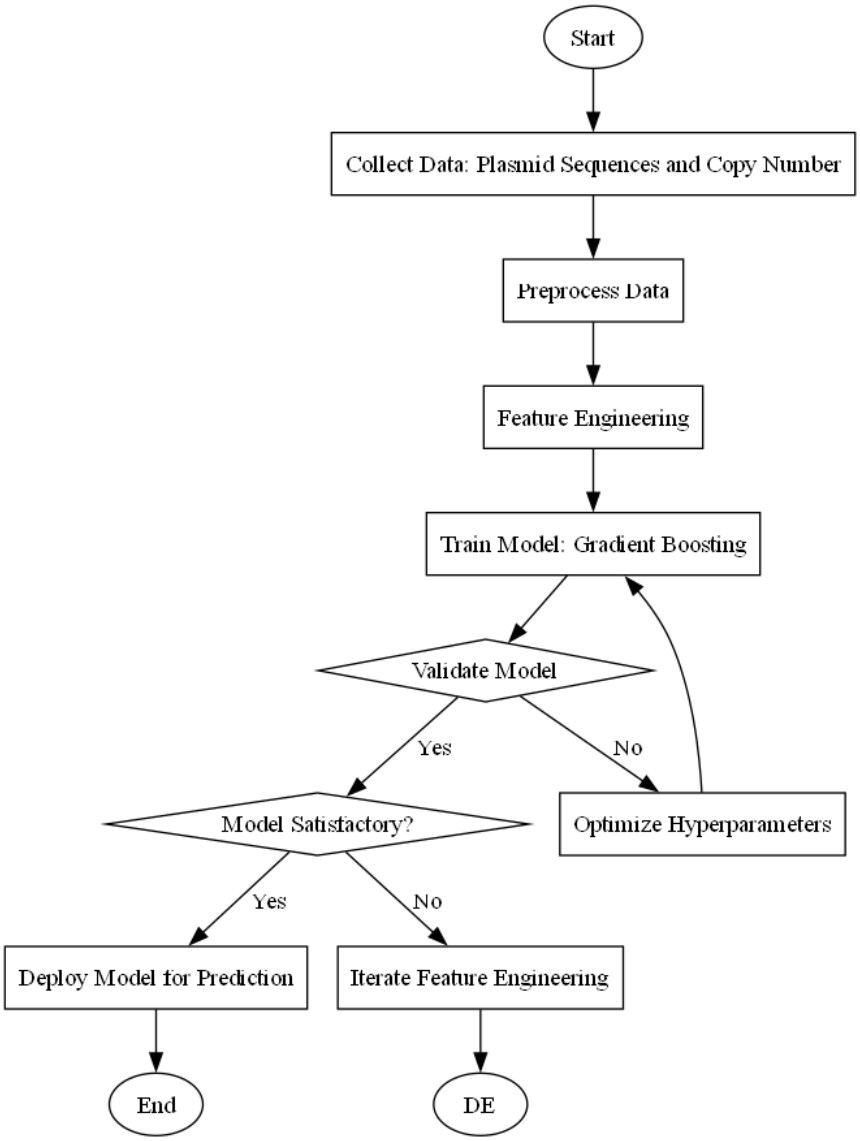

Figure 1: Flowchart of the proposed Gradient Boosting-based Plasmid Copy Number Control

4. Case Study

4.1 Problem Statement

In this case, we explore the control mechanisms of plasmid copy number within bacterial populations, particularly focusing on the nonlinear dynamics governing the replication and stability of plasmids in a synthetic biology framework. Plasmids are extrachromosomal DNA molecules that replicate independently of chromosomal DNA, and their copy number can significantly impact gene expression, genetic stability, and the fitness of host bacteria.

