Biochemical Engineering Journal 186 (2022) 108578

Contents lists available at ScienceDirect

Biochemical Engineering Journal

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

A comparative evaluation of machine learning algorithms for predicting syngas fermentation outcomes

Garrett W. Roell [a], Ashik Sathish [b,c], Ni Wan [a], Qianshun Cheng [d], Zhiyou Wen [b,c], Yinjie J. Tang [a,*], Forrest Sheng Bao [e,*]

[a] Washington University in St. Louis, Department of Energy, Environmental, and Chemical Engineering, St. Louis, MO 63130, USA
[b] Iowa State University, Department of Agricultural and Biosystems Engineering, Ames, IA 50011, USA
[c] Iowa State University, Department of Food Science and Human Nutrition, Ames, IA 50011, USA
[d] University of Illinois at Chicago, Department of Mathematics, Statistics and Computer Science, Chicago, IL, 60607, USA
[e] Iowa State University, Department of Computer Science, Ames, IA 50011, USA

ARTICLE INFO

Keywords:
Clostridium carboxidivorans
Neural network
Random forest
Support vector machine
Data transformation
Model predictive control

ABSTRACT

Clostridium carboxidivorans can use syngas to produce acids and alcohols. However, simulating gas fermentation dynamics remains challenging. This study employed data transformation and machine learning (ML) approaches to predict syngas fermentation behavior. Syngas composition and fermentative metabolite concentrations (features) were paired with the production rates (prediction targets) of acetate, ethanol, butyrate, and butanol at each time point. This transformation avoided the use of time as a feature. Data augmentation by polynomial smoothing of experimental measurements was used to create a database for supervised learning of 836 rate instances from 10 gas compositions. Seven families of ML algorithms were compared, including neural networks, support vector machines, random forests, elastic nets, lasso regressors, k-nearest neighbors, and Bayesian ridge regressors. These algorithms predicted production rates for training data with Pearson correlation coefficients ($R^2 > 0.9$), but they showed poorer performance for predicting unseen test data. Among the algorithms, random forests and support vector machines produced the most accurate predictions for the test data, which could regenerate product concentration curves ($R^2 \approx 0.85$). In contrast, neural networks had a higher risk of overfitting. Additionally, ML-based feature importance analysis highlighted the significant impacts of CO and H_2 on alcohol production, which offersguidance for model predictive control. Together, these findings can help direct future applications of ML algorithms to complex bioprocesses with limited data.

1. Introduction

Biofuels and chemicals can be produced from lignocellulosic biomass via a thermochemical decomposition to syngas (CO, CO_2, and H_2) followed by microbial syngas fermentation. Compared to the Fischer-Tropsch process, syngas fermentation has low capital and environmental costs [10,15]. Modeling approaches are necessary to optimize syngas fermentation. Fermentation engineers favor model predictive control (MPC) because its control speed and process dynamics are better than traditional proportional-integral-derivative controls [23]. However, MPC requires quality kinetic models to predict growth rates. These models are typically complex since syngas fermentation performance is affected by gas-to-liquid mass transfer, cell biosynthesis capability, syngas composition, gas flow rate, product inhibitions, and metabolic shifting between acetogenic and solventogenic stages [2,17,20,26]. Extensive experiments are required to calibrate a model's process parameters (e.g., gas solubility and consumptions). It is challenging to incorporate all influential factors and complex reaction mechanisms into kinetic-based fermentation models; therefore, semi-empirical power--law models have been developed. These models describe the nonlinear effects of syngas components on production rates, but their equations are often stiff when simulating new conditions [25].

Machine learning (ML) has emerged as a viable black-box method for discovering novel relationships in multivariate systems. ML can predict complex cellular processes without mechanistic equations that explicitly link input and output variables [13]. Some examples of machine learning models applied to biological processes include an artificial neural network that was used to optimize simultaneous hydrolysis and

* Corresponding authors.
E-mail addresses: yinjie.tang@wustl.edu (Y.J. Tang), forrest.bao@gmail.com (F.S. Bao).

https://doi.org/10.1016/j.bej.2022.108578
Received 22 April 2022; Received in revised form 15 July 2022; Accepted 4 August 2022
Available online 5 August 2022

fermentation conditions [9], a genetic algorithm that was used for model predictive control of fed-batch yeast culture [16], and a support vector machine that was used to optimize lysine fermentation [31,32]. Additionally, deep reinforcement learning has been used to control microbial co-cultures [27], a k-nearest neighbors (kNN) model was developed to predict multi-step anaerobic digester efficiency [31,32], and an artificial neural network was applied to accurately estimate the flow rate of biogas from agricultural substrates [1]. Also, ML based on yeast morphology and ultrasonic measurements was used to predict and control alcohol fermentations [3,12], and hybrid mechanistic and ML models were shown to improve predictions of fermentation pH [14] and to refine kinetic parameters in industrial-scale fermentation processes [24].

Still, the application of ML models to cases with limited training data is a recurring challenge. Two approaches that have had success in this domain are transfer learning [21] and *in silico* data augmentation [28]. Moreover, various ML studies have investigated the change rate of dynamic bioprocesses [6,7,19], and deep learning techniques, such as long short-term memory (LSTM), have been used to decipher complex systems [11]. In this research, ML was used to determine the production rate of four products from a non-acetone-butanol-ethanol fermentation *Clostridium* species [29,30]. Syngas fermentation and product analyses are costly, and therefore we used a limited experimental data set for ML training and testing. This study addressed three questions: how can "small experimental data" be used to support quality ML predictions of dynamic syngas fermentations? Which ML algorithm is the best for syngas fermentation predictions? How can our analysis guide future ML projects?

This study employed four steps to facilitate ML analysis of dynamic syngas fermentations. First, a database was compiled, and the fermentation curves were smoothed via a polynomial function which introduced additional training and testing data via interpolation [22]. Second, data transformation was performed by converting time-course product concentrations to ML features and production rates. This step removed the temporal dimension and augmented the training data. Third, seven families of ML algorithms were tested, including neural networks (NNs), support vector machines (SVMs), random forests (RFs), elastic nets (ENs), lasso regressors (LAs), k-nearest neighbors regressors (kNNs), and Bayesian ridge regressors (BRs) [39]. Fourth, the rate-based ML models were used to predict time-course concentration curves under specified fermentation conditions. In general, small experimental data and dynamic systems limit ML applicability, but proper data transformation and ML algorithms were able to effectively improve model predictions [8].

2. Materials and methods

2.1. Strain, medium, bioreactor cultures, and product analysis

All fermentations conducted in this study used *Clostridium carboxidivorans* P7 (ATCC-BAA624), and seed cultures were prepared following a method described in a previous report [38]. This method involved growing seed cultures in anaerobic serum vials with P7 medium, vitamins, cysteine-sulfide reducing agent, and CO gas. The syngas fermentation runs in this study were done following a procedure described in a previous paper [30]. In brief, the seed culture was inoculated into an Applikon Mini Bioreactor at 10 % of the total fermentation volume. The reactors initially contained P7 medium with 4 g/L glucose to shorten syngas fermentation time. The cultures were grown at 37 °C with 500 rpm agitation. Syngas flow began after glucose in the medium was

Table 1
Summary of syngas fermentation experimental outcomes. The columns on the right side of the table display the concentration of acetate, ethanol, butyrate, and butanol in mM units at the end of fermentation. Condition numbers with (*) 's indicate data from [30], while data from the other conditions is from [29].

