

Article

Computationally Efficient Optimization of Bio-Jet Fuel Supply Chains Using Machine-Learning-Assisted Mixed-Integer Programming

Krystel K. Castillo-Villar ^{1,*} , Kolton Keith ¹  and Adel Alaeddini ² 

¹ Mechanical Engineering Department, Texas Sustainable Energy Research Institute, The University of Texas at San Antonio, 1 UTSA Circle, San Antonio, TX 78249, USA; kolton.keith@utsa.edu

² Department of Mechanical Engineering, Southern Methodist University, Dallas, TX 75205, USA; aalaeddini@smu.edu

* Correspondence: krystel.castillo@utsa.edu; Tel.: +1-210-458-6702

Abstract

Bio-jet fuels produced from biomass-derived feedstocks represent a promising pathway for reducing the carbon intensity of aviation energy systems. However, designing supply chain networks for bio-jet fuel production requires solving large-scale mixed-integer linear programming (MILP) models that integrate facility location, feedstock allocation, material flows, and routing decisions. These models can become computationally expensive, particularly when evaluating multiple network configurations or large candidate sets of production and processing facilities. This study develops a hybrid machine learning and optimization framework to improve the computational efficiency of bio-jet fuel supply chain network design while preserving high-quality decision outcomes. The proposed iterative procedure uses supervised learning to approximate the relationship between facility location decisions and total supply chain cost. First, an initial set of supply chain configurations is generated by solving the optimization model while using randomly selected facility locations. These solutions are then used to train predictive models, including ridge regression, feed-forward neural networks, and ensemble neural networks, with facility-location configurations as inputs and total supply chain cost as the output. The trained learner is subsequently used to identify promising facility-location candidates through Thompson sampling and small-scale linear programming. These candidate solutions are evaluated by the original mixed-integer model, and the resulting observations are fed back into the learning process until convergence. Numerical experiments show that the proposed hybrid approach obtains near-optimal bio-jet fuel supply chain designs while substantially reducing computational time. For the linear case, the method achieves solutions within 0.23–0.29% of the objective function value while reducing computational time by 70.95–81.95%. For nonlinear learning models, the optimality gap decreases further to 0.13–0.15%, with computational time reductions of 45.37–60.36%. For the Texas case study and the modeling assumptions evaluated, the findings demonstrate that machine-learning-assisted optimization can reduce computational effort while preserving high-quality supply chain solutions. The extent of these benefits may vary with network size, candidate-facility structure, facility-capacity assumptions, demand characteristics, and the amount of information available to train the learning models.



Academic Editors: Ana Paula Soares Dias and Frederico G Fonseca

Received: 7 July 2026

Revised: 24 July 2026

Accepted: 28 July 2026

Published: 29 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: bio-jet fuel; biomass supply chain; mixed-integer linear programming; machine learning; supply chain network design; neural network; bioenergy systems

network-design decisions and total supply chain cost. The resulting procedure creates an iterative feedback loop in which optimization generates high-fidelity performance observations, while machine learning guides the search toward increasingly promising regions of the solution space.

In the application considered in this study, the strategic decision assigned to the learning component is biorefinery location. The learner is initially trained using solutions obtained from the reduced supply chain model under predefined biorefinery-location configurations. Based on this information, it estimates the cost implications of alternative facility-location patterns and proposes new configurations for evaluation. Although the method requires solving several reduced optimization models, each individual model is substantially smaller than the original full-scale formulation. As shown in the computational experiments, this learning-assisted decomposition can produce near-optimal biofuel supply chain designs while significantly reducing the total solution time.

The contribution of this paper is threefold. First, it presents a machine-learning-assisted solution strategy for computationally intensive mixed-integer biofuel supply chain design models. Second, it evaluates alternative supervised learning models, including linear and nonlinear predictors, for estimating the cost performance of facility-location configurations. Third, it demonstrates the computational value of the proposed framework in an industrial case study of the bio-jet fuel supply chain, showing that near-optimal solutions can be obtained with substantial reductions in computational time.

The remainder of this paper is organized as follows. Section 2 reviews the literature on biofuel supply chain optimization and learning-assisted optimization methods. Section 3 presents the mathematical formulation of the bio-jet fuel supply chain network design model. It also describes the industrial case study. Section 4 introduces the proposed learning-optimization procedures and discusses the numerical experimentation and computational performance. Finally, Section 5 concludes the paper with implications for biofuel supply chain design and future research directions.

2. Literature Review

Biofuel and biomass supply chain network design has been widely studied using mathematical programming models that integrate facility location, feedstock allocation, transportation, conversion, inventory, and demand satisfaction decisions. These models are especially relevant for strategic planning because biomass resources are geographically dispersed, transportation costs are distance-sensitive, and conversion facilities require substantial capital investment. Prior studies have formulated biomass and biofuel supply chain design problems using mixed-integer linear programming, stochastic programming, and hub-and-spoke network structures to account for facility opening decisions, material flows, conversion yields, and uncertainty in feedstock supply or demand [2–5]. Reviews of biomass-to-bioenergy and biofuel supply chain optimization have emphasized that cost-effective network design requires the simultaneous consideration of feedstock availability, preprocessing, transportation infrastructure, conversion technologies, and final fuel distribution [2,6].

A major challenge in this literature is the computational burden associated with large-scale facility location and network design models. Facility location problems are generally NP-hard, and the size of the solution space increases rapidly as the number of candidate facilities, transportation arcs, scenarios, or technology alternatives expands [7,8]. In biofuel applications, this challenge is amplified by the need to represent geographically detailed feedstock availability, multimodal transportation options, biomass quality variability, and capacity constraints. Several studies have addressed these issues through stochastic programming and decomposition-based methods. For example, Marufuzzaman et al. [3]

developed a two-stage stochastic programming model for biodiesel production, while Castillo-Villar et al. [4] and Aboytes-Ojeda et al. [5] incorporated biomass quality variability into supply chain optimization models. Other work has considered price-dependent biomass availability, yield uncertainty, and climate-related storage losses [9–12]. These studies provide important modeling foundations for biofuel supply chain design, but large candidate sets and repeated scenario evaluation remain computationally demanding.

To improve computational tractability, researchers have also proposed heuristic, metaheuristic, and hybrid decomposition approaches for biomass and biofuel supply chain problems. Metaheuristics are commonly used for hard combinatorial optimization problems when exact methods become computationally expensive [13]. In supply chain and hub-location contexts, genetic algorithms, simulated annealing, and robust metaheuristic approaches have been used to search large solution spaces more efficiently [14,15]. In the biomass supply chain literature, Aboytes-Ojeda et al. [16] proposed a hybrid metaheuristic and exact-method approach for a two-stage stochastic biofuel hub-and-spoke network problem. Such approaches can reduce computational time and improve scalability, but they often require problem-specific algorithm design, parameter tuning, or specialized search operators.

More recently, machine learning has been introduced as a complementary tool for optimization and supply chain decision-making. Rather than replacing mathematical optimization, machine learning can be used to approximate expensive model components, identify promising regions of the feasible space, support policy selection, or learn decision rules from optimization outputs. Defourny et al. [17] used machine learning to support policy selection in multistage stochastic programming, while Oroojlooyjadid et al. [18] demonstrated how deep learning can integrate prediction and optimization in inventory decision-making. In biomass supply chain design, Goettsch et al. [19] developed a hybrid machine learning and stochastic MILP approach to select potential depot locations for a biomass co-firing supply chain. Their method used machine learning classifiers to screen candidate depot locations before solving the optimization model, thereby reducing the computational burden of the stochastic MILP. This work is closely related to the present study because it demonstrates the value of combining machine learning with mathematical programming in biomass supply chain network design.

Recent reviews further demonstrate the growing integration of machine learning and mathematical optimization. Scavuzzo et al. [20] reviewed machine-learning-enhanced mixed-integer linear programming methods and showed that learning-based approaches have been applied to several components of branch-and-bound algorithms, including branching, node selection, cutting-plane selection, primal heuristics, and solver configuration. Their review emphasizes that machine learning is increasingly used as a complementary mechanism for improving the efficiency of established optimization procedures rather than as a replacement for mathematical programming. Others have also expanded the methodological understanding of sustainable aviation fuel (SAF) supply chains. Liang et al. [21] systematically reviewed advances in SAF supply chain modeling and optimization, including mathematical programming, machine learning, and multi-objective approaches. They identified feedstock availability, production scalability, logistics, cost competitiveness, and policy uncertainty as continuing challenges and highlighted the need for computational methods capable of supporting integrated and scalable supply chain decisions. These recent developments reinforce the relevance of learning-assisted optimization approaches for computationally demanding SAF network-design problems.

The present study differs from prior machine-learning-assisted biomass supply chain research in how the learning component interacts with the optimization model. In Goettsch et al. [19], machine learning is used primarily as a candidate-screening mechanism to

reduce the set of potential depot locations before optimization. By contrast, the proposed framework uses machine learning as an iterative decision-making component. The learner selects candidate biorefinery-location configurations, the reduced supply chain MILP evaluates those configurations, and the resulting objective-function values are incorporated into subsequent learning iterations. Thus, the proposed method treats the supply chain MILP as an expensive evaluator and reframes facility selection as a sequential learning problem. Building on preliminary work in machine-learning-assisted biofuel supply chain optimization [22], this study expands the computational analysis by evaluating additional solution procedures and comparing alternative learning-based search strategies.

