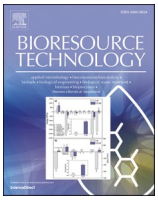
Bioresource Technology 370 (2023) 128528

Contents lists available at ScienceDirect

Bioresource Technology

journal homepage: www.elsevier.com/locate/biortech



Machine learning for surrogate process models of bioproduction pathways

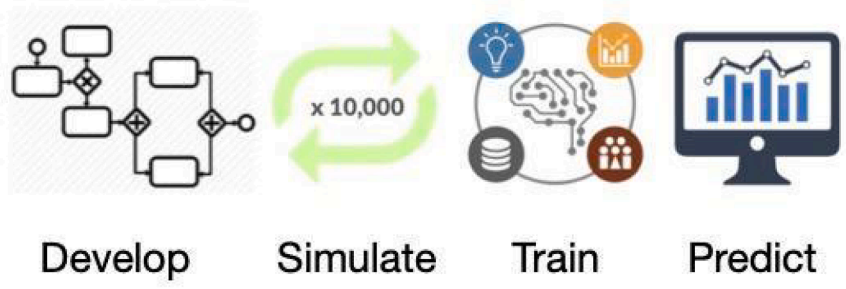
Tyler Huntington [a,b], Nawa Raj Baral [a,b], Minliang Yang [a,b], Eric Sundstrom [b,c], Corinne D. Scown [a,b,d,e,*]

[a] Life-cycle, Economics, and Agronomy Division, Joint BioEnergy Institute, 5885 Hollis Street, Emeryville, CA 94608, USA
[b] Biosciences Area, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA
[c] Advanced Biofuels and Bioproducts Process Development Unit, 5885 Hollis Street, Emeryville, CA 94608, USA
[d] Energy Technologies Area, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA, 94720, USA
[e] Energy & Biosciences Institute, University of California, Berkeley, 282 Koshland Hall, Berkeley, CA 94720, USA

HIGHLIGHTS

- Machine learning can be used to develop surrogate models from process simulations.
- Surrogate models make technoeconomic models more accessible and fast to run.
- Surrogate models are most useful when further design changes will not be made.
- Automated design strategies can be complementary to machine learning approaches.
- Advanced sampling strategies may yield further performance improvements.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Biofuels
Bioproducts
Technoeconomic analysis
Life-cycle assessment
TPOT

ABSTRACT

Technoeconomic analysis and life-cycle assessment are critical to guiding and prioritizing bench-scale experiments and to evaluating economic and environmental performance of biofuel or biochemical production processes at scale. Traditionally, commercial process simulation tools have been used to develop detailed models for these purposes. However, developing and running such models can be costly and computationally intensive, which limits the degree to which they can be shared and reproduced in the broader research community. This study evaluates the potential of an automated machine learning approach to develop surrogate models based on conventional process simulation models. The analysis focuses on several high-value biofuels and bioproducts for which pathways of production from biomass feedstocks have been well-established. The results demonstrate that surrogate models can be an accurate and effective tool for approximating the cost, mass and energy balance outputs of more complex process simulations at a fraction of the computational expense.

1. Introduction

Technoeconomic analysis and life-cycle assessment are powerful analytical tools for evaluating novel bioproduction processes, identifying key cost bottlenecks, and drivers of greenhouse gas emissions and other environmental impacts (Mahmud et al., 2021; Scown et al., 2021).

* Corresponding author at: Life-cycle, Economics, and Agronomy Division, Joint BioEnergy Institute, 5885 Hollis Street, Emeryville, CA 94608, USA.
E-mail address: cdscown@lbl.gov (C.D. Scown).

https://doi.org/10.1016/j.biortech.2022.128528
Received 29 October 2022; Received in revised form 20 December 2022; Accepted 21 December 2022
Available online 24 December 2022

For a wide range of chemical and biological processes, the critical first step is to design and simulate a production facility. This process simulation provides process equipment lists and sizing requirements, as well as mass and energy balances that are key inputs to cash flow analyses and life-cycle environmental assessments (LCAs). The demand for process design and simulation is vast, with applications in academic research, due diligence for investors, strategy development and fundraising within startups, and evaluation of investments made by multinational corporations (Burk, 2022). However, completing these analyses is resource intensive, requiring specialized commercial software packages and engineering domain expertise to devise realistic facility design (s) (Scown et al., 2021). This barrier to entry can limit applications of technoeconomic analysis (TEA) in early stage technical decision making, where preliminary analysis can often deliver maximum value. A natural question is whether any parts of the process can be automated and broadly shared without inadvertently delivering inaccurate or otherwise flawed results. Machine learning, in combination with conventional process models, offers an opportunity to pursue this goal through the development of surrogate technoeconomic analysis models and other automated design tools. (see Fig. 1.)

Applications of machine learning in the chemical industry have become increasingly widespread over the last several decades (Lee et al., 2018). Examples in the literature range from the use of machine learning to predict chemical reaction properties (Marcou et al., 2015), screen and design catalysts (Li et al., 2017) and optimize the performance of chemical process unit operations (Ochoa-Estopier et al., 2013; Verma et al., 2018; Zheng et al., 2009). More relevant to this study, machine learning has played an important role in process system engineering with a focus on environmental sustainability (Negny et al., 2012). Liao et al. (2020) combined machine learning with kinetics-based process simulation to develop life-cycle inventory data for activated carbon production. Their approach yielded a modeling framework capable of predicting greenhouse gas (GHG) emissions associated with activated carbon production from 73 types of woody biomass (Liao et al., 2020). Song et al. (2017) demonstrated the use of deep artificial neural networks (ANNs) to estimate life-cycle impacts of industrially produced chemicals based on their molecular structure (Song et al., 2017). Similarly, Kaab et al. (2019) developed ANNs and another machine learning technique called adaptive neuro fuzzy inference to predict cradle-to-grave environmental impacts and output energy of sugar cane production (Kaab et al., 2019). Romeiko et al. (2019) demonstrated the use of two other machine learning techniques–support vector machines and gradient boosting regression trees– to make spatially explicit predictions of life cycle global warming and eutrophication from corn production. They found that the gradient boosting regression model outperformed the support vector machine approach, despite requiring longer training time to achieve this superior level of predictive accuracy (Romeiko et al., 2019). Applications of machine learning specific to biofuels have mainly focused on yield forecasting of feedstock crops (Huntington et al. 2020), and prediction of fuel physicochemical properties (Aminian and ZareNezhad, 2018; Özgür and Tosun, 2017) and combustion characteristics (Baghban et al., 2018; Comesana et al., 2022).

Fewer studies have evaluated the potential of machine learning techniques as an alternative to traditional process simulation models for use in end-to-end technoeconomic analysis and life-cycle assessment. This study seeks to fill this research gap by presenting a modeling framework that leverages automated machine learning (auto-ML) to develop surrogate models trained on process simulation data for a range of biorefinery configurations, with the ultimate goal of predicting cost and mass/energy balances. While the machine learning models are not trained to directly predict life-cycle assessment outputs (i.e. greenhouse gas emissions and water consumption), this study demonstrates their ability to predict mass and energy balances, which can serve as inputs to conventional life-cycle assessments. To complement the surrogate modeling approach, a tool was also developed for automating the design and cost analysis of downstream separation, product recovery and purification strategies, with the goal of providing more flexibility on the costly downstream process (typically between 20 % and 60 % of total costs (Martínez-Aragón et al., 2009; Wang et al., 2016)). Both the surrogate model and downstream separation tool are demonstrated using ethanol production as a case study.