No	Gas Comp. % $CO/CO_2/H_2/N_2$	Flow Rate (mL/min)	Trial	Number of Time Points	Final Concentration (mM)			
					Acetate	Ethanol	Butyrate	Butanol
1	20/15/5/60	20	1	11	20.0	32.7	4.6	16.1
			2	11	24.1	29.7	5.8	25.5
2	20/15/25/40	20	1	10	32.8	78.4	7.0	24.5
			2	10	27.8	83.0	5.9	26.1
3	20/35/25/20	20	1	11	44.2	76.1	5.9	15.3
			2	11	44.1	108.9	3.5	10.0
4	40/15/25/20	20	1	11	17.8	63.9	1.4	3.7
			2	11	25.4	57.6	3.1	11.9
5	20/15/5/60	5	1	7	43.1	9.2	10.1	3.6
6*	50/37.5/12.5/0	1	1	12	30.1	36.7	8.2	14.8
			2	12	49.0	13.7	12.5	4.9
7*	50/37.5/12.5/0	10	1	10	26.3	74.0	2.8	11.2
			2	10	27.1	61.4	2.2	7.2
8	40/15/25/20	5	1	12	38.3	28.0	6.8	6.6
9	20/20/0/60	20	1	7	50.5	38.4	12.3	10.7
10*	50/37.5/12.5/0	20	1	10	30.3	21.9	5.1	4.1
			2	10	32.2	68.6	3.6	7.6

exhausted. Syngas composition and gas flow rate varied based on the experimental conditions described in Table 1. Artificial oxygen-free syngas was provided from gas cylinders, and a system of Alicat mass flow controllers mixed the gases. A built-in Applikon Bioreactor condenser and cold trap placed in an ice bath prevented the loss of the vaporized alcohols in the exhaust gas. The captured alcohols were accounted for in the total alcohol production measurements. The growth of the P7 strain was measured using optical density of the culture (OD_{660}) that was correlated with the dry cell weight. Syngas fermentation products (acetate, butyrate, ethanol, and butanol) were determined using a gas chromatography (GC) system equipped with a flame ionization detector (FID).

2.2. Fermentation data collection and transformation

Ten different syngas fermentation conditions were used to test the productivity of *Clostridium carboxidivorans* P7 (Table 1). The data is sourced from two previous reports: Conditions 1–5, 8, and 9 [29]; Conditions 6, 7, and 10 [30]. Compositions 5, 8, and 9 each had one trial, while the others had two trials. In sum, this work used extracellular product concentration data from 17 fermentation runs (data/experimental_data.csv in this project's GitHub repository: https://github.com/garrettroell/SyngasMachineLearning).

The time-course concentration data was preprocessed to smooth concentration curves, to increase the number of time points for training and testing, and to calculate production rates at each time point. Since the time intervals between experimental measurements were not uniform, the time-course data was smoothed and interpolated to get concentrations every 0.1 days using a second-order Savitzky–Golay filter [22]. A parity plot comparing measured and smoothed data can be seen in Supplemental Fig. 1. Points from the first 24 hours of fermentation were not considered for training or testing data since cell growth in this period was driven by glucose and not syngas [30]. From the beginning of day two, the rate of concentration change for a given metabolite (unit: mM/L/day) was calculated by subtracting its current concentration from its concentration 0.1 days earlier and dividing by 0.1. After these transformations, the number of pairs of input and output instances increased from 176 to 836. Conditions 1–7 were used as training data, and "unseen" conditions 8–10 were used as test data. Data leakage was avoided by preventing data from a single condition from being in both training and testing data sets.

2.3. ML model construction and parameter estimations

We formulate our problem as a regression problem and thus the ML models are regressors. Seven families of them , namely, neural networks (NNs), support vector machines (SVMs), random forests (RFs), elastic nets (ENs), lasso regressors (LAs), k-nearest neighbors regressors (kNNs), and Bayesian ridge regressors (BRs), were used to make predictions for the rate of production of ethanol, acetate, butanol, and butyrate. The input features for these models were the gas condition (flow rate in mL/min, % CO, % CO_2, % H_2, and % N_2) and the current concentration of the measured extracellular metabolites (ethanol, acetate, butanol, and butyrate in mM, and biomass in g/L). Twenty-eight ML models that each had ten features were trained (4 outputs × 7 types of regressors) (Fig. 1).

Well-tuned hyperparameters are important for ensuring strong performance from machine learning models. Following the common practice, we did cross-validation-based grid searches to find the optimal set of hyperparameters for each of the seven machine learning approaches. For each run of the cross-validation-based grid search, a model's performance was evaluated based on its predictions on a subset of training data that was held out as a validation data set. Any two runs of the cross-validation were independent and hence there was no information leakage. The hyperparameters used for each model were:

- For NNs, the number of layers ranged from 1 to 4 with a step of 1, the number of neurons in any layer ranged from 20 to 100 with a step of 20, the activation functions included tanh (tangent hyperbolic) and ReLU (rectified linear unit), and the maximum number of iterations was set to 5000.
- For SVMs, the kernel was RBF (radial basis function), and the configurable parameters, C, epsilon, and gamma, had values that ranged from 10^{-5} to 10^{5} logarithmically with a step of 10.
- For RFs, the number of estimators ranged from 10 to 200 with a step of 10, the maximal depth ranged from 2 to 40 with a step of 2, and the maximum number of samples per tree ranged from 5 % to 50 % of total samples.

