Lawrence Berkeley National Laboratory
Recent Work

Title
Tree-Based Automated Machine Learning to Predict Biogas Production for Anaerobic Co-digestion of Organic Waste

Permalink
https://escholarship.org/uc/item/1gs17016

Journal
ACS Sustainable Chemistry and Engineering, 9(38)

ISSN
2168-0485

Authors
Wang, Y
Huntington, T
Scown, CD

Publication Date
2021-09-27

DOI
10.1021/acssuschemeng.1c04612

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

ACS Sustainable Chemistry & Engineering

pubs.acs.org/journal/ascecg

Research Article

Tree-Based Automated Machine Learning to Predict Biogas Production for Anaerobic Co-digestion of Organic Waste

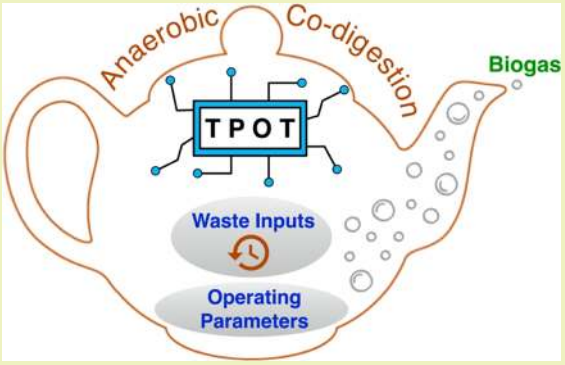
Yan Wang, Tyler Huntington, and Corinne D. Scown*

Cite This: *ACS Sustainable Chem. Eng.* 2021, 9, 12990−13000

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The dynamics of microbial communities involved in anaerobic digestion of mixed organic waste are notoriously complex and difficult to model, yet successful operation of anaerobic digestion is critical to the goals of diverting high-moisture organic waste from landfills. Machine learning (ML) is ideally suited to capturing complex and nonlinear behavior that cannot be modeled mechanistically. This study uses 8 years of data collected from an industrial-scale anaerobic co-digestion (AcoD) operation at a municipal wastewater treatment plant in Oakland, California, combined with a powerful automated ML method, Tree-based Pipeline Optimization Tool, to develop an improved understanding of how different waste inputs and operating conditions impact biogas yield. The model inputs included daily input volumes of 31 waste streams and 5 operating parameters. Because different wastes are broken down at varying rates, the model explored a range of time lags ascribed to each waste input ranging from 0 to 30 days. The results suggest that the waste types (including rendering waste, lactose, poultry waste, and fats, oils, and greases) differ considerably in their impact on biogas yield on both a per-gallon basis and a mass of volatile solids basis, while operating parameters were not good predictors of yield at this facility.



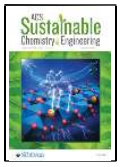
KEYWORDS: *TPOT, machine learning, biogas, anaerobic digestion, bioenergy, wastewater treatment, organic waste*

Downloaded via 104.186.254.98 on November 5, 2021 at 20:08:35 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

■ INTRODUCTION

Anaerobic digestion (AD) has been used to generate combustible fuel from organic wastes since the 1800s and, although advancements in synthetic biology have resulted in more targeted routes to producing specific fuel molecules, AD remains one of the most efficient strategies for converting mixed organic waste to renewable energy and nutrient-rich residual solids. In the U.S., there are over 2200 biogas production sites, of which over 1200 are industrial-scale AD facilities aligning with water resource recovery and additional 263 operate on livestock farms.[1−3] Ambitious "zero waste" policies from local and state governments across the U.S. will require substantial new investments in infrastructure to divert organic waste from landfills, and AD is an essential part of any viable strategy.[4,5] However, different organic wastes may be more or less suitable for use as an AD feedstock and, ideally, facilities could establish prioritization and pricing structures based on a waste's impact on digester performance. However, the development of reliable predictive models that estimate biogas yield as a function of feedstock type/composition has proved challenging. Unlike industrial bioreactors where a single microbial host utilizes a pure substrate (e.g., glucose), AD is most effective with heterogeneous organic feedstocks, as the use of a single feedstock (monodigestion) often results a nutritional imbalance of substrates.[6−9] Anaerobic co-digestion (AcoD) of multiple substrates has been widely employed to achieve the right nutrient balance and dilute inhibitory substances in the digester, improving biogas production and stability.[6−11] Co-digestion can increase biogas yield by 25−400% compared to monodigestion.[6,10] AcoD is an effective practice for wastewater treatment plants (WWTPs), which receive organics for co-digestion with their sewage sludges in exchange for tipping fee revenue and combust the resulting biogas onsite to provide heat, electricity, or mechanical power.[12,13] Approximately, 20% of AD facilities at U.S. WWTPs co-digest offsite wastes, with combined heat and electricity (CHP) as the dominant biogas use.[13] The complexity associated with utilizing a microbial community grown on heterogeneous, variable substrates means that mechanistic modeling is usually impractical; the data required for a mechanistic model is vast and impossible to collect on a regular basis. Thus, biogas yield and composition prediction has remained a largely empirical exercise. Advanced regression

Received: July 8, 2021
Published: September 16, 2021



ACS Publications

© 2021 The Authors. Published by American Chemical Society

https://doi.org/10.1021/acssuschemeng.1c04612
ACS Sustainable Chem. Eng. 2021, 9, 12990−13000

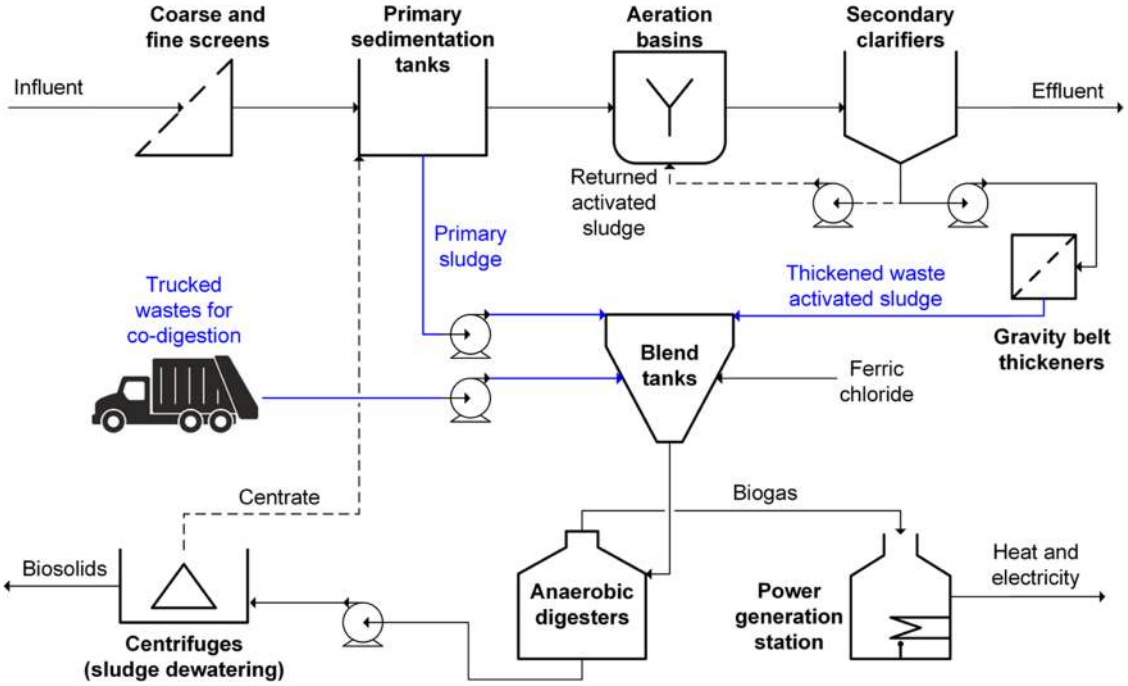


Figure 1. Simplified process flow diagram for the integrated full-scale WWTP-AcoD system. The waste streams fed into the AcoD facility are indicated in blue.

techniques, under the umbrella of machine learning (ML), offer an opportunity to improve upon current practices.