2. Materials and methods

2.1. Process model development

A commercial process modeling software package—SuperPro Designer— was used to develop a baseline process model for bioethanol from lignocellulosic biomass. Although this study makes use of a model in SuperPro, the approach presented here could be implemented with any process modeling software package, including AspenPlus. We used biomass sorghum as a lignocellulosic biomass feedstock. The process model includes several subsequent bioethanol production stages, including biomass production, supply, and preprocessing (milling), biomass pretreatment, sequential hydrolysis and bioconversion, recovery and separation, wastewater treatment, onsite energy generation (heat and power), and utilities (process water, cooling water, and clean-in-place system). Detailed descriptions of these production stages are documented in prior studies (Baral et al., 2020, 2019; Humbird et al., 2011; Yang et al., 2020) and are summarized only at a high level in this section.

Biomass production and supply includes biomass sorghum cultivation, field operations (windrowing, field drying, baling, and stacking bales at the field edge), trucking bales to the biorefinery, and outdoor storage next to the biorefinery (Baral et al., 2020). Biomass sorghum cultivation includes establishment land, materials, fuel, labor, and machineries as well as fertilizer and herbicide applications (Baral et al., 2020).

The biorefinery process model starts with conveying biomass sorghum bales from the outdoor storage to the preprocessing unit at the biorefinery. At the preprocessing unit, biomass sorghum bales are shredded, milled, and stored for a short time before transfer to the pretreatment unit (Aden et al., 2002). The milled biomass is mixed with water and a biocompatible ionic liquid (Cholinium Lysinate), and then sent to the pretreatment reactor (Magurudeniya et al., 2021). Sulfuric acid is mixed with the pretreated biomass to adjust pH (Magurudeniya et al., 2021). After pH adjustment, enzyme and water are mixed with the pretreated biomass slurry and sent to the batch sequential enzymatic hydrolysis and bioconversion reactor. For bioconversion, nitrogen sources (corn steep liquor and diammonium phosphate) and inoculum

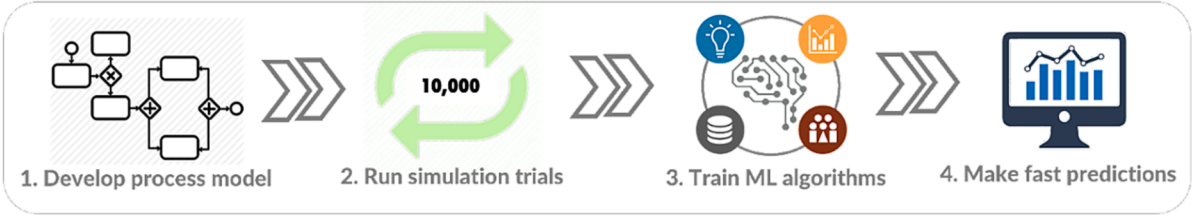


Fig. 1. Flowchart of methodological approach for developing surrogate process models using machine learning.

are added (Humbird et al., 2011). Following the bioconversion, the whole slurry is sent to the recovery and separation unit.

The recovery and separation unit first separates the solid and liquid fraction of the fermented slurry. The solid fraction is routed to the onsite energy generation unit. Ethanol and subsequently ionic liquid are recovered from the liquid fraction by distillation (Humbird et al., 2011) and pervaporation (Sun et al., 2017), respectively. Ethanol is further dried using molecular sieve adsorption (Humbird et al., 2011) and stored onsite. The recovered ionic liquid is recycled back to the pretreatment unit and reused. Following the recovery of ethanol and ionic liquid, the liquid fraction is routed to the wastewater treatment unit.

The wastewater treatment unit includes a subsequent anaerobic and aerobic treatment processes, which are consistent with prior studies (Aden et al., 2002; Humbird et al., 2011). Biogas generated in the anaerobic digester from the remaining carbon sources, such as glucose and xylose, is sent to the onsite energy generation unit. The solid biomass generated following the anaerobic and aerobic treatments is separated and routed to the onsite energy generation unit. Water is recovered and reused.

The remaining solid fraction of biomass (mainly lignin), biogas, and sludge (mainly cell mass) generated during wastewater treatment are combusted in a boiler to generate steam (Humbird et al., 2011). Natural gas is used as a supplemental energy source for the boiler when the biomass-derived energy sources available at the biorefinery are not sufficient to generate heat and power required for the facility. Steam is fed to the multistage turbine to generate power required for the facility. Part of steam extracted from the turbine is used for the upstream processes. Excess electricity is assumed to be sold to the grid.

The utilities stage provides process water, cooling water, chilled water, and hot cleaning and sterilization chemicals (clean-in-place (CIP) system). Table 1 summarizes the major operating parameters used to develop the ethanol biorefinery model for analysis in this study.

Once the material and energy balance analyses as well as equipment sizing and cost analyses were computed, the model was run for 10,000 simulation trials, randomly selecting values of each input process parameter assuming uniform distribution from a range of reasonable estimates of their minimum and maximum values. The range of each parameter was determined considering prior studies and are listed in Table 1. Upon completion of each simulation trial, key technoeconomic outputs of the model were recorded. These include the results of material and energy balances as well as capital and operating costs of each major process stage. The minimum selling price of bioethanol was computed by using a standard discounted cash flow rate of return analysis (Humbird et al., 2011) achieving a net present value of zero. Basic input parameters for the cash flow analysis include an internal rate of return of 10 %, income tax rate of 35 %, annual operating time of 7920 h (24 h/day and 330 days/year), and a plant service life of 30 years (Humbird et al., 2011; Yang et al., 2020).

2.2. Training the machine learning models

To evaluate the potential for machine learning models trained on SuperPro simulation data to accurately predict mass and energy balances, we used the process simulation model for lignocellulosic ethanol production as a case study. Twenty-one key process parameters were selected to serve as training features for the surrogate machine learning models (Table 1) based on their impacts on the production cost and life-cycle carbon footprint of the final fuel. We established a range for each input parameter and then sampled randomly from those ranges, assuming a uniform distribution for all parameters within the defined ranges.

Randomly sampling data inputs and their corresponding SuperPro-generated results ensures that the machine learning model does not "learn" correlations between input parameters that are simply an artifact of the sampling method. However, random sampling does create its own practical challenges because it requires large changes to input

Table 1
Key process parameters used as training features for the surrogate models.