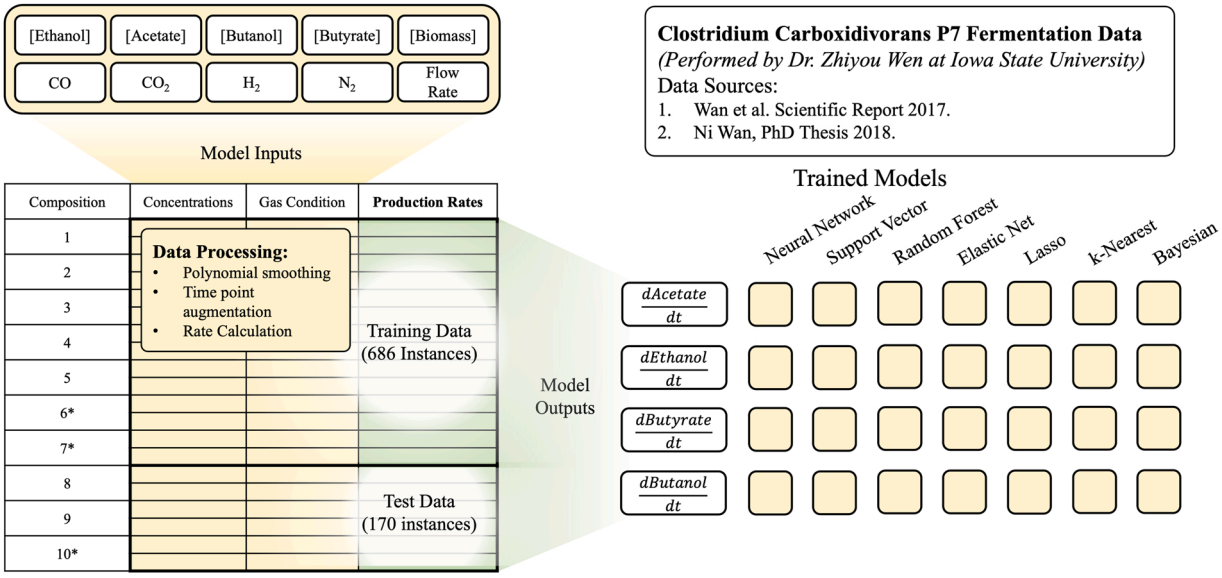

Fig. 1. Outline of machine learning methods. Data from 10 gas conditions (17 trials) is split into training and testing data. The model inputs are shown on top (brackets indicate concentrations), and the model outputs are on the right. Each box on the right represents a model trained using a specific ML algorithm for a specific output. Condition numbers with (*) 's indicate data from [30], while data from the other conditions is from [29].

- For ENs, the alpha value ranged from 10^{-10} to 10^{10} logarithmically with a step of 10, and the L1 ratio ranged from 10 % to 100 % with a step of 10 %.
- For LAs, the alpha value ranged from 10^{-5} to 10^{5} logarithmically with a step of 10.
- For kNNs, the number of neighbors ranged from 1 to 30 with a step of 1, the size of the leaves ranged from 5 to 50 with a step of 5, the nearest neighbors were computed using Ball Tree and KD Tree, and samples were weighted based on distance.
- For BRs, the number of iterations was either 300 or 500. The shape parameter for the alpha parameter of the Gamma distribution, the inverse scale parameter for the alpha parameter of the Gamma distribution, the shape parameter for the lambda parameter of the Gamma distribution, and the inverse scale parameter for the lambda parameter of the Gamma distribution all ranged from 10^{-1} to 10^{1} logarithmically with a step of 10.

All other hyperparameters used the default settings from the scikit-learn package. The models were evaluated using R^2 as reported by SciPy's linregress module. The code to run these algorithms is written in Python using the scikit-learn library [39]. The code used to generate the data in this paper can be found at https://github.com/garrettroell/SyngasMachineLearning. In summary, the seven ML approaches were chosen for analyzing syngas fermentation data because they are highly representative of many types of ML algorithms that are widely used [8]. In particular, RFs, SVMs, and NNs are the three families considered the most accurate by the ML community.

2.4. Concentration curves generated from concentration rate predictions

The following method was used to generate time-course concentration curves from models that predict concentration rates. Concentrations of acetate, ethanol, butyrate, and butanol were determined using the formula $C_{t=1} = C_{t=0} + 0.1 * R_c$, where C_0 is the current concentration, C_1 is the concentration 0.1 days in the future, and R_c is the ML-determined rate of concentration change with units (mM/day). This process was repeated until the generated curves reached the last time point of the experimental data. For conditions with two trials, the average of the two initial concentrations was used as the starting concentration for each metabolite. The time step of 0.1 days was chosen to match the time step of the smoothed data. The plots and calculated metrics compare the average experimental concentrations with the predicted concentrations.

2.5. Concentration curves generated from direct concentration predictions

For reasons explained in Section 2.2., our approach predicts concentration rates rather than concentrations directly. However, we also perform ML for direct concentration using time as model input:

$f(N_2, CO, H_2, CO_2, flowrate, time) = (biomass, acetate, butanol, butyrate, ethanol)$. The results are presented in Section 3.2. The predictions in this section are made with models trained on smoothed data.

3. Results and discussion

3.1. Time-course extracellular metabolite concentrations

Fig. 2 illustrates a typical profile of metabolite accumulation in syngas fermentation with *Clostridium carboxidivorans* P7. For the first 24 hours, glucose was the initial primary carbon source, and in this growth stage, small amounts of acetate and ethanol were produced. After 24 hours, syngas became the primary carbon source, and the product profile changed considerably. The fermentation was acetogenic for approximately 1 day, and in this phase a large amount of acetic acid was produced which caused the pH of the culture to drop. The solventogenic stage began roughly 48 hours after the start of the fermentation. Consistent with previous reports, under solventogenic conditions, the bacteria reassimilated the acetic acid and converted it into ethanol, butyrate, and butanol [33–37].

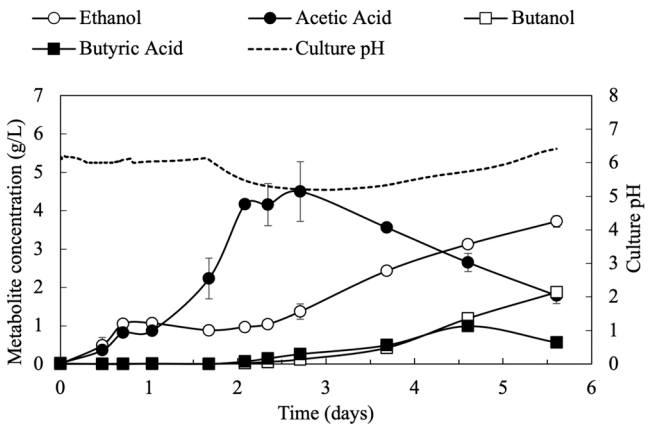

Fig. 2. Example time-course fermentation data. There were two time-course data sets for composition 2 (20 % CO, 15 % CO_2, 25 % H_2, and 40 % N_2) (Table 1). The points represent the average concentration of each metabolite, and the error bars are the standard deviation between the measurements.

3.2. Direct concentration predictions using time as a feature

We first used time along with gas composition and gas flow rate to predict product concentrations directly (Fig. 3). After training ML models using seven families of algorithms (NNs, SVMs, RFs, ENs, LAs, kNNs, and BRs), we found these models had poor performance on the unseen test data, with the R^2 values on average around 0.2. These results indicate that the direct use of ML to predict syngas fermentation time-course data is not feasible due to the variable length of lag phases and differences in inhibitory responses (i.e., the use of time as an ML feature gave unreliable results). Specifically, this method requires full time-course concentration predictions, so it could not account for batch-to-batch fermentation variations such as differences in the length of lag phases. In the next section, we transform the concentration data into rate instances and develop ML models to predict product formation rates rather than product concentrations to address this issue.