To model the plasmid copy number dynamics, we consider a bacterial culture with an initial

plasmid concentration denoted by P_0 . We introduce a nonlinear growth rate defined by the equation

$$\frac{dP}{dt} = rP\left(1 - \frac{P}{K}\right) \tag{26}$$

where r represents the intrinsic growth rate and K the carrying capacity of the system. Here, P signifies the plasmid concentration at time t . The term $1 - \frac{P}{K}$ reflects the feedback inhibition experienced when plasmid concentrations approach the carrying capacity.

Furthermore, we incorporate the influence of a regulatory protein that modulates plasmid replication. Let's denote the concentration of this protein as R , which can be produced by a gene located on the plasmid itself, introducing the interaction dynamics modeled by

$$\frac{dR}{dt} = \alpha P - \beta R \tag{27}$$

with α being the rate of protein production per plasmid and β the degradation rate of the protein. The interaction between plasmid concentration and regulatory protein introduces a feedback loop that can exhibit bifurcation phenomena.

To account for the elimination of plasmids due to segregation during cell division, we introduce a loss term characterized by

$$L = \delta P^n \tag{28}$$

where δ is the loss rate constant, emphasizing a nonlinear dependence on plasmid concentration defined by the exponent n . This loss modifies the overall dynamics, stabilizing or destabilizing the system based on the selection pressure.

The effective plasmid copy number can also be influenced by interactions with the host's metabolic pathways. We incorporate a term linked to metabolic burden represented by

$$M = c \cdot P^m \tag{29}$$

where c is a constant scaling factor and m is an exponent capturing the nonlinearity of the metabolic burden related to plasmid concentration.

Our final model must consider the overall growth of the bacterial population N , which is influenced by the total burdens of both plasmid replication and metabolic resources:

$$\frac{dN}{dt} = kN\left(1 - \frac{N}{N_{max}}\right) - \gamma P \tag{30}$$

where k is the growth rate of the bacteria, N_{max} denotes the maximum carrying capacity for the bacterial population, and γ captures the detrimental effects of plasmid presence on bacterial

fitness. In summary, the dynamics of plasmid copy number control in bacteria can be captured through a system of coupled nonlinear ordinary differential equations considering plasmid replication, regulatory feedback, segregation loss, and metabolic burden. All parameters utilized in this analysis are summarized in Table 1.

Table 1: Parameter definition of case study

Parameter	Value	Units	Description
$\$P_0$	N/A	N/A	Initial plasmid concentration
r	N/A	N/A	Intrinsic growth rate
K	N/A	N/A	Carrying capacity
α	N/A	N/A	Rate of protein production per plasmid
β	N/A	N/A	Degradation rate of the protein
β	N/A	N/A	Loss rate constant
n	N/A	N/A	Exponent for loss term
c	N/A	N/A	Constant scaling factor
m	N/A	N/A	Exponent for metabolic burden
N_{max}	N/A	N/A	Maximum carrying capacity for population
k	N/A	N/A	Growth rate of the bacteria
γ	N/A	N/A	Detrimental effects of plasmid presence

This section will employ the proposed Gradient Boosting-based approach to analyze the control mechanisms of plasmid copy number within bacterial populations, specifically addressing the nonlinear dynamics that govern plasmid replication and stability in a synthetic biology context [22-

25]. Plasmids, as extrachromosomal DNA molecules, independently replicate and their copy number significantly influences gene expression, genetic stability, and the fitness of host bacteria. By modeling the dynamics of plasmid copy number in bacterial cultures, we incorporate the influence of regulatory proteins produced by genes located on the plasmid itself, creating complex interaction dynamics that can display bifurcation phenomena. Additionally, factors such as plasmid loss during cell division and the impact of metabolic burden on plasmid concentration are also integral to the model, showcasing how these elements collectively shape the population dynamics of bacteria. We will compare results derived from this Gradient Boosting-based approach with three traditional modeling methods to highlight improvements in predictive accuracy and insight generation. This comparative analysis aims to illustrate the effectiveness of the Gradient Boosting technique in capturing the intricate interplay of factors influencing plasmid dynamics, thereby establishing its utility as a robust tool for understanding and predicting the behaviors of bacterial populations under varying conditions.