Another relevant stream of literature concerns the embedding of trained machine learning models within optimization models. In particular, trained neural networks with rectified linear unit activation functions can be represented using mixed-integer linear constraints, allowing an optimization model to search over inputs that maximize or minimize the prediction of a trained neural network [23–25]. This capability is important for the present study because the trained neural network is not only used to estimate the relationship between biorefinery-location decisions and supply chain cost; it is also converted into an optimization model that selects the predicted best-performing combination of biorefineries.

The contribution of the present paper is therefore twofold. First, it develops an iterative machine-learning-assisted MILP framework for a bio-jet fuel hub-and-spoke supply chain network design problem. Second, it compares four learning-assisted solution procedures—ridge regression, ridge regression with Thompson sampling, neural network optimization, and ensemble neural network optimization—to evaluate the tradeoff between computational efficiency, solution quality, and model complexity. By doing so, the paper extends the biofuel supply chain optimization literature beyond exact MILP, stochastic programming, metaheuristics, and pre-optimization candidate screening, and it demonstrates how supervised learning can be used to guide facility-location decisions in a computationally expensive supply chain design problem.

To further clarify the positioning of the present study, Table 1 summarizes the main literature streams related to biofuel supply chain optimization, facility-location modeling, hybrid solution methods, and machine-learning-assisted optimization. The table highlights how prior studies have addressed uncertainty, biomass quality, hub-and-spoke network design, computational complexity, and neural-network reformulations. It also identifies the specific gap addressed by the present work: the development of an iterative ML-MILP framework in which machine learning does not merely screen candidate facilities before optimization, but actively guides the sequential selection of biorefinery location configurations based on observed MILP outcomes.

Table 1. Positioning of the current study relative to the prior literature.

Literature Stream	Representative Studies and Typical Approach	Gap Addressed by the Current Study
Bioenergy and biofuel supply chain optimization	Yue et al. [2]; Atashbar et al. [6]. These studies review and classify bioenergy supply chain models, planning issues, and optimization challenges.	Provides broad background on bioenergy supply chain optimization but does not propose an iterative ML-MILP solution framework.
Stochastic biofuel supply chain design	Marufuzzaman et al. [3]; Castillo-Villar et al. [4]; Aboytes-Ojeda et al. [5]. These studies use two-stage stochastic programming and decomposition-based optimization.	Models uncertainty and biomass quality variability, but does not use machine learning to guide facility-location decisions.

Table 1. *Cont.*

Literature Stream	Representative Studies and Typical Approach	Gap Addressed by the Current Study
Biofuel network design under uncertainty	Üster and Memişoğlu [9]; Memişoğlu and Üster [10]; Geismar et al. [11]; Nur et al. [12]. These studies address biomass pricing, yield uncertainty, storage losses, and preprocessing decisions.	Addresses important sources of uncertainty, but computational acceleration through iterative learning is not the central focus.
Facility location and hub-location optimization	Owen and Daskin [7]; ReVelle and Eiselt [8]; Alumur and Kara [26]; Farahani et al. [27]. These studies establish the foundations of facility-location and hub-location modeling.	Establishes the computational difficulty of location problems, motivating learning-assisted search over candidate-facility configurations.
Metaheuristic and hybrid solution methods	Dréo et al. [13]; Xiao and Konak [14]; Ghaderi and Rahmaniani [15]; Aboytes-Ojeda et al. [16]. These studies use heuristic and hybrid methods for large-scale optimization.	Improves tractability through heuristic search, whereas the current study uses supervised learning to guide sequential biorefinery selection.
Machine-learning-assisted supply chain optimization	Defourny et al. [17]; Oroojlooyjadid et al. [18]; Goettisch et al. [19]; Keith [22]. These studies use machine learning for policy selection, inventory optimization, or candidate-location screening.	Closest prior stream. The current study goes beyond candidate screening by using machine learning iteratively to select and update biorefinery configurations based on observed MILP outcomes.
Optimization over trained neural networks	Rössig and Petkovic [23]; Tsay et al. [24]; Anderson et al. [25]. These studies develop MILP reformulations of trained ReLU neural networks.	Provides the basis for converting a trained neural network into an optimization model for selecting predicted high-performing refinery combinations.
This paper	This paper develops an iterative ML-MILP framework using ridge regression, Thompson sampling, neural network optimization, and ensemble neural network optimization.	Develops and tests an iterative-learning-guided solution framework for a Texas bio-jet fuel hub-and-spoke supply chain.

3. Mathematical Model and Industrial Case Study

The proposed formulation models the bio-jet fuel supply chain as a hub-and-spoke network design problem that involves facility-opening and material-flow decisions. Figure 2 illustrates the structure of the supply chain. Raw switchgrass is supplied at the county level and transported by truck to candidate depot locations, where it is aggregated and densified prior to long-distance shipment. The densified biomass is then moved by rail to candidate biorefineries, where it is converted into bio-jet fuel. Finally, the produced fuel is delivered by truck to airport demand nodes. Transportation duration and traffic congestion are not explicitly modeled. All nine airport demands are represented simultaneously within a single integrated optimization problem. The model determines which depots and biorefineries to open from the available candidate sets while simultaneously optimizing biomass procurement, material flows, and fuel distribution across the entire network.

3.1. Biorefinery System, Switchgrass Feedstock, and Depot Operations

A biorefinery is an integrated processing facility in which biomass is converted into fuels, energy carriers, chemicals, and other value-added products through coordinated physical, chemical, thermochemical, or biochemical operations. Biorefineries may be differentiated by feedstock, conversion platform, product slate, and level of process integration. Within this broader context, the system considered in the present study is a biochemical lignocellulosic biorefinery supplied with switchgrass. The modeled pathway includes feedstock preprocessing, storage of biomass, the conversion process, and the final customers. The pathway is represented at the strategic supply chain level using aggregate

conversion yields, facility capacities, capital costs, operating costs, and co-product credits, rather than detailed unit-operation decision variables. This representation is consistent with the scope of the study, which focuses on supply chain network design and computational performance rather than detailed process design [2,6,28].

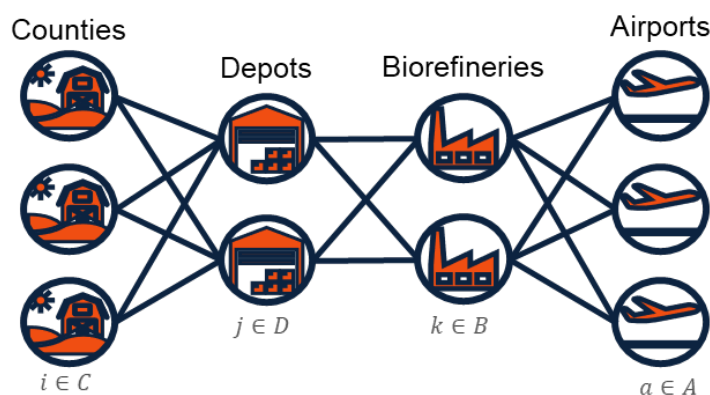


Figure 2. Hub-and-spoke network.

Switchgrass was selected because it is a perennial lignocellulosic energy crop with favorable characteristics for large-scale bioenergy systems in Texas and the broader United States. Its broad climatic adaptability, relatively high biomass productivity, compatibility with marginal or lower-quality land, and comparatively modest input requirements after establishment make it a relevant candidate for regional biofuel production. In addition, switchgrass can contribute to soil conservation, long-term carbon storage, and rural economic activity while reducing direct competition with major food crops. Its suitability for a Texas supply chain also depends on the spatial distribution of feedstock resources, access to agricultural and transportation infrastructure, proximity to preprocessing and conversion facilities, and the evolving policy and market environment for low-carbon aviation fuels [2,29,30].

The model uses county-level dry biomass availability as the exogenous feedstock-supply input. Raw, non-densified switchgrass is transported from county supply nodes to selected depots, and the depot mass balance accounts for the conversion from incoming biomass to the outgoing dry, densified stream. The present formulation does not explicitly represent spatial or temporal heterogeneity in moisture content, chemical composition, ash content, particle size, or degradation during transportation and storage. Instead, these characteristics are embedded in the aggregate supply and conversion assumptions. Biomass quality variability, storage losses, and transportation-related changes in feedstock properties are important extensions for future work, particularly because they may affect usable supply, preprocessing requirements, conversion yield, and operating costs [4,5,12].

The biochemical pathway was selected because the structural carbohydrates in lignocellulosic switchgrass can be converted into intermediates that are subsequently upgraded into bio-jet fuel and co-products. The reference pathway incorporates hydrolysis, anaerobic conversion, catalytic upgrading, and lignin utilization through the 2,3-butanediol route. The process parameters used in the network model are drawn from the underlying process-design study, including a production capacity of 657,288 dry Mg/year and an overall conversion yield of 0.1323 Mg of bio-jet fuel per dry Mg of biomass [28]. Detailed pretreatment conditions, enzyme loadings, fermentation residence times, and specific microbial or yeast strains are not modeled as independent variables; rather, their combined effects are captured in the aggregate yield, capacity, capital cost, and operating cost parameters adopted from the reference design. Accordingly, the present study evaluates the

supply chain implications of the selected pathway rather than optimizing its individual biochemical unit operations.

The modeled pathway should also be interpreted in light of the implementation challenges associated with lignocellulosic biochemical fuel production. Commercial deployment depends on reliable feedstock supply, effective pretreatment and hydrolysis, stable biological conversion, catalytic upgrading performance, capital availability, and the ability to manage variability at a large scale. The maturity of individual unit operations does not eliminate the integration and scale-up challenges associated with the complete pathway. For this reason, the model uses a consistent reference configuration for all candidate biorefineries and treats technology-specific process optimization and technology-readiness assessment as outside the present scope.

