Chemical Engineering Journal Advances 21 (2025) 100699

Contents lists available at ScienceDirect

Chemical Engineering Journal Advances

journal homepage: www.sciencedirect.com/journal/chemical-engineering-journal-advances

Development of a novel physics-informed machine learning model for advanced thermochemical waste conversion

Surika van Wyk

Biobased and Circular Technology Group, Energy & Materials Transition Unit, The Netherlands Organization for Applied Scientific Research (TNO), Petten, the Netherlands

ARTICLE INFO

Keywords:
Physics-informed
Machine learning
Fluidized bed
Biomass
Plastics
Waste

ABSTRACT

A physics-informed machine learning (ML) model, which incorporates the conservation of carbon mass, was developed to predict the product gas yield and composition for indirect gasification of waste in a fluidized bed. A dataset was compiled from experimental data of an in-house reactor, encompassing a wide range of feedstocks characteristics (biomass to plastics) and process conditions, which served as input for the model. Four data-driven models were trained and evaluated, with the XGBoost model having the best predictive accuracy (RMSE = 1.1 & R^2 = 0.99) and being adapted for the physics-informed model. The optimum physics contribution was 30 % (70 % data contribution) to maintain predictive accuracy (RMSE = 2.7 & R^2 = 0.95) and improve carbon closure. Feedstock properties were shown to have a higher feature importance compared to the operating conditions. The developed physics-informed model demonstrated the potential of ML models for the modelling of gasification of various waste streams. This is a promising first step towards improving data-driven ML models for application to thermochemical systems.

1. Introduction

The steady increase of waste, growing energy demand and rising emissions from increased fossil fuel consumption have necessitated the development of waste management strategies and reduction of fossil fuel demand by identifying renewable energy and fuel sources [1,2]. Currently most waste streams (plastic-rich or otherwise) are either landfilled or incinerated (to retrieve heat), which contributes to pollution of the soil, water, and air. However, these streams are a source of carbon which can be utilized for the production of new chemicals and fuels [3]. Biomass and biogenic waste in turn have been identified as attractive renewable fuel and energy sources [1,4]. Thermochemical conversion processes, such as pyrolysis and gasification, offer promising routes for the valorization of these heterogenous waste streams as well as producing renewable fuel and energy from biomass. These technologies help to reduce waste production, CO_2 emissions and fossil fuel dependence [2,5,6].

An important step towards further development and optimization of these thermochemical processes is modelling. Models provide information on product yields and insights on process performance and optimization, as well as enabling scale-up and real-time process control [7–9]. Modelling these processes is challenging due to the numerous complex reactions and extensive range of products. There are various modelling approaches for predicting the thermochemical behaviour and product distribution based on thermodynamics, kinetics, computation fluid dynamics (CFD) and statistical/empirical approaches (data-driven modelling). Thermodynamic models are relatively simple to implement (based on directly measured properties such as temperature, pressure and composition) and are applicable to any system (independent of reactor design). These models are however not suited for processes that are kinetically and hydrodynamically controlled such as fluidized bed reactors and processes operating at lower temperatures (750 – 1000 °C) [8,10–12]. Kinetic models are in turn suited for non-equilibrium conditions and provide more accurate predictions of product gas composition. These models' are, however, restricted to one specific system and require an extensive range of data to determine and validate the kinetic parameters [10–12]. CFD models deliver accurate predictions on the temperature and species profiles as well as product gas yield. As with the kinetic models, these models are limited to a specific reactor design, and are complex and computationally expensive to solve [7,8,10,11].

In recent years, machine learning (ML) models (data-driven) based on regression and neural networks (NN) have been developed for predicting product gas composition and yields as well as gasification efficiencies. These models aim at finding correlations between a set of input

E-mail address: surika.vanwyk@tno.nl.

https://doi.org/10.1016/j.ceja.2024.100699

Available online 22 December 2024

(features) and output (targets) variables, without knowledge on the behaviour of the physical system. These correlations are determined by employing various adaptive statistical and numerical models, which accurately describe linear relationships [8,10]. For thermochemical processes, the inputs are usually defined as the feedstock composition and process conditions, while the outputs are the product yields, compositions, and process efficiencies. ML models have been applied to various gasification and pyrolysis processes for different feedstocks and were shown to have predictive capabilities comparable to, or even better than, conventional models [7,13,14].