Process Parameter	Units	Range	Description
Biorefinery size	Bone-dry metric tons (bdt)/day	1000–4000	Bone-dry metric tons of biomass feedstock supplied to the biorefinery per day.
Biomass feedstock cost	USD/bdt	65–150	Cost (USD) per bone-dry metric ton of delivered biomass feedstock.
Cellulose	Weight fraction (%) of bone-dry biomass	20–44	Cellulose present in the delivered feedstock (percent by mass). This determines the maximum glucose available in the biomass feedstock.
Hemicellulose	Weight fraction (%) of bone-dry biomass	14–30	Hemicellulose present in the delivered feedstock (percent by mass). This determines the maximum xylose (mainly) available in the biomass feedstock.
Lignin	Weight fraction (%) of bone-dry biomass	9–24	Total lignin content (soluble and insoluble) in delivered biomass (percent by mass). The insoluble lignin represents the maximum lignin that can be sent to the combined heat and power (CHP) unit for combustion (the actual value depends on types of pretreatment method).
Ionic liquid loading rate	Weight fraction (%) of whole slurry	2.5–15.0	Fraction of ionic liquid in the pretreatment reactor (percent by mass-based on the whole slurry). Changing this parameter alone does not alter the predicted sugar yields. Glucose and xylose yields must be adjusted accordingly.
Ionic liquid cost	USD/kg	0.5–5.0	Cost (USD) per kg of ionic liquid (which is used as the solvent for pretreatment) delivered to the biorefinery.
Sulfuric acid loading rate	kg/kg-ionic liquid	0.1–0.2	Mass (kg) of sulfuric acid per kg of ionic liquid, which is required for pH adjustment after pretreatment.
Sulfuric acid cost	USD/kg	0.03–0.28	Cost (USD) per kg of sulfuric acid delivered to the biorefinery.
Solid loading rate for pretreatment	Weight fraction (%) of whole slurry	20–40	Fraction of biomass solids in the pretreatment reactor (percent by mass-based on the whole slurry).
Enzyme loading rate	mg-protein/g-glucan	7–20	Mass of enzyme (mg of protein) required per g of glucan (cellulose), which is essential to release sugars (mainly glucose and xylose) from the pretreated biomass.
Enzyme cost	USD/kg-protein	4–6	Cost (USD) per kg of enzyme delivered to the biorefinery or produced at the biorefinery.
Cellulose to glucose conversion rate	%	70–96	Combined (pretreatment and hydrolysis)

(*continued on next page*)

Table 1 (*continued*)

Process Parameter	Units	Range	Description
			percentage of cellulose converted into glucose
Xylan to xylose conversion rate	%	50–90	Combined (pretreatment and hydrolysis) percentage of hemicellulose converted into xylose
Aeration rate	volume of air sparged per unit volume of growth medium per minute (vvm)	0	Volume of air under standard conditions per volume of the whole slurry per minute (oxygen should be sufficient at least to meet the cell redox balancing.)
Glucose conversion rate	%	90–95	Total product yield during bioconversion (percent by mass of the initial glucose in a bioreactor)
Xylose conversion rate	%	80–90	Total product yield during bioconversion (percent by mass of the initial xylose in a bioreactor)
Bioconversion time	hours	36–72	Time required for the complete utilization of sugars (or retention time) in the bioconversion reactor (except setup time)
Biofuel recovery rate	%	95–99	Overall recovery of biofuel during recovery and separation processes (percent by mass of product after bioconversion).
Ionic liquid recovery rate	%	80–99	Overall recovery of ionic liquid during the recovery process (percent by mass of the initial required ionic liquid: the difference between the initial ionic liquid and the recovered ionic liquid results in the makeup ionic liquid).

parameters between individual runs of the original process simulation model. In contrast, the SuperPro model performs best—easily converging material and energy balances— when input parameters are gradually increased or decreased with each model run. In the future, more advanced sampling strategies could be implemented to avoid problems with the process model failing to converge. Sampling more frequently where combinations of input parameters result in inflection points or other nonlinear effects can also improve the accuracy of the resulting surrogate model (Bhosekar et al., 2018).

To determine the amount of training data required, we used the event per variable (EPV—a rule of thumb) (100 + EPV × number of independent variables) assuming an EPV value of 50 (Bujang et al., 2018). EPV is determined on a case-by-case basis and the minimum recommended value is 10 (Vittinghoff and McCulloch, 2007). This study relied on a statistical sampling method following the technique proposed by Cochran (1977), which resulted in about 9,000 training data inputs to achieve a 95 % likelihood of the desired error of 1 %. Since 10 % of the data is required for testing, a total of 10,000 SuperPro trials were needed to satisfy the minimum training data needs.

The 10,000 simulation trials of the SuperPro models were randomly split into a training set of 9,000 and a test set of 1,000 samples respectively. Next, 20 separate surrogate machine learning models for 20 different outputs of the SuperPro simulation model were developed. These outputs included six mass balance model outputs (diammonium phosphate (DAP) input, enzyme input, ionic liquid input, corn steep liquor (CSL) input, sulfuric acid input, and total product output), six energy model outputs (electricity use during ionic liquid pretreatment, electricity use during recovery and separation, electricity use for wastewater treatment, electricity use for bioconversion, electricity consumption in the combined heat and power section, and total onsite electricity demand), and eight section-specific cost contributions to the minimum selling price (feedstock handling and supply, ionic liquid pretreatment, enzymatic hydrolysis and bioconversion, utilities, recovery and separation, wastewater treatment, onsite energy generation, and the overall minimum selling price). Recognizing that the input process parameters used for training would differentially affect these various outputs, the 20 surrogate models were trained and evaluated separately in order to account for this variability in input/output relationships and ensure the highest possible model specificity and performance.

The Tree-Based Pipeline Optimization Tool (TPOT) was used to build and train the surrogate models. TPOT is an auto-ML library written in Python which utilizes a genetic programming algorithm to identify and optimize the best-fitting "pipeline" of machine learning algorithms for a given classification or regression problem. Since the output variables targeted for prediction (costs, mass balances and energy balances) in this analysis are quantitative and continuous in nature, the TPOTRegressor class was selected to determine the optimal regressor pipelines for the data. The TPOTRegressor class allows for numerous parameters to be manually defined, such as the number of generations, population size, offspring size, mutation rate and crossover rate, all of which affect how its genetic programming algorithm navigates the search space of candidate pipelines. TPOT provides reasonable default values for all of these parameters, and this study relied upon these defaults to streamline the pipeline selection process, this procedure had to be repeated 20 times to obtain 20 separate surrogate model pipelines for the different mass, energy and cost outputs of the SuperPro simulation model.

During the optimization process for each surrogate model, TPOT performed a search over supervised regression models, implemented in scikit-learn (Pedregosa et al., 2011), the underlying library upon which TPOT is built (Le et al. 2020). Specifically, the scikit-learn regressor functions considered by TPOT included the RandomForestRegressor, StackingRegressor, ExtraTreesRegressor, and GradientBoostingRegressor. To determine which pipeline of these functions would yield the best performance for each of the key output variables, TPOT performed internal k-fold cross validation, using mean squared error (MSE) as the objective function to be minimized. MSE is generally regarded as a suitable metric for this purpose and is also TPOT's default setting (Le et al. 2020). Once an optimal pipeline is determined, TPOT subsequently automated the process of tuning hyperparameters of the constituent machine learning models to further minimize prediction error. Once hyperparameters were tuned, pipelines were fitted to the entire training dataset and evaluated for their predictive accuracy on the 1,000 test data samples withheld from training.

TPOT was chosen as the preferred auto-ML library for this analysis due to the intuitive, thorough and well-documented application programming interface (API). While deep learning approaches such as artificial neural networks and convolutional neural networks can be effective alternatives, these techniques can be more "black box" in nature with respect to interpreting potential mechanistic relationships between input parameters and model output (Rudin, 2018). The TPOTRegressor function was applied to the training set of 9,000 SuperPro simulation trials to identify optimal machine learning pipelines for predicting minimum selling prices, mass balances and energy balances of each major process stage in the ethanol production pathway. TPOT was run separately for all output variables of interest (24 in total) to obtain specifically optimized pipelines for predicting each one. After running TPOTRegressor for each output variable, the optimal pipeline was extracted and evaluated based on its performance on the test set of 1,000 SuperPro trials withheld from training. Since TPOT is built on top

of scikit-learn, the candidate regressor functions are sourced from the scikit-learn API. Different sets of regressors were selected by TPOT for each of the separate models for prediction mass balances, energy balances and costs.