Depots serve as intermediate hubs for aggregation and preprocessing between geographically distributed biomass sources and centralized biorefineries. Switchgrass is transported by truck from county supply nodes to selected depots, where it is received, consolidated, preprocessed, and densified before being shipped by rail over long distances. Densification increases bulk density and improves transportation efficiency, while aggregation enables the depot to coordinate dispersed feedstock supply with the biorefinery's larger, more stable throughput requirements. In the mathematical model, these activities are represented by depot investment costs, preprocessing capacity constraints, biomass mass-balance relationships, and the conversion of incoming non-densified biomass into an outgoing densified stream. Detailed equipment scheduling, inventory dynamics, storage degradation, and depot-level energy consumption are not explicitly modeled and are identified as opportunities for future research.

3.2. Bio-Jet Fuel Supply Chain Network Design MILP Formulation

The mixed-integer linear programming formulation presented in Equations (1)–(15) was developed for the present bio-jet fuel supply chain case study using established biomass supply chain network-design and hub-and-spoke modeling principles. The model structure builds on prior formulations that integrate geographically distributed biomass supply, facility-location decisions, preprocessing and conversion capacities, multimodal transportation, material-flow conservation, and demand satisfaction [3–5,16]. The present formulation adapts these general modeling concepts to the Texas bio-jet fuel system considered in this study by explicitly representing county-level switchgrass supply, intermediate depot selection and preprocessing, rail transportation of densified biomass, biorefinery-location decisions, biochemical conversion, and delivery of the resulting fuel to airport demand nodes. Accordingly, Equations (1)–(15) are not reproduced directly from a single previous study; rather, they constitute the integrated network formulation developed for the present application using parameters and process assumptions drawn from the case-study data sources described in the following subsections.

The model includes three principal groups of decision variables. The continuous variables x_{ij} , y_{jk} , and d_{ka} represent, respectively, the flow of non-densified switchgrass from county i to depot j , the flow of densified biomass from depot j to biorefinery k , and the flow of bio-jet fuel from biorefinery k to airport a . The facility-location variables W_j and β_k indicate whether candidate depot j and candidate biorefinery k are opened, while A_{jk} represents activation of the rail connection between the selected depot and biorefinery locations. The variables π_a and π_c represent third-party supply used to satisfy unmet fuel and co-product requirements and are penalized in the objective function. The objective in Equation (1) minimizes the combined annualized facility-investment, transportation, rail-connection, and shortage-penalty costs. Constraint (2) limits procurement by county-level biomass availability; Constraint (3) enforces the depot mass balance between incoming

non-densified biomass and outgoing densified biomass; Constraints (4) and (5) impose demand satisfaction for the primary fuel and co-products; Constraints (6)–(8) impose rail, depot, and biorefinery capacity limits; Constraint (9) preserves the conversion mass balance between biomass input and fuel output; Constraint (10) links rail-connection activation to the opening of the corresponding depot and biorefinery; and Constraints (11)–(15) define the non-negativity, integrality, and binary domains of the decision variables.

Sets:

- $\mathcal{C}$: Set of counties/suppliers, indexed by $i \in \mathcal{C}$.
- $\mathcal{D}$: Set of potential depot locations, indexed by $j \in \mathcal{D}$.
- $\mathcal{B}$: Set of potential biorefinery locations, indexed by $k \in \mathcal{B}$.
- $\mathcal{A}$: Set of airports, indexed by $a \in \mathcal{A}$.
- $\mathcal{T}$: Set of arcs from counties to depots, where $(i, j) \in \mathcal{T}$.
- $\mathcal{R}$: Set of arcs from depots to biorefineries, where $(j, k) \in \mathcal{R}$.
- $\mathcal{U}$: Set of arcs from biorefineries to airports, where $(k, a) \in \mathcal{U}$.

Decision Variables:

- x_{ij} : Volume of non-densified switchgrass shipped along arc $(i, j) \in \mathcal{T}$.
- y_{jk} : Flow of pre-processed, densified switchgrass shipped along arc $(j, k) \in \mathcal{R}$.
- d_{ka} : Flow of bio-jet fuel shipped along arc $(k, a) \in \mathcal{U}$.
- A_{jk} : Integer variable representing the number of unit trains connecting depot $j \in \mathcal{D}$ to biorefinery $k \in \mathcal{B}$.
- β_k : Binary variable equal to 1 if location $k \in \mathcal{B}$ is selected as a biorefinery, and 0 otherwise.
- W_j : Binary variable equal to 1 if location $j \in \mathcal{D}$ is selected as a depot, and 0 otherwise.
- π_a : Third-party bio-diesel supply for airport $a \in \mathcal{A}$.
- π_c : Third-party caustic and adipic acid supply.

Parameters:

- ξ_j : Investment cost to open a depot at node $j \in \mathcal{D}$.
- q_k : Investment cost to open a biorefinery at location $k \in \mathcal{B}$.
- ψ_{jk} : Fixed cost of loading/unloading a unit train along arc $(j, k) \in \mathcal{R}$ every week for one year.
- $c_{ij}^{\mathcal{T}}$: Unit cost per metric ton shipped along arc $(i, j) \in \mathcal{T}$.
- $c_{jk}^{\mathcal{R}}$: Unit cost per metric ton shipped along arc $(j, k) \in \mathcal{R}$.
- $c_{ka}^{\mathcal{U}}$: Unit cost per metric ton shipped along arc $(k, a) \in \mathcal{U}$.
- α_1 : Penalty cost for demand shortage of bio-diesel.
- α_2 : Penalty cost for demand shortage of caustic and adipic acid.
- $s_i(o)$: Available supply in county $i \in \mathcal{C}$ under scenario $o \in \Omega$.
- δ_1 : Biomass-to-bio-jet-fuel conversion factor under anaerobic biochemical conversion.
- δ_2 : Biomass-to-caustic-and-adipic-acid conversion factor under anaerobic biochemical conversion.
- v_{jk} : Maximum capacity of a unit train along arc $(j, k) \in \mathcal{R}$.
- u_j : Pre-processing capacity of depot facility $j \in \mathcal{D}$.
- q_k : Production capacity of biorefinery $k \in \mathcal{B}$.
- D_a : Total demand for bio-diesel at airport $a \in \mathcal{A}$.
- D_c : Total demand for caustic and adipic acid.

The mathematical model minimizes the total costs associated with the BCS design, as shown in Equation (1). The total costs include investment costs to open depots, biorefineries, and their connecting arcs, as well as various transportation costs and demand penalty terms.

The model is formulated as a cost-minimization problem with an exogenously specified bio-jet fuel demand target rather than as a profit-maximization problem. Bio-jet fuel

therefore retains inherent economic value even though sales revenue is not included explicitly in the objective function. Because the airport-demand constraints prescribe the quantity of fuel to be supplied, applying a constant selling price to every unit of required bio-jet fuel would subtract the same revenue from all solutions that satisfy the same demand and would not change the relative ranking of feasible network configurations. The objective function therefore focuses on identifying the least-cost network capable of satisfying the prescribed demand. The third-party fuel variable and its associated penalty continue to carry an economic consequence for insufficient internal production.

The secondary outputs are treated differently because their quantities are linked to biomass conversion and their market values reduce the effective net cost of the selected pathway. Accordingly, adipic acid and sodium hydroxide are represented as co-product credits rather than as evidence that the primary bio-jet fuel product lacks value.

$$\begin{aligned} \text{Min} : & \sum_{j \in \mathcal{D}} \xi_j W_j + \sum_{k \in \mathcal{B}} q_k \beta_k + \sum_{j \in \mathcal{D}} \sum_{k \in \mathcal{B}} \psi_{jk} A_{jk} + \\ & \left[\sum_{i \in \mathcal{C}} \sum_{j \in \mathcal{D}} c_{ij}^T x_{ij} + \sum_{j \in \mathcal{D}} \sum_{k \in \mathcal{B}} c_{jk}^R y_{jk} + \sum_{k \in \mathcal{B}} \sum_{a \in \mathcal{A}} c_{kp}^U d_{ka} + \sum_{a \in \mathcal{A}} \alpha_1 \pi_a + \alpha_2 \pi_c \right] \end{aligned} \quad (1)$$

Subject to:

$$\sum_{j \in \mathcal{D}} x_{ij} \leq S_i \quad \forall i \in \mathcal{C}, \quad (2)$$

$$\sum_{i \in \mathcal{C}} (1 - e_i) x_{ij} = \sum_{k \in \mathcal{B}} y_{jk} \quad \forall j \in \mathcal{D}, \quad (3)$$

$$\sum_{k \in \mathcal{B}} d_{ka} + \pi_a = D_a \quad a \in \mathcal{A} \quad (4)$$

$$\sum_{k \in \mathcal{B}} \sum_{j \in \mathcal{D}} \delta_2 y_{jk} + \pi_c = D_c \quad (5)$$

$$y_{jk} \leq v_{jk} A_{jk} \quad \forall j \in \mathcal{D}, \quad k \in \mathcal{B}, \quad (6)$$

$$\sum_{i \in \mathcal{C}} \sum_{z \in \mathcal{Z}} x_{ijz} \leq u_j W_j \quad \forall j \in \mathcal{D}, \quad (7)$$

$$\sum_{j \in \mathcal{D}} y_{jk} \leq q_k \beta_k \quad \forall k \in \mathcal{B}, \quad (8)$$

$$\sum_{a \in \mathcal{A}} d_{ka} = \sum_{j \in \mathcal{D}} \delta y_{jk}, \quad \forall k \in \mathcal{B}, \quad (9)$$

$$A_{jk} \leq 0.5(\beta_k + W_j), \quad \forall k \in \mathcal{B}, \quad j \in \mathcal{D}, \quad (10)$$

$$x_{ij} \geq 0 \quad \forall i \in \mathcal{C}, \quad j \in \mathcal{D}, \quad (11)$$

$$y_{jk} \geq 0 \quad \forall j \in \mathcal{D}, \quad k \in \mathcal{B}, \quad (12)$$

$$d_{ka} \geq 0 \quad \forall k \in \mathcal{B}, \quad a \in \mathcal{A}, \quad (13)$$

$$\pi_a \geq 0 \quad \forall a \in \mathcal{A}, \quad (14)$$

$$W_j \in \{0, 1\} \quad \forall j \in \mathcal{D}; \quad \beta_k \in \{0, 1\} \quad \forall k \in \mathcal{B}; \quad A_{jk} \in \{0, 1\} \quad \forall j \in \mathcal{D}, \quad k \in \mathcal{B}, \quad (15)$$