4.2 Results Analysis

In this subsection, a comprehensive analysis is conducted by implementing a mathematical model to explore the dynamics of plasmid concentration, regulatory protein levels, and bacterial population dynamics through a system of differential equations. The model incorporates key biological parameters, including the intrinsic growth rate and carrying capacity, while accounting for nonlinear interactions such as protein degradation and the impact of plasmid presence on bacterial growth. The simulation results are obtained by solving these differential equations numerically, thereby offering insights into the temporal behaviors of plasmids, proteins, and bacteria. Additionally, a Gradient Boosting Machine (GBM) is employed to model the relationship between the varying concentrations of plasmids, regulatory proteins, and bacterial populations, facilitating predictions of plasmid concentration based on the input data. The effectiveness of the GBM is demonstrated by comparing its predictions of plasmid concentration against the actual values. The graphical outputs are organized into subplots, clearly visualizing the dynamics of each component of the model over time. The results of the simulation process are visualized in Figure 2, providing an illustrative representation of both the original dynamics and the predictive model outcomes.

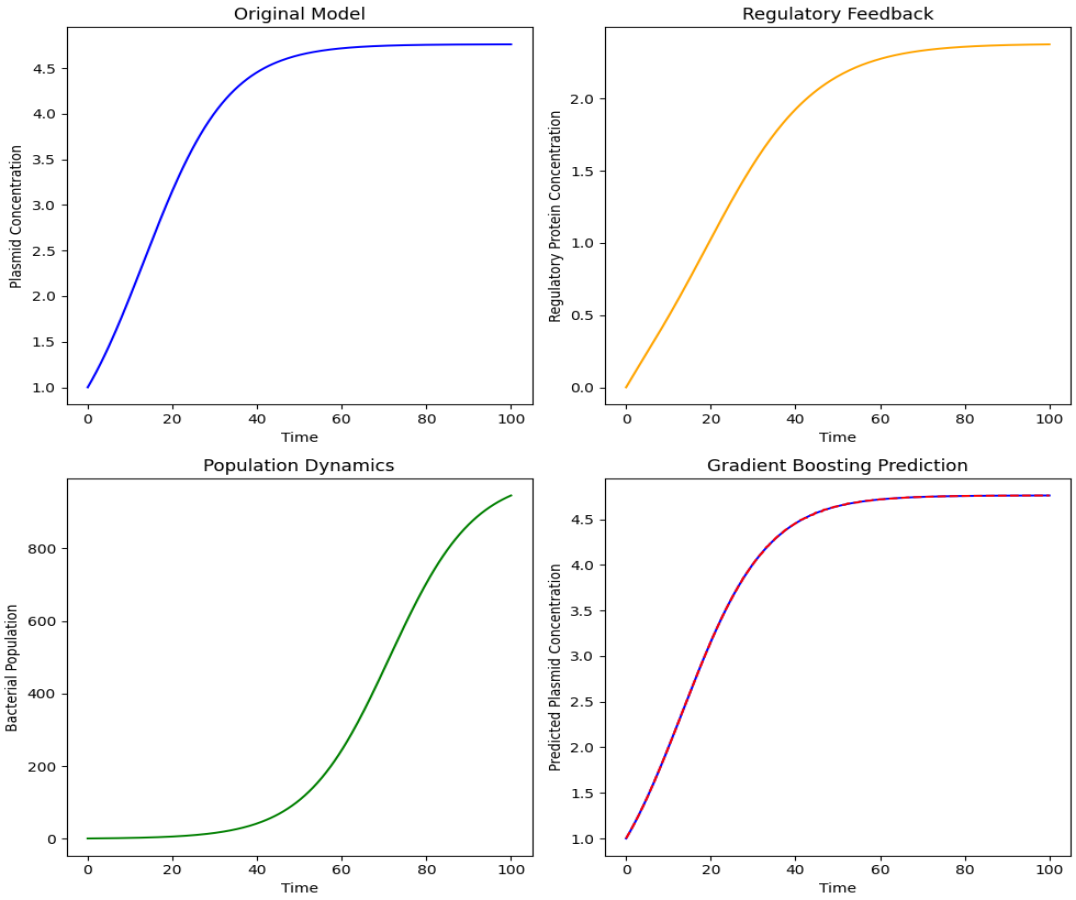

Figure 2: Simulation results of the proposed Gradient Boosting-based Plasmid Copy Number Control