Mean absolute percent error (MAPE), as shown in eq. (1), was used to evaluate the performance of the surrogate models. This error metric accounts for discrepancies in units across models outputs (i.e. US dollars for cost outputs, and physical units for mass and energy balances) so that the relative performance of these models can be more easily compared in a standardized way.

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - P_t}{A_t}\right| \quad (1)$$

In the eq. (1), n represents the number of test samples used for validation, A_t represents the actual value of a given sample, and P_t represents the predicted value for the sample.

2.3. Adding flexibility for downstream separation and purification

An important limitation in surrogate models like those developed here is their lack of flexibility beyond the selected input parameters. Once the training data is generated and the surrogate model has been trained and tested, any further changes to the process configuration or operating conditions (beyond the selected parameters) are not possible without restarting the entire process beginning with generation of new training data. Thus, combining surrogate models with mechanistic modeling could ostensibly result in a balance between the performance of machine learning models and the flexibility of simplified mechanistic models. To complement the surrogate modeling approach described in previous sections, this study explores the potential for automated design of downstream separation and product recovery in biorefineries, based on the type of product and how it is generated by host microbes. While not fully integrated with the machine learning approach in the present study, this type of automated design and cost modeling has the potential to be combined with surrogate models to offer greater flexibility.

Although designing downstream separation and product recovery processes requires deep engineering expertise and some unavoidable trial-and-error, it is possible to generate guesses that reflect configurations that process engineers would likely test initially. To accomplish this goal, we developed a decision tree capable of automating the initial design of separation and product recovery. The decision tree starts from the bioreactor (see Fig. 5).

Microbially produced products can be accumulated within microorganisms (known as intracellular products) or in the fermentation broth (known as extracellular products) (Petrides, 2015; Seader et al., 2010). Intracellular products, such as antibiotics and proteins, need to be harvested from the cell biomass prior to product purification processes. Briefly, the cells are harvested out of the fermentation broth by a solid–liquid separation process; then cells must be lysed to expose bioproducts inside the cell. After cell disruption, cell debris becomes the main impurity needed for removal. Once cell debris is discarded, based on the purity requirement or applications of the desired product, product purification process is conducted to obtain the final product for short-term on-site storage.

For extracellular products, the first step is to remove cells from the fermentation broth through a solid–liquid separation and then conduct an initial concentration of the product in the broth. After getting rid of the large impurities, depending on the desired purity of the final product and its target market, a product purification step is needed to reach the required purity level of the final product and stored onsite. Aside from intracellular and extracellular products, industries are also interested in harvesting cells such as probiotics (Grand View Research, 2022) used in the food and beverage industries. In this case, cell harvesting is performed as the first step and followed by a drying process to obtain high purity of the cell itself (Fenster et al., 2019). Lastly, some renewable gaseous products, such as isoprene require a different separation and recovery processes. For gaseous products, the product is recovered from the bioconversion reactor, condensed, and then sent to a purification process, depending on the purity requirement. Detailed selection of the unit process for different product categories can be found in the Supplementary Information. When calculating the minimum separation cost, we assume an optimal recovery efficiency (>90 % depending on the unit process) of bioproducts and no side reaction will happen during downstream separation and recovery process.

3. Results and discussion

The machine learning surrogate models for predicting mass and energy balances of bioethanol production trained on SuperPro process simulation trials achieved very high performance on the 1,000 test samples withheld from training. Since individual surrogate models were developed for separate cost, mass and energy outputs estimated by the underlying SuperPro model, we aggregated performance metrics of the models in these three categories and report categorical summary statistics here. Of all three types of outputs, the surrogate models for predicting mass balances achieved the highest accuracy. Surrogate models for predicting mass balance outputs from SuperPro simulation trials for bioethanol also performed very well on the test set. The performance of these models in terms of MAPE ranged from 0.002 to 0.013 with a mean of 0.007 and standard deviation of 0.003 (Fig. 2). Surrogate models for predicting energy balance outputs from SuperPro simulation trials for bioethanol displayed a similarly high level of performance. The MAPE values of these models ranged from 0.006 to 0.08 with a mean of 0.022 and standard deviation of 0.028 (Fig. 3). The largest MAPE of 0.08 was found for predicting onsite electricity generated. Lastly, the surrogate models for predicting minimum selling price exhibited slightly higher errors than those for mass and energy balances with a mean MAPE of 0.064, range of 0.03 to 0.12 and standard deviation of 0.034 (Fig. 4). However, the discrepancies in performance among all these models are relatively small, especially with respect to the generally low magnitude of their error rates.

Adding flexibility in the downstream separation and recovery process can help scientists gain a better understanding of the complexity of the downstream processes and adjust their time and budgets accordingly. A decision tree approach like the one presented in this paper can be used as a standalone tool, or in combination with conventional or machine learning-based models of an entire process. Users can input the physical and chemical properties of the desired product, and key process-related parameters, such as yield/tier and purity, in the additional separation tool; we then use the above-mentioned separation decision tree to recommend a strategy with total capital investment and annual operating costs associated with the process. A sample user interface of the separation tool (publicly available at lead.jbei.org) can be found in SI-Fig. S3.

Although an automated design tool is not a viable substitute for pilot and demonstration runs, users can vary the input parameters, such as target product purity or product titer, to better understand how separation costs will be impacted. Users can also compare the separation costs by selecting alternative unit processes. Taking ethanol separation as an example, the default titer is 5.4 wt% and final product purity is 97 % based on NREL's report (Humbird et al., 2011). When separating ethanol after fermentation, the separation tool suggests starting from centrifugation and filtration before distillation to remove large impurities first; then, a molecular sieve is added to eliminate water. A drying process is further added to reduce moisture content in the product before stored onsite. The total capital investment of this ethanol separation process is ~$30 million and the operating cost is ~$2 million. The estimated separation cost of ethanol is $0.24/gal according to the tool's calculation. For comparison, Amornraksa et al. evaluated three different separation processes (conventional molecular sieve, extractive distillation, and pervaporation) of bioethanol produced from corn stover and