The case study retains all 254 Texas counties as potential biomass supply nodes [29]. County-level dry biomass availability for 2022 was obtained from the Billion-Ton Report [30]. The complete county set is retained because county-level biomass availability is an exogenous resource input rather than a facility-location decision. Inclusion in the set does not require a county to supply biomass; the optimization determines county participation endogenously according to available supply, transportation distance, transportation

cost, depot accessibility, and the requirements of the selected network. Consequently, counties with limited biomass availability or unfavorable logistics may provide little or no biomass in the optimal solution.

The treatment of counties differs from the screening applied to candidate depots, biorefineries, and airports. Depot and biorefinery locations represent infrastructure investment decisions, and their candidate sets were narrowed using biomass proximity, transportation access, and location suitability to preserve practical relevance and computational tractability. Airports were limited to the nine largest demand centers included in the case study. Thus, the complete statewide feedstock resource base is preserved, while candidate infrastructure and final-demand nodes are screened according to their distinct roles in the network.

3.3. Depots

A shapefile provided by Oak Ridge National Laboratory was used to identify potential depot locations [31]. Train stations in the shapefile were treated as candidate depot locations, and this set was further narrowed by considering proximity to counties with relatively high biomass availability. A total of 33 potential depot locations are included in the model. The total capital investment required to open a depot is \$30,178,945, with an equivalent annual cost (EAC) of \$3,073,792 based on an 8% interest rate and a 20-year project life [32]. A standardized maximum preprocessing capacity of 330,000 Mg/year is used for each candidate depot to provide a consistent, representative facility scale across locations and to match the case study's demand scope.

The standardized capacity is an upper bound rather than a required operating level. Actual depot throughput remains an endogenous model outcome determined by upstream biomass availability, transportation costs, the selected network configuration, and downstream biorefinery requirements. Therefore, an opened depot may operate below its maximum capacity. Endogenous facility sizing, discrete capacity alternatives, modular expansion, and capacity-investment decisions are outside the present scope and are identified as future extensions.

3.4. Biorefineries

Potential biorefinery locations were obtained from the Bioenergy Atlas and narrowed through a suitability analysis that emphasized proximity to railways and highways, resulting in 167 candidate locations [33]. The total capital investment for a biorefinery using hydrolysis, anaerobic conversion, catalytic upgrading, and lignin conversion through the 2,3-butanediol route is estimated at \$823,698,095 in 2022 USD. The corresponding EAC plus operating cost is \$208,543,998, assuming an 8% interest rate and a 20-year project life [28]. The representative production capacity is 657,288 dry Mg/year, and the conversion yield is 0.1323 Mg of bio-jet fuel per dry Mg of biomass [28]. The conversion pathway also produces monetizable co-products, namely, adipic acid and sodium hydroxide, with values of \$296.5 and \$31.3 per dry ton of biomass, respectively [28].

A standardized maximum biorefinery scale is used because the present study evaluates facility-location decisions under a common reference process configuration rather than optimizing facility size. As with depot capacity, the specified biorefinery capacity is an upper bound; actual throughput is determined endogenously by biomass availability, transportation costs, airport demand, and the overall network configuration. This assumption isolates the computational effects of biorefinery-location selection, which is the strategic decision assigned to the learning component. Future work may extend the formulation to include discrete technology scales, modular capacity expansion, or continuous facility-sizing decisions.

3.5. Airports

The nine most active airports in Texas, based on Federal Aviation Administration flight records for 2020, are represented as simultaneous destination nodes for the bio-jet fuel produced by the supply chain [34]. Total Texas jet-fuel consumption in 2019 was estimated at 330.2 TBtu, or 8,049,851 Mg of biofuel equivalent, according to the EIA [35]. The present supply chain is designed to satisfy 10% of this statewide demand, corresponding to 804,985 Mg of biofuel equivalent. This total is allocated among the nine airports in proportion to their respective flight activity, as shown in Table 2. Each airport-specific requirement is imposed concurrently through the demand constraints; the airports are not optimized independently or in separate model runs.

Table 2. Airport demands.

Airport	Flights (2020)	Bio-Diesel Demand (Mg)
Dallas/Fort Worth (DFW)	620,831	275,493.7
George Bush t (IAH)	354,045	157,107.4
Dallas Love Field (DAL)	184,533	81,866.5
Austin-Bergstrom (AUS)	118,340	52,513.4
William P. Hobby Airport (HOU)	153,840	68,266.5
San Antonio (SAT)	165,548	73,461.9
El Paso (ELP)	83,438	37,025.6
Midland (MAF)	58,010	25,741.9
Lubbock Preston Smith (LBB)	75,466	33,488.0
Total	1,814,051	804,985

3.6. Transportation and Other Costs

Street distances were used for arcs $(i, j) \in \mathcal{T}$ and arc $(k, a) \in \mathcal{U}$ [36]. For arc $(j, k) \in \mathcal{R}$, railway distances were used. The cutoff point in the arcs that connect the set of counties with the depots is 170 km, as it is not recommended to ship biomass using trucks for long distances [37]. The transportation costs come in the form of linear regression results and are shown in Table 3.

Table 3. Transportation costs.

Arc	Mode	Payload	Fixed Cost (\$/Mg)	Variable Cost (\$/(Mg*Km))	Source
$\mathcal{T}$	Truck	Raw Biomass	4.82	0.07	[38]
$\mathcal{R}$	Train	Densified Biomass	2.55	0.02	[39]
$\mathcal{U}$	Truck	Bio-Diesel	5.50	0.07	[40]

Of note is that the fixed transportation cost for arc $\mathcal{R}$ is very low. This is in part to utilizing unit trains in lieu of trucks but also because the fixed cost for loading and unloading the unit trains appears elsewhere in the model. This practice is in agreement with the work done by Roni [39]. In addition, the value of the co-products generated through biomass conversion, including adipic acid and sodium hydroxide, is represented as a credit that reduces the effective net cost associated with the conversion pathway and the corresponding material flows.

4. Machine Learning Solution Procedures and Computational Results

This section presents the proposed machine-learning-assisted solution procedures together with their implementation, experimental design, and comparative computational performance. Refer to Figure 1 for a visual overview of the proposed procedures.

The algorithm starts with an initial sample of solutions to the supply chain optimization problem where the refineries are randomly selected. The choice of nine biorefineries is a simplifying hypothesis specific to this case study rather than a general property of the algorithm. It is supported by the demand-to-capacity structure and cost parameters of the instance. Each biorefinery has an annual capacity of (657,88) dry Mg, which, at a conversion yield of (0.1323) Mg of fuel per dry Mg, corresponds to approximately (86,960) Mg of bio-jet fuel per facility. Given an aggregate demand of (804,985) Mg, satisfying demand exclusively through the biorefinery network would require $804,985/86,960 = 10$ facilities. Nine facilities, however, supply approximately (97.2%) of total demand, with the remaining (2.8%) covered through third-party supply at a penalty cost. Because the annualized cost of opening an additional biorefinery substantially exceeds the penalty associated with covering this limited residual demand, the cost-minimizing solution opens nine facilities. This result is consistent with the optimal facility count obtained from the exact full-scale MILP benchmark. The facility cardinality enters the proposed method only through the parameter Ψ in the selection step. The framework is therefore not restricted to $\Psi = 9$. For instances with different demand, capacity, and cost structures, the ratio (D/q) may be used to identify an economically plausible neighborhood of candidate cardinalities. When the optimal cardinality is uncertain, the method can be applied over a narrow integer range of Ψ , and the solution yielding the best exact-MILP objective value can be retained.

From there, a regression-type model is fit to the collection of solutions, and the best refineries for the next run are determined via the fit regression model. The selected refineries are then fed into the supply chain optimization model to generate an additional feature-response pair, which is then used to fit a new regression. This process is iterated on to achieve the final result. Four learning-assisted procedures are evaluated: L1 (ridge regression), L2 (ridge regression with Thompson sampling), NL1 (neural network), and NL2 (ensemble neural network). Table 4 summarizes the common initial sample, model design, and selection strategy for each procedure according to the framework shown in Figure 3.