Yang et al. [6] compared various ML models to predict yields and composition for the gasification of municipal solids waste. The developed models accurately predicted, the char, tar, product gas yields and composition based on the feedstock characteristics and gasification conditions. Insights were gained into which parameters were key for optimizing the process in an economic and environmentally friendly manner. Wang et al. [15] employed various ML models to predict syngas composition and yields for biomass chemical looping gasification. With the models the key parameters were identified, which in turn reduced the number of experiments and led to an improved understanding of the design and optimization for a commercial process. To determine optimum gasification pathways for biomass to energy, Gil et al. [16] developed a ML model to predict the gas composition and yield for fluidized bed systems. The model enabled the selection of optimum process conditions for specific applications of the product gas. Xue et al. [17] trained five different ML models to predict the syngas composition during biomass gasification with steam in a fluidized bed reactor. The best model was selected and assisted with understanding the influence of various inputs and optimizing the process for the production of hydrogen-rich syngas.

The drawbacks of ML models in general are that they require a large amount of experimental data with sufficient variability for training and that interpretability is sometimes difficult for complex models due to the black-box approach [7,10]. Additionally, these models are purely data-driven and do not consider physical boundaries or constraints, such as conservation of energy and mass or governing equations, which could result in inaccurate predictions and scientific inconsistencies.

To overcome these limitations, models are being developed that combine physical mechanisms and constraints with ML algorithms to create physics-informed models (also known as physics-infused or hybrid models). The use of physics-informed models reduces the data requirement, as some complex relationships between variables are already provided, assists with model interpretability, and ensures scientific consistency. Furthermore, physics-informed models utilize the predictive capability of conventional ML models to describe relationships that are not always captured by physical models, and can be adapted and updated with new datasets to become more robust [18,19]. These models have been developed and applied for various chemical processes such as modelling the anaerobic digestion of palm oil [20], predicting the productivity of oil wells [21] and predicting NO_x emissions from coal-fired boilers [22]. For the modelling of gasification processes, Ren et al. [19] developed a physics-informed neural network (PINN) model to predict the concentration of the main biomass gasification products. The physical constraint was incorporated in the form of physical monotonicity between the syngas composition and the parameters equivalence ratio, moisture content and temperature. The developed PINN showed superior prediction performance compared to data-driven models and ensured that the results aligned with the established scientific principle of monotonicity.

Apart from Ren et al. [19] other studies on ML modelling for gasification processes are focussed on the development of data-driven models only. The aim of this work is to develop a novel physics-informed ML model, to predict the product gas yield and composition, based on feedstock properties and operating conditions, while adhering to the conservation of mass for carbon entering and exiting the system. The physics-informed model includes a wide range of feedstocks, ranging from biomass to plastics-rich, making it applicable to a diverse array of waste streams. The model will be tested with various contributions of the physical constraint, and the optimal contribution will be selected based on the predictive and carbon closure accuracy. Furthermore, the performance of the physics-informed model will be compared with the data-driven model based on predictive accuracy and model interpretability. The developed model offers the potential of smart experimental design for future tests and optimizing the thermochemical process conditions for treating biomass or plastic waste streams, while ensuring scientific consistency.

2. Methodology

2.1. Experimental set-up and data collection

The data used for the development of the models, was generated during experiments in an in-house lab-scale bubbling fluidized bed reactor located at the TNO site in Petten, The Netherlands. The reactor consisted of two zones, namely the bottom zone where the bed was located (internal diameter (ID) of 74 mm and height of 500 mm) and the top zone which was the freeboard (ID of 108 mm and height of 600 mm). The reactor was equipped with a screw feeding system and had a maximum feeding capacity of 1000 g/h. The system operated under atmospheric pressure and was equipped with electrical heating at the walls to provide external heating, when required (maximum operating temperature is 1100 °C). Fluidization gases namely nitrogen (N_2), steam and/or air were introduced in the bottom of the reactor. Additionally, tracer gases (neon or argon) were added with the fluidization gases to determine the product gas flow rate (through analysis of the tracer gas concentration at the outlet). The product gas exited at the top and passed through a cyclone to remove ash, chars, and entrained bed material, before the gas was analysed. The reactor bed could be filled with different bed materials (olivine, sand, sepiolite etc.), but for this study only sand as bed material was considered. Silica sand (d_{p50} = 310 μm) was used and the ratio of the particle diameter and internal diameter of the reactor was small enough to minimize the wall-effects.

