



Comparison of machine learning methodologies for predicting kinetics of hydrothermal carbonization of selective biomass

Mohammadreza Aghaaminiha¹ · Ramin Mehrani² · Toufiq Reza³ · Sumit Sharma¹

Received: 21 May 2021 / Revised: 5 August 2021 / Accepted: 7 August 2021 / Published online: 22 August 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract

We have examined performance of various machine learning (ML) methods (artificial neural network, random forest, support vector-machine regression, and K nearest neighbors) in predicting the kinetics of hydrothermal carbonization (HTC) of cellulose, poplar, and wheat straw performed under two different conditions: first, isothermal conditions at 200, 230, and 260 °C, and second, with a linear temperature ramp of 2 °C/min from 160 to 260 °C. The focus of this study was to determine the predictability of the ML methods when the biomass type is not known or there is a mixture of biomass types, which is often the case in commercial operations. In addition, we have examined the performance of ML methods in interpolating kinetics results when experimental data is available for only a handful of time-points, as well as their performance in extrapolating the kinetics when the experimental data from only a few initial time-points is available. While these are stringent tests, the ML models were found to perform reasonably well in most cases with an averaged mean squared error (MSE) and R^2 values of 0.25 ± 0.06 and 0.76 ± 0.05 , respectively. The ML models showed deviation from experimental data under the conditions when the reaction kinetics were fast. Overall, it is concluded that ML methods are appropriate for the purpose of interpolating and extrapolating the kinetics of the HTC process.

Keywords Supervised machine learning · Hydrothermal carbonization · Hydrochar · Biomass · Reaction kinetics

1 Introduction

Hydrothermal carbonization (HTC) is a thermochemical conversion process in which hot, pressurized, and subcritical form of residual moisture is used as a reaction medium for the conversion of wet biomass and sludges to carbon and a nutrient-rich solid product called hydrochar, process liquid with high total organic content, and a gaseous product containing primarily carbon dioxide [1–3]. Hydrochar, the main product of HTC, has been used as a solid fuel—due

to high percentage of fixed carbon [4], additive for biofuel pellets [5], adsorbent medium — due to high density of surface functional groups [6], carbon materials for anaerobic bioreactor [7], electrode materials — due to their high conductivity [8], fertilizer — due to condensed nutrients [9], and many more. Meanwhile, HTC process liquid can be converted to energy by anaerobic digestion [10] and wet air oxidation [11], and into liquid fertilizer [12], depending on the feedstock type.

In the HTC process, the reaction temperatures range from 180 to 280 °C and the pressure is usually increased above saturation pressure to keep water in its liquid state during HTC. The reaction times vary from 1 min to 18 h [2, 7, 13]. HTC process is understood to be slightly exothermic, releasing 0.74–1.1 MJ/kg of heat depending on the type of biomass [14]. During the HTC process, hydrolysis of biopolymers occurs first to produce sugar monomers, which then undergo dehydration into sugar derivatives followed by a series of complex reactions namely condensation, aromatization, and polymerization to produce hydrochar [1]. HTC reaction kinetics for an individual biopolymer (e.g., carbohydrates, proteins, hemicelluloses, cellulose, and lignin)

✉ Sumit Sharma
sharmas@ohio.edu

¹ Department of Chemical and Biomolecular Engineering, Russ College of Engineering and Technology, Ohio University, 181 Stocker Center, Athens, OH 45701, USA

² Department of Mechanical Engineering, Russ College of Engineering and Technology, Ohio University, Athens, OH 45701, USA

³ Department of Biomedical and Chemical Engineering and Sciences, Florida Institute of Technology, 150 West University Boulevard, Melbourne, FL 32901, USA

vary extensively and in most cases, their reaction kinetics are interfered directly with other reaction kinetics or indirectly by secondary products produced from separate reactions [15]. Moreover, like any other reaction, HTC reaction kinetics are affected by catalysts [16]. In fact, homogeneous catalysts (e.g., organic acids or bases) have been used to increase the selectivity of a specific product from HTC [17, 18]. A catalyst was not used in the experiments that are considered for machine learning (ML) modeling in this work. Since catalysts alter the reaction kinetics, an ML-based predictive model should be re-trained on new experimental data if a catalyst has been added or replaced in the experiments.

An impediment in the scale-up and commercialization of the HTC process is that feed from different sources varies in composition, and so there is a lack of predictability of the optimum conditions for carrying out the HTC process and the properties of the hydrochar thus formed [19]. HTC process involves a large number of complex chemical reactions, as well as physical transformations of the biomass constituents. These changes, along with the different partitioning tendencies of chemical species between the solid and the liquid phases, render mechanistic modeling of the process quite challenging. In this work, we have applied various machine learning methodologies to model the experimental kinetics of HTC of three different biomass types at different reaction conditions.

ML refers to a class of algorithms that have the ability to learn from experience (data) to perform a task (e.g., predicting the outcome of an experiment), and improve their performance on the task as they gain more experience [20, 21]. Supervised ML algorithms are first trained on some data, called the training set, and then the trained ML algorithms are used to make predictions for unseen data [21]. Some popular supervised ML algorithms are random forest (RF), artificial neural network (ANN), support vector-machine regression (SVR), and K nearest neighbors (KNNs). Li et al. implemented SVR and RF methods to develop ML models to predict fuel characteristics (yield, heating value, energy recovery, energy density, etc.) of various feedstocks after HTC and pyrolysis treatment [22]. In another work, a multi-task prediction model using ANN was developed to predict hydrochar properties based on feedstock properties as well as HTC conditions [23]. Kardani et al. modeled the conversion of feedstock as a function of feedstock composition and operating conditions using various ML techniques [24].

The goal of this study was to evaluate various ML approaches to predict the kinetics of feedstocks hydrothermally carbonized at different reaction temperatures and times. Experimental data previously published by our group is the basis of this study [15]. The experimental data includes three feedstocks namely cellulose, poplar, and wheat straw that were hydrothermally carbonized in a 5-gal reactor in two different modes: isothermal and dynamic. In the isothermal

runs, three reaction temperatures (200, 230, and 260 °C) were studied for 0–480 min. In the dynamic runs, the experiments were performed with a temperature ramp of 2 °C/min going from 160 to 260 °C, and samples were collected every 5 min. Although various physicochemical analyses were reported in the published paper [15], here we have focused on the elemental analysis (CHNS) for this ML study. Ultimate analysis of hydrochars provides elemental carbon, hydrogen, sulfur, and nitrogen content. Oxygen content is calculated by the difference method. None of the feedstocks that were studied in these experiments contain significant amounts of sulfur and nitrogen. Performing kinetics based on hydrogen is expected to be difficult, as subcritical water is used for HTC, which facilitates hydrolysis and dehydration reactions. Therefore, carbon is the best choice for the kinetics study, as carbon content was found as high as 70% of the hydrochar. However, reaction order cannot be established based on elemental composition, and so this study did not predict the specific reaction orders.