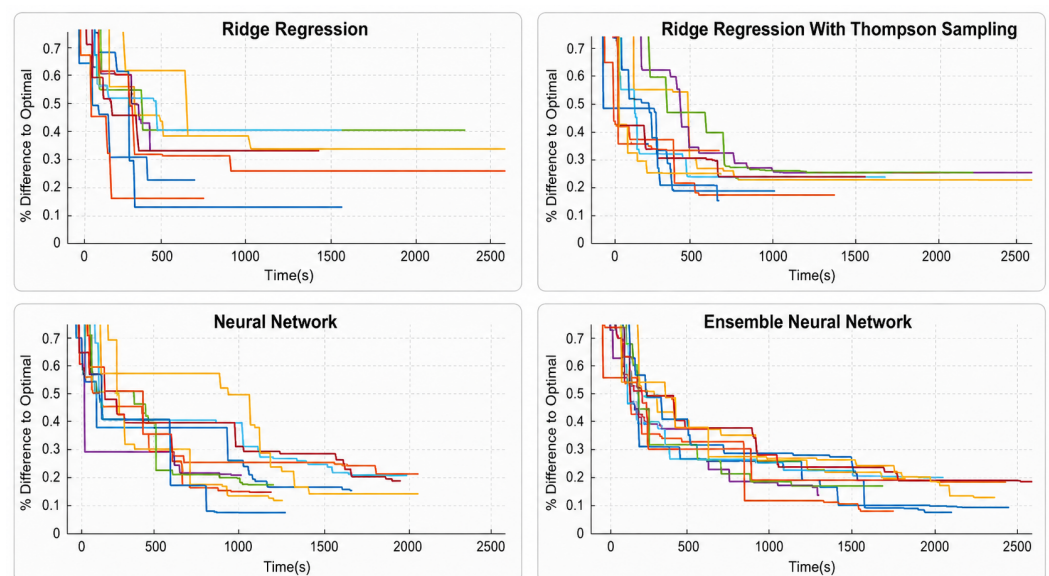


Figure 3. Best observed percent difference from the benchmark solution as a function of elapsed computational time for L1 (ridge regression), L2 (ridge regression with Thompson sampling), NL1 (neural network), and NL2 (ensemble neural network). Curves summarize the repeated experimental runs for each method.

Table 4. Overview of machine learning solution procedures.

Regression Algorithm	Initial Sample	Model and Hyperparameter Design	Selection Strategy
L1: Ridge Regression	30 SC optimization solutions with randomly selected refineries	Ridge regression with $\lambda = 0.1$	Select refineries corresponding to maximal regression coefficients
L2: Ridge Regression with Thompson Sampling	30 SC optimization solutions with randomly selected refineries	Ridge regression with $\lambda = 0.1$	Vary refineries via Thompson sampling then select refineries corresponding to maximal post-varied regression coefficients
NL1: Neural Network	30 SC optimization solutions with randomly selected refineries	NN with one hidden layer with 15 nodes	MILP used to select the best performing combination of refineries
NL2: Ensemble Neural Network	30 SC optimization solutions with randomly selected refineries	Ensemble NN with one hidden layer with 15 nodes and 10 folds	MILP used to select the best performing combination of refineries based on average performance across all folds

4.1. Regression Overview

Of paramount importance to the suggested solution procedure is how to frame the results of the supply chain optimization problem as a regression algorithm suitable for machine learning. For the regression algorithms in this work, the I/O is as follows:

Features: X

- x_{ij} takes the value 1 if refinery j was selected for run i of the supply chain optimization model and 0 otherwise.

Responses: y^*

- y_i^* denotes the savings generated by the combination of selected refineries for run i of the supply chain optimization model. It is computed according to the following formula $y_i^* = C - y_i$, where C is the solution to the supply chain optimization problem if no refineries are opened and y_i is the solution to the supply chain optimization problem for the selection of refineries used for run i .

The problem is framed in this way to generate an environment where the maximal regression coefficients are desirable as they denote the refineries that generate the most savings. For linear regression, this is the mathematical equivalent of imposing a known intercept.

What follows is a discussion of the development and implementation of the four machine-learning-based algorithms in the present work.

4.2. Linear Methods

4.2.1. L1: Ridge Regression

The linear methods utilize ridge regression in order to determine the regression coefficients for the potential biorefinery locations. While as much data as desirable are available to be generated since the problem is such that data are created via runs of the optimization model, the desired experimental outcome is a reduction in computational burden; thus, the algorithms to be utilized should strive to perform well in a limited data environment.

Each configuration is represented by 167 binary location variables, whereas each trial contains 300 feature-response pairs. In addition, all feasible configurations satisfy a fixed-cardinality constraint, which induces dependence among the predictors. This relatively low sample-to-feature ratio and the resulting multicollinearity motivate the use of ridge regression rather than ordinary least squares. Ridge regularization stabilizes the coefficient estimates and reduces their sensitivity to correlated predictors through the estimator. The problem setup coupled with numerical experimentation indicates that the regression coefficients can be applied analytically via the formula below:

$$\beta = (X^T X + \lambda I)^{-1} X^T Y^* \quad (16)$$

The penalty term λI ensures a stable and uniquely defined estimate even when the predictor matrix is poorly conditioned and shrinks weak or unstable coefficients. The ridge model is therefore used as a regularized decision-support model rather than as an unrestricted global predictor of supply-chain cost.

Once the coefficients have been obtained the next step in the solution procedure is to determine the predicted maximal performing set of biorefineries. For this, we leverage problem knowledge derived from supply and capacity considerations as well as numerical experiments to impose that a good solution should contain 9 biorefineries. With this in mind, the maximal performing set according to the regression will simply be the refineries corresponding to the nine largest regression coefficients. This regression and maximal set determination scheme constitutes the approach for the first solution procedure. This procedure is outlined in Algorithm 1.

Algorithm 1 L1: Ridge-regression-powered solution procedure:

Initialize: β_0
for $t = 0, \dots, T - 1$ **do**
 Compute: $x_{t+1} \leftarrow \operatorname{argmax}_{\beta' \subset \beta, |\beta'|=9} \sum_{i \in \beta'} \beta_i$
 Observe: y_{t+1}
 Update: β_{t+1}
end for

4.2.2. L2: Ridge Regression with Thompson Sampling

As will be demonstrated, this formulation is primed to suffer from an information exploitation problem where we start with a limited amount of data using multiple linear regression to predict the best-performing input and iterate from there. To combat this, L2 introduces Thompson Sampling to promote exploration while retaining the same ridge-regression structure. The scale parameter η was set based on a small preliminary grid search over values on either side of unity (e.g., $\eta \in \{0.25, 0.5, 1, 2, 4\}$), evaluated by the resulting optimality gap and its variance across trials. The second solution procedure begins the same as the first with the analytical computation of regression coefficients via ridge regression. From here, a normal distribution is imposed centered around each coefficient with a standard deviation related to the mean squared error (MSE) of the regression and the variance of the input. New coefficients are randomly sampled from these distributions via the equation below, where η is a scale parameter:

$$\beta_i^* = \operatorname{normrand}(\beta_i, \eta \sqrt{\frac{\text{MSE}}{\operatorname{diag}(X^T X)_i}}) \quad (17)$$

Once the new coefficients have been obtained, the process to select the predicted maximal performing set of refineries is the same as above in that the 9 largest coefficients indicate

the refineries that the regression thinks should be selected. This approach is outlined in Algorithm 2.

Algorithm 2 L2: Ridge Regression with Thompson sampling-powered solution procedure

Initialize: β_0
for $t = 0, \dots, T - 1$ **do**
 Vary: $\beta_t^* \leftarrow \beta_t$
 Compute: $x_{t+1} \leftarrow \arg \max_{\beta^{*'} \subset \beta^*, |\beta^{*'}|=9} \sum_{i \in \beta^{*'}} \beta_i^*$
 Observe: y_{t+1}
 Update: β_{t+1}
end for

4.3. Non-Linear Methods

The biorefinery-location problem is combinatorial because the selected facilities operate jointly to satisfy regional demand. The additive structure of the linear procedures does not fully capture these interactions. The nonlinear procedures therefore use neural networks to represent interactions among selected biorefineries and predict the savings associated with complete facility configurations.

4.3.1. NL1: Neural Network Optimization Algorithm

NL1 uses a relatively small feed-forward neural network because the training data are intentionally limited. The network contains one fully connected hidden layer with 15 nodes and rectified linear unit (ReLU) activation functions. The hidden layer is fully connected to an output node that predicts supply chain savings. Because the model is intended to capture nonlinear interactions, selecting facilities independently according to a simple ranking would not preserve the full learned relationship. Instead, the trained ReLU network is reformulated as an MILP to select the combination of biorefineries that yields the highest predicted savings [23].

Sets:

- $\mathcal{K}$: Set of hidden-layer nodes, indexed by $k \in \mathcal{K}$.
- $\mathcal{I}$: Set of features corresponding to potential biorefinery locations, indexed by $i \in \mathcal{I}$.
- $\mathcal{C}$: Set of feasible biorefinery selections, defined as $\mathcal{C} = \{X = (x_1, x_2, \dots, x_n) \mid \sum_{i \in \mathcal{I}} x_i = \Psi\}$.
- $\mathcal{O}$: Set of folds, indexed by $o \in \mathcal{O}$.