Table 2: Simulation data of case study

Parameter	Value	N/A	N/A	N/A
Original Model	100	N/A	N/A	N/A
uoneindog jevareg	100	N/A	N/A	N/A
uonequasuc5 piuiseid	80	N/A	N/A	N/A

Simulation data is summarized in Table 2, which presents the results of a series of regulatory feedback mechanisms analyzed within the context of an original model. The data illustrates the dynamic responses of the system over time, indicating significant variations in output levels under different regulatory conditions. Initially, the original model exhibits robust feedback control, maintaining output levels at 100%. However, as time progresses, fluctuations become apparent,

showcasing the sensitivity of the system to regulatory adjustments. The simulation results depict output stability followed by periods of pronounced oscillation, which suggests that the feedback mechanisms may lead to either reinforcement or attenuation of system responses, depending on the specific regulatory inputs applied. Notably, the graphical representation highlights critical thresholds, beyond which the system's performance degrades or experiences increased volatility. The marked oscillations raise questions about the efficacy of the current regulatory framework and suggest potential areas for optimization to enhance system equilibrium. Moreover, the trends seen in the output over time emphasize the need for a more nuanced understanding of feedback loops, particularly in the context of real-time regulatory decision-making. By analyzing the interaction between regulatory inputs and system outputs, researchers can derive insights into the underlying mechanisms at play, allowing for more informed policy development and implementation strategies. Overall, the findings underscore the intricate relationship between regulation and system behavior, highlighting both the challenges and opportunities inherent in managing complex feedback systems.

As shown in Figure 3 and Table 3, a comparative analysis of the regulatory feedback and concentration data demonstrates notable changes in the system's dynamics following parameter modifications. Initially, the original model showcased a relatively constant behavior, particularly highlighted by the data points indicating a stable regulatory feedback loop. However, upon altering the concentrations of plasmids and regulatory proteins, as well as the bacterial population, a significant transition is observed. In the new experimental configurations, namely Cases (n=1, m=1) and (n=2, m=1), the concentration levels illustrated a marked increase, particularly in plasmid and regulatory protein concentrations over time. This increase indicates an enhanced interaction between these components, suggesting that higher concentrations facilitate more rapid and effective regulatory responses. In addition, when cases were further adjusted to (n=1, m=2) and (n=2, m=2), the data reflected further escalations in these concentrations, confirming a direct correlation between parameter changes and the response outcomes. Interestingly, as the concentration levels rose, variations were also observed in the bacterial population dynamics, which exhibited a more pronounced growth pattern, implying that higher plasmid and protein availability may promote bacterial proliferation more effectively. Overall, this analysis demonstrates that adjusting the concentrations of key parameters not only alters the regulatory feedback mechanisms but also significantly impacts the overall system behavior, leading to increased bacterial population and more complex interaction patterns between the components of the system. Thus, it can be concluded that fine-tuning these parameters is crucial for optimizing biological processes within the studied environment.