Generally, AD involves four successive steps that each relies on different communities of microbes, including hydrolysis, acidogenesis, acetogenesis, and methanogenesis.[7,14] The resulting biogas product consists primarily of CH_4 (50−75%) and CO_2 (25−50%), with trace amounts of other components present as well such as H_2O, O_2, N_2, H_2S, NH_3, siloxanes, and halogenated hydrocarbons.[15−18] Biogas can be an attractive renewable fuel, particularly given the U.S. Environmental Protection Agency (USEPA) 2014 ruling that qualified biogas under the expanded Renewable Fuel Standard (RFS2) to claim D3 or D5 Renewable Identification Number (RIN) credits, depending on the type of organic wastes utilized.[19] Biogas cleaning is necessary to remove trace contaminants, at which point it can be combusted for heat and/or electricity generation, used in gas fermentation processes, or upgraded and compressed or liquified for higher-value uses in transportation applications or steam methane reforming.[20,21] Facilities employing AD to generate biogas must make decisions regarding tipping fees for, and willingness to accept, specific incoming waste types based on digester stability, biogas yield, and the likelihood of problematic contaminants (e.g., cutlery or other items that may clog or damage equipment). These decisions are often based on qualitative judgements and anecdotal observations rather than rigorous data analysis. Predictive modeling capabilities can provide a more quantitative basis to support decision-making for AcoD and improve resource utilization efficiency in the future.[6]

Mechanistic models, such as the well-known Anaerobic Digestion Model 1 (ADM1), have made important strides in the scientific community's ability to predict digestion performance. However, the ADM1 model requires knowledge of many concentration state variables (i.e., the concentrations for detail components of substrates), which necessitates extensive ongoing analysis of substrates, thus limiting its applicability in industrial facilities where this data is not regularly collected.[22−24] Also, the complicated microbial and physicochemical process of AD substantially affects the prediction accuracy of mechanistic models.[25] Given the fact that AD is often a nonlinear process, traditional statistical models (e.g., linear regression) have shown deficient performance for a generalized prediction of biogas production.[26] When mechanistic modeling is not feasible or sufficient, and training data is available, machine learning (ML) can be the best option for developing predictive models and developing insights into the influence of key parameters. In the past decade, different ML techniques (summarized in Table S1) have been leveraged to predict biogas production, including connectivism learning (e.g., artificial neural network, ANN) and statistical learning (e.g., random forest, extreme gradient boosting, support vector machine).[22,23,27−37] Previous research either employed a single technique or aimed to compare several techniques to select the best-performing approach. Table S1 summarizes 13 prior studies using ML to predict biogas production from AD as compared to this study. These prior studies were also limited by the training data available to them.

Most ML studies have used fairly limited datasets, based on digesters operating at the lab scale or larger digesters fed with only a few substrates (Table S1). To the best of our knowledge, only the study from De Clercq et al. used a comparatively large dataset, with 4 years of operational data from an industrial AcoD facility treating 16 types of organic wastes (total dataset size of 1398 entries including feedstock inputs but no operating conditions).[23] The team analyzed their data using elastic net, random forest, and extreme gradient boosting models to predict biomethane production and the study placed an emphasis on comparing model performance across different built-in time lags between feedstock input and predicted biogas production.[23] Their best-performing models achieved an R^2 between 0.8 and 0.88 for the test dataset,

https://doi.org/10.1021/acssuschemeng.1c04612
ACS Sustainable Chem. Eng. 2021, 9, 12990−13000

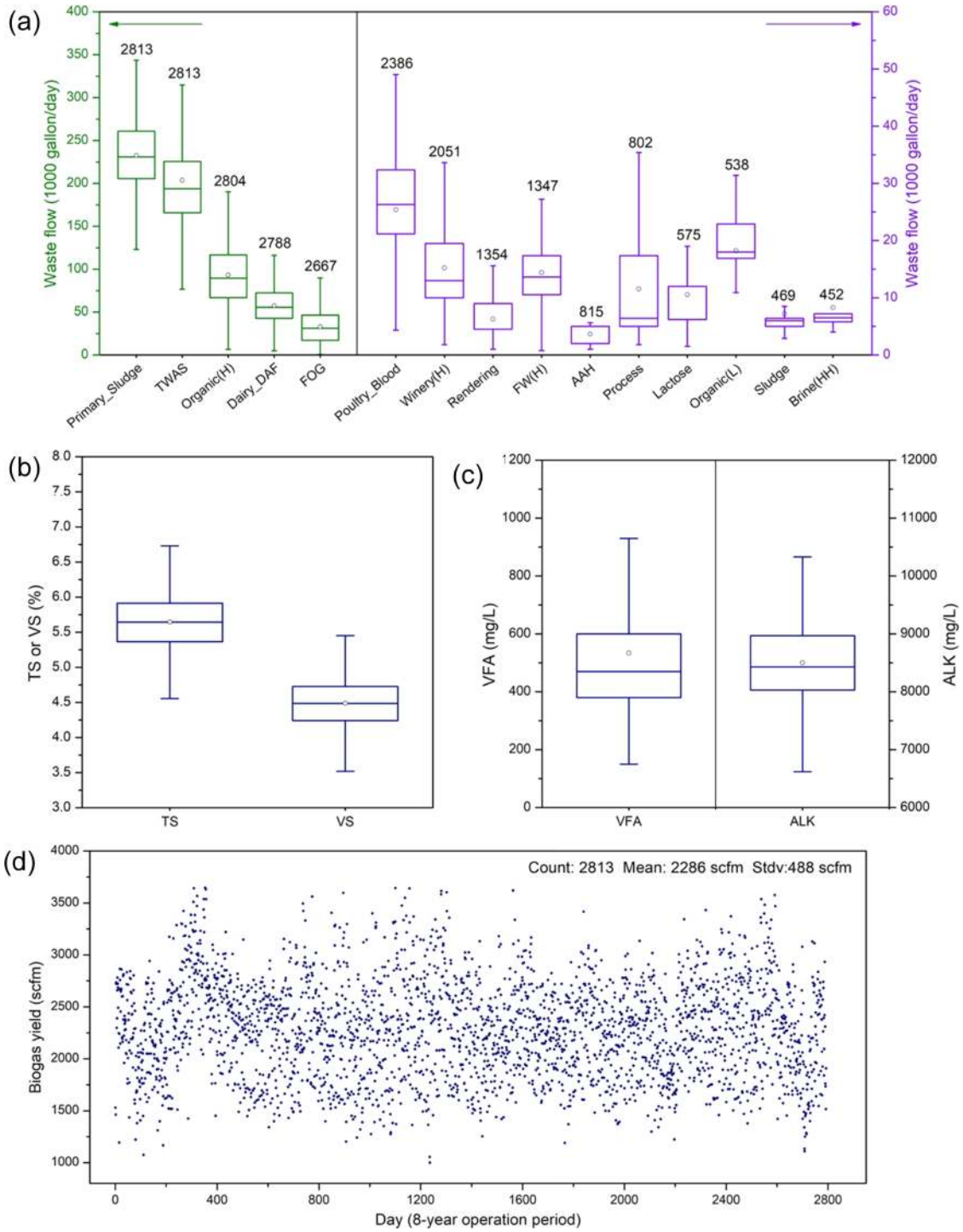


Figure 2. (a) Box plots (minimum, 25th percentile, median, 75th percentile, maximum, and mean by circles) displaying the data distribution for daily input volume (gallon/day) of 15 waste types that are in significant quantities. The flows for Primary_Sludge, TWAS, Organic (H), Dairy_DAF, and FOG are shown in the left *y*-axis (green), while those for other waste types are shown in the right *y*-axis (purple). The numeric values above the box refer to the number of collected data points, i.e., the number of days the digester accepts each waste type over an 8-year operation period. The name acronyms of wastes used along the *x*-axis are defined in Table S2. (b, c) Data distribution for digester operating parameters, including total solid (TS, %), volatile solid (VS, %), the content of volatile fatty acids (VFA, mg/L), and alkalinity (ALK, mg/L). (d) Evolution of the biogas production (scfm) during an 8-year operation period.

although the study relied on a fairly small test dataset (136 entries resulting from a 0.9/0.1 train−test split ratio).