Specifically, we have studied three different scenarios wherein the application of ML modeling is envisioned to be practically useful. The first scenario, termed *Kinetics*, is the simplest case where the entire kinetics of nitrogen, sulfur, and hydrogen are included as features to predict the kinetics of carbon during the HTC process. The second scenario, termed *Interpolation*, is a more stringent test of ML modeling. In *Interpolation*, kinetics data of only the initial, middle, and last time-points of an experiment are included as features in the model and the kinetics data for all the intermediate time-points are predicted. The rationale for this scenario is that since the HTC process occurs at high temperatures and pressures, it is not feasible to collect a large number of samples at frequent time points. Therefore, the *Interpolation* modeling methodology will help in enhancing the kinetics data by providing estimates of the weight percent of carbon at the intermediate time points. The third scenario, termed *Extrapolation*, is another stringent test of ML modeling. In this scenario, the kinetics data of the first three time-points are included as features in the model and the kinetics for the entire HTC experiment are then predicted. The rationale is that for biowastes of different compositions, it is costly and time-consuming to run the HTC experiment for the entire duration. Therefore, the *Extrapolation* modeling methodology will be useful for predicting the kinetics of an HTC process provided a short duration experimental run has been performed.

2 Material and methods

2.1 Data preprocessing

The first step is to ensure that there are no missing and incorrect values in the data [25, 26]. As shown in Table 1,

our dataset includes 1 categorical and 5 numerical features. The experimental data has three different biomass types: cellulose, poplar, and wheat straw. For each biomass type, there are four different experiments: three experiments are done in isothermal conditions and one experiment is done with a constant temperature ramp. Overall, there are 12 experiments. The dataset has 132 data-points from all the experiments. The ranges of the numerical features are listed in Table 1. The categorical feature specifies the type of the experiment performed: isothermal or dynamic. The categorical feature was converted to an indicator or dummy feature. All numerical features were scaled to the z-score of a standard normal distribution. Z-score for a feature value is obtained by subtracting the mean of the feature and then dividing the result by the standard deviation of the feature [26]. To capture the Arrhenius effect on the kinetics, $\exp\left(\frac{273+230}{273+T}\right)$ was used as a feature in lieu of T , where T is the temperature in K. Biomass type was removed from the dataset so that the trained ML model is agnostic to the type of biomass.

2.2 ML modeling strategies

The weight percent of elemental carbon in the HTC system was considered as the target variable or the label for ML modeling. The experimental data included three different types of biomasses: cellulose, poplar, and straw. Generally, biowaste is a mixture of different biomass types. So, a successful ML strategy should be able to make good predictions of the kinetics of the HTC process without the knowledge of the constituent biomass type. Therefore, the information of biomass type was removed, that is, was not included as a feature during the ML modeling. We studied three different scenarios of ML modeling: *Kinetics*, *Interpolation*, and *Extrapolation*. These scenarios, and rationale for studying them, are discussed in detail in the introduction section.

Table 1 Details of input variables (features) used in the modeling of kinetics of biomasses. Nitrogen, sulfur, and hydrogen content were hidden in the prediction data for the *Interpolation* and the *Extrapolation* scenarios for all but three time-points in the data set

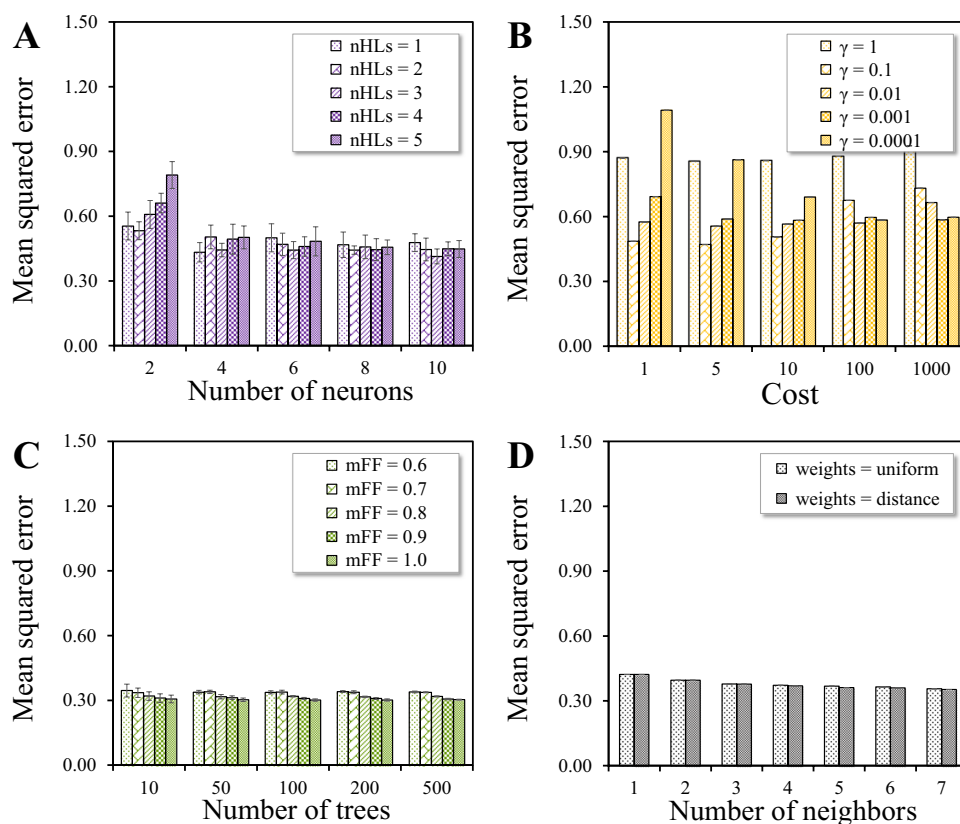
Description	Range	Unit	Type
Experiment type	(Isothermal, Dynamic)	–	Categorical
Temperature	[160–260]	°C	Numerical
Time	[0–480]	min	Numerical
Nitrogen content	[0.0–2.0]	%	Numerical
Sulfur content	[0.0–1.0]	%	Numerical
Hydrogen content	[4.0–7.0]	%	Numerical

2.3 ML implementation details

The performance of supervised ML models is a function of the selected hyperparameter values [27]. So, the first step was to determine the optimum set of hyperparameters for each of the ML models studied (SVR, RF, ANN, KNN). Once each ML model was optimized, the performance of the optimized ML models was compared to select the best model for a given application. The selected ML model was then used to make predictions of the experimental data. The implementation of these steps is discussed in detail below.

