Energy Conversion and Management: X 24 (2024) 100756

Contents lists available at ScienceDirect

Energy Conversion and Management: X

journal homepage: www.sciencedirect.com/journal/energy-conversion-and-management-x

One techno-economic analysis to rule them all: Instant prediction of hydrothermal liquefaction economic performance with a machine learned analytic equation

Muntasir Shahabuddin, Nikolaos Kazantzis, Andrew R Teixeira, Michael T. Timko *

Worcester Polytechnic Institute, Department of Chemical Engineering, Worcester MA 01609, United States

A B S T R A C T

Hydrothermal liquefaction (HTL) has remarkable potential for efficient conversion of abundant, decentralized organic wastes into renewable fuels. Because waste is a highly distributed resource with context-dependent economic viability, selection of optimal deployment sites is slowed by the need to develop detailed techno-economic analyses (TEA) for the thousands of potential deployment locations, each with their own unique combinates of scale, proximity to infrastructure/markets, and feedstock properties. An economic modeling framework that requires only easily obtainable inputs for assessing economic performance would therefore allow multiplexed analysis of many thousands of cases, whereas traditional TEA would not be possible for more than a handful of cases. Within such a context, the present study uses machine learning to guide development of a TEA and modeling framework which provides accurate cost predictions using three key inputs – feedstock cost, biocrude yield, and process scale – to estimate the minimum fuel selling price (MFSP) that an HTL process can achieve. The structure of the proposed framework is informed and based on empirical observations of cost projections made by a detailed TEA over a wide range of feedstock costs, biocrude yields, and process scales. A machine learning guided process was used to identify, train, and test a series of models using auto-generated data for training and independently reported data for testing. The most accurate model consists of three terms and requires 6 adjustable parameters to predict independently published values of MFSP (N = 28) to within an average value of ± 20.4%. It is demonstrated that the reduced-order model's predictions fall within 40% of the corresponding published values 95% of the time, and in the worst case, the associated discrepancy is 45.9%, suggesting that the accuracy of the machine learned model is indeed comparable to the TEAs that were used to build it. Moreover, the terms in the model are physically interpretable, conferring greater reliability to the use of its predictions. The model can be used to predict the dependence of MSFP on biocrude yield, scale, and feedstock cost; interestingly, MFSP is insensitive to biocrude yield and/or scale under many situations of interest and identifying the critical value for a given application is crucial to optimizing economic performance. The proposed model can be also extended to evaluate economic performance of newly developed HTL-based processes, including catalytic HTL, and the methodological framework used in this study is deemed appropriate for the development of machine learned TEA models in cases of other similar waste-to-energy technologies.

1. Introduction

Collection, composting, landfilling of waste and attendant management practices collectively constitute a major (~2%) source of greenhouse gas emissions, [1] with landfilling being the most common form of waste disposal [2]. When organic material decomposes anaerobically, the carbon it contains is released as methane—a greenhouse gas with a global warming potential approximately 20 times greater than that of CO_2—intensifying the urgency to address these coupled issues at the waste-climate nexus. On the other hand, the transportation sector is the single largest contributor to greenhouse gas emissions [3]. With the growing awareness of the risks associated with a "business as usual" scenario, lowering the emissions intensity of our energy sources and decarbonizing waste management strategies are both high priorities.

The parallel problems of waste management and transportation can be co-addressed by diverting organic waste from landfills and using it to produce alternative fuels suitable as replacements for fossil fuels. Biomass conversion technologies can be repurposed for waste conversion to make use of a variety of renewable carbon sources, such as wastewater, yard waste, and food scraps—materials traditionally destined for landfills—as viable substitutes for conventional fossil fuels. Processes like anaerobic digestion, fermentation, pyrolysis, and hydrothermal liquefaction (HTL) can all be utilized to convert wastes into energy forms such as ethanol, biogas, or biocrude oil, all of which can be used as fuels or fuel precursors [4].

Of the existing technologies, anaerobic digestion is the one most often associated with waste utilization in commercial practice [5]. While anaerobic digestion is a familiar and reliable technology option that is appropriate for wet wastes [6,7], it is hampered by slow reaction rates that result in incomplete conversion of organic materials and large reactor sizes [8]. Pyrolysis is a rapid process [9] that can achieve high

* Corresponding author.
E-mail address: mttimko@wpi.edu (M.T. Timko).

https://doi.org/10.1016/j.ecmx.2024.100756
Received 3 July 2024; Received in revised form 20 September 2024; Accepted 10 October 2024
Available online 14 October 2024

Nomenclature

Abbreviation	Term
HTL –	Hydrothermal Liquefaction
MFSP	Minimum Fuel Selling Price
GGE	Gallons Gasoline Equivalent
USD	US Dollar
AD	Anaerobic Digestion
TEA	Techno-economic Analysis
FCI	Fixed Capital Investment
DCFROR	Discounted Cash Flow Rate of Return
TDC	Total Direct Costs
NPV	Net Present Value
DTPD	Dry Tons per Day
HHV	Higher Heating Value
AIC	Akaike Information Criterion
RMSE	Root Mean Squared Error
LCA	Life-Cycle Assessment
WRRF	Waste Resource Reclamation Facility

conversion, [10] but it generates substantial char byproducts and requires an intensive drying process that can be energy-prohibitive for utilization of wet waste streams [11]. HTL fills the space between anaerobic digestion and pyrolysis, as it is compatible with wet wastes, achieves high conversion, and results in high carbon efficiency while being a relatively rapid process requiring a small footprint [12]. Drying is not required for most HTL feeds, [13] or the primary biocrude product [14–16]. Furthermore, char yields obtained using HTL are generally less than those observed for pyrolysis [17–19].

Although the aforementioned benefits present a compelling case for the use of HTL in waste management, HTL has not yet made commercial inroads [20]. A primary reason for industry hesitancy to adopt HTL is its perceived front-end risk; [21] HTL is a complex, [22] capital-intensive process due to its requirement of expensive reactors capable of handling the intense high pressure and temperature reaction conditions [16,23,24]. Despite numerous examples of HTL economic studies pointing to fossil-fuel-competitive minimum fuel selling prices (MFSPs) in the range of 2.00–4.00 US dollars (USD) per gallon gasoline equivalent (GGE) [14,20,25–33 34], the initial capital risk renders HTL a less attractive venture when compared with other more proven technologies such as anaerobic digestion (AD) [8].

The problem of estimating costs is exacerbated given that waste is a distributed resource, with thousands of point sources to consider [35]. Each of these locations corresponds to a unique combination of scale, access to infrastructure and markets, as well as feedstock properties and availability [36]. Traditional techno-economic analyses (TEAs) require hundreds of inputs that require significant resources to accurately measure, meaning that no tool exists that can be rapidly applied to the thousands of cases where HTL might be economically feasible. Quantifying the investment risk associated with early deployments is fundamental to advancing its learning curve and ultimately achieving broader adoption [37–39]. Furthermore, models that can provide accurate cost estimates for thousands of cases to prioritize investments can promote the early adoption required to demonstrate technological reliability, paving the way for more widespread commercialization [40]. Lastly, published TEAs have been performed for different feedstocks and scales, making inter-comparison difficult. What is needed is a unified approach for inter-comparison and decision making.

Techno-economic analysis (TEA) models can serve a critical function as they aim to quantify project economic outcomes, assess the risks associated with technological deployment under a realistic range of market and process conditions, as well as guide resource allocation decisions. Many TEAs have already been performed for a range of generalized scenarios where HTL apparently carries minimal risk and promises competitive MFSPs [33,41–49]. However, developing TEA models for actual scenarios expected during deployment is resource-intensive, requiring significant time investment and potentially hundreds of inputs, some of which have speculative or uncertain values and many of which are case-specific to a given deployment scenario [33,46,50]. Given their granularity and number of input parameters, these detailed TEA models can be termed "high-dimensional" [51].

In contrast to granular TEAs, machine learned and low dimensional TEA models can complement existing high-dimensional models for analysis of new scenarios for which granular input data are lacking. The challenge is to construct models that retain the accuracy of high-dimensional models by identifying the key inputs that underly the variability observed in actual practice. The obvious analogy is between detailed chemical models, which may consist of many thousands of reactions, [52] and reduced order chemical models, which focus on key reactions to retain accuracy while minimizing computational burden [53]. Detailed chemical models are required, and have been used to construct reduced order models — for example in prediction of HTL biocrude yields [54]— and in turn the reduced order models are used for applications such as optimizing reactor design where detailed models are too computationally expensive to be practical [55]. In recent years, machine learning has emerged as a tool for accurate yet computationally efficient predictions of the complex phenomena occurring in HTL systems [56]. Unfortunately, machine learning regressions are typically "black box" and as such are not easily interpreted by users [57]. The solution, proposed here, is to use machine learning methods to guide discovery of user interpretable relationships for accurate, computationally efficient TEA that can be used by the entire community, not just by experts in TEA or machine learning.

The objective of this study is to use machine learning to develop a user interpretable TEA and modeling framework for prediction of MFSP in cases of HTL conversion of organic waste to fuel. An existing high-dimensional model, published by Pacific Northwest National Laboratory [23], is used to identify which inputs have the greatest effect on predicted values of MFSP. Within the context of the present study, a family of machine learned TEA models is proposed to fit MFSP values predicted using the high-dimensional model and a handful of fitting parameters. The most robust machine learned model is identified by an iterative process consisting of regression to the training data set as predicted by the high-dimensional model followed by a comparison with a test data set consisting of MFSP values reported in independent TEA studies of HTL. The accuracy of the low-dimensional model is quantified, evaluated by comparison with the above test data, and its implications examined. It can be finally inferred that the resulting machine learned TEA model used as a reliable evaluative tool that binds previous studies; as it predicts the economic viability prospects of future projects it has the potential to accelerate adoption of HTL as an efficient waste-to-energy technology option.

The paper is organized as follows: Section 2 encompasses the structure of the proposed machine learning guided TEA and modeling framework, followed by Section 3 where the presentation of the proposed iterative model development, performance evaluation and the accompanying discussion can be found. Finally, a few concluding remarks are provided in Section 4.