Our study aims to significantly improve the state of the art relative to prior studies. First, we used the most extensive dataset documented to-date, collected from an integrated full-scale WWTP-AcoD system spanning an 8-year operation period and accepting 31 different waste streams to produce biogas (approximately double the size of the De Clercq et al. study, at 2813 entries). By using a larger, more diverse dataset, the resulting model should provide greater insights and predictive capability, given ML's usefulness for interpolation and the challenges in using ML models trained on limited

https://doi.org/10.1021/acssuschemeng.1c04612
ACS Sustainable Chem. Eng. 2021, 9, 12990−13000

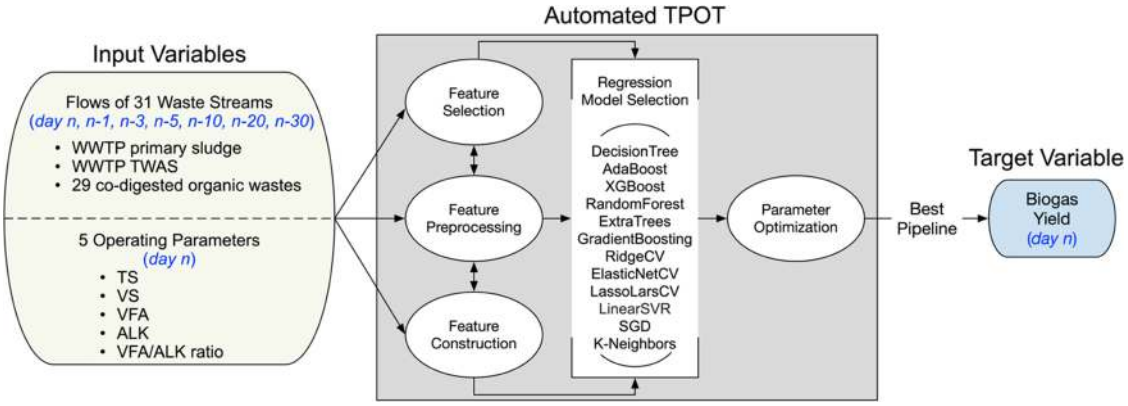


Figure 3. Schematic methodology using TPOT for biogas yield prediction based on 222 input variables, including 31 waste streams across different time lags (0, 1, 3, 5, 10, 20, 30 days) and 5 operating parameters at the current day. Typical steps in the ML pipeline automated by TPOT involve data transformation (feature selection, feature preprocessing, feature construction), model selection, and parameter optimization. The supervised learning regression models in the default TPOT configuration include decision trees, ensemble models (AdaBoost, XGBoost, forests of randomized trees, gradient tree boosting), cross-validated linear models (Ridge, Elastic Net, and Lasso using LARS algorithm), linear support vector regression (SVR), stochastic gradient descent (SGD), and k-nearest neighbors.

datasets for extrapolation. Second, this study demonstrates a newer ML modeling approach—automated ML pipelines—that can be more easily replicated by practitioners and nonexperts. One of the most successful automated ML systems is Tree-based Pipeline Optimization Tool (TPOT), which relies on genetic programming (GP) to recommend an optimized analysis pipeline including supervised classification/regression operators, feature preprocessing operators, and feature selection operators.[38−41] We compared the performance of our TPOT model with more traditional ML techniques to understand how the results may vary and what insights about industrial AcoD operations can be gleaned from these approaches. Finally, while prior studies have focused on comparing the performance of different ML models, this study places an emphasis on using the best-performing model to generate interpretable results and actionable information for AcoD facility operators.

■ METHODS

Data Collection and Structure. Data used in this study were collected from East Bay Municipal Utility District (EBMUD)'s WWTP (Oakland, California, U.S.) spanning an 8-year operation period (equivalent to 2813 days). The EBMUD service area previously included multiple large industrial facilities that contributed to high biological oxygen demand (BOD) load at the WWTP, including a dog food factory and numerous canneries across Berkeley, Emeryville, and Oakland, California. The Resource Recovery program was developed to compensate for reduced BOD load as these industrial facilities shut down by accepting various high-strength nonhazardous organic trucked wastes to increase onsite energy production. A simplified process flow diagram for this integrated WWTP-AcoD system is shown in Figure 1. The liquid wastewater treatment processes mainly include coarse and fine screens, primary sedimentation tanks, aerated activated sludge basins, and clarifiers. Treated effluent is disinfected, dechlorinated, and discharged to the San Francisco Bay. The solid treatment processes include activated sludge thickeners, blend tanks (for solid blending of primary sludge, thickened waste activated sludge (TWAS), and trucked wastes), low thermophilic anaerobic digestion, and digested biosolids dewatering. Ferric chloride ($FeCl_3$) is added to blend tanks (digester feed) for sulfide control, which reduces H_2S concentrations in the biogas. The biogas is combusted onsite to provide heat and electricity via the CHP system. The biosolids are applied to agricultural lands growing nonedible crops and, during the rainy season, used as alternative daily cover at landfills.

The AcoD facility accepts 29 types of trucked wastes, including brine, dairy, fats, oils, and greases (FOG), protein, process water, septage, sludge, food waste, winery, and general category for high-COD (chemical oxygen demand) and low-COD organics [denoted as Organic (H) and Organic (L), respectively]. Two additional waste types are sourced from the WWTP itself, for a total of 31 waste inputs. Definitions and brief descriptions of all waste streams fed into the AcoD facility are included in Table S2. Over the 8-year operation period, the daily input volume of each waste stream was recorded. Two hundred and thirty-nine thousand gallon/day of wastes on average were co-digested in the facility, while the average total feed reached 668 thousand gallons/day including WWTP sludge wastes (primary sludge and TWAS).

In addition to waste inputs, digester operating conditions are routinely monitored at the EBMUD facility, including total solid (TS, %), volatile solid (VS, %), the content of volatile fatty acids (VFA, mg/L), alkalinity (ALK, mg/L), and the VFA/ALK ratio. Operating parameters, while not truly independent of other input variables like feedstock inputs, have the potential to improve the model performance. These five operating parameters are essential factors in determining digester design and ensuring process stability. TS, representing the percentage of the dry matter (organic or inorganic), is an important attribute of digester design and operation. For example, a higher TS usually results in a smaller-sized digester and lower heating demand.[42] VS is typically regarded as a measurement of organics in the digester, serving as the basis for determining the digester organic loading rate. VFA is generated from the acidogenesis stage, comprising a class of organic acids (e.g., acetic acid, propionic acid, butyric acid), and is important to monitor because a neutral pH is optimal for methanogens.[14] Acidification (high VFA) is widely considered to be a cause of digester failure because methanogens are sensitive to low pH, which has an inhibitory effect on their growth.[14,42] As a result, ALK is needed to provide buffer capacity to neutralize VFA, thus controlling pH. A balance between VFA production and consumption by ALK (reflected by VFA/ALK ratio) can ensure a stable AD process. TS and VS in the digester were calculated based on the daily input volume of each waste stream and its specific TS or VS level obtained from the composition analysis (Figure S1). Note that VFA and ALK were not measured every day, so the missing values of VFA and ALK were imputed with the median value (raw data with non-normal distribution) and the mean value (raw data with normal distribution), respectively.

Using all feedstock and operating information provided, our model aims to predict the biogas production, thus the output (i.e., the target variable) is represented by biogas yield in standard cubic feet per

https://doi.org/10.1021/acssuschemeng.1c04612
ACS Sustainable Chem. Eng. 2021, 9, 12990−13000

minute (scfm). The input variables (i.e., features) to construct the dataset include the daily input volumes of 31 waste streams (primary sludge, TWAS, and 29 types of trucked wastes) and 5 operating parameters (TS, VS, VFA, ALK, and VFA/ALK ratio). Considering the digestion time, additional input variables were created by creating a time lag between the daily input volumes of 31 waste streams and the date of the target variable measurement, namely, 0 (no lag), 1, 3, 5, 10, 20, and 30 days, respectively. In this way, a total of 222 input variables (31 waste streams × 7 time lags + 5 operating parameters) were created. This model configuration enables the following questions to be answered: which waste stream(s) have the greatest impact on biogas yield? On what timescale do these waste inputs impact the biogas yield (measured in days)? Figure 2 displays the raw data distribution for primary inputs and output. The model data structure is presented in Figure 3.