Parameters:

- $m_{ik}(o)$: First-layer weight from feature i to hidden node k for fold $o \in \mathcal{O}$.
- $b_k(o)$: Hidden-layer bias for node k for fold $o \in \mathcal{O}$.
- $n_k(o)$: Output-layer weight from hidden node k for fold $o \in \mathcal{O}$.
- Ψ : Number of refineries to be selected.
- $U_k(o)$: Upper bound for the ReLU reformulation, defined as

$$U_k(o) = \sup_{X \in \mathcal{C}} \left\{ \sum_{i \in \mathcal{I}} m_{ik}(o) x_i + b_k(o), 1 \right\}, \quad \forall k \in \mathcal{K}, \forall o \in \mathcal{O}.$$

- $L_k(o)$: Lower bound for the ReLU reformulation, defined as

$$L_k(o) = \inf_{X \in \mathcal{C}} \left\{ \sum_{i \in \mathcal{I}} m_{ik}(o) x_i + b_k(o), -1 \right\}, \quad \forall k \in \mathcal{K}, \forall o \in \mathcal{O}.$$

Decision Variables:

- $a_k(o)$: Value passed to hidden node k for fold $o \in \mathcal{O}$.
- $h_k(o)$: ReLU-corrected value passed from hidden node k for fold $o \in \mathcal{O}$.
- x_i : Binary variable equal to 1 if refinery $i \in \mathcal{I}$ is selected and 0 otherwise.
- $y_k(o)$: Binary variable used for the ReLU reformulation for hidden node k and fold $o \in \mathcal{O}$.

$$\max R = \frac{1}{|\mathcal{O}|} \sum_{o \in \mathcal{O}} \sum_{k \in \mathcal{K}} n_k(o) h_k(o) \quad (18)$$

Subject to:

$$\sum_{i \in \mathcal{I}} x_i = \Psi, \quad (19)$$

$$\sum_{i \in \mathcal{I}} m_{ik}(o) x_i + b_k(o) = a_k(o), \quad \forall k \in \mathcal{K}, \forall o \in \mathcal{O}, \quad (20)$$

$$a_k(o) \leq h_k(o), \quad \forall k \in \mathcal{K}, \forall o \in \mathcal{O}, \quad (21)$$

$$h_k(o) \leq U_k(o) y_k(o), \quad \forall k \in \mathcal{K}, \forall o \in \mathcal{O}, \quad (22)$$

$$h_k(o) \leq a_k(o) - L_k(o)(1 - y_k(o)), \quad \forall k \in \mathcal{K}, \forall o \in \mathcal{O}, \quad (23)$$

$$L_k(o) \leq a_k(o) \leq U_k(o), \quad \forall k \in \mathcal{K}, \forall o \in \mathcal{O}, \quad (24)$$

$$y_k(o) \in \{0, 1\}, \quad \forall k \in \mathcal{K}, \forall o \in \mathcal{O}, \quad (25)$$

$$x_i \in \{0, 1\}, \quad \forall i \in \mathcal{I}. \quad (26)$$

Thus, the solution approach for the NL1 neural network solution procedure can be seen in Algorithm 3. The procedure begins with a set of refinery/solution pairs and a fit neural network denoted by θ_0 . At each step, the next set of refineries (x_{t+1}) is selected by solving the above optimization problem. Then the SC solution given that set of refineries (y_{t+1}) is determined by solving the supply chain MILP. This refinery/solution pair is then added to the dataset, and the neural network is refit θ_{t+1} before the next iteration.

Algorithm 3 NL1: Neural-network-powered solution procedure:

```

Initialize:  $\theta_0$ 
for  $t = 0, \dots, T - 1$  do
    Optimize:  $x_{t+1} = \operatorname{argmax} R(\theta_t)$ 
    Observe:  $y_{t+1}$ 
    Update:  $\theta_{t+1}$ 
end for

```

4.3.2. NL2: Ensemble Neural Network

Extending this solution procedure in the same vein as the linear case with Thompson sampling is not without its challenges. Computing the posterior distribution of the fit neural network can be challenging or even intractable. For this reason, NL2 adopts the approach of Lu and Roy in that an ensemble neural network is used instead of Thompson sampling [41]. In this approach, we apply 10 folds to the data and omit a random 10% of the data each step to generate 10 distinct datasets that are each fitted with a neural network structured as above. In this scenario, the next set of refineries constitutes those that have the average maximal performance across the 10 networks. The algorithm for this ensemble approach is shown below in Algorithm 4.

Algorithm 4 NL2: Ensemble-neural-network powered solution procedure:

```

Initialize:  $\theta_0^1, \dots, \theta_0^{10}$ 
for  $t = 0, \dots, T - 1$  do
  Optimize:  $x_{t+1} = \operatorname{argmax} R(\theta_t^1, \dots, \theta_t^{10})$ 
  Observe:  $y_{t+1}$ 
  Update:  $\theta_{t+1}^1, \dots, \theta_{t+1}^{10}$ 
end for

```

4.4. Software Implementation and Solver Selection

The optimization model is implemented in Julia using JuMP and solved with IBM CPLEX under a 0.5% optimality-gap tolerance. Regression model estimation and refinery-location selection are performed in MATLAB R2023b, version 23.2 using the Statistics and Machine Learning Toolbox and the Deep Learning Toolbox. The complete iterative algorithm is coordinated in Julia, with MATLAB routines called through the MATLAB.jl package for the learning and selection components.

MATLAB was selected because it provides an integrated environment for numerical experimentation, matrix operations, machine-learning implementation, visualization, reproducible scripting, and direct interaction with external optimization workflows. CPLEX was selected because the study requires repeated solutions of large MILP formulations and a reliable benchmark for exact solvers. Its mature branch-and-cut algorithms, presolve capabilities, numerical diagnostics, solver controls, and established integration with algebraic modeling environments make it suitable for evaluating the computational performance of the proposed procedures. Commercial software also has disadvantages, including licensing costs, access restrictions, and potential barriers to reproducibility. Open-source alternatives, including Python 3.11-based scientific and machine-learning libraries, together with solvers such as HiGHS, SCIP, and CBC, may improve accessibility and reproducibility, although computational performance, supported features, licensing terms, and integration effort may differ across platforms. The proposed ML-MILP framework is platform- and solver-independent in principle; MATLAB and CPLEX are implementation choices rather than methodological requirements.

4.5. Experimental Design and Benchmark

All proposed procedures are benchmarked against the full supply chain optimization model, which yields an objective-function value of \$2,316,428,200 and requires 3599 s of computational time.

The proposed method follows an online, active-learning design rather than an offline supervised-learning pipeline, which governs how the data are handled. Each trial begins with an initial design of 30 feature–response pairs obtained by solving the exact supply chain MILP for randomly selected refinery configurations. At each subsequent iteration the current surrogate proposes one configuration, that configuration is evaluated by the exact MILP, and the resulting pair is appended to the accumulated data on which the surrogate is refit; 270 iterations therefore yield 300 pairs per trial, repeated over 10 independent trials. Because configurations are generated and labeled on the fly, there is no fixed held-out test set because all newly generated observations are incorporated sequentially into model updating. The exact MILP instead provides full-fidelity evaluation of every recommended configuration and supports assessment of downstream decision quality. Solution quality is accordingly certified by the optimality gap of the returned design relative to the exact benchmark (Tables 5 and 6) rather than by surrogate fit on auxiliary data. Generalization is controlled within each surrogate: the linear models use ridge regularization ($\lambda = 0.1$) to control coefficient instability and multicollinearity in the limited-data setting, and the ensemble neural network uses a resampling scheme equivalent to

cross-validation, training ten networks each on a subset omitting a random 10% of the data and selecting configurations by their average predicted performance across folds.

Table 5. Detailed performance of the algorithms' solutions.

Algorithm	Mean Solution Value	Difference to Optimal	Standard Deviation
L1: Ridge Regression	\$2,323,060,838	0.29%	\$2,254,306
L2: Ridge Regression with Thompson Sampling	\$2,321,666,012	0.23%	\$1,184,592
NL1: Neural Network	\$2,319,825,893	0.15%	\$1,072,780
NL2: Ensemble Neural Network	\$2,319,507,373	0.13%	\$1,181,481

Table 6. Detailed performance of algorithms' computational time.

Algorithm	Mean Time Obtained	Time Reduction	Standard Deviation
L1: Ridge Regression	650 s	81.95%	217 s
L2: Ridge Regression with Thompson Sampling	1045 s	70.95%	518 s
NL1: Neural Network	1427 s	60.36%	313 s
NL2: Ensemble Neural Network	1966 s	45.37%	362 s

The surrogate models are not intended to provide globally accurate predictions of supply-chain cost across the entire feasible space. Their purpose is to rank or select promising refinery configurations that are subsequently evaluated at full fidelity using the original MILP. Accordingly, the primary performance criterion is the optimality gap between the best MILP-evaluated configuration returned by each procedure and the exact benchmark, as reported in Tables 5 and 6. Conventional predictive metrics such as MSE, MAE, and R^2 assess pointwise predictive accuracy on unseen observations, whereas the optimality gap directly evaluates the quality of the optimization decisions produced by the surrogate-assisted procedure.

The effective learning problem is also more structured than the nominal feature count alone suggests. Feasible observations are restricted to configurations containing exactly Ψ active locations, ridge regularization shrinks unstable coefficients, and the sequential design progressively concentrates evaluations in more promising regions of the solution space. The linear models therefore use ridge regularization with $\lambda = 0.1$ to control coefficient instability under the limited-data setting. The ensemble neural-network procedure further reduces sensitivity to individual training samples by fitting ten models to resampled subsets of the accumulated data and selecting configurations according to their average predicted performance.

Figure 3 contains the best solution represented as a percent difference from the optimal solution observed in the process up to the current time. The benchmark objective value is not available to the learning algorithms and does not influence model fitting, candidate selection, or stopping. Instead, every proposed configuration is evaluated by the original supply chain MILP, and the gap to the benchmark is calculated ex post solely to compare the quality and convergence behavior of the four procedures. The linear methods converge faster than the nonlinear ones as expected due to the additional computational complexity of fitting a NN as opposed to analytical linear coefficients. Additionally, the introduction of Thompson Sampling does slow the convergence rate of the ridge regression but appears to greatly reduce solution variance. Both linear methods converge rather quickly but we do observe some cases where the Thompson Sampling kicks the solutions out of a steady state and obtains a lower solution value.