2. Methodology

The goal of this study is to use machine learning to guide development of an accurate, computationally efficient, and user interpretable model for estimating values of MFSP for HTL processes. The desired end result can be termed a machine learning TEA model, since machine learning was used as a crucial tool during model development. Fig. 1 shows that the approach consists of a series of six steps: 1) A published high-dimensional TEA model was selected. 2) Key impactful variables/inputs are identified through a sensitivity analysis to be retained in a

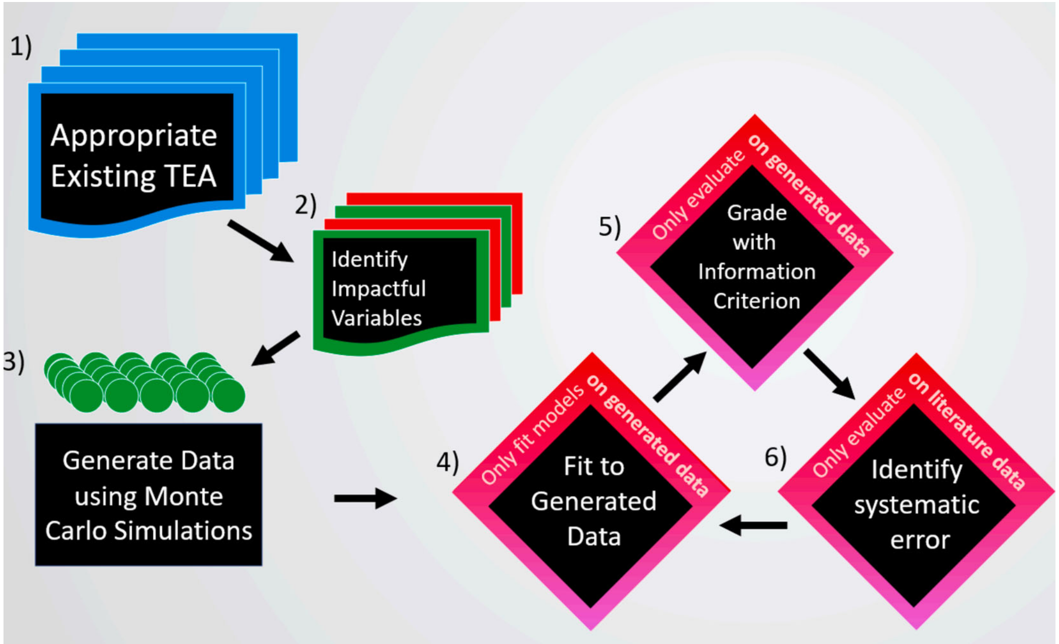

Fig. 1. Workflow used in this study for the studied TEA, generation of data, and iterative development and scoring of models.

machine learned model. 3) The high-dimensional TEA model was used with varying values of these key inputs over pre-determined ranges to generate a training data set consisting of random variations of the key inputs and the corresponding values of the MFSP. 4) A functional form for the machine learned TEA model was selected or perturbed using the proposed process and its empirical parameters regressed using the training data set generated in step 3. Predictions made by the new machine learned TEA model were 5) evaluated for their "goodness of fit" using a standard information criterion (discussed in Section 3), and then 6) compared with MFSP values reported in the literature from independent studies of HTL, with the literature data playing the role of a test data set in the overall model development process. Steps 4–6 were iterated to refine the functional form of the model to obtain satisfactory performance (r^2 of near unity against literature data) as determined by accuracy (+/- ~50% of the actual value in terms of mean absolute error) and other pertinent qualitative criteria and quantitative constraints, including a parity against literature data with passing through the origin, and a slope equal to unity.

2.1. Selection of an existing High-Fidelity Techno-Economic analysis (TEA) model.

Many TEA studies on HTL economic performance assessment have been published in the literature, [16,23,41,42,58] and of these the process economics model developed by Pacific Northwest National Lab (PNNL) was adopted as a starting point [23]. The PNNL model is representative of a well-developed and documented HTL process that operates continuously, includes heat recovery, and relies on ubiquitous natural gas infrastructure for heating. Many HTL TEAs have explored deviations from this base process design, which confers to it a status of community acceptance [42,59,60]. A simplified version of the process flow diagram in Snowden-Swan et al. can be found in SI-Fig. 2. The PNNL model has been widely accepted as a benchmark analysis of the HTL process to quantify improvements and predictive models [23].

The selected high-fidelity TEA follows a standard methodology: a mass and energy balance is developed following the process flow diagram, where the magnitude of each material and energy flow has monetary value attached to it to quantify variable operating costs and are used in capital cost scaling factors for each unit operation to calculate Fixed Capital Investment (FCI). Fixed operating costs are comprised of labor and maintenance. These operating costs populate annual cash flows which account for capital investment interest, taxes, overall plant life, and investor equity which modify Fixed Capital Investment (FCI) and operating cost estimates and are then embedded into the structure of a standard Discounted Cash Flow Rate of Return (DCFROR) modeling framework at a given discount rate (10% at base case). The key DCFROR economic parameters are held constant at the values used by Snowden-Swan et al. [23] and are listed in Table 1. Additional discussion of the DCFROR formulation and detailed TEA parameters is provided in SI-Table 1, with example calculations and validation values for capital and operating costs in SI-Tables 2 and 3. The sum of the annual cash flows from the DCFROR are then added to calculate the project's "Net Present Value" (NPV), which quantifies the project's value over its lifetime in present dollars. The Minimum Fuel Selling Price (MFSP) is then iteratively calculated such that the project NPV is equal to zero. In essence, this calculates the MFSP required to hit the rate of return (discount rate) as set by the project parameters. [23] The baseline model (and thus the fitted models) is reported in $USD 2016 equivalent. A constant value of 1.10 USD/GGE was added as a correction factor to the calculated MFSP to account for the cost of upgrading, following the recommendation of Zhu et al.[61] The cost of biocrude can be recovered by subtracting 1.10 USD/GGE from the MFSP

Table 1
Discounted cash flow rate of return (DCFROR) economic model parameters, as reported in [23].

Variable	Value
Equity	40%
Loan Interest	8%
Loan Term (yrs)	10
Working Capital Investment (as a % of FCI)	5%
Depreciation Period (yrs)	10
Construction Period (yrs)	3
% Spent in Year 1	8%
% Spent in Year 2	60%
% Spent in Year 3	32%
Internal Rate of Return	10%
Income Tax Rate	21%
Plant Life (yrs)	30

Table 2
Variables and bounds for Monte Carlo simulations.

Variable	Lower Bound	Upper Bound	Source/ Justification
Higher Heating Value (HHV, MJ/kg)	30	40	[18,23,42,63,64]
HTL Scale (DTPD Feedstock)	3	500	[15,23,50,65,66]
Feedstock Cost (USD/dry ton)	−200	200	[41,42,44,45,61]
Yield (wt%)	10%	60%	[49,67–71]
Capital Cost Factor (%Total Direct Costs)	−50%	50%	+/-50% TDC from [23]
Electricity Cost (USD/kWh)	3	12	[72]
Natural Gas (USD/scf)	1	10	[23,42,44,45]
Ash Content in dry solids (wt %)	0%	30%	+/- 100% from [23]
Feedstock Solids loading, ash free (wt%)	5%	45%	+/- 80% from [23]
Overhead/Maintenance Factor	0%	180%	+/-100% from [29]

Table 3
Error metrics of this study's model against published MFSPs.

Statistical Metric	Value	Unit
Number of Points Considered	28	–
Parity R^2	0.960	–
Mean absolute Error (MAE)	1.32	USD/GGE
Mean Squared Error (MSE)	3.04	$(USD/GGE)^2$
Root Mean Squared Error (RMSE)	1.74	USD/GGE
Mean % Absolute Deviation	20.4%	–
Std. Dev. of % Abs Dev.	12.2%	–
Max % Absolute Deviation	45.9%	–
Min % Absolute Deviation	1.4%	–

predicted by the models presented here. More details are reported by Snowden-Swan et al. [23].

2.2. Sensitivity analysis to identify key input parameters using generated data

The next step (comprising of steps 2 and 3 in Fig. 1) involved a sensitivity analysis to identify key TEA model inputs that determine the MFSP output. A series of 100,000 probabilistic simulations was performed using the PNNL model, where all key model inputs were considered as random variables following appropriately chosen probability distributions bounded by the values summarized in Table 2. In these simulations, the Monte Carlo method [41,42,44] was used to sample key input values from the aforementioned distributions and over their prescribed respective numerical ranges selected to represent the full range of MFSPs observable – including studies published in the pertinent literature (as shown in Table SI-3).

One of the most important inputs included in the sensitivity analysis is process scale. Simple power-law relationships were used in Snowden-Swan et al. to model the dependence of individual equipment capital costs on scale, with values ranging between 0.3–0.8 used for estimating the costs of various pieces of equipment. [23,62] In the present study these scaling factors are unchanged from Snowden-Swan et al.'s model. Operating costs, on the other hand, vary linearly with mass and energy balances. Further details regarding operating cost calculations and process parameters are detailed in the SI. The result of the Monte Carlo simulations was 100,000 sets of model inputs with their corresponding MFSP output values.

The input–output data set arising from sensitivity analysis was analyzed using a Pearson (linear) correlation plot, in which a Pearson (linear) correlation analysis was performed for each variable against MFSP and the corresponding value of the correlation constant (r^2) was determined. The Pearson correlation coefficient measures the linearity between two variables [73], which this study uses to estimate the relative contribution of a given input variable to MFSP output. Accordingly, MFSP prediction outputs are most sensitive to inputs which correspond to the highest r^2 values.

Following identification of key (most impactful on MFSP) parameters, new Monte Carlo simulations were performed to generate a set of 10,000 data points involving the identified key model inputs and the corresponding MFSP outputs. In these simulations, biocrude yield was varied between 10% – 60%, process scale between 3 and 500 DPTD, and feedstock between –200 – +200 USD/dry ton. Uniform probability distributions covering the above ranges were assumed for the above key inputs, and each input was allowed to vary independently of the others. The effect of each variable on capital and operating expenses was modeled as described previously for the sensitivity analysis, following economies of scale and linear adjustments to equipment sizing and material/energy balances respectively.

2.3. Techno-Economic analysis model development via regression

The resulting input–output data set constitutes a set of training data derived from 10,000 input–output test cases that was then used to regress a series of combinatorially generated models (described in Section 3.3, representative of step 4 in Fig. 1). The construction of each of these models was accomplished using Python's symbolic math library, SymPy (ver. 1.12)[74], and each model's respective parameters were fitted to the generated data using the curve fit optimization methods present in the SciPy Python library (ver. 1.11.2).[75] SymPy was used to filter mathematically equivalent models further described in Eq. (4) while the SciPy library was used to fit generated data from the model described in Section 2.2 to each of the hypothesized models using the Trust Region method.

To assess the relative performance of each model, the Akaike Information Criterion (AIC) was utilized, as described by Eq. (2) [76,77]:

$$AIC = N^{*}LN(MSE) + 2k \tag{2}$$

where N is the number of data points included in the regression, MSE is the mean squared error performance of the model against the data it was fit to (i.e. the generated data), and k is the number of fitted parameters. AIC is used to balance model accuracy with the number of fitting parameters to reduce the natural tendency to overfitting.

2.4. Model evaluation and identification of systematic error

The accuracy of empirical models selected based on the AIC analysis was determined by comparison with 28 literature data points published independently of the PNNL data set (in search of systematic error – step 6 of Fig. 1), detailed in Table SI-4. These data points represent all TEA studies published as of 2022, with the only exception being a mobile study that was deemed to differ qualitatively from the other data due to significantly different process operation conditions and the unique inclusion of fuel subsidies.[26] The literature data constituted HTL TEAs processing a variety of wet wastes, including lignocellulosic yard wastes, wastewater sludge, microalgae, and food waste, with broadly similar processes but with substantive differences in reaction conditions such as residence times and processing temperatures/pressures. Scales considered by these studies range from 1 to 2000 DTPD of feed processed, and feedstock costs varied from − 50–1800 USD/dry ton. For studies that perform uncertainty analyses, we include the median case and the confidence intervals reported each as separate points in the validation. Collectively, the studies in the test set represent multiple economic conditions in the US and various European countries. For studies that report MFSPs in alternative currencies, exchange rates were used at the date of publication. A comprehensive table of considered TEAs is included in the SI – also reported primarily in USD 2016, though European studies converted to USD were left in their reported years to avoid over manipulation of the original values.