TPOT Overview. TPOT uses the dataset as input and recommends a best-performing ML pipeline with a series of operations related to feature selection, feature preprocessing, feature construction, and ML modeling (Figure 3). GP is used to optimize the pipelines. In this case, the population consists of a set of randomly generated pipelines to be evaluated; copies of the best-performing pipelines from each iteration (known as a generation) of the optimization process are created and imposed with random changes (e.g., the addition or removal of an operation or the parameter tuning of an operation), enabling the development of new pipelines that are never explored. The worst-performing pipelines are removed from the population at the end of each generation before starting the next generation. TPOT was run with a default configuration and considered the following supervised learning regression models during the optimization process: decision trees, ensemble models (AdaBoost, XGBoost, forests of randomized trees, gradient tree boosting), cross-validated linear models (Ridge, Elastic Net, and Lasso using LARS algorithm), linear support vector regression (SVR), stochastic gradient descent (SGD), and k-nearest neighbors. Prior to the TPOT analysis, the dataset was randomly partitioned into a training set and a test set with a 0.75/0.25 train−test split ratio. The pipeline was trained on the training set and evaluated on the test set. The following TPOT parameter settings were used to generate a regression predictive model on the training set: the number of generations was 100, the size of the population was 100, it used 5-fold CV, and the scoring function (performance metric) was mean squared error (MSE). Further increasing the number of generations and the size of the population did not improve the internal CV score.

Extensive detail on the TPOT algorithm and its implementation can be found in the previously published literature.[38−41] Like other automated ML systems, TPOT minimizes user intervention by automating the process from end to end. However, like other commonly used ML models, the pipeline built by TPOT can overfit the data (i.e., the model learns the detail and noise in the training data too well so that its generalization ability to new data is negatively affected). While the pipeline optimization procedure used here employs *k*-fold CV to reduce overfitting, alternative methods such as multiobjective and Pareto optimization could further prevent TPOT models from overfitting.[38] A downside of the GP process is the high computational costs associated with exploring and optimizing over a vast space of ML pipelines and solutions. Incorporating metalearning techniques into TPOT, which inject domain knowledge in the form of preranking pipelines explored in the GP process, could potentially lower the running time without influencing performance.[43] Additionally, ML approaches (automated and conventional models) might not significantly outperform basic statistical forecasters in time-series predictions.[44] For example, Huntington et al. evaluated a "temporally binned" train−test split in their sorghum yield model using an extremely randomized trees (ExtraTrees) model and found decreased model performance.[45]

Model Evaluation. The generalization performance of the trained model was evaluated on the test dataset. The metrics used to examine the precision and accuracy of the model include the coefficient of determination (R^2) and the root mean square error (RMSE). While R^2 offers a relative measure of model fit, RMSE provides an error metric in the same unit as the target variable, making it highly interpretable. Higher R^2 and lower RMSE of the predicted-versus-observed plots for the test dataset refer to higher precision and accuracy of a model for predicting biogas yield. Additionally, the relative RMSE (%) for both training and test datasets, determined by dividing RMSE with the average value of observed data, was used to compare the prediction ability between different ML models. Model accuracy can be considered excellent for a relative RMSE smaller than 10%, good if between 10 and 20%, fair if between 20 and 30%, and poor if greater than 30%.[46]

Feature Importance and Partial Dependence (PD) Calculations. The two most widely used model interpretation techniques are feature importance and partial dependence (PD) plots. The permutation feature importance was calculated for the TPOT model using the scikit-learn library.[47] Feature importance quantifies the relative importance of a feature (i.e., an input variable) by calculating the change in prediction error after the values of a given feature are randomly permuted. A larger increase in error suggests that the model relies on this feature to predict the target variable and thus has higher importance. MSE was used as the scoring function. The mean and standard deviation of feature importance were calculated over 50 permutations of a given feature in the training dataset. PD plots and individual conditional expectation (ICE) plots were produced using the PDPbox library.[48] Using a previously fit model, PD plots visualize the predicted response as a function of the chosen feature, while the effects of all other features in the model are averaged out. ICE plots show the functional relationship between the feature and the predicted response separately for each instance. Specifically, an ICE plot produces one line per instance, while the PD plot is simply the average of all lines in the ICE plot.

■ RESULTS AND DISCUSSION

A critical first question to be answered through this analysis is whether machine learning can be effectively used to predict biogas yield based on detailed operating data. If the model achieved acceptable performance, a follow-on question is what insights the analysis provides into potential strategies for optimizing the facility studied here and other similar AD facilities.

Prediction of Biogas Yield Using the TPOT Regression Model. To address the question of whether machine learning can be utilized to predict biogas yield, the TPOT model prediction performance was assessed by comparing the predicted values with the observed (measured) biogas yield in the test dataset (Figure 4). R^2 and RMSE were used as the metrics to examine the precision and accuracy of the model, respectively. The regression model selected in the best-performing TPOT pipeline was ExtraTrees. The ExtraTrees pipeline from TPOT performed well on the test dataset with an R^2 of 0.72, which generally represents a good predictive capacity for a model trained on real-world industrial data.[25] For the sake of comparison, an alternative model using the ANN model was also developed. ANN has been the most commonly used technique to predict the performance of AD facilities (Table S1). In this case, the most popular ANN type, multilayer perceptron (MLP), was employed (Figure S2) as the baseline for comparison with the TPOT model. TPOT (R^2 = 0.72, RMSE = 247 scfm) outperforms MLP (R^2 = 0.56, RMSE = 327 scfm). Also, the relative RMSE is around 10% for TPOT and 14% for MLP in the test dataset. Although the performance of the TPOT model is higher in training than in testing, it outperforms MLP in both cases (Figure S3). Note that it is not unusual for an ML model to have better performance on the training dataset than the test dataset.[49] Notably, while many ANN parameters require manual tuning,

https://doi.org/10.1021/acssuschemeng.1c04612
ACS Sustainable Chem. Eng. 2021, 9, 12990−13000

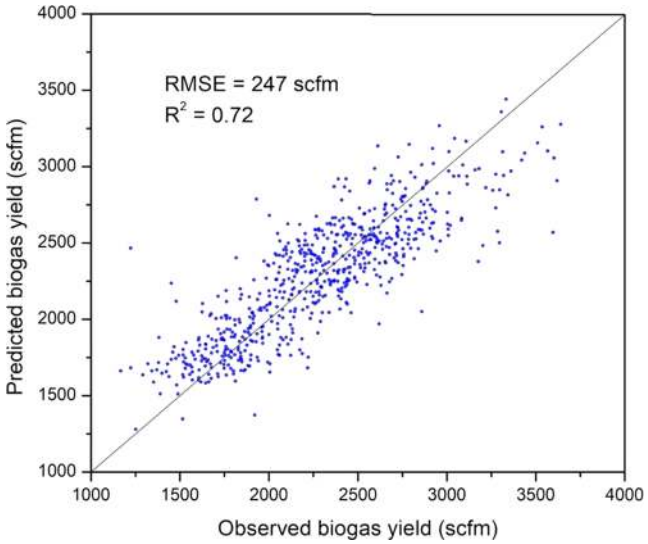


Figure 4. Comparison between observed and predicted biogas yield (scfm) for the test dataset using TPOT. ExtraTrees regression model was selected for the best prediction performance. Model prediction performance is evaluated by RMSE and R^2 and also visually revealed by the extent of data clustering around the identity line ($y = x$, in black).

TPOT automates the parameter tuning process, making it a more practical ML approach for nonexperts.

Importance of Input Variables Influencing Biogas Yield. Model interpretability is critical, whether it is developed as a research tool or to guide decision-making for facility operation. This study relies on expertise in bioprocesses and experience in the operation of AD facilities to guide the model development and interpretation such that the results answer scientifically interesting and operationally relevant questions. To investigate the influence of input variables (daily input volumes of 31 waste streams each evaluated over different time lags and 5 operating parameters) on biogas yield, permutation feature importance was generated for the TPOT model (Figure 5). The important scores provide insights into which input variables have the greatest influence on biogas yield.