The set-up was equipped with various thermocouples inside and above the bed to monitor the temperature profiles and to ensure that the temperature distribution inside the bed was uniform. Pressure sensors were present to measure the pressure drops over the bed. With the temperature and pressure drop profiles, the fluidization behaviour of the bed could be monitored to ensure that proper fluidization and conversion of feedstock were occurring during the experiments. In Fig. 1, a simple schematic of the reactor set-up is given.

The minimum fluidization velocity was calculated before each test and the flow settings were set so that the velocity in the bed was well above the minimum velocity. Prior to each experiment the reactor was heated to the desired temperature and fluidization gases (and tracer) were introduced to fluidize the bed. To prevent the backflow of hot gases

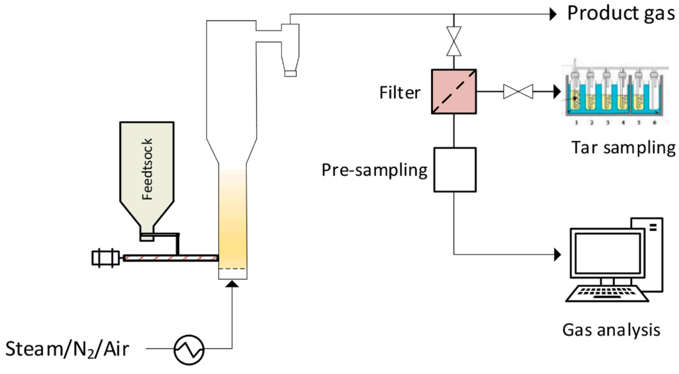

Fig. 1. Bubbling fluidized bed installation.

from the reactor to the feeding screw, 1 NL/min of N_2 was added to the feeding screw. Once stable conditions were reached, which were confirmed by the temperature and pressure drop profiles in the reactor, the feedstock was introduced via the feeding screw and the flow of fluidization gases was reduced as it was substituted by the product gas being formed. The product gas was analysed both online and through offline sampling and analysis to quantify the composition. The aim of the experiments was to measure the main components in the product gas and all carbon containing species to calculate the complete carbon balance over the system. For the online analysis an ABB gas analyser was used to continuously monitor the permanent gases (CO, CO_2, CH_4, H_2 and O_2). Furthermore a Varian micro-gas chromatographer (GC) analyser was used to analyse the product gas (N_2, Ne, Ar/O_2, CO, CO_2, CH_4, C_2H_2, C_2H_4, C_2H_6, C_3H_6, C_3H_8, benzene, toluene, H_2S and COS) at six minute intervals. For the offline analysis, gas bags were taken during steady state operation and analyzed using a Thermo Scientific Trace 1310 GC with a flame ionisation detector (GC-FID) to measure the concentrations of the C_1-C_6 hydrocarbons as well as H_2. Wet chemical sampling (tar sampling in isopropanol) of the gas was done according to the standardized tar measurement protocol and the samples were analysed afterwards using a GC-FID to determine the concentrations of benzene, toluene and tar components (molecular weight $>$ 92.1 g/mol) [23]. The full description of the analysis methods can be found in [24]. After each experiment (once feeding was stopped), the remaining char in the bed was combusted by introducing air to the system. The CO_2 (and CO) concentration was measured online and used to quantify the char yield.

With all the analysis information, the carbon mass flow of each component was calculated (based on concentration and product gas flow rate), and the complete carbon balance of the process was calculated based on the analysed carbon content of the feedstock, as shown in Eq. (1).

$$Carbon\ closure = \left(\frac{\sum_i C_{i,product\ gas} + \sum_i C_{i,tar} + C_{char}}{C_{feedstock}}\right) \times 100 \tag{1}$$

Considering the experimental and analytical errors of the set-up and analysis equipment, as well as sampling errors, the carbon balance was expected to close within 100 ± 10 %. Numerous test campaigns have been performed with the installation, investigating different feedstocks and conditions including test campaigns of the European project BRISK 2 [24–27].

From the data of these test campaigns, a dataset consisting of 231 samples was compiled. The feedstocks ranged from biomass-rich (beechwood, lignin, olive pomace, miscanthus) to plastic-rich (DKR 350, virgin polypropylene (PP) and polyethylene (PE)). The wide range of feedstocks ensures variability for both the feedstock characteristics as well as the product gas composition. For biomass-rich feedstocks, CO, CO_2, H_2 and CH_4 were the main products, while for plastic-rich feedstocks such a PE and PP, the main products were small hydrocarbons such as C_2H_4 and C_3H_6 as well as aromatics such as benzene. The variability ensures the model's robustness and validity across a wide range of feedstocks. In Fig. 2, a breakdown of the feedstock distribution for the compiled dataset is given.