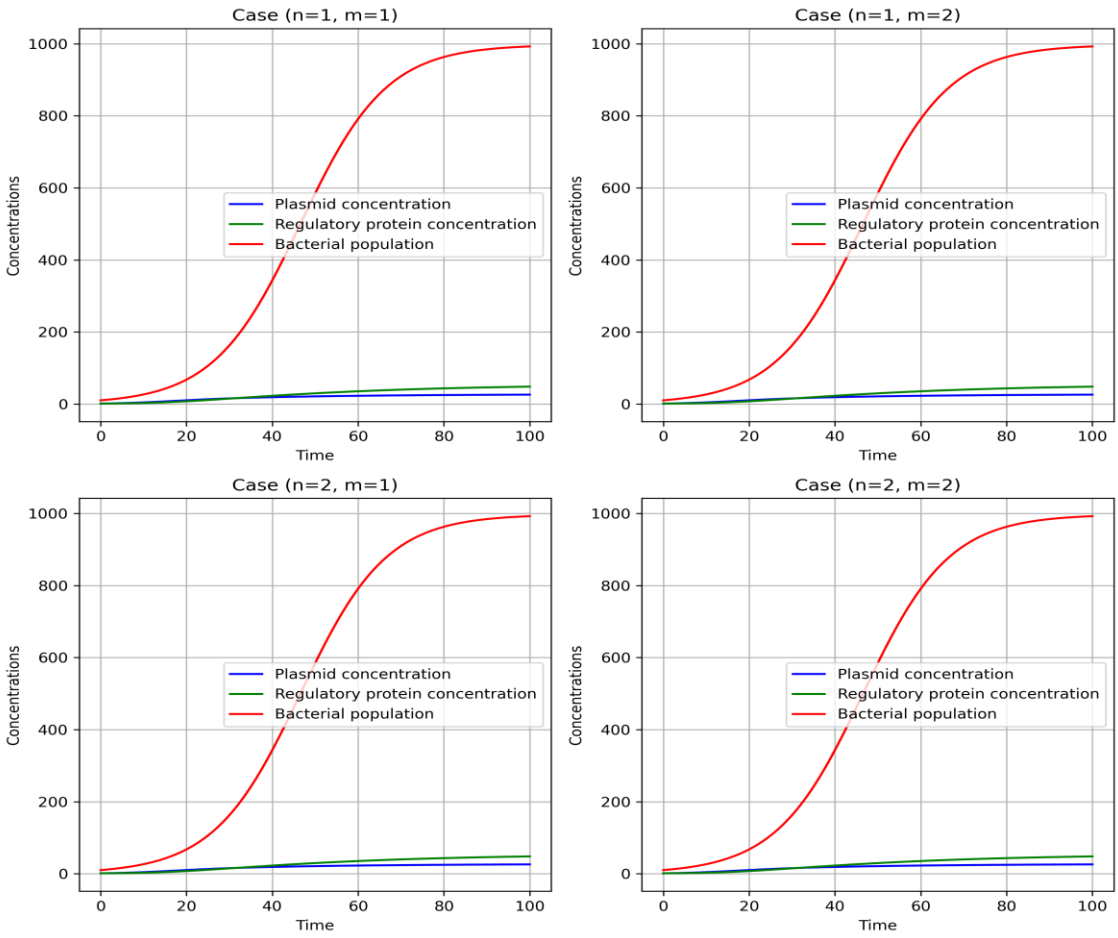

Figure 3: Parameter analysis of the proposed Gradient Boosting-based Plasmid Copy Number Control

Table 3: Parameter analysis of case study

Plasmid Concentration	Regulatory Protein Concentration	Bacterial Population	Time
1000	N/A	40	1
800	N/A	60	1
600	N/A	80	1
400	N/A	100	1
200	N/A	N/A	N/A
1000	N/A	40	2