For the nonlinear methods, it is visually difficult to distinguish between the best solutions; however, the ensemble NN exhibits more uniform convergence behavior. The

progressive reduction in optimality gap indicates that the sequential procedures identify increasingly competitive configurations; however, this pattern should not be interpreted as an independent measure of out-of-sample predictive accuracy.

For a deeper dive into performance as it pertains to solution integrity, Table 5 details statistics for the best solutions obtained from each algorithm over each of the 10 trials.

From a mean solution value standpoint, it can be seen that, with each increase in algorithmic complexity, we see a reduction in the gap to the optimal solution. The ability of the NN approaches to capture the nonlinear effects associated with the combinatorial optimization problem that is the selection of biorefinery locations really is highlighted here. From a solution variance perspective, we see that the basic ridge regression has roughly double the solution variance of the other methods.

4.6. Computational Performance

The computational performance is summarized in Table 6. The reported time is an end-to-end measure beginning with the generation of the initial random MILP sample and ending at the first occurrence of the best solution obtained during a run. Consequently, the reported values include all computational activities required by each procedure, including surrogate-model fitting, refinery-configuration selection, evaluation of selected configurations using the original supply chain MILP, iterative data updating, and implementation overhead. The values should therefore not be interpreted as the execution time of a single iteration or of an individual model component.

This phenomenon is crucial for practitioners that seek to implement this algorithm, as early termination of the algorithm may lead to a lower quality solution than what was possible. As for the nonlinear methods, we see an increase in time from the linear ones as not only do we have to fit a NN, or 10, for each iteration but also have to solve a MILP to determine the predicted optimal set of biorefineries. With that being said, we still observe significant time savings as opposed to solving the full problem.

The comparisons reported in Tables 5 and 6 are interpreted as descriptive measures of computational performance rather than as evidence of statistical superiority among the proposed methods. The primary objective of the analysis is to characterize the practical tradeoff between solution quality and computational effort under the conditions of the Texas case study. Accordingly, emphasis is placed on the magnitude of the observed optimality gaps, computational-time reductions, and run-to-run variability. Statistical significance alone does not measure the magnitude or practical relevance of an observed difference [42]. For example, NL2 achieved a mean optimality gap of 0.13%, compared with 0.15% for NL1, while requiring a greater mean computational effort. Similarly, L1 produced the largest mean optimality gap among the four methods, 0.29%, but achieved the greatest computational-time reduction, 81.95%. These results indicate a tradeoff between solution quality and computational efficiency rather than a universal ranking of the methods. The observed differences should therefore be interpreted within the scope of the evaluated case study, algorithm settings, and 10 independent experimental trials.

Stringing all of this together, the basic Ridge-Regression-powered solution procedure is able to produce a near-optimal solution with very little computational effort. This approach, however, has high variance in solutions due to the under-determined nature of the problem. To remedy this, Thompson Sampling can be employed at the cost of the computational burden. This implementation will lower the solution variance and mean, thus yielding better performance from an optimality standpoint. The next step in increasing performance can be realized from the introduction of an NN that is able to capture the nonlinear effects associated with the combinatorial optimization problem of biorefinery location selection. Additionally, an ensemble NN has the potential to even further reduce the optimality gap.

5. Conclusions and Future Work

This study proposed a hybrid machine learning and mixed-integer linear programming (ML-MILP) solution framework for computationally intensive bio-jet fuel supply chain network design problems. The proposed approach reduces the computational burden of the full MILP formulation by transferring part of the facility-location search process to a supervised learning algorithm. Instead of solving the complete optimization problem directly, the method iteratively evaluates selected facility-location configurations through a reduced MILP model, uses the resulting objective function values to train a predictive model, and then identifies new candidate configurations with strong expected performance.

Four algorithmic variants were developed and evaluated: two based on linear learning models and two based on nonlinear learning models. These approaches were applied to a hub-and-spoke bio-jet fuel supply chain design problem in the state of Texas, where counties serve as biomass supply nodes, depots perform densification, biorefineries convert densified biomass into bio-jet fuel, and airports represent demand locations. The computational results show that all proposed methods were able to generate near-optimal supply chain designs while reducing solution time relative to solving the full optimization problem through exact methods. The nonlinear methods achieved smaller optimality gaps than the linear approaches, although this improvement came with higher computational requirements. This trade-off suggests that the selection of a learning model should depend on the decision context: linear models may be preferable when computational speed and interpretability are priorities, whereas nonlinear models may be more appropriate when marginal improvements in solution quality justify additional computational effort.

From a managerial perspective, the proposed framework provides a practical decision-support tool for strategic biofuel infrastructure planning. Bio-jet fuel supply chains involve high-capital, long-horizon decisions, including where to locate depots and biorefineries, how to allocate biomass resources, and how to route materials through the network. In practice, decision-makers often need to evaluate multiple scenarios involving feedstock availability, demand growth, transportation costs, policy incentives, technology assumptions, or regional development objectives. The economic deployment of bio-jet fuel is also affected by fuel-price volatility and policy dependence. The effective market value of bio-jet fuel may vary with feedstock and conversion costs, conventional jet-fuel prices, carbon-intensity valuation, tax incentives, production credits, low-carbon fuel programs, blending requirements, and other regional or federal policy mechanisms. These factors can materially affect commercial feasibility and investment timing, even when the physical supply chain is evaluated using a fixed-demand cost-minimization model. Solving a full-scale MILP model for each scenario can be computationally expensive and may limit the usefulness of optimization models in time-sensitive planning environments. The proposed ML-MILP framework helps address this limitation by reducing computation time while preserving high-quality solutions, allowing planners to explore a broader set of alternatives more efficiently.

The results also have implications for the deployment of sustainable aviation fuel systems. By accelerating the evaluation of candidate facility-location configurations, the proposed method can support more agile planning for emerging biofuel markets. Public agencies, energy firms, airlines, and infrastructure developers could use this type of approach to compare alternative investment strategies, identify robust regions for biorefinery development, and assess the cost implications of different feedstock and transportation assumptions. More broadly, the framework illustrates how machine learning can complement, rather than replace, mathematical optimization in industrial energy systems. Optimization remains essential for producing rigorous cost and flow evaluations, while machine learning helps guide the search process toward promising regions of the solution space.

The computational findings should be interpreted within the scope of the Texas case study and the assumptions used in the present formulation. The analysis considers one feedstock, one biochemical conversion pathway, predetermined candidate locations, standardized facility capacities, deterministic annual supply and demand, and a fixed number of selected biorefineries. Consequently, the relative performance of L1, L2, NL1, and NL2 may differ for networks with substantially different candidate-set sizes, transportation structures, endogenous facility sizing, multiple conversion technologies, uncertain or price-responsive demand, stochastic biomass availability, or stronger feedstock-quality heterogeneity.

Several opportunities remain for future research. First, incorporating facility-level attributes into the regression process could improve the predictive accuracy of the learning models and reduce the number of MILP evaluations required. Such attributes may include feedstock proximity, transportation accessibility, regional biomass density, conversion capacity, labor availability, infrastructure readiness, or policy incentives. Second, future work could extend the learning component beyond biorefinery-location decisions to include additional first-stage decisions, such as depot opening, technology selection, capacity expansion, multimodal transportation infrastructure, and the location of emerging airports, aviation hubs or fuel distribution terminals. Third, the framework could be tested under uncertainty by incorporating stochastic or robust optimization elements related to biomass supply, fuel demand, conversion yields, and transportation costs. Finally, future studies could evaluate the transferability of the proposed approach to other bioenergy and biorefinery systems, including biodiesel, hydrogen, or waste-to-energy networks.

A direct sensitivity analysis on a constant bio-jet fuel selling price was not included because, under the current fixed-demand cost-minimization formulation, the resulting revenue term would shift the net economic outcome without changing facility location, transportation, or material flow decisions. Future extensions could instead formulate the problem as a profit- or net present value maximization model with price-responsive demand, variable production, policy incentives, carbon credits, and alternative market price scenarios. Under those conditions, fuel-price sensitivity would directly affect production levels, infrastructure investment, and optimal network design.

Overall, the findings demonstrate that hybrid ML-MILP approaches can improve the computational tractability of large-scale biofuel supply chain design problems. By combining the rigor of mathematical optimization with the adaptive search capabilities of supervised learning, the proposed framework offers a promising pathway for supporting strategic decisions in bioenergy infrastructure planning and industrial decarbonization.

Author Contributions: Conceptualization, K.K.C.-V.; methodology, K.K., K.K.C.-V. and A.A.; formal analysis, K.K.; investigation, K.K., K.K.C.-V. and A.A.; resources, K.K.C.-V.; data curation, K.K., K.K.C.-V. and A.A.; writing—original draft preparation, K.K. and K.K.C.-V.; writing—review and editing, A.A.; funding acquisition, K.K.C.-V. All authors have read and agreed to the published version of the manuscript.