Feedstocks with a plastic content of $\geq$ 50 % were regarded as plastic-rich feedstocks, while those with a lower plastic content ($<$ 50 %) were regarded as biomass-rich. For some feedstocks the plastics content could be easily determined, as these were model mixtures containing virgin plastic pellets such as PE and PP, mixed with beechwood, while for real waste streams such as refuse derived fuel (RDF) and textile waste it was difficult to determine the exact plastic content. Since the composition of these feedstocks can vary greatly depending on the biogenic content, these were presented as in-between feedstocks. Another real waste stream included in the dataset was DKR-350, which is known to be plastic-rich waste streams and was included under plastics.

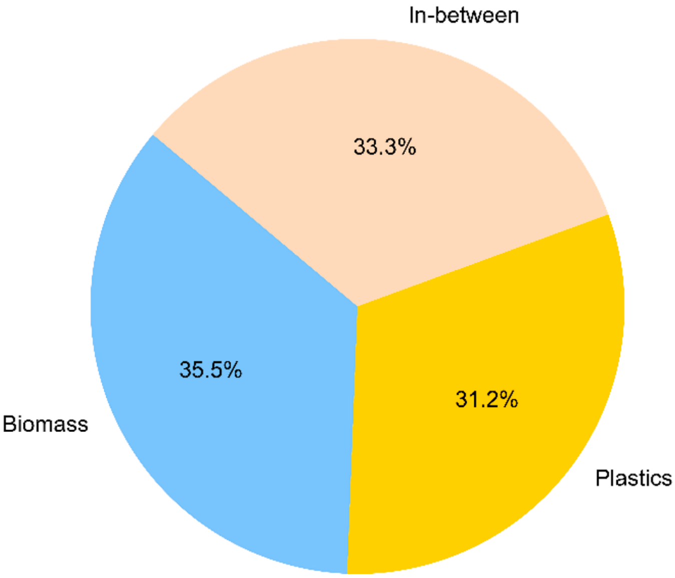

Fig. 2. Distribution of feedstock in the dataset.

Apart from the variation in feedstock, variations in temperatures and steam flows were also included in the dataset. The dataset was limited to only include indirect (or allothermal) gasification/thermal cracking tests with silica sand as bed material. Furthermore, only data with a carbon closure of 90 – 110 % was included, as the closure falls within the limits of the experimental error.

2.2. Data analysis and pre-processing

The first step in data analysis and pre-processing was to divide the complied data into input features and output targets. The input features were operating parameters (temperature, steam-to-carbon ratio (StC) and feed rate) and feedstock characteristics (elemental composition and ash yield). The output targets were the product gas yield and concentrations (vol.%, dry, tar-free, N_2-free basis) of the main components in the product gas (H_2, CO, CO_2, CH_4, C_2H_4, C_2H_6, C_3H_6 and benzene). In order to calculate the carbon balance for the system the carbon flow rates of the remaining C_2-C_6 hydrocarbons, char and tar (including toluene) were also added as output features. For each input feature and output target the statical parameters (namely mean, median, maximum, and minimum) were calculated to have an indication of the distribution and variation in the data.

The Spearman's correlation coefficients were calculated, to have an indication of the linear correlation between any two parameters. The correlation coefficient is a non-parametric statistic and does not require the data to be normally distributed as it utilizes monotonic functions to determine the correlation. The correlation coefficient ranges from −1 to 1, with a negative value indicating a negative linear correlation and positive value indicating a positive linear correlation. Strong correlations are considered to have values of $<$ 0.8 for theoretical data [6,28].

The correlations will not be used for the feature selection as the inputs have already been selected. The correlations will be used in support of the explainability of the model i.e. explaining the relations between certain inputs and outputs based on the corelations with other inputs. The heatmap with the correlation coefficients can be seen in the supporting information (SI).

After pre-processing (filter data for the mass balance closure and normalizing of the data for the support vector regression (SVR) model (see below)) and evaluation, the dataset was randomly divided into a training and test set. The training set consised of 80 % of the data and was used to train different ML models (see Section 2.3.1) for the comparison purposes and also the physics-informed model. The test set, consisting of the remaining 20 % of the data, was used for model evaluation.