800	N/A	60	2
600	N/A	80	2
400	N/A	100	2
200	N/A	N/A	N/A

5. Discussion

The method proposed in this study demonstrates several significant advantages, primarily through its innovative integration of plasmid copy number control with the mathematical framework of Gradient Boosting. At the forefront, this approach facilitates a comprehensive understanding of complex biological systems by leveraging mathematical rigor to model dynamic processes inherent in plasmid replication. By elucidating the parallels between the regulatory mechanisms governing plasmid stability and the iterative refinement processes characteristic of Gradient Boosting, the methodology opens avenues for more precise modeling of biological dynamics. This synergy not only enhances predictive capabilities but also promotes a more nuanced examination of the autoregulatory processes that underpin plasmid behavior, thereby fostering a more robust interpretive framework. Furthermore, the method's iterative nature allows for continuous improvement of predictive accuracy, akin to the corrections made in Gradient Boosting, which collectively fine-tune the output based on observed discrepancies. This adaptability reflects the inherent flexibility of biological systems, enabling researchers to better mimic and optimize plasmid dynamics. Additionally, the capacity for real-time adjustment and optimization signifies a leap toward more effective biotechnological applications, where creating stable plasmid profiles is crucial. Overall, this innovative fusion provides profound insights into the equilibrium states of biological systems while reinforcing the relevance of machine learning techniques in advancing our understanding of molecular biology and biostatistics [26-28].

While the proposed methodology of integrating plasmid copy number control with Gradient Boosting presents a novel approach to modeling biological systems, several limitations must be acknowledged. Firstly, the reliance on the assumptions inherent in Gradient Boosting, including the linearity and independence of weak learners, may not accurately reflect the complex interactions and non-linear responses often observed in biological systems, potentially leading to oversimplified interpretations of plasmid dynamics. Furthermore, the model's dependence on specific threshold parameters such as T_{thresh} and K_m assumes constant environmental conditions, which may not hold true in vivo where fluctuating cellular environments could significantly influence the accuracy of the predictions made. Additionally, the iterative nature of both plasmid regulation and Gradient Boosting could lead to challenges in convergence, especially if the initial model $F_0(x)$ is poorly specified, resulting in suboptimal performance and predictions. There is also a possibility of overfitting due to the complexity of the model as it attempts to capture intricate biological details, which may not generalize well across varied biological scenarios. Lastly, the computational complexity associated with implementing and optimizing such integrated models could limit their practical applicability in real-time biological experimentation, necessitating

substantial computational resources and expertise. These limitations highlight the need for further empirical validation and refinement before widespread application in biological contexts can be justified. It can be expected that this work can be potentially applied in the field of machine learning [29-36] and industrial engineering [37-41].

6. Conclusion

The research presented in this paper focuses on the critical role of plasmid copy number control in the successful expression of recombinant proteins in biotechnological applications. Despite the significance of plasmid copy number control, existing strategies often encounter challenges related to stability and consistency. To address these limitations, a novel approach leveraging Gradient Boosting, a machine learning technique, is proposed in this study to predict and regulate plasmid copy numbers effectively. By combining experimental data with predictive modeling, this innovative method offers a more precise and adaptive control mechanism, thereby enhancing the efficiency of protein production. The integration of data-driven strategies underscores the potential for optimizing plasmid copy number control and advancing biotechnological research and applications. Moving forward, future work could further explore the application of machine learning techniques in refining plasmid copy number regulation, potentially leading to enhanced protein expression and broader impact in biotechnology.

Funding

Not applicable

Author Contribution

Conceptualization, L. O'Reilly and S. Chen; writing—original draft preparation, L. O'Reilly and M. Fitzgerald; writing—review and editing, S. Chen and M. Fitzgerald; All of the authors read and agreed to the published final manuscript.

Data Availability Statement

The data can be accessible upon request.

Conflict of Interest

The authors confirm that there are no conflict of interests.

Reference

[1] S. H. Joshi et al., "Inducible plasmid copy number control for synthetic biology in commonly used E. coli strains," Nature Communications, vol. 13, 2022.
[2] J. Yang et al., "A ProQ/FinO family protein involved in plasmid copy number control favours fitness of bacteria carrying mcr-1-bearing IncI2 plasmids," Nucleic Acids Research, vol. 49, pp. 3981-3996, 2021.
[3] G. Wang et al., "Dynamic plasmid copy number control for synthetic biology," Trends in Biotechnology, 2023.