To compare a given model's performance against available literature data, the reported scale, yield, and feedstock costs were collected from each TEA study appearing in the literature and used to predict MFSP values. This "predicted" MFSP was then compared against the available literature values – ideally resulting in a model predicting values as close to the "real" literature values as possible. Literature data are solely involved in model evaluation and deliberately isolated from model formulation to avoid fitting a model with "knowledge" of values it will be evaluated against. Accordingly, the approach described in Sections 2.3 and 2.4 can be considered as consisting of model training (or regression) and model testing procedures respectively. Model performance is examined qualitatively to guide further refinement in an iterative process.

2.5. Iterative model improvement

The TEA model structure development, training, and testing/evaluation steps (steps 4–6 in Fig. 1) were performed over multiple iterative cycles. In each of these cycles, the quantitative and qualitative performance (described in Section 2.4, and further in the discussion) of the regression models were considered for generation of new model forms. The new models were regressed using the training data and evaluated using the test data. When tested, models with systematic error (i.e. low accuracy, high precision) are evaluated and the iterative process is continued. The iterative process was concluded when r^2 approximately converged to unity and the qualitative criteria of slope equaling 1.0 and intercept equaling 0.0 were satisfied, within reasonable bounds of uncertainty.

3. Results and discussion

The objective of this work is to develop a machine learned yet accessible techno-economic analysis (TEA) and modeling framework that uses readily available inputs to predict HTL MSFP for different options of feedstock type and process scale. The starting point was a sensitivity analysis of a published TEA [23] to identify the inputs with the greatest effect on overall process economics, as captured by the predicted minimum fuel selling price (MFSP). Several different types of empirical models were then developed using the inputs identified from sensitivity analysis; this step can be considered as model training. The accuracy of these models was then evaluated using 28 MFSP values that have appeared in the published literature; this step can be considered equivalent to model testing. The training–testing process resulted in a machine learned TEA model capable of estimating MFSP with minimal measurable inputs. The ensuing Discussion is organized in a similar structure as shown in Section 2: 1) sensitivity analysis; 2) training and testing of several generations of models; 3) identification and recommendation of the most accurate and robust model for future use.

3.1. Sensitivity analysis

The first step was sensitivity analysis of the TEA published previously by PNNL to determine the key variables to include in further consideration in model development [23]. The Monte Carlo simulation method, with relevant variables, distributions, and simulation parameters described in Table 2, was used to generate a family of predicted MFSP values for many different values of 10 key input parameters. Fig. 2 shows the results of a Pearson correlation analysis of these simulations, indicating that feedstock cost, HTL plant scale (or feedstock dry flow rate), and process biocrude yield have the greatest effect on MFSP. Several published studies corroborate the finding that these three variables are the most important inputs for determining MFSP [14,23,42,78]. Conceptually, these three inputs are orthogonal to one another: process scale directly impacts the capital investment and material flow rates (operating cost); feedstock cost is a key (and often expensive) component of operating cost; and yield defines the amount of product sold – in the context of a TEA, these are each independent degrees of freedom. Any accurate model will therefore require these three inputs.

In addition to plant size, feedstock cost, and biocrude yield, the sensitivity analysis identified overhead/maintenance, solids loading, and ash content as the next three most important variables determining MFSP. Of these, overhead/maintenance might reasonably be expected to correlate with plant size in a machine learned model, reflected in a weak (but observable) positive correlation. Accordingly, as a first approximation and in the effort to keep the model as simple, general, and interpretable as possible, overhead/maintenance was not considered as

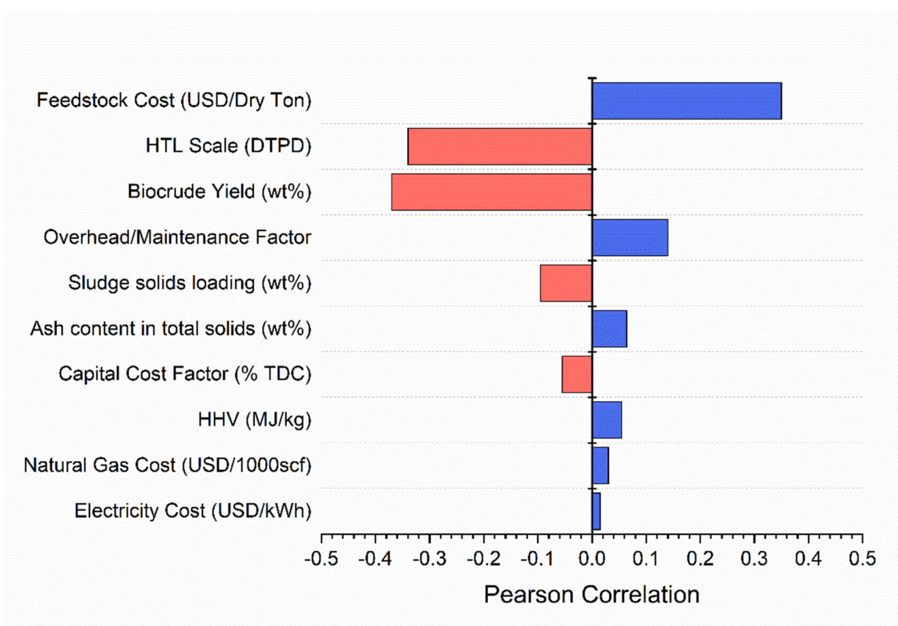

Fig. 2. A Pearson correlation plot detailing the linear correlation of each variable under study against MFSP. Feedstock cost, HTL Scale, and Yield have the highest absolute correlation values, and thus the greatest expected impact on MFSP.

an input parameter for initial model development. We revisit this decision later.

Unlike overhead/maintenance, solids loading, and ash content are nearly orthogonal to plant size, feedstock cost, and biocrude yield due to their association with feedstock composition, which indicates that including these variables in a machine learned model might improve accuracy. On the other hand, the values of the correlation constants between solids loading or ash content and MFSP are much less (< 0.1) than those between feedstock cost, plant size, or biocrude yield and MFSP. Accordingly, including solids loading and ash content in a machine learned model runs the risk of undesirable overfitting. Furthermore, adding more input requirements detracts from model simplicity. A machine learned model consisting of feedstock cost, plant capacity, and biocrude yield was deemed suitable for displaying the requisite balance between accuracy, over fitting, and complexity. All further model development work used these three input parameters.

3.2. Machine learned TEA models consisting of linear combinations

Having identified feedstock cost, plant capacity, and biocrude as the three required model inputs, the next step was to develop accurate models for estimating MFSP values. Many different functional forms are conceivable as postulates. MFSP values resulting from the Monte Carlo simulation runs were used to narrow down the functional form that might best capture the relationships between input and output variables. As a starting point, Fig. 3 plots predicted values of MFSP varying each of the three key inputs alone, with all other input variables held constant at their baseline values. Fig. 3 demonstrates that the relationships between biocrude yield or plant scale and MFSP are power laws; the relationship between MFSP and feedstock cost is satisfactorily depicted as linear. The empirical observations from Fig. 3 were used to guide all subsequent models.

The analysis shown in Fig. 3 encourages a simple "First-Order" model consisting of the following structure/terms:

$$MFSP = a^{*}scale^{b} + c^{*}yield^{d} + e^{*}feedcost \quad (3)$$

where *a*, *b*, *c*, *d*, and *e* are empirically fit parameters with values defined from regression analysis in Fig. 3. Eq. (3) can be considered as a linear combination of single-variable monomials of plant scale, biocrude yield, and feedstock cost with degrees b, d, and 1 respectively. The predictions of Eq. (3) were then evaluated/tested against 28 values of MFSP published in the literature (as listed in the SI). Fig. 4 is the resulting parity plot, which ideally should result in a linear line with slope equal to 1.0, a correlation constant (r^2) approaching 1.0, and minimal outlying points. Unfortunately, the correlation between Eq. (3) and published MFSP values suffers from two major problems that are apparent in the best-fit line: 1) the slope is substantially greater than 1.0 and 2) there are numerous outlying data points. Scaling the entire response by the

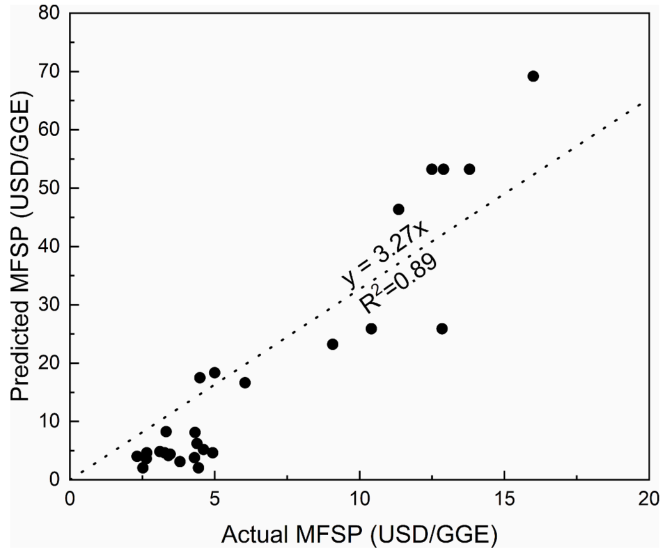

Fig. 4. The parity of the model depicted in eq. (3) (a linear combination of the "elementary" relationships of yield, scale, and feedstock cost against literature MFSPs.

deviation in the test data shown in Fig. 4 would "fix" the first of these problems; however, that *ad hoc* solution has no physical basis and would undoubtedly result in model overfitting while not addressing the outlying data points.[76] Accordingly, the model represented by the linear combination of feedstock cost, biocrude yield, and plant capacity as shown in Eq. (3) is inadequate for use as an accurate and predictive machine learned model.