Hyper-parameter tuning. The ranges of the hyperparameters for each ML algorithm are listed in Table S1 (Supporting Information). Performance of the ML algorithms was evaluated by calculating the mean squared error (MSE) between the predicted and the experimental values of the label as the evaluation metric. Generally for implementing ML methodologies, dataset is split randomly into training and testing sets. In this work, we are interested in predicting the entire kinetics of an experiment. So, the approach of randomly splitting the dataset in training and testing set is not suitable. Since the experimental data comprises of 12 different experiments, the best ML model was evaluated by performing a 12-fold cross-validation with each HTC experiment considered as a fold. So, one experiment is completely hidden, and the remaining 11 experiments are utilized for training the ML models. In the case of stochastic ML algorithms, ANN, and RF, the 12-fold cross-validation was iterated 50 times to ensure robustness of results. Only the hyperparameters that are understood to have the largest effect on the performance of the ML algorithms were studied while choosing suggested default values of the other hyperparameters. As an example, for ANN, the activation function was set to *Relu*; *Adam* solver was used for the optimization of weight vectors; and the learning rate was set to 0.001 [28]. Figure 1 compares the performance of the ML algorithms with different sets of hyperparameters for the *Kinetics* modeling scenario. Similar comparison of the performance for the *Interpolation* and *Extrapolation* scenarios are shown in Fig. S1 and S2 of the supplementary material, respectively. In the case of ANN (Fig. 1A), it was observed that increasing the number of hidden layers, *nHLs*, worsened the performance when the number of neurons, *nNeurons*, in each layer was 2. When the number of neurons was increased beyond 2, no improvement in the MSE was observed with the increase in *nHLs*. A similar trend was observed in Fig. S1(A). In Fig. S2(A), the performance of ANN was found to worsen with the increase in *nHLs*. This is expected as the total number of neurons and the connections between them geometrically increase with the increase in the *nHLs*, and so, the connections get sub-optimally trained when the training dataset is not large. The SVR model was implemented using the Gaussian Radial Basis Function kernel. The hyperparameter

Fig. 1 Performance of **A** artificial neural network (ANN), **B** support vector-machine regression (SVR), **C** random forest (RF), and **D** K nearest neighbors (KNN) for the prediction of carbon content for different values of the hyperparameters for the *Kinetics* scenario. The error bars are standard deviation of predictions on 50 independent iterations for ANN and RF. Similar figures for *Interpolation* and *Extrapolation* scenarios are provided in the supporting information



γ , γ is the inverse of the standard deviation of the Gaussian functions used in the kernel. For the SVR model (Fig. 1B), the performance was found to be worst for the smallest and largest values of γ (0.001 and 1.0), which is understandable as a small value of γ makes the Gaussian spread large, so that accuracy of the prediction decreases. A large value of γ makes the Gaussian function sharply peaked so that the training data is overfitted. Figure 1B shows that in general, the performance worsened with the increase in the *cost*, which is the regularization parameter of SVR. A large value of the *cost* may result in overfitting of the training data. $\gamma = 0.1$ and *cost* = 5 were found to be the optimum values of the hyperparameters. In the case of *Interpolation* and *Extrapolation* scenarios, a smaller value of γ (= 0.001) and a larger *cost* (= 100) were found to be optimum. In the case of RF model (Fig. 1C), the best performance, that is lowest MSE, was observed when the fraction of features to be split for making decision trees was set to 1.0. Beyond 100 trees, no improvement in the performance was observed and so the RF model with 100 trees was optimum. For the *Interpolation* and *Extrapolation* scenarios, the performance of the RF model did not improve beyond 10 trees. In the case of KNN (Fig. 1D), a monotonically decreasing trend of the MSE was obtained, implying that the number of neighbors = 7 is the optimum choice. For the *Interpolation* and *Extrapolation* scenarios, the number of neighbors = 3 and

7 respectively were found to be optimum. Furthermore, it was found that applying distance-based *weights* to the label values of the neighbors resulted in a slight improvement in the performance. Once the optimum set of hyperparameters for each ML model was determined, the performance of these models was compared against each other for the three scenarios (Fig. 2). Table 2 lists the hyperparameters of the optimum ML models. It was found that RF is the best model for the *Kinetics* scenario, whereas SVR is the best model for both *Interpolation* and *Extrapolation* scenarios. The best performing ML model for each scenario was chosen for further data analysis.

3 Results and discussion

In all the ML models studied, we hid the information of biomass type from the models. The rationale for hiding this information is that in practical applications, we may not have a single biomass type in the biowaste but a mixture of biomass types. Different biomass types have significant differences in their microscopic structure and composition, and thus, they are expected to demonstrate different kinetics behavior. Therefore, hiding the information of biomass type make it quite challenging to predict the kinetics.