The most important input variables that affect biogas yield are waste inputs from high-COD organics, FOG, dairy, protein, lactose, and sludge categories, while operating parameters did not have high importance scores (Figure 5, Table S3, also see Table S2 for the waste categories and the definitions of waste name acronyms). The less significant role of operating parameters in influencing biogas yield can be attributed to the fact that the EBMUD facility's operations have remained stable over the time period studied here. Because conditions do not generally frequently deviate from the acceptable range, biogas production is primarily dependent on the types of waste being fed into the digester. At a facility in which pH, for example, is allowed to become sufficiently acidic to inhibit microbial growth, the resulting model might show a stronger relationship between VFA and biogas yield. The top 8 waste types include Organic (H), FOG, Rendering, Poultry_Blood, Dariy_DAF, Primary_Sludge, TWAS, and Lactose. The appropriate time lag to assign for each waste type was determined, representing the duration between when it is delivered to the facility and when it has the greatest impact on biogas yield. Figure 5a shows the important results for each waste type and time lag. Poultry blood, rendering waste, and

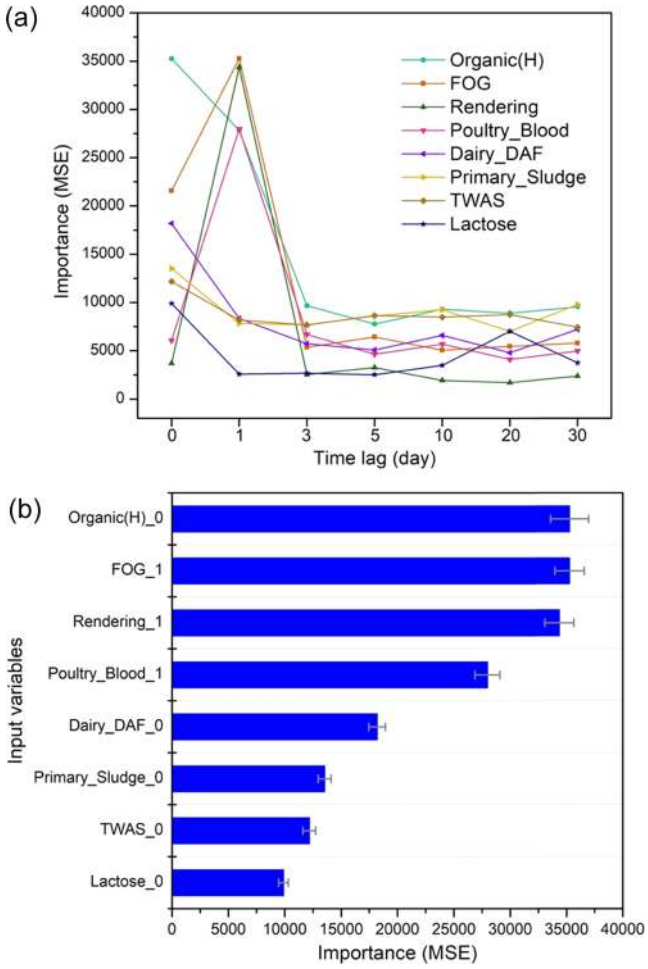
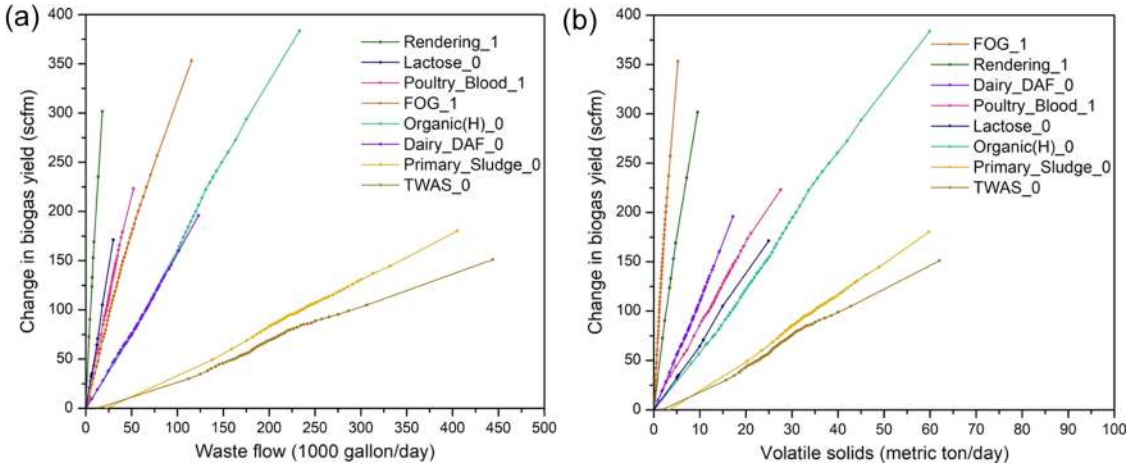


Figure 5. (a) Input variable importance of top 8 waste types in the TPOT model across different time lags (0, 1, 3, 5, 10, 20, 30 days). The important values of all input variables are compiled in Table S3. The relative importance scores are calculated using the permutation feature importance technique. An input variable that leads to a higher increase in the mode prediction error (mean squared error, MSE) upon randomly permuted is more important in the model. The name acronyms of wastes are defined in Table S2. (b) Importance ranking for the 8 most influential waste types, of which each shows the highest importance score from the time-lag variable set (data compiled from (a)). The suffix on waste names represents the specific number of lagged days. Error bars denote the standard deviations over 50 times of permuting an input variable.

FOG all appear to have the greatest impact on biogas yield with a one-day time lag, whereas all other waste types have the largest importance scores with no time lag (zero days). This agrees with the facility operators' observations that the biogas yield is obviously boosted within a matter of hours after feeding the more sugar-rich wastes into the digesters, while blood and other protein and lipid-rich wastes typically boost yields within a day.

Figure 5b compiles the best-performing version of each input variable (based on the time-lag analysis in Figure 5a) and their relative performance was compared to determine which waste types are the primary drivers of biogas yield. Organic (H), while not the largest input by volume (Figure 2a), results in the highest importance score. This is a general category used by the facility to denote waste streams with a COD greater than 20 000 mg/L. Organic (H) might include, for example,

https://doi.org/10.1021/acssuschemeng.1c04612
ACS Sustainable Chem. Eng. 2021, 9, 12990−13000



Input variables	Slope values (Figure 6a) 10^{-3} scfm/(gallon•day^{-1})	Slope values (Figure 6b) 10^{-3} scfm/(kg VS•day^{-1})
Rendering_1	17.71	33.48
Lactose_0	5.66	6.80
Poultry_Blood_1	4.48	8.47
FOG_1	3.47	76.46
Organic(H)_0	1.63	6.36
Dairy_DAF_0	1.55	11.05
Primary_Sludge_0	0.48	3.24
TWAS_0	0.38	2.73

Figure 6. Partial dependence (PD) plots depicting the quantitative relationships between (a) daily input volume (gallon/day) or (b) daily volatile solid (VS) load (metric ton/day) of the 8 most influential waste types and the resulting change in biogas yield (scfm). The table provides the numeric values of line slopes determined by linear curve fitting. The legends are arranged in descending order based on the approximated slopes of the lines. Each line is for the input variable that shows the highest importance score in the time-lag set (0, 1, 3, 5, 10, 20, 30 days) of each waste type. The suffix on waste names denotes the specific number of lagged days. The dots on the lines refer to 50 percentile points across the whole value range of each input variable; note that some percentile points can be the same thus each line does not necessarily contain 50 points. The dispersion of dots visually reveals the data point distribution of each input variable. Individual conditional expectation (ICE) plots for these 8 input variables are shown in Figure S4.

carbohydrate-rich waste from beverage processing facilities. Anecdotally, sugar-rich wastes have been noted to produce a near-immediately evident impact on microbial activity in the digesters. Followed by Organic (H) in terms of impact on biogas yield are the three protein and lipid-rich waste types, all of which are well known to be desirable supplemental inputs in wet AD.[50,51] In addition, the results indicate that, although WWTP sludge wastes (primary sludge and TWAS) have considerably higher daily input volumes than the trucked co-digested wastes (Figure 2a), their contributions to biogas yield are less significant than the wastes in the categories of high-COD organics, FOG, dairy, and protein. Compositional analysis on the co-digested wastes shows that the wastes in these categories typically have higher TS, VS, COD, and total nitrogen than others (Figure S1). Combined, the waste types with higher organic matter contents and faster digestion rates are more contributory to biogas yield.