The model corresponding to Eq. (3) is limited at least in part because the linear combination of the three inputs does not allow for interactions between them, a limitation arising from the restriction placed on Fig. 3 that only one input was allowed to vary at a time. To investigate the effects of this restriction, results from the Monte Carlo analysis were used to construct plots of predicted MFSP with respect to values of each of these three key inputs while easing the restriction that the values of the other two must be held constant at their respective baseline values. Fig. 4 summarizes the result of this test. The dashed lines provided in Fig. 4 are the trend lines expected from the simple analysis shown in Fig. 3. Unlike for the simpler analysis, the power law and linear relationships that are obvious for independent variation of each input variable on its own no longer persists for the general case that allows all three inputs to vary at the same time. Instead, MFSP only very roughly follows the expected power law (biocrude yield or plant scale) or linear (feedstock costs) relationships expected from predictions made when only one input is varied at a time. Since the three key inputs are

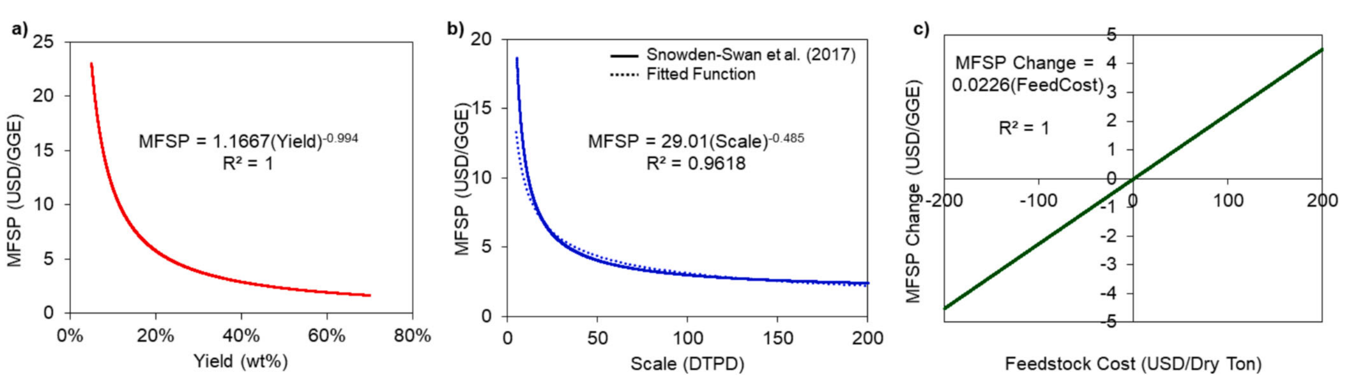

Fig. 3. The behavior of a) yield b) scale and c) feedstock cost on MFSP when varied independently for the generated dataset. Yield and scale follow inverse power law behavior, while feedstock cost adjusts MFSP linearly.

presumed to be independent of one another – and hence allowed to vary independently – Fig. 4 establishes that any accurate model must capture interactions that arise from variation of the three inputs with respect to one another. Variation in each input should, in aggregate, be expected to capture the costs accounted for in a standard TEA: the capital costs associated with equipment, the fixed operating costs (maintenance, insurance, etc.) and variable operating costs (chemical/feedstock costs incurred to make the product), each presented in marginal costs per unit product.

3.3. Machine learned TEA models consisting of linear combinations and interaction terms

Fig. 5 establishes that none of the three independent variables used as inputs can predict MFSP on their own. Eq. (3) shows that accounting for each in a linear combination does not result in an accurate model either – which implies that each variable must interact with one another, giving rise to cross terms with coupled interdependence. The approach for a rational selection of cross terms for inclusion in a TEA reduced-order model is not obvious *a priori*, and so the validity of mathematically equivalent permutations of each term can be examined as follows. Within the context of the present study, the following structure-postulate is introduced for the pertinent functional form:

$$MFSP = f(scale, yield, feedcost) = \sum_{|i|=1}^{3} f^{i}$$

where: $f = [f_1; f_2; f_3]$, $f^i = f_1^{i_1} f_2^{i_2} f_3^{i_3}$ with $|i| = i_1 + i_2 + i_3$ and $i_1, i_2, i_3 \in \{0, 1, 2, 3\}$, that encompasses first, second and third-degree monomials/cross terms of the constitutive elementary functions:

$$f_1(scale) = a_{1,i} * scale^{b_{1,i}}$$

$$f_2(yld) = a_{2,i} * yld^{c_{2,i}}$$

$$f_3(feedcost) = a_{3,i} * feedcost$$

Therefore, to capture all possible contributions associated with first, second and third-order cross terms (i.e., terms including the impacts of multiple independent variables), the combinatorial method yields multiple potential distinct models whose empirical parameters are appropriately fitted, with details of equation forms provided in SI-Table 5. For each model, two figures of merit were calculated – the root mean squared error (*RMSE*) and the Akaike Information Criterion (AIC) value [77]. The use of the RMSE represents a familiar criterion in the development of empirical models. However, one drawback of the RMSE is that minimizing this value could lead to overfitting through the introduction of a high number of (possibly extraneous) fitting parameters. Overfitting detracts from the use of empirical models in new situations, an outcome which should be avoided. Consequently, AIC can be used to identify overfitting. Unlike RMSE, AIC penalizes a model for overuse of parameters that lead to incremental accuracy benefits – as such, AIC can be used to identify and avoid overfitting. Quantitatively, lower values of the AIC denote models that perform better statistically while minimizing the number of fitted parameters. As a further note, AIC values can be used only for comparison of one model to another as their absolute values are meaningless on their own.

AIC values are plotted in Fig. 6 as a function of the number of fitted parameters included in the TEA model. As a general trend, AIC values tend to decrease when the number of parameters is increased. Interestingly, each value of the number of fitting parameters (i.e., 1, 2, 3, or more fitting parameters) corresponds to a set of potential models, with widely varying AIC values. Clearly, not all possible models are equally useful. Accordingly, special attention is directed to the “best models” identified by the combinatorial approach, where the best models correspond to the ones with the lowest AIC values. Here, the AIC values of three models stand out from the rest. Results from those three models, which have 5, 6, and 8 fitting parameters, are highlighted in Fig. 6 as the red, purple, and green points, respectively. Their corresponding AIC values are roughly 600, 700, and 1300, respectively. These values are much less than those of competing models with the same number of fitting parameters, by as much as a factor of 50. Because of their

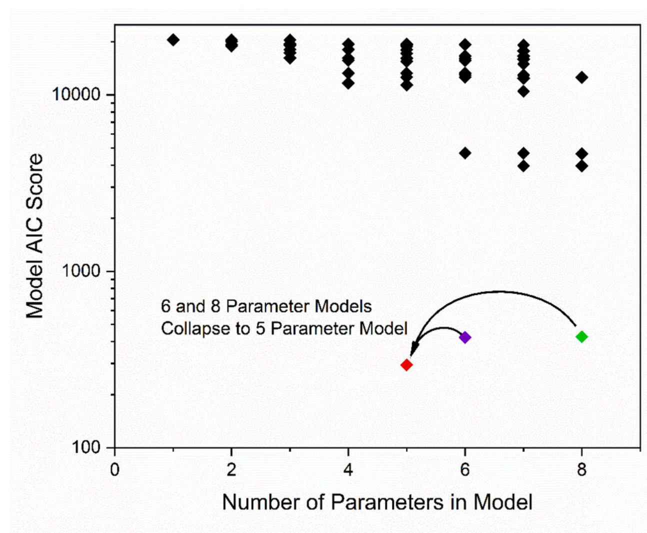

Fig. 6. AIC scoring of combinatorically developed models plotted against the number of fitted parameters for each model. The best performing models have a low overall score.

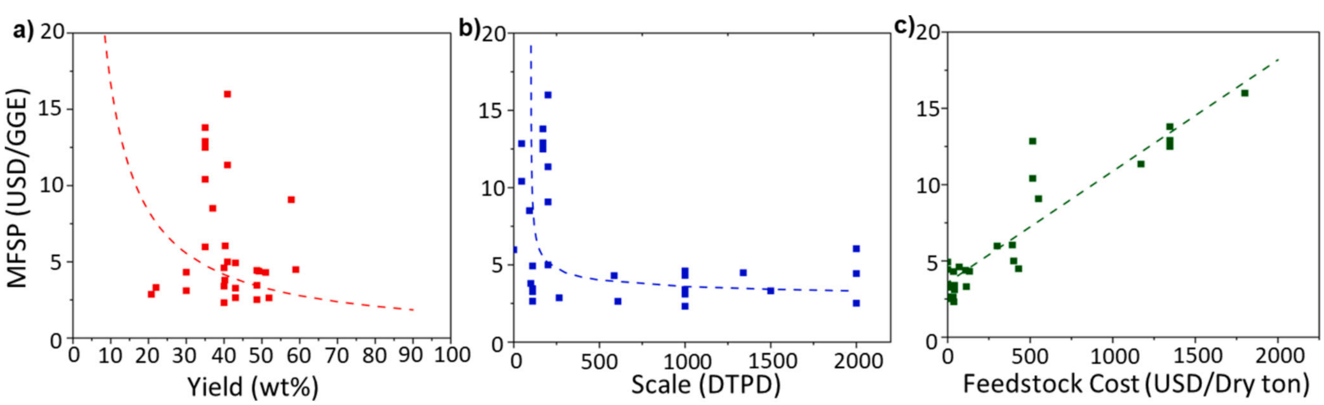

Fig. 5. The MFSP of literature data when plotted against a) yield b) scale and c) feedstock cost. The expected qualitative relationships between MFSP and the variables under study (as developed in Fig. 3) are shown in dotted lines.

promising performance, these three models were examined more closely, and all other models were excluded from further consideration.

Following the methodology of combinatorial treatment explained at the beginning of Section 3.3, closer examination of the three best performing models highlighted in Fig. 6 reveals that the six- and eight-parameter models both converge to the best performing five parameter model. The forms of these models are shown in Eqs. (5)-(7) (corresponding to the green, purple, and red points respectively in Fig. 6) showing that combining two parameters appearing in the eight-parameter model results in the six-parameter model; likewise, combining three of the parameters in the eight-parameter model into a single parameter results in the five-parameter model:

$$MFSP = a_1 * \frac{scale^{b_1}}{yield^{c_1}} + a_2 * \frac{feedstockcost}{yield^{c_2}} + a_3 * scale^{b_2} yield^{c_3} \quad (5)$$

$$MFSP = a_1 * \frac{scale^{b_1}}{yield^{c_1}} + a_2 * \frac{feedstockcost}{yield^{c_2}} + a_3 * feedcost \quad (6)$$

$$MFSP = a_1 * \frac{scale^{b_1}}{yield^{c_1}} + a_2 * \frac{feedstockcost}{yield^{c_2}} \quad (7)$$

Accordingly, comparing Eqs. (5)–(7) indicates that the three models identified as the most promising by the AIC analysis are equivalent to one another. They all reduce to the five-parameter version, which incidentally has the lowest AIC value of any generated by the combinatorial method used in this study. For these reasons, the six- and eight-parameter models were rejected from further consideration so that the five-parameter model could be developed further.

The next step was to evaluate the performance of the best five-parameter model using the MFSP values in the test data retrieved from the literature (listed in Table SI-4). Recall from Fig. 1 that the values in the model were fit using data generated using the PNNL model, meaning that comparison with literature values is an independent test of model accuracy. Fig. 7 compares MFSP values predicted by the best five-parameter model with literature values as a parity plot. Overall, the correlation constant obtained using the new five-parameter model is comparable to the earlier version consisting of the linear combination of the power law and linear terms. The error of the most extreme outlying point decreases from 200% to 125% using the new model, which is a promising trend but not entirely satisfactory. Moreover, the five-parameter model tends to systematically underpredict MFSP for values greater than 12 USD/GGE and overpredicts at values less than 6 USD/GGE – which are undesirable results. The value of the best-fit slope decreases from roughly 3 for the simpler model to 2, which is promising but not yet the ideal slope for parity, which should equal 1.0. Collectively, these observations point to a source of persistent systematic error left uncaptured by the approach thus far.