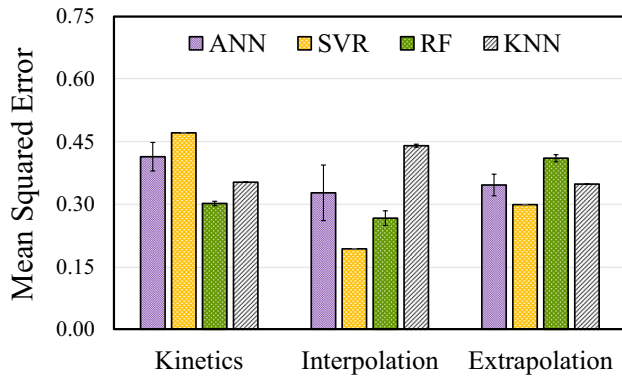


Fig. 2 Comparison of the performance of the best ML models for each scenario. The best ML model for *Kinetics* was RF (*nTR* 100, *mFF* 1.0), for *Interpolation* was SVR (γ 0.001, *cost* 100) and for *Extrapolation* was SVR (γ 0.001, *cost* 10). The error bars are standard deviation of predictions on 50 independent iterations for ANN and RF

Figure 3 compares the ML predictions of weight percent of carbon with the experimental results of the HTC process

for the three biomass types wherein the temperature of the HTC was ramped up at a constant rate of 2 °C/min from 160 to 260 °C. The *Kinetics* scenario, modeled using the RF model, matches the experimental results quite well for all the biomass types studied, except in the case of HTC of straw at temperatures above 230 °C where the prediction is above the experimental values (Fig. 3C). The RF model overpredicts the results of straw at high temperatures because the isothermal HTC experiments performed at temperatures of 230 °C and 260 °C (discussed later) show a higher weight percent of carbon as compared to the experimental results in Fig. 3C. Compared to cellulose, which does not contribute to the HTC reaction until 230 °C [13], extractives and hemicelluloses in straw start the HTC reaction as early as 180 °C [29]. Therefore, the reaction kinetics of straw showed an increase of reactivity from 180 °C. For the same reason, the *Interpolation* scenario, modeled using SVR, slightly overpredicts the weight percent of carbon at higher temperatures for the HTC of cellulose (Fig. 3A), whereas its prediction is quite good for the poplar biomass (Fig. 3B). In the case of straw biomass, the prediction of the *Interpolation* scenario is quite

Table 2 Optimum values of the hyperparameters along with their MSE are reported. Note that the reported MSE are the average of 50 independent iterations for the ANN and RF models

Scenarios ↓	ANN		SVR		RF	KNN		
	<i>nHLs</i>	<i>nNeurons</i>	γ	<i>cost</i>	<i>nTR</i>	<i>mFF</i>	<i>K</i>	<i>weights</i>
<i>Kinetics</i>	3	10	0.1	5	100	1.0	7	distance
MSE	0.414 ± 0.034		0.471		0.301 ± 0.005		0.353	
<i>Interpolation</i>	3	8	0.001	100	10	0.8	3	uniform
MSE	0.327 ± 0.067		0.192		0.266 ± 0.018		0.440	
<i>Extrapolation</i>	1	6	0.001	10	50	0.7	7	distance
MSE	0.346 ± 0.026		0.299		0.410 ± 0.009		0.348	

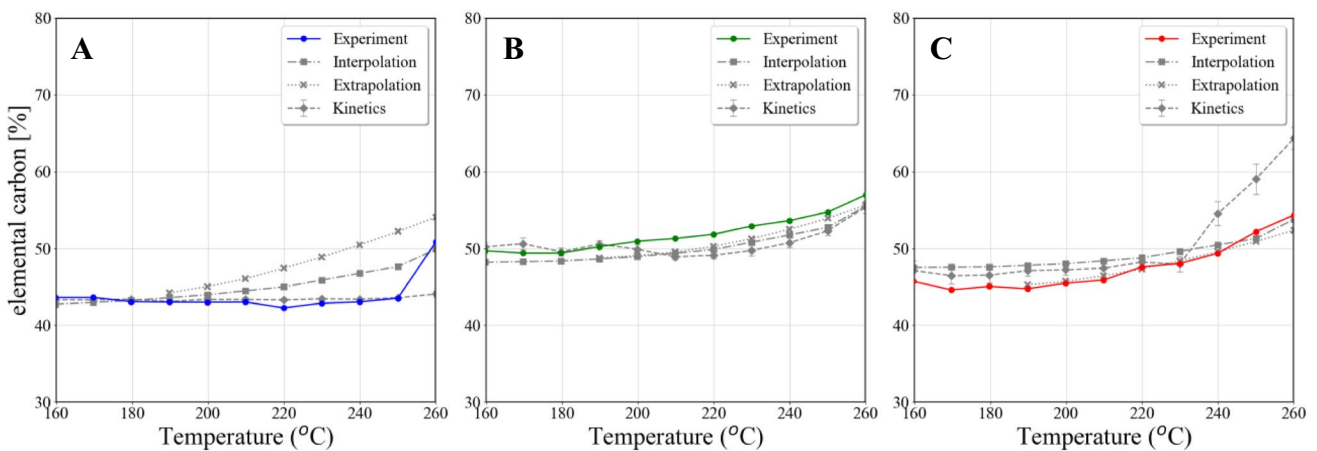


Fig. 3 Comparison between the predicted and the experimentally reported carbon content during the dynamic HTC of **A** cellulose, **B** poplar, and **C** straw. In these experiments, temperature was varied at a constant rate of 2 °C/min. ML models used for *Kinetics*, *Interpolation*, and *Extrapolation* scenarios were RF (*nTR* 100, *mFF* 1.0), SVR (γ 0.001, *cost* 100), and SVR (γ 0.001, *cost* 10), respectively. The

error bars for kinetics scenario are standard deviation of predictions on 50 independent iterations. Overall MSEs of the ML predictions are as follows: **A** *Kinetics*: 0.052, *Interpolation*: 0.056, *Extrapolation*: 0.320; **B** *Kinetics*: 0.044, *Interpolation*: 0.035, *Extrapolation*: 0.026; and **C** *Kinetics*: 0.204, *Interpolation*: 0.050, *Extrapolation*: 0.009

good. Interestingly, the *Extrapolation* scenario, modeled via SVR, matches the experimental results of poplar and straw quite well. In the case of cellulose, it slightly overpredicts the experimental results for higher temperatures. It should be noted that since the biomass type is hidden from the model, the ML model is unable to recognize that cellulose is relatively inert until 230 °C and then reacts with very fast kinetics [15], and therefore, the ML model predicts a continuous increase in carbon with the increase of HTC temperature that causes the observed discrepancy between the experimental results and ML predictions. It should be mentioned here that the trained ML model is applicable for the specific heating rate of 2 °C/min, and we have not tested the performance of this model on a different heating rate. For a different heating rate, one may need to re-train the ML model.