Funding: This material is based upon work that is supported by the National Institute of Food and Agriculture (NIFA), U.S. Department of Agriculture (USDA), under the Hispanic Serving Institutions Education Grants Program, award no. 2020-38422-32258.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Charles, D. The Impact of Oil Prices on Guyana's Real Exchange Rate: An AlphaFold-Decomposition Analysis. *FinTech Sustain. Innov.* **2026**, *2*, A3.
- Yue, D.; You, F.; Snyder, S.W. Biomass-to-bioenergy and biofuel supply chain optimization: Overview, key issues and challenges. *Comput. Chem. Eng.* **2014**, *66*, 36–56. [\[CrossRef\]](#)
- Marufuzzaman, M.; Eksioglu, S.D.; Huang, Y.E. Two-stage stochastic programming supply chain model for biodiesel production via wastewater treatment. *Comput. Oper. Res.* **2014**, *49*, 1–17. [\[CrossRef\]](#)
- Castillo-Villar, K.K.; Eksioglu, S.D.; Taherkhorsandi, M. Integrating biomass quality variability in stochastic supply chain modeling and optimization for large-scale biofuel production. *J. Clean. Prod.* **2017**, *149*, 904–918. [\[CrossRef\]](#)
- Aboytes-Ojeda, M.; Castillo-Villar, K.K.; Eksioglu, S.D. Modeling and optimization of biomass quality variability for decision support systems in biomass supply chains. *Ann. Oper. Res.* **2022**, *314*, 319–346.
- Atashbar, N.Z.; Labadie, N.; Prins, C. Modeling and optimization of biomass supply chains: A review and a critical look. *IFAC-PapersOnLine* **2016**, *49*, 604–615. [\[CrossRef\]](#)
- Owen, S.H.; Daskin, M.S. Strategic facility location: A review. *Eur. J. Oper. Res.* **1998**, *111*, 423–447. [\[CrossRef\]](#)
- ReVelle, C.S.; Eiselt, H.A. Location analysis: A synthesis and survey. *Eur. J. Oper. Res.* **2005**, *165*, 1–19. [\[CrossRef\]](#)
- Üster, H.; Memişoğlu, G. Biomass logistics network design under price-based supply and yield uncertainty. *Transp. Sci.* **2017**, *52*, 474–492.
- Memişoğlu, G.; Üster, H. Design of a biofuel supply network under stochastic and price-dependent biomass availability. *IIE Trans.* **2021**, *53*, 869–882. [\[CrossRef\]](#)
- Geismar, H.N.; McCarl, B.A.; Searcy, S.W. Optimal design and operation of a second-generation biofuels supply chain. *IIE Trans.* **2021**, *57*, 322–335. [\[CrossRef\]](#)
- Nur, F.; Aboytes-Ojeda, M.; Castillo-Villar, K.K.; Marufuzzaman, M. A two-stage stochastic programming model for biofuel supply chain network design with biomass quality implications. *IIE Trans.* **2021**, *53*, 845–868. [\[CrossRef\]](#)
- Dréo, J.; Pétrowski, A.; Siarry, P.; Taillard, E. *Metaheuristics for Hard Optimization: Methods and Case Studies*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2006.
- Xiao, Y.; Konak, A. A genetic algorithm with exact dynamic programming for the green vehicle routing & scheduling problem. *J. Clean. Prod.* **2017**, *167*, 1450–1463. [\[CrossRef\]](#)
- Ghaderi, A.; Rahmaniani, R. Meta-heuristic solution approaches for robust single allocation p-hub median problem with stochastic demands and travel times. *Int. J. Adv. Manuf. Technol.* **2016**, *82*, 1627–1647.
- Aboytes-Ojeda, M.; Castillo-Villar, K.K.; Roni, M.S. A decomposition approach based on meta-heuristics and exact methods for solving a two-stage stochastic biofuel hub-and-spoke network problem. *J. Clean. Prod.* **2020**, *247*, 119176. [\[CrossRef\]](#)
- Defourny, B.; Ernst, D.; Wehenkel, L. Scenario trees and policy selection for multistage stochastic programming using machine learning. *Inf. J. Comput.* **2013**, *25*, 488–501. [\[CrossRef\]](#)
- Oroojlooyjadid, A.; Snyder, L.V.; Takáč, M. Applying deep learning to the newsvendor problem. *IIE Trans.* **2020**, *52*, 444–463. [\[CrossRef\]](#)
- Goettsch, D.; Castillo-Villar, K.K.; Aranguren, M. Machine-Learning Methods to Select Potential Depot Locations for the Supply Chain of Biomass Co-Firing. *Energies* **2020**, *13*, 6554. [\[CrossRef\]](#)
- Scavuzzo, L.; Aardal, K.; Lodi, A.; Yorke-Smith, N. Machine Learning Augmented Branch and Bound for Mixed Integer Linear Programming. *Math. Program.* **2024**. [\[CrossRef\]](#)
- Liang, Y.Y.; Shahabuddin, M.; Ahmed, S.F.; Tan, J.X.; Ali, S.M. Optimizing Sustainable Aviation Fuel Supply Chains: Challenges, Mitigation Strategies and Modeling Advances. *Fuel* **2025**, *402*, 135972. [\[CrossRef\]](#)
- Keith, K. A neural network powered solution approach for computationally expensive mixed integer programs for bio jet-fuel supply chain network design. In *Proceedings of Cyber-Physical Systems and Internet of Things Week 2023*; Association for Computing Machinery: New York, NY, USA, 2023; pp. 373–377.
- Rössig, A.; Petkovic, M. Advances in verification of ReLU neural networks. *J. Glob. Optim.* **2021**, *81*, 109–152.
- Tsay, C.; Kronqvist, J.; Thebelt, A.; Misener, R. Partition-based formulations for mixed-integer optimization of trained ReLU neural networks. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 3068–3080.
- Anderson, R.; Huchette, J.; Ma, W.; Tjandraatmadja, C.; Vielma, J.P. Strong mixed-integer programming formulations for trained neural networks. *Math. Program.* **2020**, *183*, 3–39. [\[CrossRef\]](#)
- Alumur, S.; Kara, B.Y. Network hub location problems: The state of the art. *Eur. J. Oper. Res.* **2008**, *190*, 1–21. [\[CrossRef\]](#)
- Farahani, R.Z.; Hekmatfar, M.; Arabani, A.B.; Nikbakhtsh, E. Hub location problems: A review of models, classification, solution techniques, and applications. *Comput. Ind. Eng.* **2013**, *64*, 1096–1109. [\[CrossRef\]](#)

28. Davis, R.E.; Grundl, N.J.; Tao, L.; Biddy, M.J.; Tan, E.C.; Beckham, G.T.; Humbird, D.; Thompson, D.N.; Roni, M.S. *Process Design and Economics for the Conversion of Lignocellulosic Biomass to Hydrocarbon Fuels and Coproducts: 2018 Biochemical Design Case Update; Biochemical Deconstruction and Conversion of Biomass to Fuels and Products via Integrated Biorefinery Pathways*; National Renewable Energy Laboratory (NREL): Golden, CO, USA, 2018. [CrossRef] [PubMed]
29. U.S. Census Bureau. U.S. Gazetteer: 2010, 2000, and 1990. 2012. Available online: <https://www.census.gov/geographies/reference-files/time-series/geo/gazetteer-files.html> (accessed on 23 January 2023).
30. Langholtz, M.H.; Stokes, B.J.; Eaton, L.M.; Brandt, C.C.; Davis, M.R.; Theiss, T.J.; Turhollow, A.F., Jr.; Webb, E.; Coleman, A.; Wigmosta, M.; et al. *Billion-Ton Report: Advancing Domestic Resources for a Thriving Bioeconomy, Volume 1: Economic Availability of Feedstocks*; Technical report; Oak Ridge National Laboratory (ORNL): Oak Ridge, TN, USA, 2016; Volume 1.
31. Center of Transportation Analysis. Railroad Network. 2017. Available online: <http://cta.ornl.gov/transnet/RailRoads.html> (accessed on 15 January 2023).
32. Lamers, P.; Roni, M.S.; Tumuluru, J.S.; Jacobson, J.J.; Cafferty, K.G.; Hansen, J.K.; Kenney, K.; Teymouri, F.; Bals, B. Techno-economic analysis of decentralized biomass processing depots. *Bioresour. Technol.* **2015**, *194*, 205–213. [CrossRef] [PubMed]
33. National Renewable Energy Laboratory. BioEnergy Atlas. 2017. Available online: <https://maps.nrel.gov/bioenergyatlas> (accessed on 15 November 2022).
34. Federal Aviation Administration. Airport Data and Information Portal. 2020. Available online: <https://adip.faa.gov/agis/public/public> (accessed on 17 November 2022).
35. Energy Information Administration. Transportation Sector Energy Consumption Estimates. 2020. Available online: https://www.eia.gov/state/seds/sep_sum/html/pdf/sum_btu_tra.pdf (accessed on 15 November 2022).
36. Project-OSRM. Project-OSRM/osrm-backend. 2017. Available online: <https://github.com/Project-OSRM/osrm-backend> (accessed on 21 November 2022).
37. Huang, Y.; Chen, C.W.; Fan, Y. Multistage optimization of the supply chains of biofuels. *Transp. Res. Part E Logist. Transp. Rev.* **2010**, *46*, 820–830. [CrossRef]
38. Abbas, D.; Handler, R.; Dykstra, D.; Hartsough, B.; Lautala, P. Cost Analysis of Forest Biomass Supply Chain Logistics. *J. For.* **2013**, *111*, 271–281. [CrossRef]
39. Roni, M.S. *Analyzing the Impact of a Hub and Spoke Supply Chain Design for Long-Haul, High-Volume Transportation of Densified Biomass*; Mississippi State University: Starkville, MS, USA, 2013.
40. Searcy, E.; Flynn, P.; Ghafoori, E.; Kumar, A. The relative cost of biomass energy transport. *Appl. Biochem. Biotechnol.* **2007**, *137*, 639–652. [CrossRef] [PubMed]
41. Lu, X.; Van Roy, B. Ensemble sampling. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 3260–3268.
42. Wasserstein, R.L.; Lazar, N.A. The ASA Statement on p-Values: Context, Process, and Purpose. *Am. Stat.* **2016**, *70*, 129–133. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.