3.3. Metabolite production rate predictions using current metabolite concentrations

The seven ML approaches were used to predict metabolite production rates from ten input features (i.e., gas flow rate, gas composition, product concentrations, and biomass concentration). Generally, the models fit training data better than testing data. NNs, RFs, and kNNs fit acetate and ethanol training data very well (R^2 >0.93), while SVMs, ENs, LAs, and BRs were less accurate ($R^2 < 0.63$) (Fig. 4). For unseen test data, RF offered the most accurate predictions for acetate production rate ($R^2 = 0.62$), while NNs had the lowest accuracy with an R^2 of 0.14 (Table 2). The other five algorithms showed similar accuracy with test set R^2 values ranging from 0.23 to 0.43. For ethanol production rates, NNs gave the best predictions ($R^2 = 0.21$), and kNNs made the least accurate predictions ($R^2 = 0.03$). The other five models had similar performance on the test set ($R^2 \approx 0.15$). The predictions for ethanol were considerably less accurate than acetate due to the complex factors governing ethanol synthesis. Compared to acetate, ethanol synthesis requires extra enzyme steps and an additional reducing equivalent (Supplemental Fig. 2). Additionally, ethanol is mainly synthesized during a later fermentation stage (solventogenesis) when cells are reassimilating acetate into ethanol. Since this stage is dependent on the amount of acetate produced in acetogenesis, it was more difficult to predict.

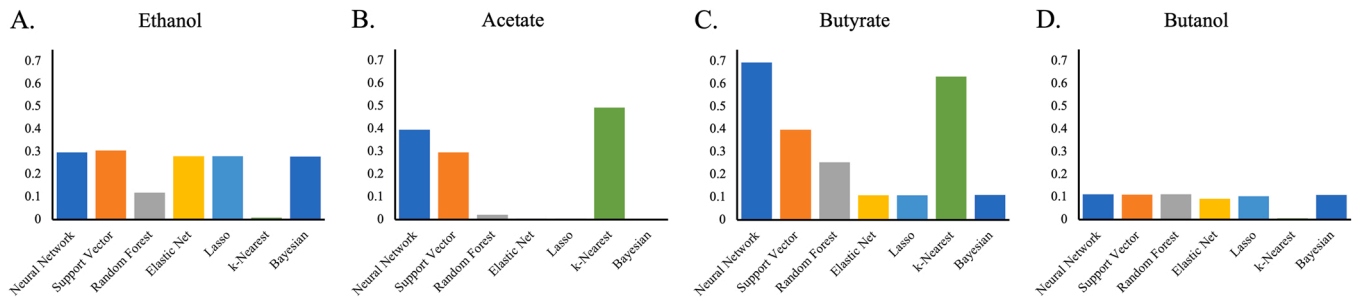

Fig. 3. Test set R^2 values from the direct prediction of product concentrations using time as a variable.

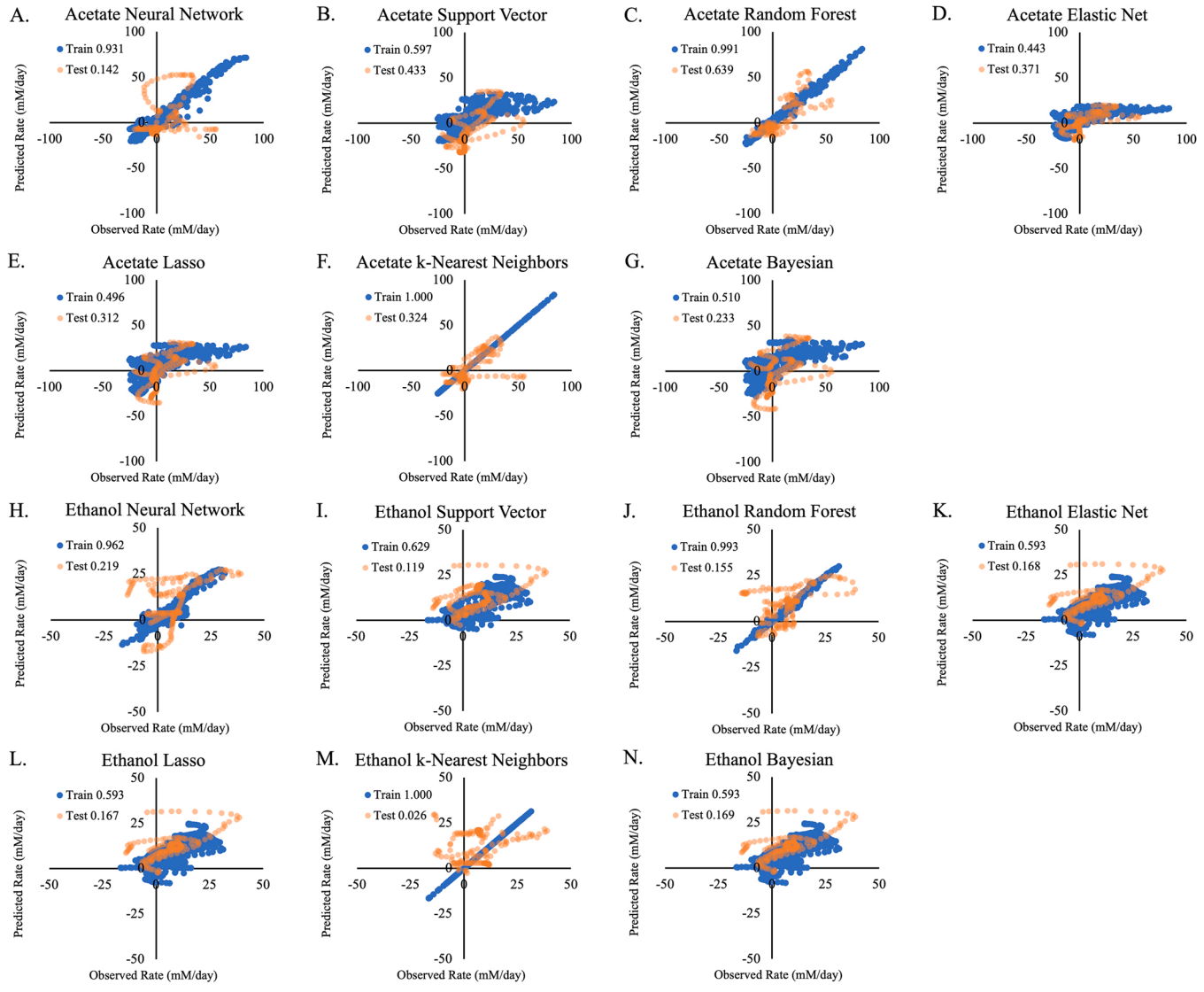

Fig. 4. C2 product synthesis rate predictions (observations vs. predictions). The blue dots represent predictions of training set rates, and the orange dots represent predictions of testing set rates. The x-axis is the observed rate, and the y-axis was the predicted rate, so points that fall on the 45° line through the origin represent accurate predictions. The unit in all scatterplots is mM/day. The values in the legends of each plot are the R^2 value for that data series. RMSE values for each plot could be seen in Supplemental Fig. 3.

The seven families of algorithms were also applied to predict the production rates of butanol and butyrate (Fig. 5). The predictions of C4 production rates by NNs, RFs, and kNNs reached $R^2 > 0.91$ for the training data, while SVMs, ENs, LAs, and BRs had lower quality fits. For butyrate production rate prediction, SVMs performed the best for unseen test data ($R^2 = 0.53$), and kNNs performed the weakest ($R^2 = 0.37$). All models provided relatively poor predictions for butanol synthesis. SVMs were the top performer of all the models with an R^2 of 0.29 (Table 2). The trend of acid prediction being more accurate than alcohol prediction continued for the four-carbon products.