Figure 4 compares the ML predictions and results of isothermal HTC of cellulose at different temperatures. In the experiments, it is observed that the reaction proceeds slowly at 200 °C so that not much increase in the weight fraction of carbon is observed as a function of time (Fig. 4A). The ML prediction from the *Kinetics* scenario matches the experimental results quite well. The ML prediction from the *Interpolation* scenario also shows a good match to the data, but predicts a slightly faster kinetics. The *Extrapolation* scenario overpredicts the weight fraction of carbon. For the extrapolation scenario, there is no information on the kinetics at later times and with the biomass type hidden, the predictions are solely based on the kinetics observed in the training data. Cellulose being relatively inert at 200 °C justifies this observation. In Fig. 4B, the kinetics of HTC of cellulose at 230 °C is shown. A significant increase in the weight percent of carbon is observed after 40 min, implying that the reaction gets

initiated at this temperature. The *Kinetics* scenario is able to capture the initiation of the reaction as the information about the weight percent of other elements (N, S, H) are provided as inputs. However, both the *Interpolation* and *Extrapolation* scenarios fail to capture the initiation of the reaction and so deviate from the observed kinetics. Figure 4C shows the kinetics of HTC of cellulose at 260 °C. At 260 °C, the HTC reaction occurs rapidly and so the weight percent of carbon is close to 70% within the first few minutes of the reaction and then gradually increases with time. The *Kinetics* scenario captures the kinetics quite well whereas the *Interpolation* and *Extrapolation* scenarios underpredict the weight percent of carbon. Figure 5 shows the kinetics of isothermal HTC of poplar at 200 °C, 230 °C, and 260 °C. At 200 °C, the reaction has sluggish kinetics and a gradual increase in the weight percent of carbon is observed over time. All the three scenarios are able to capture the kinetics quite well at this temperature. At 230 °C and 260 °C, much faster kinetics are observed and some mismatch in the predictions of the ML models are observed even though the correct trends are captured. Figure 6 shows the isothermal HTC of straw at the three temperatures. The slow kinetics at 200 °C is captured well by the ML models. At higher temperatures, the ML models show the correct trend but there are some deviations from the experimental results. Figure 7 shows parity plots of the three scenarios for all the experiments. The overall R^2 for the *Kinetics*, *Interpolation*, and *Extrapolation* scenarios is 0.70, 0.82, and 0.71, respectively. This shows that the trained ML model worked reasonably well in the predictions for these three scenarios.

In Fig. 8, the relative importance of kinetics of N, S, and H, time, and temperature for predicting the weight percent of

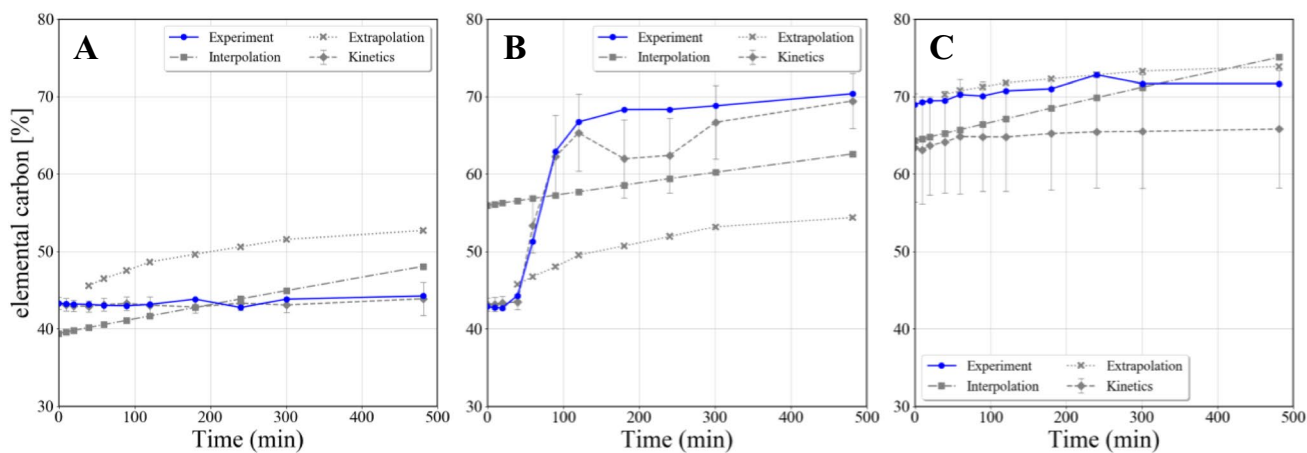


Fig. 4 Comparison between the predicted and the experimentally reported values of carbon content during the isothermal HTC of cellulose biomass at **A** 200 °C, **B** 230 °C, and **C** 260 °C. ML models used for *Kinetics*, *Interpolation*, and *Extrapolation* scenarios were RF (nTR 100, mFF 1.0), SVR (γ 0.001, $cost$ 100), and SVR (γ 0.001, $cost$ 10), respectively. The error bars for kinetics scenario are standard

deviation of predictions on 50 independent iterations. Overall MSEs of the ML predictions are as follows: **A** *Kinetics*: 0.002, *Interpolation*: 0.084, *Extrapolation*: 0.431; **B** *Kinetics*: 0.098, *Interpolation*: 1.205, *Extrapolation*: 2.365; and **C** *Kinetics*: 0.403, *Interpolation*: 0.166, *Extrapolation*: 0.018