Four machine learning models namely, decision tree (DT), random forest (RF), SVR and extreme gradient boosting (XGBoost) were trained and compared, before the physics-informed model was developed (see Section 2.3). The DT, RF and XGBoost models are insensitive to the variable scale and thus no data scaling was applied for these models. For the SVR model, the training and test sets were normalized using the Z-score standardization [15].

2.3. Modelling

2.3.1. ML model development

Four ML models were evaluated namely DT, RF, SVR and XGBoost. These models have been applied in previous studies for ML modelling of gasification and pyrolysis of both biomass and waste [6,7,16,17,29–31]. DT, RF and XGBoost are decision tree-based methods that can be used for classification and regression problems. The advantages of decision tree-based methods include simple implementation, minimal data pre-processing, and the ability to fit complex non-linear relationships, which is the case for thermochemical conversion processes. Tree-based models are also less sensitive to outliers which is the case with some experimental data points due to experimental errors. Depending on the complexity, the interpretability of the models is also easier compared to artificial neural networks (ANNs). Clear interpretability is important when trying to understand the influence of input parameters on the resulting product gas composition and optimizing conditions for maximum yield. It should, however, be mentioned that for ensemble methods the interpretability can also become complicated.

DT models are prone to overfitting, but with the use of ensemble methods the models are more robust as these algorithms have built-in regularization mechanisms. Two key ensemble methods are RF and XGBoost. The RF algorithm is based on a combination of tree predictors, with each tree being trained through bootstrapped samples from the training data. The trees are thus trained independently and combined by averaging the prediction to form the final model. RF models are able to handle complex datasets with interactions among variables and provides insights into feature importance which in turn assists with understanding the key drives in the thermochemical process. XGBoost is a gradient boosting algorithm where several weak leaners (small trees which are shallow) are created and combined to form one strong learner. With this algorithm (based on the greedy method) the fitting of each new tree is done to correct for the errors (residuals) of the combined previous trees, thus improving the overall model in a sequential manner [6,7]. XGBoost has a high predictive accuracy (compared to DT and RF), can handle complex interactions, which is the case for complex thermochemical processes and has built-in mechanisms to prevent overfitting which could happen with smaller datasets that include noise due to experimental errors [32].

The SVR model is the regression form of the support vector machine (SVM) model which is used for classification problems. The SVR algorithm constructs a hyperplane (or set of hyperplanes) in a higher-dimensional space to predict the continuous output of values and is applicable to both linear and non-linear regression problems. The advantage of this algorithm is that it yields good prediction accuracy for both small and large datasets, it is robust against outliers and can capture complex interactions between variables. It does, however, require more data pre-processing than decision tree models [7,33].

The ML models were programmed in Python using the Scikit learn library. The hyperparameters were tuned using Optuna which is an automatic hyperparameter optimization framework and is computationally less expensive compared to Grid Search as well as allowing for better integration with some ML frameworks [34].

The models' performance was evaluated and compared by comparing statistical performance parameters namely, the coefficient of determination (R^2) and the root-mean-square-error (RMSE). The model with the best performance will have a high R^2 value (close to one) and a low RMSE. The parameters were calculated for both the test and training sets and compared.

2.3.2. Physics-informed model

Following the evaluation, the XGBoost was selected to continue with the development of the physics-informed model (see Section 3.2). XGBoost is based on the gradient boosting algorithm, but is optimized by having built-in regularization to avoid overfitting and performs parallelization for faster computing time. Furthermore it has better handling of missing data, has various tree pruning methods and built-in cross-validation techniques [21,32]. XGBoost also supports custom loss functions, which is important when incorporating the physics-guided section to the model [35].

Physics were incorporated in the form of conservation of carbon mass. For training the model, the objective function is defined and optimized, which combines the loss function (L) and the regularization term (Ω) and is written as follow [32]:

$$Obj = \sum_i L_i(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{2}$$

The loss function is the measure of the predictive capability of the model and the mean squared error (MSE) was chosen as the loss function in this study:

$$L_i = (y_i - \hat{y}_i)^2 \tag{3}$$

Eq. (3) depicts the loss function for a data-driven model, where y_i is the actual value and $\hat{y}_i$ is the predicted value. For the physics-informed model, the loss function is modified in the following manner (see Eq. (4)):

$$L_i = (1-\lambda)(y_i - \hat{y}_i)^2 + \lambda\left(C_{i,in_{(experimental)}} - C_{i,out\,(predicted)}\right)^2 \tag{4}$$