In addition to R^2, root mean squared error (RMSE) and mean absolute percent error (MAPE) were used to evaluate the accuracy of rate

Table 2
Comparison of performance of the six ML algorithms. The average column shows the algorithm's average R^2 ± standard deviation of R^2.

Algorithm	Explanation of Method		Performance on Test Sets				
		Metric	Acetate	Ethanol	Butyrate	Butanol	Average value
Neural Nets	A collection of layers of 'neurons' that can be activated or not determines output.	R^2	0.142	0.219	0.505	0.176	0.261 ± 0.144
		RMSE	19.15	11.23	2.84	4.03	
Support Vector Machines	A high-dimensional plane, constructed using a kernel function, is used to determine outputs.	R^2	0.433	0.119	0.526	0.290	0.342 ± 0.154
		RMSE	20.48	13.04	2.71	3.29	
Random Forests	An ensemble of decision trees is used to predict outputs.	R^2	0.639	0.155	0.459	0.198	0.363 ± 0.197
		RMSE	10.28	11.15	1.53	3.29	
Elastic Nets	A regularized linear method that constrains its fitted variables using L_1 and L_2 penalties.	R^2	0.371	0.168	0.419	0.277	0.309 ± 0.096
		RMSE	12.64	12.24	1.66	3.68	
Lasso regressors	A regularized linear method that constrains the sum of its fitted variables.	R^2	0.312	0.167	0.421	0.277	0.294 ± 0.091
		RMSE	17.94	12.35	1.66	3.68	
K-Nearest Neighbors algorithms	A method that returns the average value of the *k* most similar points found in the training data.	R^2	0.324	0.026	0.367	0.177	0.224 ± 0.134
		RMSE	13.97	13.55	2.03	2.83	
Bayesian Ridge regressors	A method that uses a probabilistic approach to make regression estimations	R^2	0.233	0.169	0.413	0.279	0.273 ± 0.090
		RMSE	21.46	12.31	1.65	3.62	

predictions. Based on RMSE values, random forests showed the best accuracy as they were the most accurate for acetate and butyrate and showed above average performance for ethanol and butanol (Table 2). For acetate and ethanol, random forests performed the best with MAPE values of 10.8 % and 3.2 %, respectively (Supplemental Table 1). MAPE values were not able to be computed accurately for butyrate and butanol because when the true values were very small, minor prediction errors had very large absolute percentage errors ($>10^8$).

3.4. Using rate-based ML models to predict time-course fermentation concentrations

The production rate models were used to generate time-course concentration curves starting with the initial experimental concentrations and iteratively calculating the concentrations based on the current concentrations and rate predictions. Fig. 6 shows the curves generated by SVMs, RFs, and NNs for the test set gas compositions. These figures are a subset of the figures generated. Time-course curves for each condition generated by each algorithm can be found in Supplemental Fig. 5–14. SVMs and RFs were chosen for this figure because they were the best performing algorithms for rate prediction. kNNs were chosen because they demonstrate an example of how overfitting can lead to poor test set predictions. The curves generated by SVMs and RFs were accurate with average test set R^2 values of 0.78 and 0.87, respectively. These algorithms performed the best at predicting rates, so their accurate curve predictions are expected.

Supplemental Table 2 contains the R^2 values for each algorithm when predicting concentration curves. Interestingly, the top-performing algorithm for the test set, RFs, was more accurate with the test set than the training set. The top-performing algorithm for the training set was kNN, but kNNs were among the lowest performers for the test set. kNNs had an excellent fit for composition 10 but a poor fit for composition 9 (Fig. 6). The variability can be explained by kNN's dependence on training data closely matching the testing data. Composition 9 was unique since it did not contain hydrogen gas, and as a result, kNNs made poor predictions for this condition. Additionally, the poor performance of the NNs for composition 9 was likely because NNs were overfitted due to their large number of parameters. Both issues could be resolved with a more extensive training data set.

3.5. The comparison of different ML algorithms for syngas predictions

The seven families of regressors exhibited different performance patterns across products and prediction methods (Table 2, Supplemental Table 1, and Supplemental Table 2). For test set rate prediction, RFs performed the best, and SVMs were a close second. Both algorithms had average R^2 values of ~0.35 for rate predictions. ENs and LAs performed somewhat well with average R^2 values of ~0.30, while NNs, kNNs, and BRs had the worst average performances with R^2 values of ~0.25. ENs and LAs are relatively simple algorithms with fewer fitted variables than the other ML methods. Since the linear methods outperformed kNNs and NNs, it can be concluded that kNNs and NNs were overfitted. Despite NN's overall poor performance, it offered the best predictions of ethanol production rate. Since an algorithm could have large variability in performance across products, the selection of ML algorithms should be made only after testing multiple options [8]. Syngas fermentation is a highly nonlinear and dynamic system, so complex ML algorithms (e.g., NN) risk poor fits to unseen test data if training data is insufficient. In contrast, the simpler ML algorithms, ENs and LAs, are 'safer' options because they have less fitted variables, making them less likely to overfit. Moreover, BR showed average to below average performance relative to the other six algorithms (Table 2 and Supplemental Table 1). In general, Bayesian methods perform better for larger sets of data where each feature is assumed to independently contribute to the probability of other features. However, the data in this study was limited, and potentially some of the model features may have had a dependence on one another (e.g., acetate consumption is directly linked to ethanol production during syngas fermentation).

This research found that data transformation (concentrations → rates) improved ML quality. The rate predicting models generated time-course product curves that closely matched unseen experimental data. For example, SVMs, RFs, LAs, and ENs provided predictions with R^2