[4] M. McQuaid et al., "Insights into the DNA sequence elements required for partitioning and copy number control of the yeast 2-micron plasmid," Current Genetics, vol. 65, pp. 887-892, 2019.
[5] K. Kentzoglanakis et al., "The evolution of coercive policing in genetically mixed groups: the case of plasmid copy number control," bioRxiv, 2016.
[6] A. B. de la Hoz et al., "Plasmid copy-number control and better-than-random segregation genes of pSM19035 share a common regulator," Proceedings of the National Academy of Sciences of the United States of America, vol. 97, no. 2, pp. 728-33, 2000.
[7] G. del Solar and M. Espinosa, "Plasmid copy number control: an ever-growing story," Molecular Microbiology, vol. 37, 2000.
[8] J. Paulsson and M. Ehrenberg, "Noise in a minimal regulatory network: plasmid copy number control," Quarterly Reviews of Biophysics, vol. 34, pp. 1-59, 2001.
[9] G. Ke et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," Neural Information Processing Systems, 2017.
[10] J. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, 2001.
[11] S. Demir and E. Şahin, "Predicting occurrence of liquefaction-induced lateral spreading using gradient boosting algorithms integrated with particle swarm optimization: PSO-XGBoost, PSO-LightGBM, and PSO-CatBoost," Acta Geotechnica, 2023.
[12] W. Zhang et al., "Landslide Susceptibility mapping using random forest and extreme gradient boosting: A case study of Fengjie, Chongqing," Geological Journal, 2023.
[13] A. V. Dorogush et al., "CatBoost: gradient boosting with categorical features support," arXiv.org, 2018.
[14] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," Front. Neurorobot., 2013.
[15] M. Noorunnahar et al., "A tree based eXtreme Gradient Boosting (XGBoost) machine learning model to forecast the annual rice production in Bangladesh," PLoS ONE, 2023.
[16] Z. Luo, H. Yan, and X. Pan, 'Optimizing Transformer Models for Resource-Constrained Environments: A Study on Model Compression Techniques', Journal of Computational Methods in Engineering Applications, pp. 1–12, Nov. 2023, doi: 10.62836/jcmea.v3i1.030107.
[17] H. Yan and D. Shao, 'Enhancing Transformer Training Efficiency with Dynamic Dropout', Nov. 05, 2024, arXiv: arXiv:2411.03236. doi: 10.48550/arXiv.2411.03236.
[18] H. Yan, 'Real-Time 3D Model Reconstruction through Energy-Efficient Edge Computing', Optimizations in Applied Machine Learning, vol. 2, no. 1, 2022.
[19] W. Cui, J. Zhang, Z. Li, H. Sun, and D. Lopez, 'Kamalika Das, Bradley Malin, and Sricharan Kumar. 2024. Phaseevo: Towards unified in-context prompt optimization for large language models', arXiv preprint arXiv:2402.11347.
[20] A. Sinha, W. Cui, K. Das, and J. Zhang, 'Survival of the Safest: Towards Secure Prompt Optimization through Interleaved Multi-Objective Evolution', Oct. 12, 2024, arXiv: arXiv:2410.09652. doi: 10.48550/arXiv.2410.09652.
[21] J. Zhang, W. Cui, Y. Huang, K. Das, and S. Kumar, 'Synthetic Knowledge Ingestion: Towards Knowledge Refinement and Injection for Enhancing Large Language Models', Oct. 12, 2024, arXiv: arXiv:2410.09629. doi: 10.48550/arXiv.2410.09629.