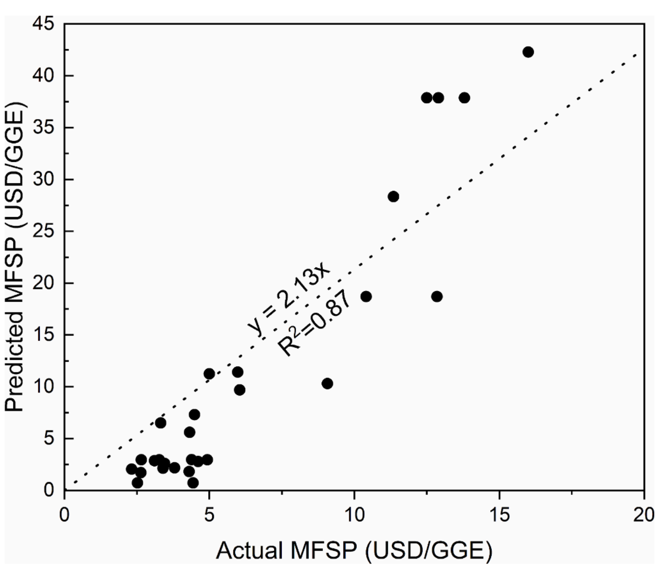

Fig. 7. Parity of the five parameter model depicted in Eq. (7), identified by combinatorial testing of potential models.

3.4. Model refinement

To this point, the analysis has been performed entirely through a multivariate regression-based methodological approach-that is, a traditional machine learning approach. That stated, a goal of this work was a user interpretable model rather than as a black box, which the machine learning guided approach has in fact achieved. To further explore this, we present Eq. (7) from the perspective of the economic "phenomena" that comprise it through Eq. (8). Specifically, the first term in Eq. (8) arises from observable phenomena associated with economies of scale. The second equality in Eq. (8) represents the five-parameter model with each term broken into the "physical phenomena" they represent. The form shown in the righthand side of Eq. (8) shows that MFSP depends on two terms. The first term depends on system scale divided by yield, which can be interpreted as the marginal "capital cost" portion of the MFSP. The second term is a function of feedstock cost divided by yield, corresponding to the "variable operating cost," at least for the limiting situation when feedstock cost dominates operating costs.

$$MFSP = a * \frac{scale^b}{yield^c} + d * \frac{feedstockcost}{yield^e} = f_1\left(\frac{CAPEX}{YIELD * scale}\right) + f_2\left(\frac{OPEX}{YIELD * scale}\right) \quad (8)$$

The simplified re-casting of Eq. (8) and its user interpretable form motivates a search for a missing term that accounts for its failures. Reconsidering Fig. 2, the variable with the next highest sensitivity after scale, yield, and feedstock cost is maintenance/overhead. When considering the cumulative effects of additional factors not included in our combinatorial model, maintenance/overhead, and remaining chemical/utility costs significantly contribute to an unaccounted term that captures *fixed* operating costs and variable operating costs other than those associated with feedstock. MFSP is only weakly responsive to several other similar costs in terms of sensitivity — like capital discounting and utilities' contribution to costs — but in aggregate these terms can have a notable impact on total MFSP. This impact would not be revealed by a sensitivity analysis such as that of Fig. 2 because it is the aggregate effect and not the isolated effect that must be considered. Since the five-parameter model does not include fixed operating costs, non-feedstock dependent variable operating costs, or discounting, (which are not implicitly captured by the most sensitive variables considered in our model), its failures become more understandable. In this case, the desire for the simplest possible model led to the formulation of a model that was too simple to explain the complexity of the phenomenon in question. In other words, the five-parameter model underfits, leading to its breakdown during testing by comparison with literature data.

Fortunately, once the problem of under fitting is diagnosed, it is easily corrected. Recalling the previous discussion of Fig. 2, a solution is realized through the addition of a sixth parameter that accounts for the previously neglected contributions of operating costs encompassing labor, overhead, maintenance, and remaining variables unaccounted for by the three variables considered in the five-parameter model. The form of the newly proposed six-parameter model is presented as Eq. (9a), where a_3 is the "remaining operating cost" term.

$$MFSP = a_1 * \frac{scale^{b_1}}{yield^{c_1}} + a_2 * \frac{feedstock\ cost}{yield^{c_2}} + [a_3] \quad (9a)$$

$$MFSP = \underbrace{6.607^* \frac{scale^{-0.6577}}{yield^{1.195}}}_{marginalCAPEX} + \underbrace{0.00321^* \frac{feedstock\ cost}{yield^{1.062}}}_{marginalOPEX_{variable}} + \underbrace{2.698}_{remainingOPEX} \quad (9b)$$

Best-fit values determined from model training are shown in Eq. (9b), where the variables of Eq. (9) have units as follows: units of scale are in dry tons per day (DTPD), units of yield are in weight fraction (ranging from 0 to 1, corresponding to 0% and 100% yields), units of feedstock cost are in USD/dry ton, and the final MFSP is returned in USD/GGE. As with the five- parameter model, the six-parameter model consists of two terms that are inversely related to yield (roughly linearly) and represent capital costs (related to scale approximately following the economies of scale “0.6 rule”[62]) and variable operating costs (linearly adjusted by feedstock cost). The third term captures remaining operating costs, that is, costs which have no dependence on feedstock and hence are independent of scale. The AIC value calculated for the regression applied to the six-parameter model to the training data set is

−3900, which is a notable improvement over the five-parameter model’s score of 500. The six-parameter model shown in Eq. (9a) resembles ones that have been proposed for levelized (discounted) costs of electricity production [79], a reassuring physical similarity given the combinatorial approach used to generate Eq. (9a).

As with the other models, the accuracy of the six-parameter model was determined by comparison of its prediction with the test data obtained from previously published TEAs. The feedstock cost, yield, and scale associated with 28 MFSPs from 15 published HTL TEAs (SI-Table 4) were collected and used in this study’s MFSP prediction model. Fig. 8 is the corresponding parity plot. Qualitatively, the six-parameter model does not systematically over- or under-predict costs at any value of MFSP, unlike the five-parameter model. Figure SI-1 confirms this qualitative observation as a percent deviation plot as a function of MFSP.

Table 3 summarizes the results derived from testing the proposed six-parameter model. Quantitatively, the correlation constant is the most robust observed in this study (r^2 = 0.9604). The mean absolute error (MAE) represents the average absolute distance of a predicted point from the true study’s value – which alone implies that on average, the model’s estimate will have an error of ± 1.32 USD/GGE. The root mean squared error (RMSE) criterion measures the modeling error with a bias towards larger deviations from the model; it has a slightly higher value at: ±1.74 USD/GGE. The percent absolute deviation normalizes for the magnitude of the predicted value: on average the model predicts a ±

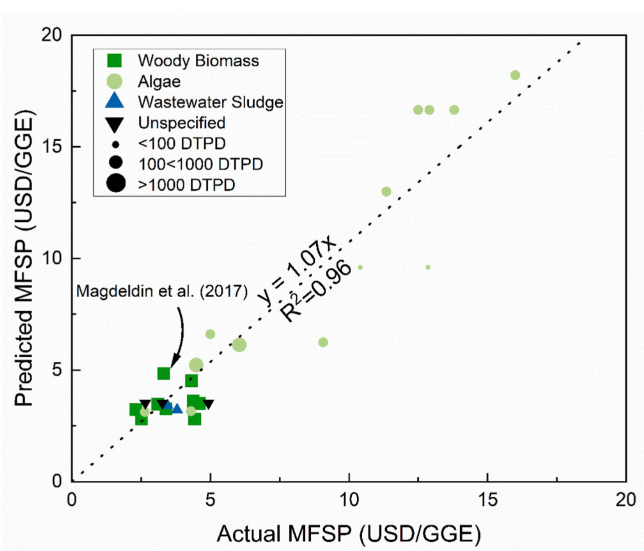

Fig. 8. A parity plot of the known MFSP taken from prior TEAs compared with predictions from the six parameter model shown in Eq. (9).

20.4% deviation with a maximum observed deviation of − 45.9%. Data points in Fig. 8 are color coded and shape selected to represent the type of biomass; their marker size is rated to process scale. No obvious trends between accuracy and feed type or process scale are evident in Fig. 8, as required of a robust correlation that is neither overfit nor underfit.

Achieving ± 20% accuracy in a machine learned model makes it useful in many situations. In fact, any TEA performed to estimate economic viability of HTL is – at present – a projection, since HTL is not yet operated at a commercial scale required for hard data. As result, the inputs to the TEA can never be known with certainty. Uncertainty in the inputs propagates to uncertainty in the TEA projections. Starting with the granular TEA used here, ±10% variation in the three key inputs (scale, biocrude yield, and feedstock cost) corresponds to + 17% and − 14% uncertainty in projected values of MFSP. Accordingly, uncertainty analysis establishes that the accuracy of the model is more than sufficient given realistic uncertainty in the input parameters, and in fact error relative to the predicted value (% absolute deviation in Table 3) is relatively consistent regardless of the actual predicted value – as seen in the weak correlation of error against predicted MFSP in SI-Fig. 1.

The highest deviation observed in Fig. 8 is 45.9%, which arises from an overprediction of MFSP reported in a study by Magdeldin et al.[33] The Magdeldin et al. study [33] includes the sale of byproduct hydrogen and biochar that are sold alongside to supplement the revenue stream provided by sale of the biocrude stream, offering a qualitative explanation for the overpredicted MFSP [33]. There are no other notable outliers – the points with deviation greater than ∼>35% comprise mainly of confidence intervals provided by studies performing uncertainty analyses. In other words, the other outliers are already regarded as uncertain in the original studies. Regardless of predicted error in the developed model, the irreducible uncertainty in input parameters equally impacts the accuracy of the machine learned model and the granular, high dimensional version, meaning that 20% accuracy is as good as can reasonably be expected.

The physical interpretability of Eq. (9) is particularly useful in accounting for process upgrades or deviations from standard practice. For example, and as mentioned in the Methods section, the study by Aierzhati et al. was excluded from the analysis as it modeled economic performance of a mobile process whereas all of the other studies modeled fixed processes [26]. Interestingly, when the fuel cost subsidies included in the Aierzhati et al. study are accounted for (as shown in the SI), Eq. 9 predicts Aierzhati et al. MFSP to within 3.5% [26]. The agreement may in part be fortuitous, but it nonetheless is an extreme stress test of the model that builds confidence in its usage.

As an additional note on usage, the correlation between the predicted MFSP values and the test data is biased towards higher fuel costs, which is a natural consequence of numerical regression – i.e., larger values can result in larger errors if not properly fit, meaning that regression will tend to overemphasize these data points. That stated, it is important to clarify that these higher MFSP points were not involved in fitting of the model. Indeed, HTL processes resulting in higher fuel costs (e.g., MFSP > 10 USD/GGE) are inherently less useful than more realistic values in the target range of 2–5 USD/GGE (though Figure SI-1 shows that the relative error of the model is insensitive to the magnitude of predicted MFSP) [23]. Accordingly, as more TEAs become available for realistic values of projected MFSP, Eq. (9) can be refit for better accuracy, especially to account for the importance of terms not explicitly captured by the six-parameter model and that may become important for economically viable MFSPs (see Fig. 2). Future improvements notwithstanding, the model in its current form will be sufficiently accurate for many applications.