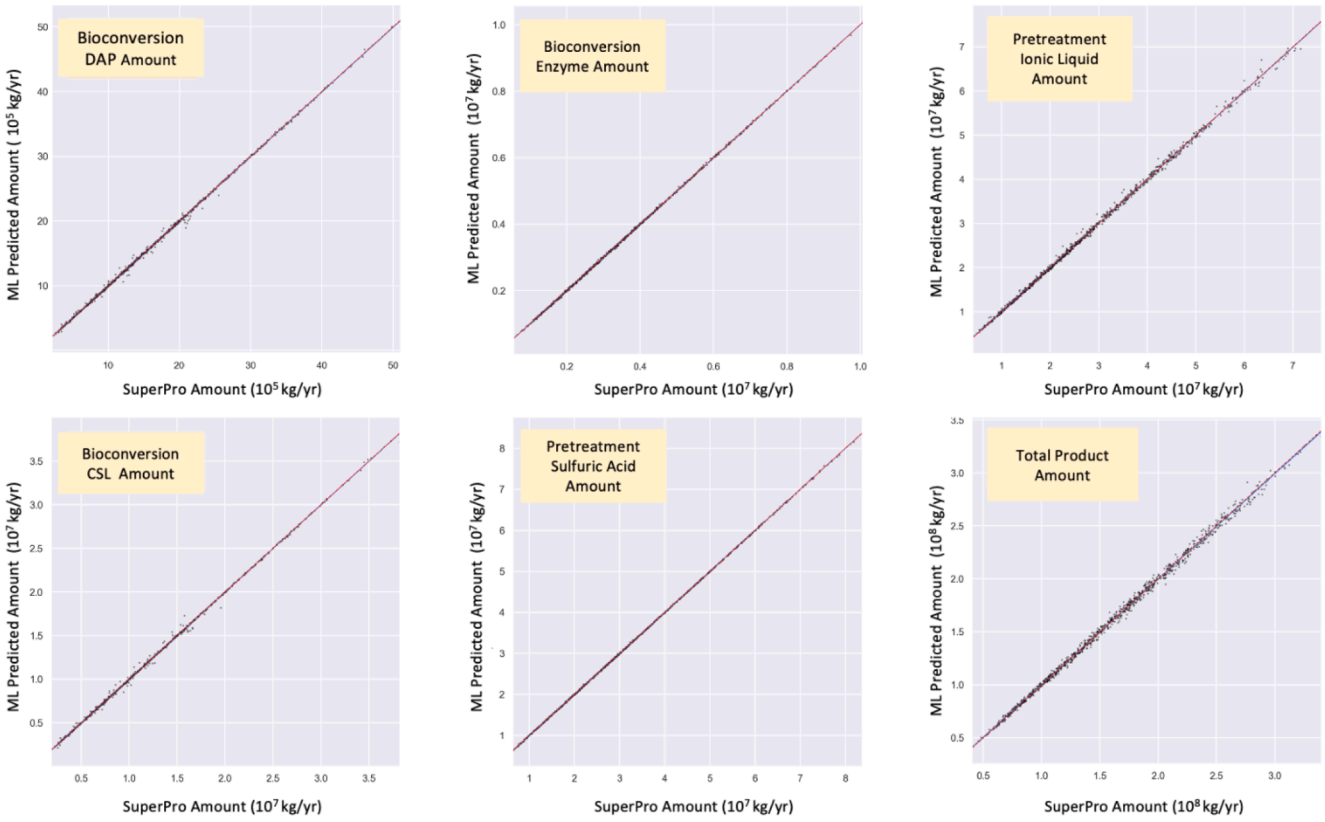


Fig. 2. Parity plots of mass balance predictions by ML surrogate models versus SuperPro Designer simulation outputs for major process stages of the ethanol production pathway. The MAPE values of these models on test data (withheld from training) ranged from 0.002 to 0.013 with a mean of 0.007 and standard deviation of 0.003. A greater clustering of points around the y = x line (dashed red) in these plots indicates closer agreement between predicted and actual values, and thus higher prediction accuracy. DAP = diammonium phosphate; CSL = corn steep liquor. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

they reported that the total capital investment varied between $22 million to $36 million, with the conventional molecular sieve being the highest (Amornraksa et al., 2020). The ethanol separation cost from the biomass sorghum to ethanol process modeled in SuperPro Designer is ~ $0.28/gal. If the product titer is decreased by 50 %, the separation costs will be increased to $0.45/gal. When the final purity requirement decreases from 97 % to 90 %, the corresponding separation cost could be reduced to $0.16/gal. In product concentration steps, instead of distillation, users can also select filtration; however, switching to filtration could increase the separation costs to $0.27/gal. The market value of the final product also has a large impact on the separation costs. In the ethanol case, if the market selling price of ethanol is high (>$100/kg), the final separation cost jumps to $4.81/gal with the large contribution from the freeze dryer, which is the last concentration step. Although mass and energy balances were not explored in this particular case study, a decision tree approach can apply to such analyses, assuming it is well integrated with the full-scale biorefinery model to account for waste heat recovery and other recycle streams.

The results show that machine learning can be used to develop fairly accurate surrogate models for complex process simulations of biofuel production pathways. The predictive accuracy of such models varies depending on the model output of interest, with mass and energy balance surrogate models generally performing better than those for predicting minimum selling price models. However, it is important to highlight the variability in surrogate model performance depending on the major process stage for which outputs are being predicted. While the mass and energy balance surrogate models had lower mean error rates, the standard deviation in performance was greatest for the energy prediction surrogate models, followed by the minimum selling price surrogates and lastly the mass balance surrogates. The largest MAPE values of 0.08 and 0.02, were for predicting onsite electricity generated and electricity required for ionic liquid pretreatment, respectively. The high errors are related to the ionic liquid recycle stream. In 141 of the 10,000 SuperPro trials run for this study, material and energy balances for the pretreatment stage fail to converge. This failure to converge resulted from the ionic liquid recovery rate changing from a low to a high value in back-to-back runs. An inaccurate prediction of electricity requirements during pretreatment also increases the error for total electricity generated onsite. The modeling error caused by the failed convergence of the SuperPro model can be minimized by adding a dummy unit to flush out the main stream where the recycling stream is connected, then re-running the material and energy balances. Further, adding more dependent variables, including material flow through the pretreatment reactor or size of the pretreatment reactor as well as parameters that determine biogenic energy sources routed to the boiler, including lignin and biogas, can minimize the error found in this study.

This study adds to a growing body of work in the biofuels or biochemicals production industry to leverage machine learning techniques for modeling tasks that have conventionally relied upon costly and complex process simulation software. In the last several decades, rapid computational advances have made machine learning tools increasingly accessible across a wide range of scientific domains, especially in the chemical engineering space (Dobbelaere et al., 2021; Lee et al., 2018; Schweidtmann et al., 2021). Examples in biofuel production span from the use of artificial neural networks to estimate the density and kinematic viscosity of biodiesel (Özgür and Tosun, 2017) to support vector machines for predicting metabolic fluxes (Wu et al., 2016). With growing interests in making industrial system processes more