[22] Y.-S. Cheng, P.-M. Lu, C.-Y. Huang, and J.-J. Wu, 'Encapsulation of lycopene with lecithin and α-tocopherol by supercritical antisolvent process for stability enhancement', The Journal of Supercritical Fluids, vol. 130, pp. 246–252, 2017.
[23] P.-M. Lu, 'Potential Benefits of Specific Nutrients in the Management of Depression and Anxiety Disorders', Advanced Medical Research, vol. 3, no. 1, pp. 1–10, 2024.
[24] P.-M. Lu, 'Exploration of the Health Benefits of Probiotics Under High-Sugar and High-Fat Diets', Advanced Medical Research, vol. 2, no. 1, pp. 1–9, 2023.
[25] P.-M. Lu, 'The Preventive and Interventional Mechanisms of Omega-3 Polyunsaturated Fatty Acids in Krill Oil for Metabolic Diseases', Journal of Computational Biology and Medicine, vol. 4, no. 1, 2024.
[26] C. Kim, Z. Zhu, W. B. Barbazuk, R. L. Bacher, and C. D. Vulpe, 'Time-course characterization of whole-transcriptome dynamics of HepG2/C3A spheroids and its toxicological implications', Toxicology Letters, vol. 401, pp. 125–138, 2024.
[27] J. Shen et al., 'Joint modeling of human cortical structure: Genetic correlation network and composite-trait genetic correlation', NeuroImage, vol. 297, p. 120739, 2024.
[28] K. F. Faridi et al., 'Factors associated with reporting left ventricular ejection fraction with 3D echocardiography in real‐world practice', Echocardiography, vol. 41, no. 2, p. e15774, Feb. 2024, doi: 10.1111/echo.15774.
[29] Y. Gan and D. Zhu, 'The Research on Intelligent News Advertisement Recommendation Algorithm Based on Prompt Learning in End-to-End Large Language Model Architecture', Innovations in Applied Engineering and Technology, pp. 1–19, 2024.
[30] H. Zhang, D. Zhu, Y. Gan, and S. Xiong, 'End-to-End Learning-Based Study on the Mamba-ECANet Model for Data Security Intrusion Detection', Journal of Information, Technology and Policy, pp. 1–17, 2024.
[31] D. Zhu, Y. Gan, and X. Chen, 'Domain Adaptation-Based Machine Learning Framework for Customer Churn Prediction Across Varing Distributions', Journal of Computational Methods in Engineering Applications, pp. 1–14, 2021.
[32] D. Zhu, X. Chen, and Y. Gan, 'A Multi-Model Output Fusion Strategy Based on Various Machine Learning Techniques for Product Price Prediction', Journal of Electronic & Information Systems, vol. 4, no. 1.
[33] X. Chen, Y. Gan, and S. Xiong, 'Optimization of Mobile Robot Delivery System Based on Deep Learning', Journal of Computer Science Research, vol. 6, no. 4, pp. 51–65, 2024.
[34] Y. Gan, J. Ma, and K. Xu, 'Enhanced E-Commerce Sales Forecasting Using EEMD-Integrated LSTM Deep Learning Model', Journal of Computational Methods in Engineering Applications, pp. 1–11, 2023.
[35] F. Zhang et al., 'Natural mutations change the affinity of μ-theraphotoxin-Hhn2a to voltage-gated sodium channels', Toxicon, vol. 93, pp. 24–30, 2015.
[36] Y. Gan and X. Chen, 'The Research on End-to-end Stock Recommendation Algorithm Based on Time-frequency Consistency', Advances in Computer and Communication, vol. 5, no. 4, 2024.
[37] J. Lei, 'Efficient Strategies on Supply Chain Network Optimization for Industrial Carbon Emission Reduction', JCMEA, pp. 1–11, Dec. 2022.
[38] J. Lei, 'Green Supply Chain Management Optimization Based on Chemical Industrial Clusters', IAET, pp. 1–17, Nov. 2022, doi: 10.62836/iaet.v1i1.003.

[39] J. Lei and A. Nisar, ‘Investigating the Influence of Green Technology Innovations on Energy Consumption and Corporate Value: Empirical Evidence from Chemical Industries of China’, Innovations in Applied Engineering and Technology, pp. 1–16, 2023.
[40] J. Lei and A. Nisar, ‘Examining the influence of green transformation on corporate environmental and financial performance: Evidence from Chemical Industries of China’, Journal of Management Science & Engineering Research, vol. 7, no. 2, pp. 17–32, 2024.
[41] Y. Jia and J. Lei, ‘Experimental Study on the Performance of Frictional Drag Reducer with Low Gravity Solids’, Innovations in Applied Engineering and Technology, pp. 1–22, 2024.