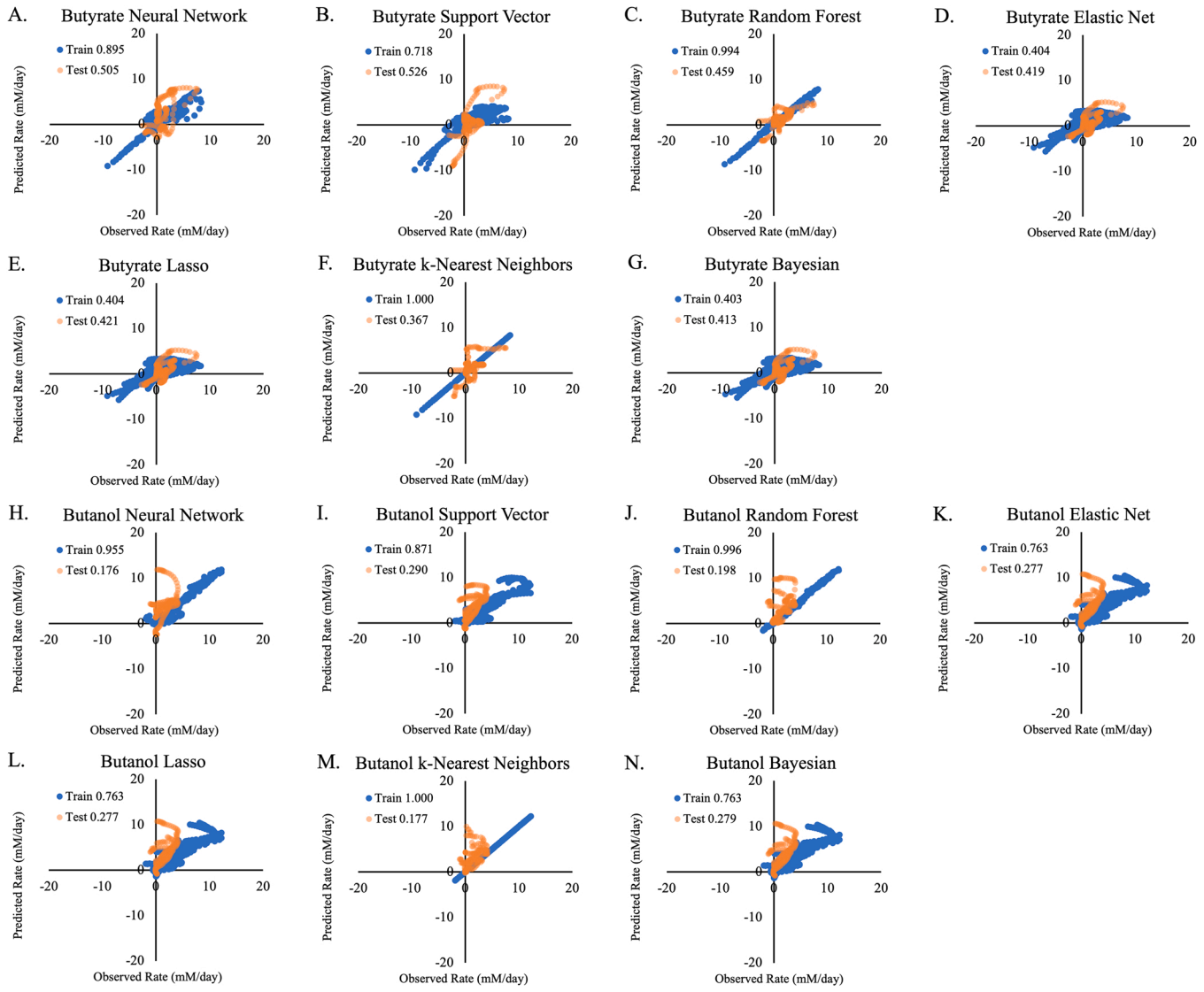

Fig. 5. C4 product synthesis rate predictions (observations vs. predictions). The blue dots represent predictions of training set rates, and the orange dots represent predictions of testing set rates. The x-axis is the observed rate, and the y-axis is the predicted rate, so points that fall on the 45° line through the origin represent accurate predictions. The unit in all scatterplots is mM/day. The values in the legends of each plot are the R^2 value for that data series. RMSE values for each plot could be seen in Supplemental Fig. 4.

values of ~0.8. This improvement could be explained by three reasons. First, the rate predictions used product concentrations as features, so inhibition effects and product status could be incorporated into ML training and predictions. Second, fermentation processes had different lag phase time lengths, so the data transformation avoided time as ML input and reduced its feature uncertainty. Third, by taking advantage of the known base concentrations at the starting time, the rate-based ML models could predict time-course concentrations well.

This study had two limitations. One issue was that data transformation caused dependence among rate values from temporally close data points because production rates were derived from smoothed time series curves. As a result, the same prediction error was often repeated, leading to low R^2 values. For example, in kNN's predictions of acetate production rates, underestimations were observed for ~20 consecutive samples (Fig. 4). Second, the ML approach had poor predictions of alcohol production rate, suggesting additional features were needed to capture the metabolic phase shifting during fermentation.

3.6. Feature importance analysis and biological insights

The RF models were used to determine the relative weight of the gas components when predicting the production rates of the four products (Fig. 7). The importance of a gas's concentration on a metabolite's production rate was determined by averaging the impurity reduction when the gas value (as a percentage of total gas composition) was used to split the decision trees [39]. In this context, feature importance offers guidelines for steering syngas fermentations towards desired products. Specifically, the metabolic network in syngas fermentation was controlled by complex host-product-substrate interactions with multiple input substrates and output products [33], and large amounts of energy molecules are required to make the products (Supplementary Fig. 2) [5]. In industrial applications, fermentation engineers prefer to produce C4 products over C2 products since butyrate and butanol are more valuable than acetate and ethanol. Feature importance analysis shows that butyrate's production rate was heavily dependent on the concentration of CO in the feed gas and that butanol's production rate was mainly dependent on the concentration of H_2. These findings offer control strategies to optimize syngas compositions to create specific products