Quantitative Relationships between the Most Influential Waste Inputs and Biogas Yield. While feature importance calculations identify the most influential input variables, it is necessary to explore the functional relationships between these important input variables and the target variable. PD (Figure 6) or ICE (Figure S4) plots enable us to isolate the quantitative effect of adjusting the values of an input variable on future prediction outcomes. PD plots visualize the marginal (average) effect of an input variable of interest on the target variable, while ICE plots demonstrate the heterogeneity or dispersion of the effect.[52] The relative impacts of the most influential waste inputs (in daily volume) on biogas yield are illustrated by the PD plots (Figure 6a). As expected, all of these waste inputs positively impact biogas yield. The slopes of PD lines reveal the increase in biogas yield on a unit basis of each waste input throughout its whole value range, following the order: Rendering > Lactose > Poultry_Blood > FOG > Organic (H) > Dairy_DAF > Primary_Sludge > TWAS. The results between feature importance and PD calculations are consistent, considering that the overall effect of each waste on biogas yield is ascribed to its specific biogas production capacity (revealed by the PD line slope) as well as its feed availability (the feed amount). High-COD organics (Organic (H)) emerge as having the highest importance score, but after examining per-unit-flow values, rendering waste, lactose, poultry blood, and FOG all result in greater increases in biogas yield.

Because different waste streams contain varying amounts of solids, AD facility operators often characterize wastes on the basis of VS. Normalizing the results in Figure 6a with respect to the daily VS load provides the results in a more usable (and

https://doi.org/10.1021/acssuschemeng.1c04612
ACS Sustainable Chem. Eng. 2021, 9, 12990−13000

generalizable) form for other facilities (see Figure 6b). The normalized results in Figure 6a suggest considerable variation in the impact on biogas yield on a mass of VS basis. The ordering from greatest impact to least impact on biogas yield also shifts; FOG, rendering waste, dairy waste after dissolved air flotation (DAF) treatment, poultry blood, and lactose all demonstrate higher capacity for biogas production on a per-mass VS basis. Together, the feature importance ranking and PD plots provide useful insights into how different inputs impact biogas yield that can be leveraged for plant operators to manage and operate the AcoD facility. The numeric slopes in Figure 6 can be used to predict the impact of any given waste stream on biogas yield and thus inform tipping fees and priorities for which wastes to accept. The fact that some wastes impact biogas production more rapidly than others indicates that operators could attempt to manage incoming waste such that biogas production is as stable as possible, thus minimizing the need for flaring excess biogas. However, anecdotal evidence from the EBMUD facility and other local AcoD and AD facilities suggests that the greatest challenge in timing is the reduction in waste hauling during weekends. Facilities could make use of a combination of on-site liquid storage and slower-degrading waste types to partially mitigate this problem.

Opportunity to Predict Biogas Trace Compounds Using ML Techniques. Raw biogas typically contains trace undesirable compounds (contaminants), of which the amounts are highly dependent on the origin of organic sources. H_2O, H_2S, and NH_3 are commonly present, while siloxanes and halogenated hydrocarbons can also be present.[15−18] These contaminants can cause problems to the equipment (e.g., engines, pipelines, valve fittings) for biogas utilization. For example, H_2O, H_2S, NH_3, and halogenated hydrocarbons cause corrosion and SiO_2 formed from siloxanes causes abrasion on gas motor surfaces.[15,18,53] Biogas cleaning is thus normally conducted for all commonly used biogas applications. The prediction of these trace compounds, especially by employing ML techniques, can be informative if it enables operators to select waste streams that minimize their formation or simply optimize gas cleaning investments to manage the contaminants.

H_2S is the most influential trace compound to be treated in biogas for energy applications, considering the risk of sulfide stress cracking (embrittlement)[54] and the ubiquitous presence of sulfur in biological substrates (particularly those with high protein levels).[17] The methods for biogas cleaning differ according to the required quality demands for the contaminants in specific end uses of biogas. H_2S is considered as the main component for assessing the biogas quality in heating boilers and internal combustion engines, where its concentration should be lower than 1000 ppm.[17,54] However, for biogas use as vehicle fuel or natural gas, the requirements for H_2S become much stricter (e.g., <120 ppm in the U.S.) and maximum allowable concentrations vary by country.[17,54] H_2S is typically removed during digestion through a precipitation reaction with metal ions (Fe^{2+}, Fe^{3+}) or after digestion by absorption (e.g., using water for scrubbing or metal oxide), adsorption (e.g., on activated carbon), biological filters, and membrane separation.[15,16]

The EBMUD facility studied here removes H_2S by adding $FeCl_3$ to the digester, which reduces the H_2S concentration to less than 300 ppm and thus satisfies the requirement for combustion onsite. This also means that H_2S concentrations are measured in the biogas after much of the sulfur has been removed by $FeCl_3$. Furthermore, the facility monitors H_2S on an approximately weekly basis, so the dataset for H_2S is fairly limited in its granularity. As one might expect, using ML techniques to predict H_2S did not prove to be useful because the H_2S output is most strongly correlated with the dose of $FeCl_3$. Since the actual H_2S content produced from the input wastes is not measured, the minimum required amount of $FeCl_3$ cannot be accurately estimated with our modeling approach. That said, ML techniques do have the potential to predict concentrations of trace compounds in such a way that could be useful in applications requiring high biogas quality, where frequent measurement of biogas composition is conducted. First, the in situ H_2S removal methods during digestion are less efficient in achieving the required level of H_2S (and other contaminants of concern) for transport fuel or pipeline quality, where post-treatment methods after digestion are needed.[16] Also, very little work has been done to explore the potential of ML techniques for predicting these biogas contaminants. To our knowledge, only Strik et al. applied ANN (MATLAB Neural Network Toolbox) to predict H_2S and NH_3 concentrations in an experimental AD setup a decade ago.[55] This topic presents an opportunity to use computational methods to improve the efficiency of the biogas industry and gain insights into the complex dynamics of the microbial communities that break down mixed organic waste. With richer datasets from facility operations and even lab-scale experiments, researchers could harness (automated) ML techniques to develop better predictions of the concentrations of trace compounds and biogas quality for biogas utilization systems that employ the post-treatment biogas cleaning methods (e.g., for biogas upgrading technologies). This could also help with the management of air pollutant emissions, as NOx emissions from flaring have been shown to have a relationship with NH_3 concentrations in raw biogas.[56] Tracking NH_3 could also serve as a proxy for nutrient loading in the effluent, which could enable facilities to develop strategies that minimize water quality impacts of using high-nitrogen wastes.

■ CONCLUSIONS

Anaerobic digesters have an important role to play in diverting organic waste from landfills and producing renewable energy.[5] AcoD technology in particular is being embraced as a pragmatic strategy for increasing biogas production, leveraging existing infrastructure while overcoming the challenges associated with the substrate properties and system stability in the single-substrate AD process. The key to the success of AcoD processes is system optimization and the ability to manage a diverse set of incoming waste streams. ML models, which are ideally suited to capturing the behavior of systems that are too complex to model mechanistically, can improve researchers' and operators' understanding of the AcoD process and its performance as a function of varying feed substrates or operating conditions. Our work contributes to a growing field of biogas production prediction using ML techniques and the use of TPOT with a substantially larger dataset than any previously documented (based on 8-year industrial-scale operations) makes this study unique. Our work provides evidence for the robust predictive power of TPOT applied to AcoD modeling, as demonstrated by its superior prediction performance compared to the basic ANN model (MLP). The combination of feature importance and PD analyses allowed us to differentiate between waste streams that have a larger impact

https://doi.org/10.1021/acssuschemeng.1c04612
ACS Sustainable Chem. Eng. 2021, 9, 12990−13000

because of greater incoming volumes and waste streams that have a greater impact per unit of waste input to the digester. Our approach of testing different time lags also provided insights into how different wastes are broken down once loaded into the digester. By developing and improving predictive models for AcoD performance, we hope to enable more efficient facility operation, a better understanding of how microbial communities respond to different substrates and operating conditions, and ultimately a more sustainable organic waste valorization industry.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acssuschemeng.1c04612.