The six-parameter model described by Eq. (9) can be considered as the most appropriate for predictive use among those considered by this study. While additional terms could be added to improve model accuracy when compared with the collected literature data, Eq. (9) has the advantage of minimizing overfitting while using inputs that are readily available and readily interpretable. Minimizing overfitting is important to the generalization of scenarios not explicitly considered during model development. For example, adding terms that account for byproduct yields – especially char, which might require disposal, and aqueous phases, which require treating – might improve accuracy. However, byproduct yields are constrained by carbon balance and hence related to biocrude yield; as biocrude yield increases, byproduct yields must naturally decrease. For this reason, Eq. (9) must implicitly capture the effects of byproduct yields on MFSP, despite lacking explicit terms. Similar arguments apply to other would-be correction factors.

A key advantage of Eq. (9) is that it relies only on inputs that are easily obtained. Feedstock costs can be determined for different waste streams, initially based on published surveys [80] and for greater accuracy based on discussion with local waste generators. Scale is determined by the availability of waste streams. Biocrude yield can easily be measured experimentally; machine learning models now exist that predict biocrude yield based on feedstock composition – which is itself easily measured and often known in advance – and operating conditions [81,82]. Because of these features, Eq. (9) can be used for intercomparison of published TEAs, as summarized in Table SI-4 which converts published values of MFSP to a common basis of a 100 dry ton/day HTL process. Scale normalized values of MFSP shown in Table SI-4 demonstrate that studies that assume scales $> 1{,}000$ dry ton/day gain an apparent advantage in MFSP that vanishes when brought to a common scale. Unless a specific reason argues otherwise, results from TEAs will be more transparent if they are performed at a common scale of 100 dry ton/day. Stated otherwise, all that glitters is not gold.

While Eq. (9) is physically interpretable its implications are not easily visualized. To permit visualization of key trends, Fig. 9 was constructed based on contour plots of Eq. (9) predictions. Fig. 9 takes advantage of the fact that feedstock cost can be treated as a correction to MFSP, as suggested by the form of Eq. (9a). Accordingly, Fig. 9a consists of MFSP predictions for different values of yield and scale, in the absence of feedstock costs. Interestingly, Fig. 9a shows that MFSP becomes increasingly independent of yield with increasing scale; in other words, for sufficient scale, MFSP becomes nearly insensitive to biocrude yield, a result which can be used to guide future development efforts.

Once Fig. 9a is used to calculate MFSP without feedstock cost, Fig. 9b can be used to determine a feedstock dependent correction factor to calculate the expected total MFSP. The correction factor is a linear function of feedstock cost, as anticipated by Eq. (9a). As expected, negative feedstock costs reduce the initial estimated value of MFSP determined from Fig. 9a and positive values increase it. Interestingly, MFSP sensitivity to feedstock cost depends on the biocrude yield, as is apparent from consideration of the lines in Fig. 9b.

Fig. 9 validates many of the intuitively understood relationships between MFSP, feedstock cost, yield, and scale. As anticipated qualitatively by Fig. 5, scale and yield are coupled to each other and to MFSP via a power-law relationship – following the expected economies of scale relationship and the direct relationship between yield and total fuel content respectively. At a hypothetically infinite plant scale and 100% yield, the cost of production asymptotically approaches a constant value as the marginal capital costs decrease – representative of the feed-independent operating costs. When combined with the power law relationships the asymptotic behavior implies the existence of an “elbow point”, or a point of diminishing returns: notably the majority of “economies of scale” associated capital cost reductions occur at plant sizes less than 100 dry tons per day (DPTD). Furthermore, depending on scale, yield improvements beyond 40% result in diminishing returns, which has particularly important implications for strategies aimed at enhancing yield by acquiring higher value feeds, deploying catalysts, or improving reactor design.

The relationship between feedstock cost and biocrude yield in Fig. 9b deserves its own discussion. As a rule, biocrude yield of a given feed depends most strongly on lipid content [67,69]. Inasmuch as high lipid content feeds are more valuable than low lipid content versions – for example, as frequently is the case with microalgae [44,45]– the benefits of selecting a high yield, yet expensive feed compared to a low yield, yet inexpensive one can be quantified using Fig. 9. Waste feeds, which may be zero cost or even negative cost, are a different case. As an example, Fig. 10 consists of MFSP predictions for three fixed values of feedstock cost – a value representing the average landfill tipping fee in the USA (which is negative), zero cost, and the absolute value of the USA aggregate tipping fee (i.e., a positive feedstock cost equal in magnitude to the tipping fee) [80]. A moisture content of 25% is assumed for landfill waste. [83] Fig. 10 shows that MFSP is extremely sensitive to feedstock cost; in contrast, and as deduced from Fig. 9, the sensitivity of MFSP to scale and yield is correlated with one another and MFSP becomes nearly insensitive to these variables once they reach critical values. This is further shown in the reduced sensitivity to scale outlined in Figure SI-3.

In addition to the predictions shown in Figs. 9 and 10, Eq. (9) can be extended for many applications, several of which are proposed here. One obvious application is comparison of two or more different scenarios, for example for processing different feedstocks, to one another. Investment benefits from identifying the most promising option; however, selecting the most promising option can be difficult without a common basis, such

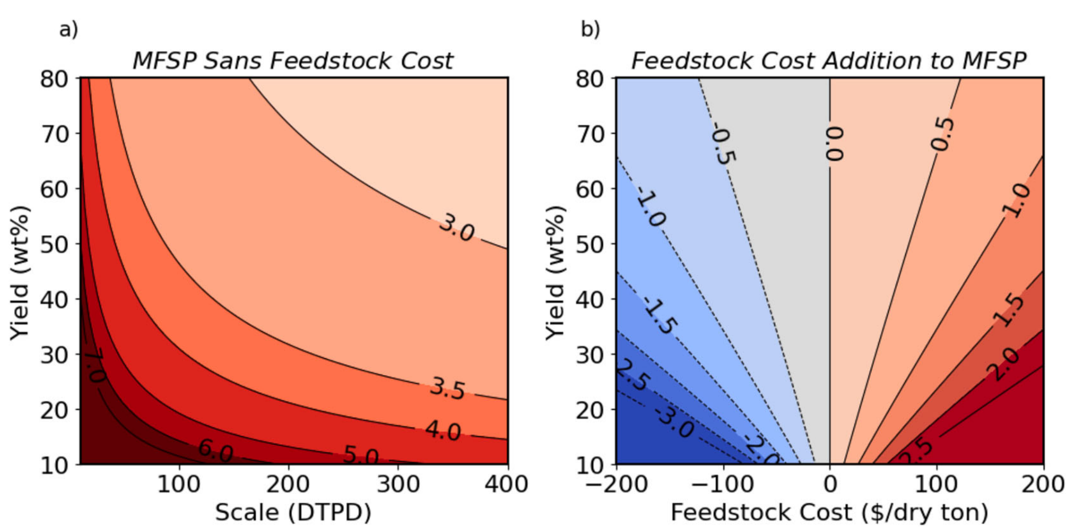

Fig. 9. Predicted MFSPs (in USD/GGE) reported in Eq. (9) broken down by a) the contribution of marginal CAPEX and marginal fixed OPEX and b) the contribution of feedstock cost to variable OPEX. The sum of the values calculated in a) and b) is the MFSP of a given deployment as predicted by Eq. (9).

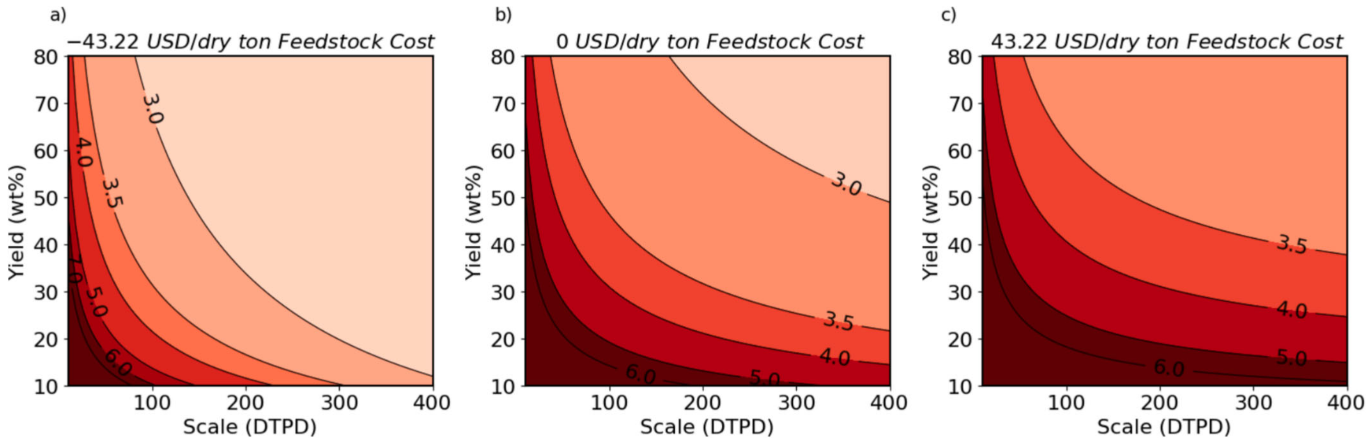

Fig. 10. MFSPs (in $USD/GGE) as predicted by Eq. (9) for three fixed feedstock costs: a) a feedstock cost of 57.63 USD/dry ton, approximately representative of an HTL plant diverting waste at the rate of the average landfill tipping fee in the USA, b) a feedstock cost of 0 USD/dry ton, and c) a feedstock cost of 57.63 USD/dry ton.

as a fixed feed or fixed scale. As a concrete example, a decision may be required to select between investing in HTL for conversion of algae grown for wastewater treatment, waste biomass or green waste, and animal manure. Biocrude yield alone cannot be used to differentiate between these three potential scenarios, since other factors – namely scale or feedstock cost – may also vary. Therefore, Eq. (9) can be used to guide investments in these cases, provided that scale and feedstock cost are known ahead of time.

Another application of the new model is for consideration of hundreds or even thousands of potential investments for a single feed, or for a feed with only modest site-to-site variation. The estimated 15,000 Waste Resource Reclamation Facilities (WRRFs) located in the U.S. are an illustrative example [36]. While all of these WRRFs produce a sewage sludge stream amenable to HTL, they differ in scale and local tipping fees that directly impact feedstock cost. Eq. (9) can be used as a first-order method to rank potential WRRF sites based on projected costs; from that point, resources can be invested to examine in detail only a small subset of potentially viable sites, conserving resources by removing non-viable sites from consideration. In both cases, Eq. (9), when used judiciously as a decision-making guide, can greatly reduce uncertainty and improve resource allocation for HTL deployment before significant resources are dedicated toward case-specific economic modeling.