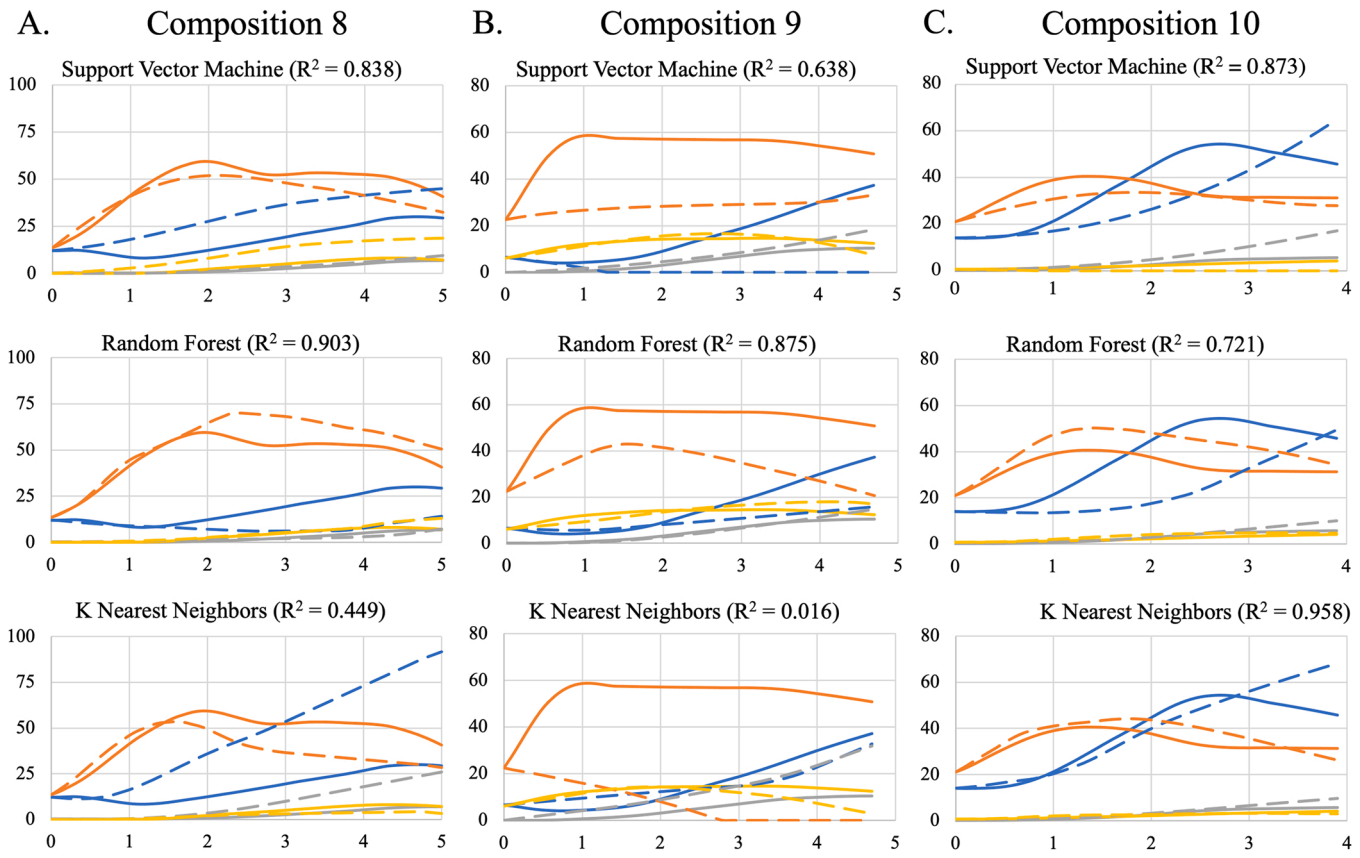

Fig. 6. Concentration curves generated from rate predictions. The solid lines represent experimental data, and the dashed lines represent predicted data. Blue = ethanol, orange = acetate, gray = butanol, yellow = butyrate. The x-axis is time in days, and the y-axis is concentration in mM units.

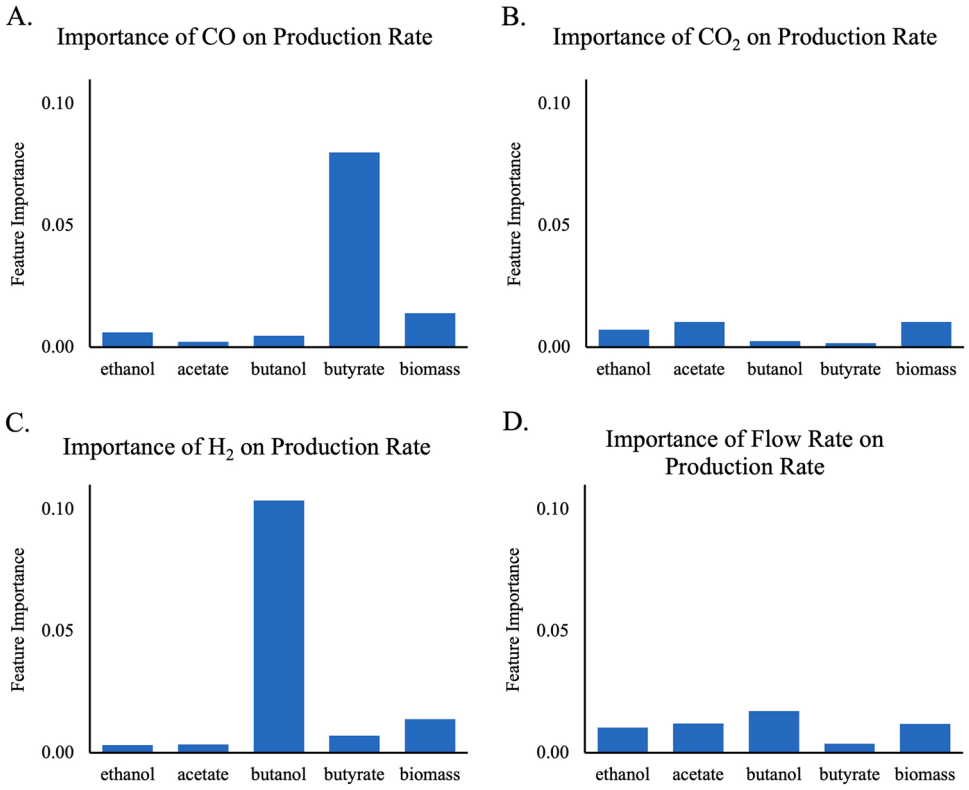

Fig. 7. Feature importance of gas composition on metabolite production. The size of the bar indicates the importance of that gas for determining the production rate of the labeled metabolite.

[35]. The feature analyses also showed how ML could use limited experimental data to 'relearn' biosynthesis patterns.

4. Conclusion

This study evaluated seven families of machine learning (ML) algorithms to predict syngas fermentation production rates and time-course product concentration curves based on limited fermentation trials. Time-course predictions based on rate predictions were more accurate than direct concentration predictions. Generally, the ML methods were more accurate for acid production rates than for alcohol production rates, indicating that the features used in this study did not capture all the factors that determine alcohol production rate (e.g., metabolic shifts or other intrinsic biological factors). Random forests and support vector machines gave the best rate predictions and generated accurate time-course concentration curves. Additionally, feature importance analysis reaffirmed guidelines for how gas composition can control product profiles. Future studies can build off this work by increasing the amount of syngas fermentation data, including new features to capture cellular regulation and stress responses to bioreactor conditions, or by applying an ensemble machine learning approach. The advancement of machine learning is a promising route for facilitating model predictive control to optimize fermentation outcomes [18].

Authorship contribution statement

GWR and YJT conceived the data transformation and ML for studying gas fermentation. FSB, QC, and GWR developed and tested the ML models. AS, WZ, and NW performed experiments. All authors wrote and proofread the paper.

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

The data and code is available at https://github.com/garrettroell/SyngasMachineLearning

Acknowledgments

This work was supported by National Science Foundation (NSF) (MCB-1821828, CNS-1817089) to FB and NSF (CBET-1438125) to ZW and YT.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.bej.2022.108578.