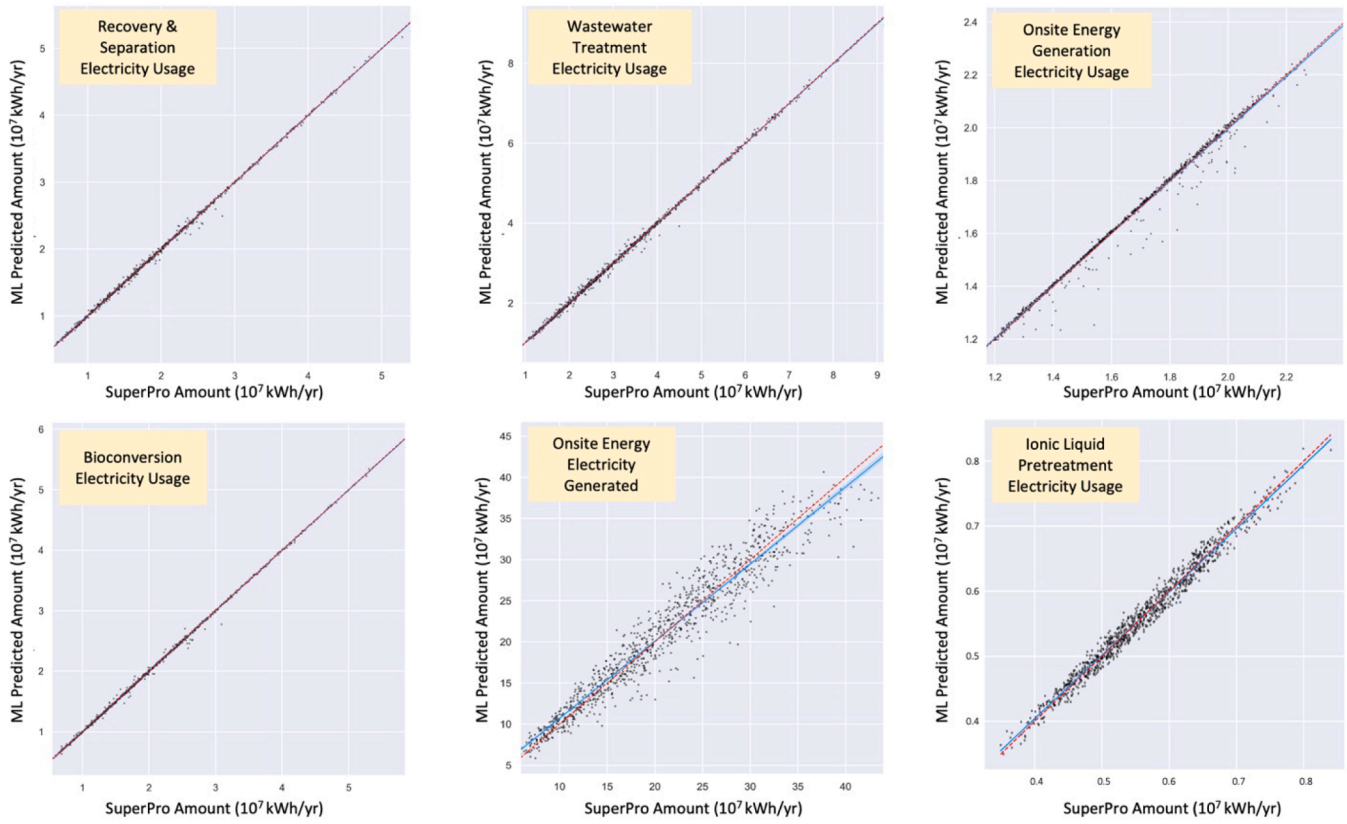


Fig. 3. Parity plots of energy balance predictions by ML surrogate models versus SuperPro Designer simulation outputs for major process stages of the ethanol production pathway. The MAPE values of these models on test data (withheld from training) ranged from 0.006 to 0.08 with a mean of 0.022 and standard deviation of 0.028. A greater clustering of points around the y = x line (dashed red) in these plots indicates closer agreement between predicted and actual values, and thus higher prediction accuracy. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

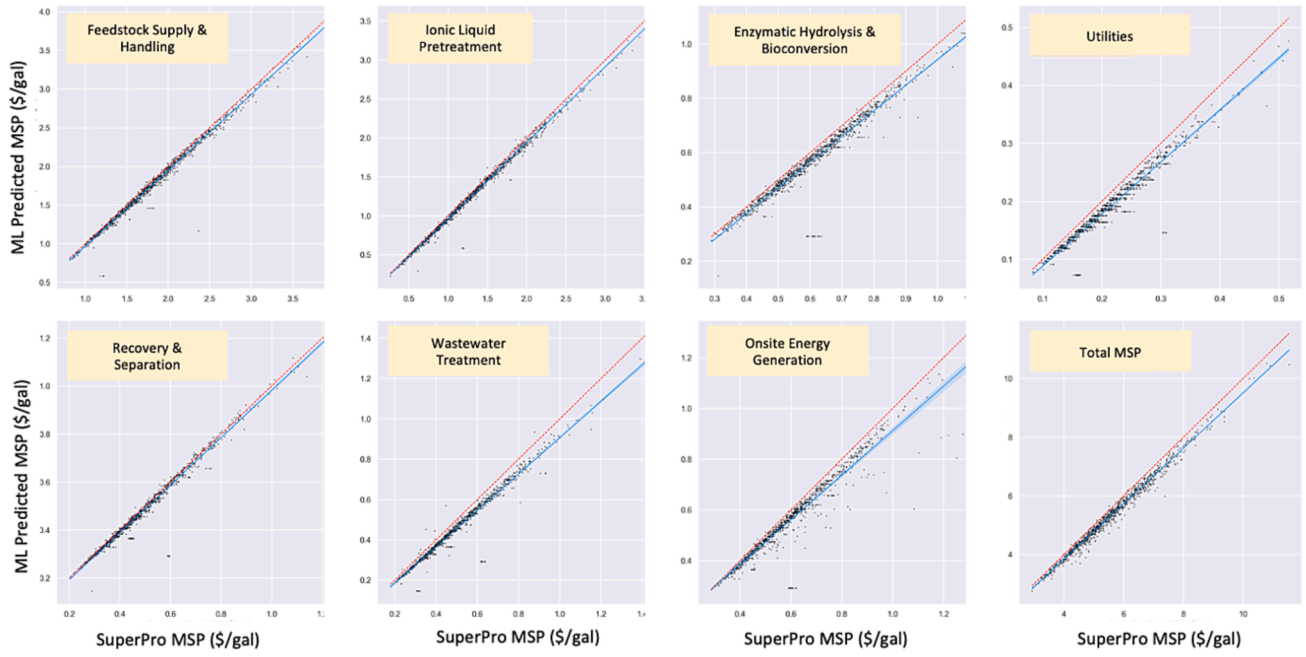


Fig. 4. Parity plots of minimum selling price (MSP) predictions by ML surrogate models versus SuperPro Designer simulation outputs for major process stages of the ethanol production pathway. The MAPE values of these models on test data (withheld from training) ranged from 0.03 to 0.12 with a mean of 0.064 and standard deviation of 0.034. A greater clustering of points around the y = x line (dashed red) in these plots indicates closer agreement between predicted and actual values, and thus higher prediction accuracy. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

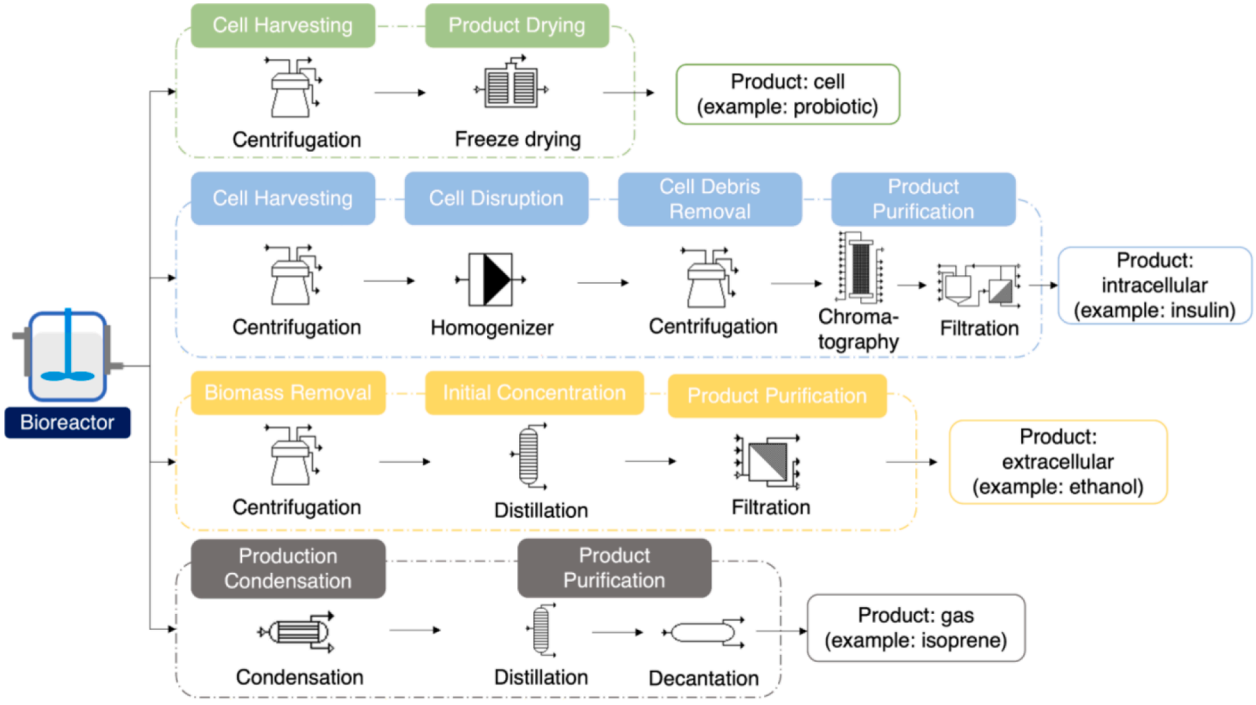


Fig. 5. General approaches of flexible downstream separation, purification and product recovery processes.