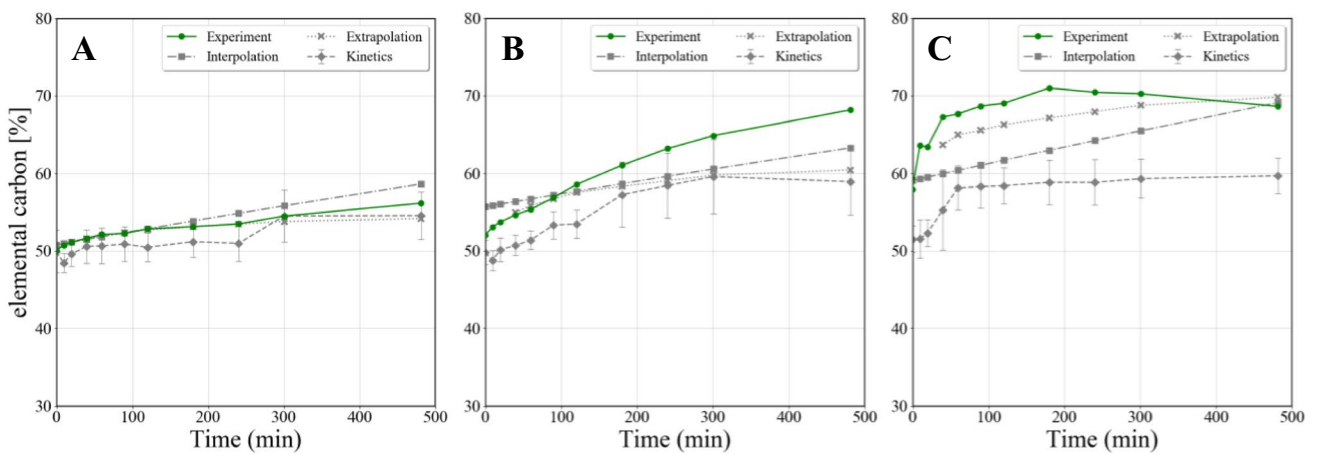


Fig. 5 Comparison between the predicted and the experimentally reported values of carbon content during the isothermal HTC of poplar biomass at **A** 200 °C, **B** 230 °C, and **C** 260 °C. ML models used for *Kinetics*, *Interpolation*, and *Extrapolation* scenarios were RF (*nTR* 100, *mFF* 1.0), SVR (γ 0.001, *cost* 100), and SVR (γ 0.001, *cost* 10), respectively. The error bars for kinetics scenario are standard devia-

tion of predictions on 50 independent iterations. Overall MSEs of the ML predictions are as follows: **A** *Kinetics*: 0.033, *Interpolation*: 0.012, *Extrapolation*: 0.007; **B** *Kinetics*: 0.272, *Interpolation*: 0.099, *Extrapolation*: 0.163; and **C** *Kinetics*: 1.325, *Interpolation*: 0.399, *Extrapolation*: 0.090

carbon for the *Kinetics* scenario is shown. To determine the significance of each feature, the value of that feature was randomly permuted to break the relationship between the model’s prediction and the feature. The importance of a feature was measured by calculating the increase in the model’s predicted MSE after permuting the feature [30]. A feature is “important” if shuffling its values increases the prediction error. This process was repeated for all features independently and the performance of the model was estimated by calculating the

MSE of the model’s prediction. For the *Kinetics* scenario, it appears that the features of *Time* and *Temperature* are relatively less important. However, such a conclusion is misleading. Both *Time* and *Temperature* affect the percentages H, N, and S. Therefore, the kinetics of H, N and S, in effect, capture the relationship of time and temperature and percentage of carbon. As shown in Fig. 8, the weight percent of hydrogen is a good predictor of the kinetics of carbon.

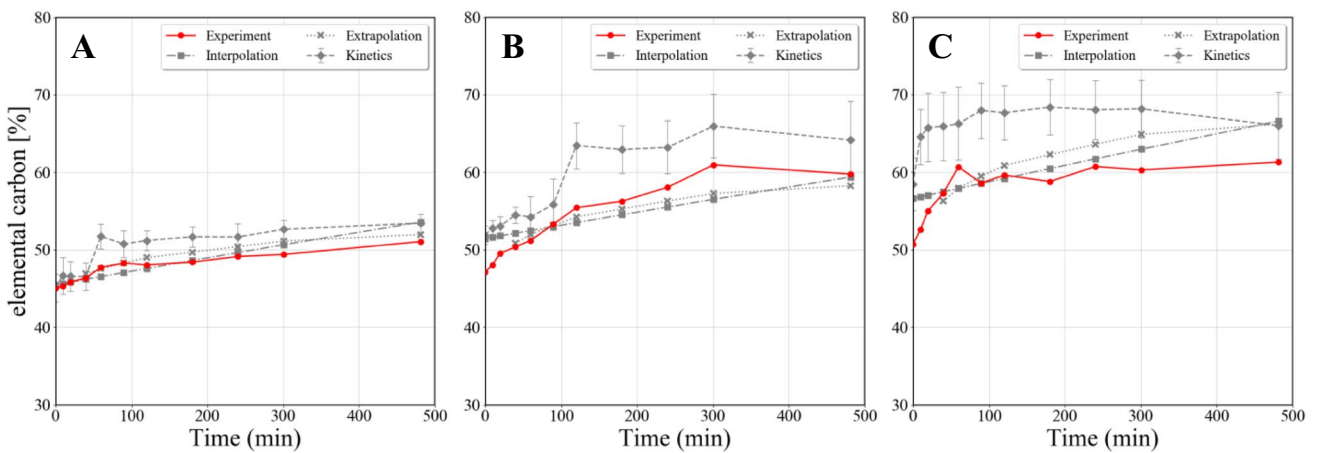


Fig. 6 Comparison between the predicted and the experimentally reported values of carbon content during the isothermal HTC of straw biomass at **A** 200 °C, **B** 230 °C, and **C** 260 °C. ML models used for *Kinetics*, *Interpolation*, and *Extrapolation* scenarios were RF (*nTR* 100, *mFF* 1.0), SVR (γ 0.001, *cost* 100), and SVR (γ 0.001, *cost* 10), respectively. The error bars for kinetics scenario are standard devia-

tion of predictions on 50 independent iterations. Overall MSEs of the ML predictions are as follows: **A** *Kinetics*: 0.072, *Interpolation*: 0.016, *Extrapolation*: 0.012; **B** *Kinetics*: 0.286, *Interpolation*: 0.079, *Extrapolation*: 0.033; and **C** *Kinetics*: 0.853, *Interpolation*: 0.111, *Extrapolation*: 0.112