> Including the literature summary of ML applications in predicting biogas production from anaerobic digestion (Table S1), brief description on the waste streams fed into the AcoD facility (Table S2), typical composition analysis data on the waste categories (Figure S1), biogas yield prediction using ANN-MLP (Figure S2), relative RMSE results for the training and test datasets using TPOT and ANN-MLP (Figure S3), ranked permutation importance values for all input variables (Table S3), and ICE plots for the 8 most influential waste types (Figure S4) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Corinne D. Scown − *Energy & Biosciences Institute, University of California, Berkeley, Berkeley, California 94704, United States; Life-Cycle, Economics, and Agronomy Division, Joint BioEnergy Institute, Emeryville, California 94608, United States; Biological Systems & Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, United States; Energy Analysis & Environmental Impacts Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, United States;* orcid.org/0000-0003-2078-1126; Phone: (510) 486-4507; Email: cdscown@lbl.gov

Authors

Yan Wang − *Energy & Biosciences Institute, University of California, Berkeley, Berkeley, California 94704, United States; Life-Cycle, Economics, and Agronomy Division, Joint BioEnergy Institute, Emeryville, California 94608, United States; Biological Systems & Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, United States;* orcid.org/0000-0002-9147-1136

Tyler Huntington − *Life-Cycle, Economics, and Agronomy Division, Joint BioEnergy Institute, Emeryville, California 94608, United States; Biological Systems & Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acssuschemeng.1c04612

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors would like to acknowledge Michael Hyatt and John Hake from the East Bay Municipal Utility District for providing access to their facility data and giving extensive feedback on the design and results of our study. This study was supported by the U.S. Department of Energy, Energy Efficiency and Renewable Energy, Bioenergy Technologies Office. This work was also enabled by tools and resources provided by 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.

■ REFERENCES

(1) U.S. Environmental Protection Agency. AgSTAR Data and Trends. https://www.epa.gov/agstar/agstar-data-and-trends (accessed Feb 19, 2021).

(2) American Biogas Council. How Many Operational Anaerobic Digesters Are There in the U.S.? https://americanbiogascouncil.org/resources/faqs/ (accessed Feb 19, 2021).

(3) Water Environment Federation. Water Resource Recovery Facilities with Operating Anaerobic Digestion. http://www.resourcerecoverydata.org/biogasdata.php (accessed Feb 19, 2021).

(4) Nordahl, S. L.; Devkota, J. P.; Amirebrahimi, J.; Smith, S. J.; Breunig, H. M.; Preble, C. V.; Satchwell, A. J.; Jin, L.; Brown, N. J.; Kirchstetter, T. W.; Scown, C. D. Life-cycle greenhouse gas emissions and human health trade-offs of organic waste management strategies. *Environ. Sci. Technol.* **2020**, *54*, 9200−9209.

(5) Satchwell, A. J.; Scown, C. D.; Smith, S. J.; Amirebrahimi, J.; Jin, L.; Kirchstetter, T. W.; Brown, N. J.; Preble, C. V. Accelerating the deployment of anaerobic digestion to meet zero waste goals. *Environ. Sci. Technol.* **2018**, *52*, 13663−13669.

(6) Hagos, K.; Zong, J.; Li, D.; Liu, C.; Lu, X. Anaerobic co-digestion process for biogas production: Progress, challenges and perspectives. *Renewable Sustainable Energy Rev.* **2017**, *76*, 1485−1496.

(7) Neshat, S. A.; Mohammadi, M.; Najafpour, G. D.; Lahijani, P. Anaerobic co-digestion of animal manures and lignocellulosic residues as a potent approach for sustainable biogas production. *Renewable Sustainable Energy Rev.* **2017**, *79*, 308−322.

(8) Zhang, C.; Xiao, G.; Peng, L.; Su, H.; Tan, T. The anaerobic co-digestion of food waste and cattle manure. *Bioresour. Technol.* **2013**, *129*, 170−176.

(9) Mostafa Imeni, S.; Pelaz, L.; Corchado-Lopo, C.; Maria Busquets, A.; Ponsá, S.; Colón, J. Techno-economic assessment of anaerobic co-digestion of livestock manure and cheese whey (Cow, Goat & Sheep) at small to medium dairy farms. *Bioresour. Technol.* **2019**, *291*, No. 121872.

(10) Shah, F. A.; Mahmood, Q.; Rashid, N.; Pervez, A.; Raja, I. A.; Shah, M. M. Co-digestion, pretreatment and digester design for enhanced methanogenesis. *Renewable Sustainable Energy Rev.* **2015**, *42*, 627−642.

(11) Siddique, M. N. I.; Wahid, Z. A. Achievements and perspectives of anaerobic co-digestion: A review. *J. Cleaner Prod.* **2018**, *194*, 359−371.

(12) Shen, Y.; Linville, J. L.; Urgun-Demirtas, M.; Mintz, M. M.; Snyder, S. W. An overview of biogas production and utilization at full-scale wastewater treatment plants (WWTPs) in the United States: Challenges and opportunities towards energy-neutral WWTPs. *Renewable Sustainable Energy Rev.* **2015**, *50*, 346−362.

(13) Pennington, M. *Anaerobic Digestion Facilities Processing Food Waste in the United States (2017 & 2018)*; Survey Results EPA/903/S-21/001; U.S. Environmental Protection Agency (EPA), 2021.

(14) Bajpai, P. *Anaerobic Technology in Pulp and Paper Industry*, 1st ed.; SpringerBriefs in Applied Sciences and Technology; Springer: Singapore, 2017; pp 7−12. DOI: 10.1007/978-981-10-4130-3.