sustainable, there has been an emerging trend in the application of machine learning to techno-economic and life-cycle impact modeling. Machine learning techniques combined with kinetic simulation modeling have been shown to be effective in predicting generating energy and greenhouse gas inventory data of activated carbon production (Liao et al., 2020) and support vector regression models have proven useful in system optimizations to lower NOx emissions from a coal-fired utility boilers (Zheng et al., 2009). While these examples illustrate that conventional simulation modeling methods in the chemical industry have started to become accompanied and enhanced by machine learning techniques, fewer studies have explored the potential of tree-based models, specifically via the use of automated machine learning algorithms, to serve as surrogates for estimating costs, mass and energy balances of process simulations for modeling the production of high value biofuels and bioproducts. Thus, the demonstration that such an approach can yield high-performing models in this study sets it apart in the literature as a novel application of these computational methods in an industrial sector where technological development is imperative to securing a clean and sustainable future. Moreover, we show that tree-based decision algorithms for determining specific equipment configurations of certain system model components can offer a powerful complement to surrogate modeling. In this study, we demonstrate the viability of this approach for downstream separation processes. Such algorithms could enable rapid development of prospective system designs that could then be used to build process simulation models that could serve as the basis for training ML-based surrogate models as we have demonstrated in this study's methodology. Thus, there is vast potential for a synergistic link between these computational approaches which could serve to accelerate the design, build and test cycle of bioprocess system development and optimization. Future work that explores the integration of these methods would be a worthwhile pursuit that stands to advance the field.

4. Conclusions

Technoeconomic analyses and life-cycle assessments are critical to estimating costs and environmental impacts of biofuel and bioproduct production pathways. The ability to run these computations rapidly is important to fully explore the design space for biorefineries. While these approaches have limitations with regard to model generalizability, the methodology presented in this study shows that machine learning and decision-tree algorithms can usefully extend the utility of conventional process simulations in order to more rapidly predict model outputs and iterate over potential equipment layouts. Future work is needed to develop these approaches further with an aim toward increasing their flexibility and accuracy.

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.

Data availability

Data will be made available on request.

Acknowledgements

This work was part of the DOE Joint BioEnergy Institute (http://www.jbei.org) supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U.S. Department of Energy. This study was also supported by the U.S. Department of Energy, Energy Efficiency and Renewable Energy, Bioenergy Technologies Office. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.biortech.2022.128528.

References

A. Aden M. Ruth K. Ibsen J. Jechura K. Neeves J. Sheehan B. Wallace L. Montague A. Slayton J. Lukas Lignocellulosic Biomass to Ethanol Process Design and Economics Utilizing Co-Current Dilute Acid Prehydrolysis and Enzymatic Hydrolysis for Corn Stover National Renewable Energy Laboratory (NREL) 2002 Golden, CO 10.2172/15001119.

Aminian, A., ZareNezhad, B., 2018. Accurate predicting the viscosity of biodiesels and blends using soft computing models. Renew. Energy 120, 488–500. https://doi.org/10.1016/j.renene.2017.12.038.

Amornraksa, S., Subsaipin, I., Simasatitkul, L., Assabumrungrat, S., 2020. Systematic design of separation process for bioethanol production from corn stover. BMC Chem. Eng. 2, 10. https://doi.org/10.1186/s42480-020-00033-1.

Baghban, A., Kardani, M.N., Mohammadi, A.H., 2018. Improved estimation of Cetane number of fatty acid methyl esters (FAMEs) based biodiesels using TLBO-NN and PSO-NN models. Fuel 232, 620–631. https://doi.org/10.1016/j.fuel.2018.05.166.

Baral, N.R., Kavvada, O., Mendez-Perez, D., Mukhopadhyay, A., Lee, T.S., Simmons, B.A., Scown, C.D., 2019. Techno-economic analysis and life-cycle greenhouse gas mitigation cost of five routes to bio-jet fuel blendstocks. Energy Environ. Sci. 12, 807–824. https://doi.org/10.1039/C8EE03266A.

Baral, N.R., Dahlberg, J., Putnam, D., Mortimer, J.C., Scown, C.D., 2020. Supply cost and life-cycle greenhouse gas footprint of dry and ensiled biomass sorghum for biofuel production. ACS Sustain. Chem. Eng. 8, 15855–15864. https://doi.org/10.1021/acssuschemeng.0c03784.

Bujang, M.A., Sa'at, N., Sidik, T.M.I.T.A.B., Joo, L.C., 2018. Sample Size Guidelines for Logistic Regression from Observational Studies with Large Population: Emphasis on the Accuracy Between Statistics and Parameters Based on Real Life Clinical Data. Malays. J. Med. Sci. 25, 122–130. https://doi.org/10.21315/mjms2018.25.4.12.

Burk, C., 2022. Techno-Economic Analysis for Hard-Tech Innovation. https://www.activate.org/learn (accessed, October 27, 2022).

Cochran, W.G., 1977. Sampling Techniques. John Wiley & Sons.

Comesana, A.E., Huntington, T.T., Scown, C.D., Niemeyer, K.E., Rapp, V.H., 2022. A systematic method for selecting molecular descriptors as features when training models for predicting physiochemical properties. Fuel 321, 123836.

Dobbelaere, M.R., Plehiers, P.P., Van de Vijver, R., Stevens, C.V., Van Geem, K.M., 2021. Machine learning in chemical engineering: strengths, weaknesses, opportunities, and threats. Engineering. https://doi.org/10.1016/j.eng.2021.03.019.

Fenster, K., Freeburg, B., Hollard, C., Wong, C., Rønhave Laursen, Ouwehand, A.C., 2019. The production and delivery of probiotics: A review of a practical approach. Microorganisms 7 (3), 83. https://doi.org/10.3390/microorganisms7030083.

Grand View Research, 2022. Probiotics Market Size, Industry Report, 2021–2030 [WWW Document]. accessed 10.12.22. https://www.grandviewresearch.com/industry-analysis/probiotics-market.

D. Humbird R. Davis L. Tao C. Kinchin D. Hsu A. Aden P. Schoen J. Lukas B. Olthof M. Worley D. Sexton D. Dudgeon Process Design and Economics for Biochemical Conversion of Lignocellulosic Biomass to Ethanol Dilute-Acid Pretreatment and Enzymatic Hydrolysis of Corn Stover. National Renewable Energy Laboratory (NREL) 2011 Golden 10.2172/1013269 CO (United States).

Huntington, T., Cui, X., Mishra, U., Scown, C.D., 2020. Machine learning to predict biomass sorghum yields under future climate scenarios. Biofuels, Bioproducts and Biorefining 14 (3), 566–577.