References

[1] T. Beltramo, C. Ranzan, J. Hinrichs, B. Hitzmann, Artificial neural network prediction of the biogas flow rate optimised with an ant colony algorithm, Biosyst. Eng. 143 (2016) 68–78.
[2] D. Beneroso, J.M. Bermúdez, A. Arenillas, J.A. Menéndez, Comparing the composition of the synthesis-gas obtained from the pyrolysis of different organic residues for a potential use in the synthesis of bioplastics, J. Anal. Appl. Pyrolysis 111 (2015) 55–63.
[3] A. Bowler, J. Escrig, M. Pound, N. Watson, Predicting alcohol concentration during beer fermentation using ultrasonic measurements and machine learning, Fermentation 7 (2021) 7.
[5] C. Cheng, W. Li, M. Lin, S.-T. Yang, Metabolic engineering of *Clostridium carboxidivorans* for enhanced ethanol and butanol production from syngas and glucose, Bioresour. Technol. 284 (2019) 415–423.
[6] Z. Costello, H.G. Martin, A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data, npj Syst. Biol. Appl. 4 (2018) 19.
[7] E.A. Del Rio-Chanona, F. Fiorelli, D. Zhang, N.R. Ahmed, K. Jing, N. Shah, An efficient model construction strategy to simulate microalgal lutein photo-production dynamic process, Biotechnol. Bioeng. 114 (2017) 2518–2527.
[8] M. Fernandez-Delgado, E. Cernadas, S. Barro, D. Amorim, Do we need hundreds of classifiers to solve real world classification problems? J. Mach. Learn. Res. 15 (2014) 3133–3181.
[9] J. Fischer, V. Lopes, S. Cardoso, U. Coutinho Filho, V. Cardoso, Machine learning techniques applied to lignocellulosic ethanol in simultaneous hydrolysis and fermentation, Braz. J. Chem. Eng. 34 (2017) 53–63.
[10] B.D. Heijstra, C. Leang, A. Juminaga, Gas fermentation: cellular engineering possibilities and scale up, Micro Cell Fact. 16 (2017) 60.
[11] S. Hochreiter, J. Schmidhuber, Long Short-Term Memory, Neural Comput. 9 (1997) 1735–1780.
[12] K. Itto-Nakama, S. Watanabe, N. Kondo, S. Ohnuki, R. Kikuchi, T. Nakamura, W. Ogasawara, K. Kasahara, Y. Ohya, AI-based forecasting of ethanol fermentation using yeast morphological data, Biosci., Biotechnol., Biochem. 86 (2022) 125–134.
[13] C.E. Lawson, J.M. Martí, T. Radivojevic, S.V.R. Jonnalagadda, R. Gentz, N. J. Hillson, S. Peisert, J. Kim, B.A. Simmons, C.J. Petzold, S.W. Singer, A. Mukhopadhyay, D. Tanjore, J.G. Dunn, H. Garcia Martin, Machine learning for metabolic engineering: A review, Metab. Eng. 63 (2021) 34–60.
[14] B. Li, Y. Lin, W. Yu, D.I. Wilson, B.R. Young, Application of mechanistic modelling and machine learning for cream cheese fermentation pH prediction, J. Chem. Technol. Biotechnol. 96 (2021) 125–133.
[15] F. Liew, M.E. Martin, R.C. Tappel, B.D. Heijstra, C. Mihalcea, M. Köpke, Gas fermentation-A flexible platform for commercial scale production of low-carbon-fuels and chemicals from waste and renewable feedstocks, Front. Microbiol. 7 (2016) 694, 694-694.
[16] J. Lin, Model predictive control of glucose feeding for fed-batch candida utilis biomass production, Res. J. Biotechnol. (2013) 8.
[17] P.C. Munasinghe, S.K. Khanal, Syngas fermentation to biofuel: evaluation of carbon monoxide mass transfer and analytical modeling using a composite hollow fiber (CHF) membrane bioreactor, Bioresour. Technol. 122 (2012) 130–136.
[18] Z.K. Nagy, Model based control of a yeast fermentation bioreactor using optimally designed artificial neural networks, Chem. Eng. J. 127 (2007) 95–109.
[19] H. Narayanan, L. Behle, M.F. Luna, M. Sokolov, G. Guillén-Gosálbez, M. Morbidelli, A. Butté, Hybrid-EKF: hybrid model coupled with extended Kalman filter for real-time monitoring and control of mammalian cell culture, Biotechnol. Bioeng. 117 (2020) 2703–2714.
[20] J.J. Orgill, H.K. Atiyeh, M. Devarapalli, J.R. Phillips, R.S. Lewis, R.L. Huhnke, A comparison of mass transfer coefficients between trickle-bed, hollow fiber membrane and stirred tank reactors, Bioresour. Technol. 133 (2013) 340–346.
[21] A.W. Rogers, F. Vega-Ramon, J. Yan, E.A. del Río-Chanona, K. Jing, D. Zhang, A transfer learning approach for predictive modeling of bioprocesses using small data, Biotechnol. Bioeng. 119 (2022) 411–422.
[22] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (1964) 1627–1639.
[23] D.E. Seborg, D.A. Mellichamp, T.F. Edgar, F.J. Doyle, Process. Dynamics and Control, John Wiley & Sons,, 2010.
[24] P. Shah, M.Z. Sheriff, M.S.F. Bangi, C. Kravaris, J.S.-I. Kwon, C. Botre, J. Hirota, Deep neural network-based hybrid modeling and experimental validation for an industry-scale fermentation process: Identification of time-varying dependencies among parameters, Chem. Eng. J. 441 (2022), 135643.
[25] L. Shampine, S. Thompson, Stiff systems, Scholarpedia 2 (2007) 2855.
[26] Y. Shen, R. Brown, Z. Wen, Syngas fermentation of Clostridium carboxidivoran P7 in a hollow fiber membrane biofilm reactor: Evaluating the mass transfer coefficient and ethanol production performance, Biochem. Eng. J. 85 (2014) 21–29.
[27] N.J. Treloar, A.J.H. Fedorec, B. Ingalls, C.P. Barnes, Deep reinforcement learning for the control of microbial co-cultures in bioreactors, PLOS Comput. Biol. 16 (2020), 1007783 e1007783-e1007783.
[28] A. Tulsyan, C. Garvin, C. Ündey, Advances in industrial biopharmaceutical batch process monitoring: machine-learning methods for small data problems, Biotechnol. Bioeng. 115 (2018) 1915–1924.
[29] Wan, N. , 2018. Application of metabolic modeling and machine learning for investigating microbial systems. PhD Thesis, Washington University in St. Louis.
[30] N. Wan, A. Sathish, L. You, Y.J. Tang, Z. Wen, Deciphering Clostridium metabolism and its responses to bioreactor mass transfer during syngas fermentation, Sci. Rep. 7 (2017) 10090.
[31] B. Wang, M. Shahzad, X. Zhu, K.U. Rehman, S. Uddin, A non-linear model predictive control based on grey-wolf optimization using least-square support vector machine for product concentration control in l-lysine fermentation, Sensors 20 (2020) 20.
[32] L. Wang, F. Long, W. Liao, H. Liu, Prediction of anaerobic digestion performance and identification of critical operational parameters using machine learning algorithms, Bioresour. Technol. 298 (2020), 122495.
[33] J. Zhang, S. Taylor, Y. Wang, Effects of end products on fermentation profiles in Clostridium carboxidivorans P7 for syngas fermentation, Bioresour. Technol. 218 (2016) 1055–1063.
[34] Á. Fernández-Naveira, M.C. Veiga, C. Kennes, H-B-E (hexanol-butanol-ethanol) fermentation for the production of higher alcohols from syngas/waste gas, Journal of Chemical Technology & Biotechnology 92 (2017) 712–731.
[35] J. Daniell, M. Köpke, S.D. Simpson, Commercial biomass syngas fermentation, Energies 5 (2012) 5372–5417.

[36] J.R. Phillips, H.K. Atiyeh, R.S. Tanner, J.R. Torres, J. Saxena, M.R. Wilkins, R. L. Huhnke, Butanol and hexanol production in Clostridium carboxidivorans syngas fermentation: Medium development and culture techniques, Bioresource Technology 190 (2015) 114–121.
[37] S. Ramió-Pujol, R. Ganigué, L. Bañeras, J. Colprim, How can alcohol production be improved in carboxydotrophic clostridia? Process Biochemistry 50 (2015) 1047–1055.
[38] Ahmed, A. (2006). Effects of biomass-generated syngas on cell-growth, product distribution and enzyme activities of Clostridium carboxidivorans P7T. PhD Thesis, Oklahoma State University.
[39] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12 (2011) 2825–2830.