(15) Chaemchuen, S.; Zhou, K.; Verpoort, F. From biogas to biofuel: materials used for biogas cleaning to biomethane. *ChemBioEng Rev.* **2016**, *3*, 250−265.
(16) Ryckebosch, E.; Drouillon, M.; Vervaeren, H. Techniques for transformation of biogas to biomethane. *Biomass Bioenergy* **2011**, *35*, 1633−1645.
(17) Rasi, S.; Läntelä, J.; Rintala, J. Trace compounds affecting biogas energy utilisation—A review. *Energy Convers. Manage.* **2011**, *52*, 3369−3375.
(18) Muñoz, R.; Meier, L.; Diaz, I.; Jeison, D. A review on the state-of-the-art of physical/chemical and biological technologies for biogas upgrading. *Rev. Environ. Sci. Biotechnol.* **2015**, *14*, 727−759.
(19) U.S. Environmental Protection Agency. Renewable Fuel Standard Program. https://www.epa.gov/renewable-fuel-standard-program (accessed Feb 19, 2021).
(20) Nguyen, D.; Nitayavardhana, S.; Sawatdeenarunat, C.; Surendra, K. C.; Khanal, S. K. Biogas Production by Anaerobic Digestion: Status and Perspectives. In *Biofuels: Alternative Feedstocks and Conversion Processes for the Production of Liquid and Gaseous Biofuels*; Elsevier, 2019; pp 763−778. DOI: 10.1016/B978-0-12-816856-1.00031-2.
(21) Kapoor, R.; Ghosh, P.; Tyagi, B.; Vijay, V. K.; Vijay, V.; Thakur, I. S.; Kamyab, H.; Nguyen, D. D.; Kumar, A. Advances in biogas valorization and utilization systems: A comprehensive review. *J. Cleaner Prod.* **2020**, *273*, No. 123052.
(22) Wang, L.; Long, F.; Liao, W.; Liu, H. Prediction of anaerobic digestion performance and identification of critical operational parameters using machine learning algorithms. *Bioresour. Technol.* **2020**, *298*, No. 122495.
(23) De Clercq, D.; Wen, Z.; Fei, F.; Caicedo, L.; Yuan, K.; Shang, R. Interpretable machine learning for predicting biomethane production in industrial-scale anaerobic co-digestion. *Sci. Total Environ.* **2020**, *712*, No. 134574.
(24) Batstone, D. J.; Keller, J.; Angelidaki, I.; Kalyuzhnyi, S. V.; Pavlostathis, S. G.; Rozzi, A.; Sanders, W. T. M.; Siegrist, H.; Vavilin, V. A. The IWA Anaerobic Digestion Model No 1 (ADM1). *Water. Sci. Technol.* **2002**, *45*, 65−73.
(25) Batstone, D. J.; Puyol, D.; Flores-Alsina, X.; Rodríguez, J. Mathematical modelling of anaerobic digestion processes: applications and future needs. *Rev. Environ. Sci. Biotechnol.* **2015**, *14*, 595−613.
(26) Dandikas, V.; Heuwinkel, H.; Lichti, F.; Drewes, J. E.; Koch, K. Predicting methane yield by linear regression models: A validation study for grassland biomass. *Bioresour. Technol.* **2018**, *265*, 372−379.
(27) Sakiewicz, P.; Piotrowski, K.; Ober, J.; Karwot, J. Innovative artificial neural network approach for integrated biogas—wastewater treatment system modelling: Effect of plant operating parameters on process intensification. *Renewable Sustainable Energy Rev.* **2020**, *124*, No. 109784.
(28) Xu, F.; Wang, Z.-W.; Li, Y. Predicting the methane yield of lignocellulosic biomass in mesophilic solid-state anaerobic digestion based on feedstock characteristics and process parameters. *Bioresour. Technol.* **2014**, *173*, 168−176.
(29) Abu Qdais, H.; Bani Hani, K.; Shatnawi, N. Modeling and optimization of biogas production from a waste digester using artificial neural network and genetic algorithm. *Resour., Conserv. Recycl.* **2010**, *54*, 359−363.
(30) Mahanty, B.; Zafar, M.; Park, H.-S. Characterization of co-digestion of industrial sludges for biogas production by artificial neural network and statistical regression models. *Environ. Technol.* **2013**, *34*, 2145−2153.
(31) Yetilmezsoy, K.; Turkdogan, F. I.; Temizel, I.; Gunay, A. Development of Ann-Based Models to Predict Biogas and Methane Productions in Anaerobic Treatment of Molasses Wastewater. *Int. J. Green Energy* **2013**, *10*, 885−907.
(32) Jaroenpoj, S.; Yu, J.; Ness, J. Development of artificial neural network models for biogas production from co-digestion of leachate and pineapple peel. *Glob. Environ. Eng.* **2015**, *1*, 42−47.
(33) Flores-Asis, R.; Méndez-Contreras, J. M.; Juárez-Martínez, U.; Alvarado-Lassman, A.; Villanueva-Vásquez, D.; Aguilar-Lasserre, A. A. Use of artificial neuronal networks for prediction of the control parameters in the process of anaerobic digestion with thermal pretreatment. *J. Environ. Sci. Health, Part A* **2018**, *53*, 883−890.
(34) Ghatak, M. D.; Ghatak, A. Artificial neural network model to predict behavior of biogas production curve from mixed lignocellulosic co-substrates. *Fuel* **2018**, *232*, 178−189.
(35) Regoa, A. S.; Leiteb, S. A.; Leiteb, B. S.; Grilloc, A. V.; Santosa, B. F. Artificial Neural Network Modelling for Biogas Production in Biodigesters. *Chem. Eng. Trans* **2019**, *74*, 25−30.
(36) Beltramo, T.; Klocke, M.; Hitzmann, B. Prediction of the biogas production using GA and ACO input features selection method for ANN model. *Inf. Process. Agric.* **2019**, *6*, 349−356.
(37) Gonçalves Neto, J.; Vidal Ozorio, L.; Campos de Abreu, T. C.; Ferreira dos Santos, B.; Pradelle, F. Modeling of biogas production from food, fruits and vegetables wastes using artificial neural network (ANN). *Fuel* **2021**, *285*, No. 119081.
(38) Olson, R. S.; Urbanowicz, R. J.; Andrews, P. C.; Lavender, N. A.; Kidd, L. C.; Moore, J. H. Automating Biomedical Data Science through Tree-Based Pipeline Optimization. In *Applications of Evolutionary Computation*; Squillero, G.; Burelli, P., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, 2016; Vol. *9597*, pp 123−137. DOI: 10.1007/978-3-319-31204-0_9.
(39) Olson, R. S.; Moore, J. H. TPOT: A Tree-Based Pipeline Optimization Tool for Automating Machine Learning. In *Automated Machine Learning: Methods, Systems, Challenges*; Hutter, F.; Kotthoff, L.; Vanschoren, J., Eds.; The Springer Series on Challenges in Machine Learning; Springer International Publishing: Cham, 2019; pp 151−160. DOI: 10.1007/978-3-030-05318-5_8.
(40) Olson, R. S.; Bartley, N.; Urbanowicz, R. J.; Moore, J. H. In *Evaluation of a Tree-Based Pipeline Optimization Tool for Automating Data Science*, Proceedings of the 2016 on Genetic and Evolutionary Computation Conference - GECCO '16; Friedrich, T., Eds.; ACM Press: New York, USA, 2016; pp 485−492.
(41) EpistasisLab. Tree-Based Pipeline Optimization Tool. http://epistasislab.github.io/tpot/ (accessed Feb 21, 2021).
(42) Meegoda, J. N.; Li, B.; Patel, K.; Wang, L. B. A review of the processes, parameters, and optimization of anaerobic digestion. *Int. J. Environ. Res. Public Health* **2018**, *15*, No. 2224.
(43) Laadan, D.; Vainshtein, R.; Curiel, Y.; Katz, G.; Rokach, L. In *MetaTPOT: Enhancing A Tree-Based Pipeline Optimization Tool Using Meta-Learning*, Proceedings of the 29th ACM International Conference on Information & Knowledge Management; ACM: New York, NY, USA, 2020; pp 2097−2100.
(44) Paldino, G. M.; De Stefani, J.; De Caro, F.; Bontempi, G. Does AutoML Outperform Naive Forecasting? *Eng. Proc.* **2021**, *5*, No. 36.
(45) Huntington, T.; Cui, X.; Mishra, U.; Scown, C. D. Machine learning to predict biomass sorghum yields under future climate scenarios. *Biofuels, Bioprod. Biorefin.* **2020**, *14*, 566−577.
(46) Despotovic, M.; Nedic, V.; Despotovic, D.; Cvetanovic, S. Evaluation of empirical models for predicting monthly mean horizontal diffuse solar radiation. *Renewable Sustainable Energy Rev.* **2016**, *56*, 246−260.
(47) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(48) PDPbox. https://pdpbox.readthedocs.io/en/latest/ (accessed Mar 28, 2021).
(49) Wang, L.; Zhou, X.; Zhu, X.; Dong, Z.; Guo, W. Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. *Crop J.* **2016**, *4*, 212−219.
(50) Salama, E.-S.; Saha, S.; Kurade, M. B.; Dev, S.; Chang, S. W.; Jeon, B.-H. Recent trends in anaerobic co-digestion: Fat, oil, and grease (FOG) for enhanced biomethanation. *Prog. Energy Combust. Sci.* **2019**, *70*, 22−42.

(51) Bayr, S.; Rantanen, M.; Kaparaju, P.; Rintala, J. Mesophilic and thermophilic anaerobic co-digestion of rendering plant and slaughterhouse wastes. *Bioresour. Technol.* **2012**, *104*, 28−36.
(52) Molnar, C. *Interpretable Machine Learning*; Lulu. com, 2020.
(53) Kapoor, R.; Ghosh, P.; Kumar, M.; Vijay, V. K. Evaluation of biogas upgrading technologies and future perspectives: a review. *Environ. Sci. Pollut. Res.* **2019**, *26*, 11631−11661.
(54) Chen, X. Y.; Vinh-Thang, H.; Ramirez, A. A.; Rodrigue, D.; Kaliaguine, S. Membrane gas separation technologies for biogas upgrading. *RSC Adv.* **2015**, *5*, 24399−24448.
(55) Strik, D. P.; Domnanovich, A. M.; Zani, L.; Braun, R.; Holubar, P. Prediction of trace compounds in biogas from anaerobic digestion using the MATLAB Neural Network Toolbox. *Environ. Modell. Software* **2005**, *20*, 803−810.
(56) Preble, C. V.; Chen, S. S.; Hotchi, T.; Sohn, M. D.; Maddalena, R. L.; Russell, M. L.; Brown, N. J.; Scown, C. D.; Kirchstetter, T. W. Air pollutant emission rates for dry anaerobic digestion and composting of organic municipal solid waste. *Environ. Sci. Technol.* **2020**, *54*, 16097−16107.

https://doi.org/10.1021/acssuschemeng.1c04612
ACS Sustainable Chem. Eng. 2021, 9, 12990−13000