Kaab, A., Sharifi, M., Mobli, H., Nabavi-Pelesaraei, A., Chau, K.-W., 2019. Combined life cycle assessment and artificial intelligence for prediction of output energy and environmental impacts of sugarcane production. Sci. Total Environ. 664, 1005–1019. https://doi.org/10.1016/j.scitotenv.2019.02.004.

Le, T.T., Fu, W., Moore, J.H., 2020. Scaling tree-based automated machine learning to biomedical big data with a feature set selector. Bioinformatics 36, 250–256. https://doi.org/10.1093/bioinformatics/btz470.

Lee, J.H., Shin, J., Realff, M.J., 2018. Machine learning: Overview of the recent progresses and implications for the process systems engineering field. Comput. Chem. Eng. 114, 111–121. https://doi.org/10.1016/j.compchemeng.2017.10.008.

Li, Z., Ma, X., Xin, H., 2017. Feature engineering of machine-learning chemisorption models for catalyst design. Catal. Today 280, 232–238. https://doi.org/10.1016/j.cattod.2016.04.013.

Liao, M., Kelley, S., Yao, Y., 2020. Generating energy and greenhouse gas inventory data of activated carbon production using machine learning and kinetic based process simulation. ACS Sustain. Chem. Eng. 8, 1252–1261. https://doi.org/10.1021/acssuschemeng.9b06522.

Magurudeniya, H.D., Baral, N.R., Rodriguez, A., Scown, C.D., Dahlberg, J., Putnam, D., George, A., Simmons, B.A., Gladden, J.M., 2021. Use of ensiled biomass sorghum increases ionic liquid pretreatment efficiency and reduces biofuel production cost and carbon footprint. Green Chem. 23, 3127–3140. https://doi.org/10.1039/D0GC03260C.

Mahmud, R., Moni, S.M., High, K., Carbajales-Dale, M., 2021. Integration of techno-economic analysis and life cycle assessment for sustainable process design – A review. J. Clean. Prod. 317, 128247 https://doi.org/10.1016/j.jclepro.2021.128247.

Marcou, G., Aires de Sousa, J., Latino, D.A.R.S., de Luca, A., Horvath, D., Rietsch, V., Varnek, A., 2015. Expert system for predicting reaction conditions: the Michael reaction case. J. Chem. Inf. Model. 55, 239–250. https://doi.org/10.1021/ci500698a.

Martínez-Aragón, M., Burghoff, S., Goetheer, E.L.V., de Haan, A.B., 2009. Guidelines for solvent selection for carrier mediated extraction of proteins. Separation and Purification Technology 65, 65–72. https://doi.org/10.1016/j.seppur.2008.01.028.

Negny, S., Belaud, J.P., Cortes Robles, G., Roldan Reyes, E., Ferrer, J.B., 2012. Toward an eco-innovative method based on a better use of resources: application to chemical process preliminary design. J. Clean. Prod. 32, 101–113. https://doi.org/10.1016/j.jclepro.2012.03.023.

Ochoa-Estopier, L.M., Jobson, M., Smith, R., 2013. Operational optimization of crude oil distillation systems using artificial neural networks. Comput. Chem. Eng. 59, 178–185. https://doi.org/10.1016/j.compchemeng.2013.05.030.

Özgür, C., Tosun, E., 2017. Prediction of density and kinematic viscosity of biodiesel by artificial neural networks. Energy Sources, Part A: Recovery, Utilization, and Environmental Effects 39, 985–991. https://doi.org/10.1080/15567036.2017.1280563.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., 2011. Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research 12, 2825–2830.

Petrides, D., 2015. Bioprocess design and economics.

Romeiko, X.X., Guo, Z., Pang, Y., 2019. Comparison of Support Vector Machine and Gradient Boosting Regression Tree for Predicting Spatially Explicit Life Cycle Global Warming and Eutrophication Impacts: A case study in corn production, in: 2019 IEEE International Conference on Big Data (Big Data). Presented at the 2019 IEEE International Conference on Big Data (Big Data), IEEE, pp. 3277–3284. doi:10.1109/BigData47090.2019.9005581.

Rudin, C., 2018. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. arXiv. doi:10.48550/arxiv.1811.10154.

Schweidtmann, A.M., Esche, E., Fischer, A., Kloft, M., Repke, J., Sager, S., Mitsos, A., 2021. Machine learning in chemical engineering: A perspective. Chemie Ingenieur Technik 93, 2029–2039. https://doi.org/10.1002/cite.202100083.

Scown, C.D., Baral, N.R., Yang, M., Vora, N., Huntington, T., 2021. Technoeconomic analysis for biofuels and bioproducts. Curr. Opin. Biotechnol. 67, 58–64. https://doi.org/10.1016/j.copbio.2021.01.002.

Seader, J.D., Henley, E.J., Roper, D.K., 2010. Separation Process Principles with Applications using Process Simulators, 3rd ed. Wiley, Hoboken, NJ.

Song, R., Keller, A.A., Suh, S., 2017. Rapid Life-Cycle Impact Screening Using Artificial Neural Networks. Environ. Sci. Technol. 51, 10777–10785. https://doi.org/10.1021/acs.est.7b02862.

Sun, J., Shi, J., Murthy Konda, N.V.S.N., Campos, D., Liu, D., Nemser, S., Shamshina, J., Dutta, T., Berton, P., Gurau, G., Rogers, R.D., Simmons, B.A., Singh, S., 2017. Efficient dehydration and recovery of ionic liquid after lignocellulosic processing using pervaporation. Biotechnol. Biofuels 10, 154. https://doi.org/10.1186/s13068-017-0842-9.

Verma, O.P., Manik, G., Suryakant Jain, V.K., Jain, D.K., Wang, H., 2018. Minimization of energy consumption in multiple stage evaporator using Genetic Algorithm. Sustainable Computing: Informatics and Systems 20, 130–140. https://doi.org/10.1016/j.suscom.2017.11.005.

Vittinghoff, E., McCulloch, C.-E., 2007. Relaxing the rule of ten events per variable in logistic and Cox regression. American J. Epidemiol. 165 (6), 710–718. https://doi.org/10.1093/aje/kwk052.

Wang, J., Lin, M., Xu, M., Yang, S.-T., 2016. Anaerobic Fermentation for Production of Carboxylic Acids as Bulk Chemicals from Renewable Biomass. Adv. Biochem. Eng. Biotechnol. 156, 323–361. https://doi.org/10.1007/10_2015_5009.

Wu, S.G., Wang, Y., Jiang, W., Oyetunde, T., Yao, R., Zhang, X., Shimizu, K., Tang, Y.J., Bao, F.S., 2016. Rapid prediction of bacterial heterotrophic fluxomics using machine learning and constraint programming. PLoS Comput. Biol. 12, e1004838.

Yang, M., Baral, N.R., Simmons, B.A., Mortimer, J.C., Shih, P.M., Scown, C.D., 2020. Accumulation of high-value bioproducts in planta can improve the economics of advanced biofuels. Proc Natl Acad Sci USA 117, 8639–8648. https://doi.org/10.1073/pnas.2000053117.

Zheng, L.-G., Zhou, H., Cen, K.-F., Wang, C.-L., 2009. A comparative study of optimization algorithms for low NOx combustion modification at a coal-fired utility boiler. Expert Syst. Appl. 36, 2780–2793. https://doi.org/10.1016/j.eswa.2008.01.088.