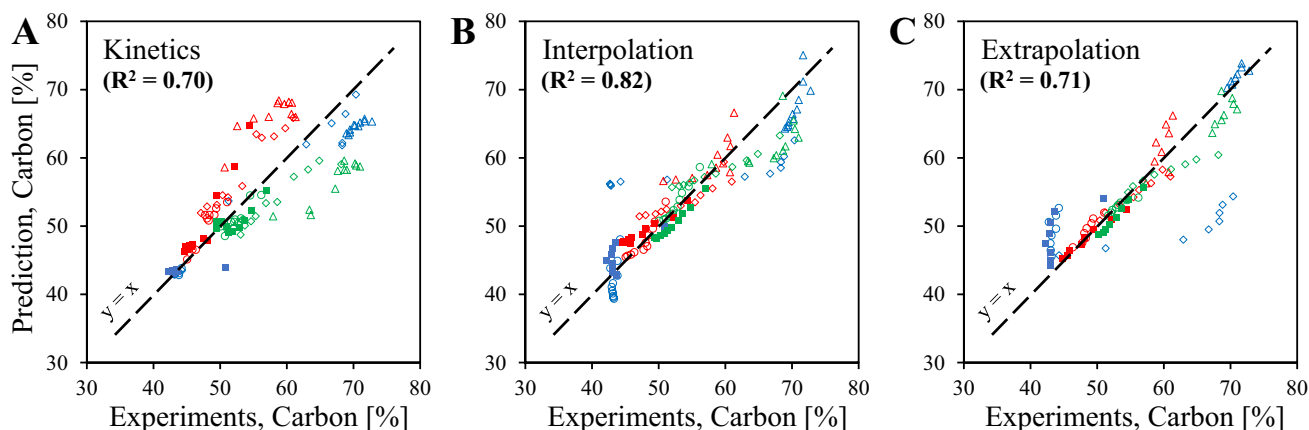


Fig. 7 Parity plots comparing the experimental and predicted values of percent carbon for all experiments for the three scenarios: **A** Kinetics, **B** Interpolation, and **C** Extrapolation. Symbols codes:

filled=dynamic experiments, hollow=isothermal experiments; blue=cellulose, green=poplar, red=straw; and circle=200 °C, diamond=230 °C, square=260 °C

4 Conclusions

In this work, we have explored the applicability of ML methods in predicting the kinetics of the HTC process of various biomasses. For this modeling, the information about biomass type was kept hidden from the ML models, so that one can represent practical scenarios where one is either interested in determining the kinetics of HTC of a new biomass type or when there is a mixture of various biomass types to be analyzed. We modeled three different scenarios, termed as *Kinetics*, *Interpolation*, and *Extrapolation*. In the *Kinetics* scenario, the weight percent of carbon was predicted, while the time-dependent weight percent of nitrogen, sulfur, and hydrogen were made available as input features. The *Interpolation* scenario was more stringent, wherein the weight percent of all elements were hidden, but their values at the initial, final, and middle time-point were made available to the ML model. The third scenario, *Extrapolation*, was even

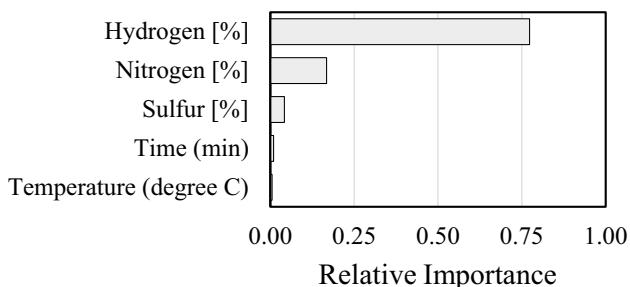


Fig. 8 Relative importance of the features in predicting the carbon content of different biomasses. Kinetics of hydrogen has the most significant impact on the performance of ML models. Since the kinetics of nitrogen, hydrogen and sulfur are hidden from the *Interpolation* and *Extrapolation* scenarios, their prediction of the experimental data showed deviations for the cases when the kinetics were fast

more stringent, wherein the weight percent of the elements were made available only for the first three time-points. We found that the trained ML models performed reasonably well for predicting the kinetics in the case of dynamic experiments, that is, the experiments in which the temperature of the reactor was increased at a fixed rate in time with an average mean squared error (MSE) of 0.09 ± 0.1 . The *Interpolation* and *Extrapolation* scenarios showed deviation in the predictions when the kinetics were fast, such as for cellulose at 230 °C, for which the MSE was found to be 1.20 and 2.36, respectively. In the case of straw, which represented a mixture of biomasses, the ML prediction showed some deviation from experimental data with the reported MSEs of 0.85, 0.11, and 0.11 for the *Kinetics*, *Interpolation*, and *Extrapolation* scenarios, respectively. To the best of our knowledge, these three ML modeling scenarios have not been investigated in any prior study of the HTC process. In a previous work, focused on predicting fuel properties of products obtained from HTC using ML, the reported R^2 was ≈ 0.90 [22]. In another work, the final yield of HTC products upon varying reaction conditions and biomass type was modeled via ML methods and the predictions were reported to have $R^2 > 0.95$ [24]. The three scenarios studied in this work are more stringent applications of ML as compared to the previous studies and so our overall R^2 range from 0.7 to 0.8. A disadvantage of this study was that the number of experimental data points was not large. Nevertheless, it is concluded that ML models are useful in modeling the HTC kinetics of various biomass types. ML models can be used for interpolating the kinetics data derived from experiments. Furthermore, these models are also useful for extrapolating the kinetics to longer times. It is recommended that for biomasses that show faster kinetics, more experimental data-points should be collected for accurate predictive modeling.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s13399-021-01858-3>.

Author contribution Conceptualization: Sumit Sharma, Toufiq Reza; Methodology: Sumit Sharma, Mohammadreza Aghaaminiha; formal analysis and investigation: Mohammadreza Aghaaminiha, Ramin Mehrani, Toufiq Reza, Sumit Sharma; writing—original draft preparation: Mohammadreza Aghaaminiha, Ramin Mehrani; writing—review and editing: Sumit Sharma, Toufiq Reza; funding acquisition: Toufiq Reza, Sumit Sharma; Resources: Sumit Sharma, Toufiq Reza; supervision: Sumit Sharma.

Funding This study is supported by the US Department of Agriculture National Institute of Food and Agriculture grant 2021–67022-34487. The computational resources were provided by NSF XSEDE grant DMR 190005.

Declarations

Competing interests The authors declare no competing interests.

Availability of data and material Experimental data previously published by our group is the basis of this study [15]. The data is available for download here: https://github.com/ssumit12/Hydrothermal_carbonization and is also made available in the Supplementary Information.

Code availability The machine learning Python code is available for download here: https://github.com/ssumit12/Hydrothermal_carbonization.