Other, perhaps less obvious, applications of the new model are possible. For example, many different variations of HTL have been published in the past several years, with specific emphasis on the use of catalysts to improve yields. [84] Eq. (9) can be modified for modeling costs associated with catalytic HTL by using measured data for the catalytic biocrude yield – obtained experimentally – and adding a term associated with lifetime catalyst costs. Lifetime catalyst costs require knowledge of the upfront cost and the usable catalyst lifetime, as described by LeClerc et al. in their study on the use of hydroxyapatite for catalytic HTL [85]. The end result of modifying the HTL cost model would be quantification of the benefits of different catalysts, rather than a less sophisticated ranking that considers only biocrude yield – which could be applied in a manner similar to the earlier discussion on accounting for biofuel subsidies in Aierzhati et al. [26] The result will be rational allocation of available resources to catalysts and processes that offer the greatest economic benefits.

The methodology presented here is easily generalized to other waste-to-energy technologies, such as pyrolysis or anaerobic digestion, and potentially to kindred analyses including life-cycle assessments (LCA). The required inputs are: 1) a detailed model that can be used to identify the most sensitive variables for determining costs and 2) published studies that report process costs and are independent of the aforementioned detailed model. After the detailed model is used to identify the most sensitive variables, Monte Carlo simulations of the detailed model over reasonable ranges of these variables are used to generate the training data. These training data are then used to regress parameters in one or more economic models. The form presented in Eq. (9) can be used as a starting point for regression analysis, or the entire process described here can be followed to discover a new form appropriate for the process under consideration. The final step is comparison of the results of the trained model to the independently reported data points to estimate the accuracy of the reduced-order model; if the accuracy is not satisfactory, additional terms can be added to the regression model to improve accuracy. The resulting models can then be used for estimating costs and subsequently allocating resources in diverse waste-to-energy applications where cost is a major driver.

4. Conclusions

In this work, we use machine learning to discover a TEA and modeling framework capable of predicting HTL fuel cost outcomes. Machine learning was used to identify the three most impactful and readily accessible process variables: feedstock cost, biocrude yield, and process scale. The relationship and interaction between these variables in determining MFSP were elucidated and developed into a single generalized equation to model MFSP outcomes for a given HTL deployment. The result is an easily interpretable, analytic equation requiring only the feedstock cost, biocrude yield, and scale of an HTL venture to estimate MFSP. When the accuracy of this model was compared against 28 published MFSP values, the model had an MAE of ± 1.32 USD/GGE, and a mean % deviation of ± 20.4%. The new model can be used to bind or constrain by converting them to a common basis for inter-comparison.

The machine learned model can be used to understand the interdependence of key variables. MFSP is sensitive to feedstock cost at all relevant values; in contrast, MFSP becomes nearly insensitive to feedstock cost and scale once they reach critical values. Identifying these critical values is imperative to optimal process design. Investing in increasing yield is not profitable when yield already exceeds its critical value. Similarly, investing in greater feedstock aggregation to boost process scale faces diminishing returns when scale exceeds its critical value.

The low dimensional model proposed here can be used for first-pass deployment decision-making with orders of magnitude less time and information resources compared with granular TEAs, thereby permitting simultaneous comparison of many deployment options to identify the most promising investments. The model can easily be modified to account for technology improvements by use of additional terms; for example, the effect of a catalyst that improves yields can be modeled by manually adjusting the yield input to the predictive model and addition of a term to account for catalyst costs. The method used here to develop the predictive model can easily be generalized to other systems in which technological performance and economic outcomes need to be evaluated with sparse data.

CRediT authorship contribution statement

Muntasir Shahabuddin: Writing – original draft, Methodology, Investigation, Formal analysis, Data curation. **Nikolaos Kazantzis:** Writing – review & editing, Supervision, Methodology. **Andrew R Teixeira:** Writing – review & editing, Supervision. **Michael T. Timko:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank Leslie Snowden-Swan and Yunhua Zhu (Pacific Northwest National Laboratory) for sharing their sewage sludge TEA upon which this work was built. Financial support was provided through the National Science Foundation NSF Research Traineeship (NRT) Award No. 2021871 and the U.S. Department of Energy through Grant DE-EEE0008513. MS gratefully acknowledges funding through the National Science Foundation Graduate Research Fellowship, under Grant No. 2038257. The authors thank the anonymous reviewers, including the one who suggested selecting a more memorable title.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecmx.2024.100756.

Data availability

Data will be made available on request.

References

[1] Mohareb EA, MacLean HL, Kennedy CA. Greenhouse gas emissions from waste management—assessment of quantification methods. J Air Waste Manag Assoc 2011;61:480–93.
[2] Chen T-C, Lin C-F. Greenhouse gases emissions from waste management practices using Life Cycle Inventory model. J Hazard Mater 2008;155:23–31.
[3] Emissions Trends and Drivers. in: C. Intergovernmental Panel on Climate, (Ed.). Climate Change 2022 - Mitigation of Climate Change: Working Group III Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, 2023. pp. 215-94.
[4] S. Sharma, R. Meena, A. Sharma. Biomass conversion technologies for renewable energy and fuels: A review note.
[5] Paul S, Dutta A. Challenges and opportunities of lignocellulosic biomass for anaerobic digestion. Resour Conserv Recycl 2018;130:164–74.
[6] Khalid A, Arshad M, Anjum M, Mahmood T, Dawson L. The anaerobic digestion of solid organic waste. Waste Manag 2011;31:1737–44.
[7] Saral JS, Satheesh AR, Ranganathan P. Economic and environmental analysis of algal biorefinery for the production of renewable fuels and co-product. Energy Convers Manage: x 2022;14:100189.
[8] Mainardis M, Flaibani S, Mazzolini F, Peressotti A, Goi D. Techno-economic analysis of anaerobic digestion implementation in small Italian breweries and evaluation of biochar and granular activated carbon addition effect on methane yield. J Environ Chem Eng 2019;7:103184.
[9] Ojha DK, Viju D, Vinu R. Fast pyrolysis kinetics of lignocellulosic biomass of varying compositions. Energy Convers Managem: x 2021;10:100071.
[10] Elhenawy Y, Fouad K, Bassyouni M, Al-Qabandi OA, Majozi T. Yield and energy outputs analysis of sawdust biomass pyrolysis. Energy Convers Manage: x 2024;22: 100583.
[11] Wright MM, Daugaard DE, Satrio JA, Brown RC. Techno-economic analysis of biomass fast pyrolysis to transportation fuels. Fuel 2010;89:S2–10.
[12] Brilovich Mosseri M, Duenyas A, Cohen EMA, Vitkin E, Steinbruch E, Epstein M, et al. Hydrothermal liquefaction of representative to Israel food waste model. Energy Convers Manage: x 2023;20:100475.
[13] Khoo CG, Hirose Y, Matsumura Y, Lam MK, Tan IS, Lee KT. Valorisation of Chlorella vulgaris biomass for multi-products synthesis via hydrothermal processing. Energy Convers Manage: x 2023;20:100399.
[14] Zhu Y, Albrecht KO, Elliott DC, Hallen RT, Jones SB. Development of hydrothermal liquefaction and upgrading technologies for lipid-extracted algae conversion to liquid fuels. Algal Res 2013;2:455–64.
[15] Cheng F, Jarvis JM, Yu J, Jena U, Nirmalakhandan N, Schaub TM, et al. Bio-crude oil from hydrothermal liquefaction of wastewater microalgae in a pilot-scale continuous flow reactor. Bioresour Technol 2019;294:122184.
[16] S.B. Jones, Y. Zhu, D.B. Anderson, R.T. Hallen, D.C. Elliott, A.J. Schmidt, et al. Process Design and Economics for the Conversion of Algal Biomass to Hydrocarbons: Whole Algae Hydrothermal Liquefaction and Upgrading. United States, 2014.
[17] Demirbas A. Effects of temperature and particle size on bio-char yield from pyrolysis of agricultural residues. J Anal Appl Pyrol 2004;72:243–8.
[18] Maag A, Paulsen A, Amendsen T, Yelvington P, Tompsett G, Timko M. Catalytic hydrothermal liquefaction of food waste using CeZrOx. Energies 2018;11.
[19] Marrone PA, Elliott DC, Billing JM, Hallen RT, Hart TR, Kadota P, et al. Bench-scale evaluation of hydrothermal processing technology for conversion of wastewater solids to fuels. Water Environ Res 2018;90:329–42.
[20] Suren Wijeyekoon KT, Corkran H, Bennett P. Commercial status of direct thermochemical liquefaction technologies. IEA Bioenergy August 2020.
[21] Ghadge R, Nagwani N, Saxena N, Dasgupta S, Sapre A. Design and scale-up challenges in hydrothermal liquefaction process for biocrude production and its upgradation. Energy Convers Manage: x 2022;14:100223.
[22] Mordechai Koskas Y, Golberg A, Gozin M, Kribus A. Process simulation for mass balance of continuous biomass hydrothermal liquefaction with reaction kinetics. Energy Convers Manage: x 2023;20:100477.
[23] L.J. Snowden-Swan, Y. Zhu, R.T. Hallen, T.R. Hart, M.D. Bearden, J. Liu, et al. Conceptual Biorefinery Design and Research Targeted for 2022: Hydrothermal Liquefaction Processing of Wet Waste to Fuels. Pacific Northwest National Laboratory. 27186 (2017) 89.
[24] M. Biddy, R. Davis, A. Dutta, A. Singh, L. Tao, E. Tan, et al. Integrated Strategies to Enable Lower-Cost Biofuels. Office of Energy Efficiency and Renewable Energy, U. S. Department of Energy, 2020.
[25] Tzanetis KF, Posada JA, Ramirez A. Analysis of biomass hydrothermal liquefaction and biocrude-oil upgrading for renewable jet fuel production: The impact of reaction conditions on production costs and GHG emissions performance. Renew Energy 2017;113:1388–98.
[26] A. Aierzhati, J. Watson, B. Si, M. Stablein, T. Wang, Y. Zhang. Development of a mobile, pilot scale hydrothermal liquefaction reactor: food waste conversion product analysis and techno-economic assessment. Energy Convers Manage: X. (2021) 100076.
[27] L.J. Snowden-Swan, Y. Zhu, S.B. Jones, D.C. Elliott, A.J. Schmidt, R.T. Hallen, et al. Hydrothermal Liquefaction and Upgrading of Municipal Wastewater Treatment Plant Sludge: A Preliminary Techno-Economic Analysis, Rev.1. United States, 2016.
[28] Savage S. Decision Making with Insight. PWS Publishers, Boston, MA: Duxbury Press; 2003.
[29] R.d.N.S. Scholtes. Flexibility in Engineering Design. MIT Press2011.
[30] Seifert T, Sievers S, Bramsiepe C, Schembecker G. Small scale, modular and continuous: A new approach in plant design. Chem Eng Process 2012;52:140–50.
[31] Juneja A, Murthy GS. Evaluating the potential of renewable diesel production from algae cultured on wastewater: techno-economic analysis and life cycle assessment. Aims Energy 2017;5:239–57.
[32] DeRose K, DeMill C, Davis RW, Quinn JC. Integrated techno economic and life cycle assessment of the conversion of high productivity, low lipid algae to renewable fuels. Algal Res 2019;38:101412.
[33] Magdeldin M, Kohl T, Järvinen M. Techno-economic assessment of the by-products contribution from non-catalytic hydrothermal liquefaction of lignocellulose residues. Energy 2017;137:679–95.
[34] Lozano EM, Løkke S, Rosendahl LA, Pedersen TH. Production of marine biofuels from hydrothermal liquefaction of sewage sludge. Preliminary techno-economic analysis and life-cycle GHG emissions assessment of Dutch case study. Energy Convers Manage: x 2022;14:100178.
[35] Imran S, Ahmad DHK. Quantum GIS based descriptive and predictive data analysis for effective planning of waste management. IEEE Access 2020;8:46193–205.
[36] Seiple T, Coleman A, Skaggs R. Municipal wastewater sludge as a sustainable bioresource in the United States. J Environ Manage 2017;197:673–80.
[37] Daugaard T, Mutti L, Wright M, Brown R, Companation P. Learning rates and their impacts on the optimal capacities and production costs of biorefineries. Biofpr 2014;9:82–94.
[38] Teplitz CJ. The learning curve deskbook: A reference guide to theory, calculations, and applications. New York, NY: Quorum Books; 1991.
[39] Shahabuddin M, Italiani E, Teixeira AR, Kazantzis N, Timko MT. Roadmap for deployment of modularized hydrothermal liquefaction: understanding the impacts of industry learning, optimal plant scale, and delivery costs on biofuel pricing. ACS Sustain Chem Eng 2023;11:733–43.
[40] A.C. Nathan Miller, Ji Eun Park, Anil Baral, Chris Malins, Stephanie Searle. Measuring and Addressing Investment Risk in the Second Generation Biofuels Industry. The International Council on Clean Transportation. ICCT, ICCT, 2013.