References

- Funke A, Ziegler F (2010) Hydrothermal carbonization of biomass: a summary and discussion of chemical mechanisms for process engineering. *Biofuels Bioprod Biorefining* 4:160–177. <https://doi.org/10.1002/bbb.198>
- Libra JA, Ro KS, Kammann C et al (2011) Hydrothermal carbonization of biomass residuals: a comparative review of the chemistry, processes and applications of wet and dry pyrolysis. *Biofuels* 2:71–106. <https://doi.org/10.4155/bfs.10.81>
- Reza MT, Andert J, Wirth B et al (2014) Review article: Hydrothermal carbonization of biomass for energy and crop production. *Appl Bioenergy* 1:11–29. <https://doi.org/10.2478/apbi-2014-0001>
- Mazumder S, Saha P, Reza MT (2020) Co-hydrothermal carbonization of coal waste and food waste: fuel characteristics. *Biomass Convers Biorefinery*. <https://doi.org/10.1007/s13399-020-00771-5>
- Reza MT, Uddin MH, Lynam JG, Coronella CJ (2014) Engineered pellets from dry torrefied and HTC biochar blends. *Biomass Bioenergy* 63:229–238. <https://doi.org/10.1016/j.biombioe.2014.01.038>
- Saha N, Volpe M, Fiori L et al (2020) Cationic dye adsorption on hydrochars of winery and citrus juice industries residues: performance, mechanism, and thermodynamics. *Energies* 13:4686. <https://doi.org/10.3390/en13184686>
- Reza MT, Rottler E, Tölle R et al (2015) Production, characterization, and biogas application of magnetic hydrochar from cellulose. *Bioresour Technol* 186:34–43. <https://doi.org/10.1016/j.biortech.2015.03.044>
- Fuertes AB, Sevilla M (2015) High-surface area carbons from renewable sources with a bimodal micro-mesoporosity for high-performance ionic liquid-based supercapacitors. *Carbon* 94:41–52. <https://doi.org/10.1016/j.carbon.2015.06.028>
- Adjuik T, Rodjom AM, Miller KE et al (2020) Application of hydrochar, digestate, and synthetic fertilizer to a *Miscanthus x giganteus* crop: implications for biomass and greenhouse gas emissions. *Appl Sci* 10:8953. <https://doi.org/10.3390/app10248953>
- Wirth B, Reza T, Mumme J (2015) Influence of digestion temperature and organic loading rate on the continuous anaerobic treatment of process liquor from hydrothermal carbonization of sewage sludge. *Bioresour Technol* 198:215–222. <https://doi.org/10.1016/j.biortech.2015.09.022>
- Reza MT, Freitas A, Yang X, Coronella CJ (2016) Wet air oxidation of hydrothermal carbonization (HTC) process liquid. *ACS Sustain Chem Eng* 4:3250–3254. <https://doi.org/10.1021/acsschemeng.6b00292>
- McGaughy K, Reza MT (2018) Recovery of macro and micro-nutrients by hydrothermal carbonization of septage. *J Agric Food Chem* 66:1854–1862. <https://doi.org/10.1021/acs.jafc.7b05667>
- Diakité M, Paul A, Jäger C et al (2013) Chemical and morphological changes in hydrochars derived from microcrystalline cellulose and investigated by chromatographic, spectroscopic and adsorption techniques. *Bioresour Technol* 150:98–105. <https://doi.org/10.1016/j.biortech.2013.09.129>
- Funke A, Ziegler F (2011) Heat of reaction measurements for hydrothermal carbonization of biomass. *Bioresour Technol* 102:7595–7598. <https://doi.org/10.1016/j.biortech.2011.05.016>
- Reza MT, Wirth B, Lüder U, Werner M (2014) Behavior of selected hydrolyzed and dehydrated products during hydrothermal carbonization of biomass. *Bioresour Technol* 169:352–361. <https://doi.org/10.1016/j.biortech.2014.07.010>
- Sztancs G, Kovacs A, Toth AJ et al (2021) Catalytic hydrothermal carbonization of microalgae biomass for low-carbon emission power generation: the environmental impacts of hydrochar co-firing. *Fuel* 300:120927. <https://doi.org/10.1016/j.fuel.2021.120927>
- Lynam JG, Coronella CJ, Yan W et al (2011) Acetic acid and lithium chloride effects on hydrothermal carbonization of lignocellulosic biomass. *Bioresour Technol* 102:6192–6199. <https://doi.org/10.1016/j.biortech.2011.02.035>
- Ischia G, Fiori L (2021) Hydrothermal carbonization of organic waste and biomass: a review on process, reactor, and plant modeling. *Waste Biomass Valorization* 12:2797–2824. <https://doi.org/10.1007/s12649-020-01255-3>
- Román S, Libra J, Berge N et al (2018) Hydrothermal carbonization: modeling, final properties design and applications: a review. *Energies* 11:216. <https://doi.org/10.3390/en11010216>
- Mitchell TM (1997) Does machine learning really work? *AI Mag* 18:11–11. <https://doi.org/10.1609/aimag.v18i3.1303>
- Géron A (2019) Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems. O'Reilly Media, Inc.
- Li J, Pan L, Suvarna M et al (2020) Fuel properties of hydrochar and pyrochar: prediction and exploration with machine learning. *Appl Energy* 269:115166. <https://doi.org/10.1016/j.apenergy.2020.115166>
- Li J, Zhu X, Li Y et al (2021) Multitask prediction and optimization of hydrochar properties from high-moisture municipal solid waste: application of machine learning on waste-to-resource. *J Clean Prod* 278:123928. <https://doi.org/10.1016/j.jclepro.2020.123928>
- Kardani N, Marzbali MH, Shah K, Zhou A (2021) Machine learning prediction of the conversion of lignocellulosic biomass during hydrothermal carbonization. *Biofuels* 0:1–13. <https://doi.org/10.1080/17597269.2021.1894780>

25. Famili A, Shen W-M, Weber R, Simoudis E (1997) Data pre-processing and intelligent data analysis. *Intell Data Anal* 1:3–23. <https://doi.org/10.3233/IDA-1997-1102>
26. Kotsiantis SB, Kanellopoulos D, Pintelas PE (2006) Data preprocessing for supervised learning. *Int J Comput Sci* 1:111–117
27. Aghaaminiha M, Ghanadian SA, Ahmadi E, Farnoud AM (2020) A machine learning approach to estimation of phase diagrams for three-component lipid mixtures. *Biochim Biophys Acta BBA - Biomembr* 1862:183350. <https://doi.org/10.1016/j.bbamem.2020.183350>
28. Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: machine learning in Python. *J Mach Learn Res* 12:2825–2830
29. Hoekman SK, Broch A, Robbins C (2011) Hydrothermal carbonization (HTC) of lignocellulosic biomass. *Energy Fuels* 25:1802–1810. <https://doi.org/10.1021/ef101745n>
30. Breiman L (2001) Random Forests. *Mach Learn* 45:5–32. <https://doi.org/10.1023/A:1010933404324>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.