[41] Jiang Y, Jones SB, Zhu Y, Snowden-Swan L, Schmidt AJ, Billing JM, et al. Techno-economic uncertainty quantification of algal-derived biocrude via hydrothermal liquefaction. Algal Res 2019;39:101450.
[42] Li S, Jiang Y, Snowden-Swan LJ, Askander JA, Schmidt AJ, Billing JM. Techno-economic uncertainty analysis of wet waste-to-biocrude via hydrothermal liquefaction. Appl Energy 2021;283:116340.
[43] E. Medina-Martos, P. Miranda-Rey, J.-L. Gálvez-Martos, J. Dufour. Techno-economic Assessment of a Hydrothermal Liquefaction Process for Energy Recovery from Food Waste. in: S. Pierucci, F. Manenti, G.L. Bozzano, D. Manca, (Eds.), Computer Aided Chemical Engineering. Elsevier2020. pp. 1729-34.
[44] Batan L, Graff G, Bradley T. Techno-economic and Monte Carlo probabilistic analysis of microalgae biofuel production system. Bioresour Technol 2016;219:45–52.
[45] Ou L, Thilakaratne R, Brown R, Wright M. Techno-economic analysis of transportation fuels from defatted microalgae via hydrothermal liquefaction and hydroprocessing. Biomass Bioenergy 2015;72:45–54.
[46] Pearce M, Shemfe M, Sansom C. Techno-economic analysis of solar integrated hydrothermal liquefaction of microalgae. Appl Energy 2016;166:19–26.
[47] Ranganathan P, Savithri S. Techno-economic analysis of microalgae-based liquid fuels production from wastewater via hydrothermal liquefaction and hydroprocessing. Bioresour Technol 2019;284:256–65.
[48] Pedersen TH, Hansen NH, Pérez OM, Cabezas DEV, Rosendahl LA. Renewable hydrocarbon fuels from hydrothermal liquefaction: a techno-economic analysis. Biofuels Bioprod Biorefin 2018;12:213–23.
[49] Bbosa D, Mba-Wright M, Brown R. More than ethanol: a techno-economic analysis of a corn stover-ethanol biorefinery integrated with a hydrothermal liquefaction process to convert lignin into biochemicals. Biofpr 2018;12:497–509.
[50] Hansen N, Pedersen T, Rosendahl L. Techno-economic analysis of a novel hydrothermal liquefaction implementation with electrofuels for high carbon efficiency. Biofpr 2019;13:660–72.
[51] Kenney DH, Paffenroth RC, Timko MT, Teixeira AR. Dimensionally reduced machine learning model for predicting single component octanol–water partition coefficients. J Cheminf 2023;15:9.
[52] Gao CW, Allen JW, Green WH, West RH. Reaction Mechanism Generator: automatic construction of chemical kinetic mechanisms. Comput Phys Commun 2016;203:212–25.
[53] Bhattacharjee B, Schwer DA, Barton PI, Green WH. Optimally-reduced kinetic models: reaction elimination in large-scale kinetic mechanisms. Combust Flame 2003;135:191–208.
[54] Cheng F, Belden ER, Li W, Shahabuddin M, Paffenroth RC, Timko MT. Accuracy of predictions made by machine learned models for biocrude yields obtained from hydrothermal liquefaction of organic wastes. Chem Eng J 2022;442:136013.
[55] Schwer DA, Lu P, Green WH. An adaptive chemistry approach to modeling complex kinetics in reacting flows. Combust Flame 2003;133:451–65.
[56] Hough BR, Beck DAC, Schwartz DT, Pfaendtner J. Application of machine learning to pyrolysis reaction networks: reducing model solution time to enable process optimization. Comput Chem Eng 2017;104:56–63.
[57] Belden ER, Rando M, Ferrara OG, Himebaugh ET, Skangos CA, Kazantzis NK, et al. Machine learning predictions of oil yields obtained by plastic pyrolysis and application to thermodynamic analysis. ACS Engineering Au 2023;3:91–101.
[58] Athaley A, Saha B, Ierapetritou M. Biomass-based chemical production using techno-economic and life cycle analysis. AIChE J 2019;65.
[59] Ou L, Li S, Tao L, Phillips S, Hawkins T, Singh A, et al. Techno-economic analysis and life-cycle analysis of renewable diesel fuels produced with waste feedstocks. ACS Sustain Chem Eng 2022;10:382–93.
[60] Shia Y-P, Yu B-Y. Development of a rigorous and generalized model on the hydrothermal liquefaction (HTL) process for bio-oil production. Process Saf Environ Prot 2023;171:541–54.
[61] Zhu Y, Biddy MJ, Jones SB, Elliott DC, Schmidt AJ. Techno-economic analysis of liquid fuel production from woody biomass via hydrothermal liquefaction (HTL) and upgrading. Appl Energy 2014;129:384–94.
[62] Peters M, Timmerhaus K, West R. Plant Design and Economics for Chemical Engineers. 5th ed. McGraw-Hill New York; 2002.
[63] Duan P, Savage P. Hydrothermal Liquefaction of a Microalga with Heterogeneous. Catalysts 2011.
[64] Gollakota ARK, Kishore N, Gu S. A review on hydrothermal liquefaction of biomass. Renew Sustain Energy Rev 2018;81:1378–92.
[65] Liu X, Saydah B, Eranki P, Colosi LM, Greg Mitchell B, Rhodes J, et al. Pilot-scale data provide enhanced estimates of the life cycle energy and emissions profile of algae biofuels produced via hydrothermal liquefaction. Bioresour Technol 2013;148:163–71.
[66] Silva Thomsen LB, Carvalho PN, dos Passos JS, Anastasakis K, Bester K, Biller P. Hydrothermal liquefaction of sewage sludge; energy considerations and fate of micropollutants during pilot scale processing. Water Res 2020;183:116101.
[67] Valdez P, Tocco V, Savage P. A general kinetic model for the hydrothermal liquefaction of microalgae. Bioresour Technol 2014;163:123–7.
[68] Ross A, Biller P, Kubacki M, Lea-Langton A, Jones J. Hydrothermal processing of microalgae using alkali and organic acids. Fuel 2010;89:2234–43.
[69] Biller P, Ross AB. Potential yields and properties of oil from the hydrothermal liquefaction of microalgae with different biochemical content. Bioresour Technol 2011;102:215–25.
[70] Minowa T, Yokoyama S-Y, Kishimoto M, Okakura T. Oil production from algal cells of Dunaliella tertiolecta by direct thermochemical liquefaction. Fuel 1995;74:1735–8.
[71] Suzuki A, Nakamura T, Yokoyama S-Y, Ogi T, Koguchi K. Conversion of sewage sludge to heavy oil by direct thermochemical liquefaction. J Chem Eng Jpn 1987;21.
[72] Real-Time Maps and Charts. ISO New England. (2020).
[73] J. Benesty, J. Chen, Y. Huang, I. Cohen. Pearson correlation coefficient. Noise reduction in speech processing. Springer, 2009. pp. 1-4.
[74] S.C. Meurer A, Paprocki M, Čertík O, Kirpichev SB, Rocklin M, Kumar A, Ivanov S, Moore JK, Singh S, Rathnayake T, Vig S, Granger BE, Muller RP, Bonazzi F, Gupta H, Vats S, Johansson F, Pedregosa F, Curry MJ, Terrel AR, Roučka Š, Saboo A, Fernando I, Kulal S, Cimrman R, Scopatz A. SymPy: symbolic computing in Python. Peer J Computer Sci 3 (2017).
[75] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020;17:261–72.
[76] Stoica P, Selen Y. Model-order selection: a review of information criterion rules. IEEE Signal Process Mag 2004;21:36–47.
[77] Akaike H. A new look at the statistical model identification. IEEE Trans Autom Control 1974;19:716–23.
[78] Nie Y, Bi X. Life-cycle assessment of transportation biofuels from hydrothermal liquefaction of forest residues in British Columbia. Biotechnol Biofuels 2018;11.
[79] Pawel I. The cost of storage – how to calculate the Levelized Cost of Stored Energy (LCOE) and applications to renewable energy generation. Energy Procedia 2014;46:68–77.
[80] Landfill Tipping Fee 2022 List. EREF Report Topics: Landfill. Environmental Research & Education Foundation, EREF, 2022.
[81] Shafizadeh A, Shahbeig H, Nadian MH, Mobli H, Dowlati M, Gupta VK, et al. Machine learning predicts and optimizes hydrothermal liquefaction of biomass. Chem Eng J 2022;445:136579.
[82] Katongtung T, Onsree T, Tippayawong N. Machine learning prediction of biocrude yields and higher heating values from hydrothermal liquefaction of wet biomass and wastes. Bioresour Technol 2022;344:126278.
[83] Krause MJ, Eades W, Detwiler N, Marro D, Schwarber A, Tolaymat T. Assessing moisture contributions from precipitation, waste, and leachate for active municipal solid waste landfills. J Environ Manage 2023;344:118443.
[84] Nagappan S, Bhosale RR, Nguyen DD, Chi NTL, Ponnusamy VK, Woong CS, et al. Catalytic hydrothermal liquefaction of biomass into bio-oils and other value-added products – a review. Fuel 2021;285:119053.
[85] LeClerc HO, Tompsett GA, Paulsen AD, McKenna AM, Niles SF, Reddy CM, et al. Hydroxyapatite catalyzed hydrothermal liquefaction transforms food waste from an environmental liability to renewable fuel. iScience 2022;25:104916.