Where $C_{i,in}$ represents the carbon entering the system (mass flow rate), which is calculated from the input parameters namely the feed rate and carbon content of the feedstock. $C_{i,out}$ is the carbon exiting the system which is calculated from the predictions made for the product gas composition and product gas yield as well as the tar, char and C_2-C_6 carbon flow rates. Lastly, λ is a parameter with a value between 0 and 1, employed to balance the data-driven part and the physics-informed part of the model. The higher the value of λ, the more physics-driven the model will be. The second term of the objective function is the regularization term that is used to penalize model complexity and avoid overfitting of the model (see, Eq. (5)):

$$\Omega(f_k) = \gamma T + \frac{1}{2}\beta \parallel \omega \parallel^2 \tag{5}$$

Where γ is the minimal loss reduction required for splitting a new leaf, $T$is the number of leaf nodes of the tree, β is the penalty term for the weight values of ω. During training these hyperparameters are tuned to optimize the trade-off between bias and variance of the model [18]. To train the model the following form of the objective function is optimized (see Eq. (6)):

$$Obj = -\frac{1}{2}\sum_{j=1}^{T} \frac{\left(\sum_{i\epsilon I_j} g_i\right)^2}{\sum_{i\epsilon I_j} h_i + \beta} + \gamma T \tag{6}$$

Where I_i is the total set of leaf nodes. The gradient g_i (first derivative of the loss function) and hessian h_i (second derivative of the loss function) also needs to be estimated. Based in the loss function defined in Eq. (4), the gradient and hessian are given as follow (see Eqs. (7) and (8)):

$$g_i = -2(1-\lambda)(y_i - \hat{y}_i) - 2\lambda\left(C_{i,in_{(experimental)}} - C_{i,out\,(predicted)}\right) \tag{7}$$

$$h_i = 2 \tag{8}$$

The derivation of the traditional XGBoost model as well as the explanation of the algorithm steps can be found in [32]. The same statistical parameters defined in Section 2.3.1 were also used for evaluating the performance of the physics-informed model. The procedure for fitting the physics-informed models is shown in Fig. 3.

Input features for the model are the elemental composition, feed rate and operating conditions, while the outputs are the product gas yield, concertation (vol.%) of the main gas components and the carbon flows of the char, tar and C_2-C_6 hydrocarbons. These input features were chosen as they are easily measured and can be directly used as input. The outputs can be easily compared with the product gas composition measured with the online analysers. For the training and testing of the physics-informed model, the inlet and outlet carbon flows were calculated beforehand. This saves computational time and simplifies the fitting and hyperparameter tuning, as the carbon outlet flows do not have to be calculated iteratively. The carbon entering the system was however calculated during the training procedure from the input parameters. From Fig. 3, it is seen that carbon outlet flows, based on the gas composition and product gas yield, are calculated prior to splitting the data into training and test sets. The training was done by minimizing the MSE for the carbon flows of the individual components (with the exception of the hydrogen which is kept as a concentration in vol.% and the product gas yield) and the total carbon entering and exiting the system (Eq. (4)).

The hyperparameters were tuned once more using the Optuna optimization framework and were tuned separately for each selected λ. The hyperparameters tuned were; gamma, subsample, min_child_weight, max_depth and eta. For the training and tuning, the multi_output_tree was selected as the multi_strategy parameter. This is a recent addition to the XGBoost model, where a single tree predicts all outputs simultaneously instead of the default method where a single tree is built for each strategy [35]. This method is useful for datasets where the outputs are correlated as in the case of gasification/pyrolysis products i.e. the presence of some components in the product gas promotes/inhibits the formation of others. Also, for the yields of the products i.e. more char can lead to lower tar and gas yields etc. The multi_output_tree was also selected when comparing the different data-driven models. In Fig. 4a simple layout of both strategies are compared.

After the models were trained and evaluated, the predicted carbon flows of the major components were converted back to volume concentrations using the predicted product gas yield and carbon flow rates. The carbon flows of the test and training sets were also converted back to product gas concentrations and compared with the predicted values.

The next step was to explore the model interpretability, and to gain insight into feature importance. For this, a Shapley Additive Explanations (SHAP) analysis was performed, which is a flexible technique developed from game theory, to assess both the global and local interpretability of the models. For the global interpretability, the sum of absolute SHAP values gives the overall importance of each feature. For the local interpretability, SHAP values are calculated for individual predictions, which shows the contribution of each feature and indicates

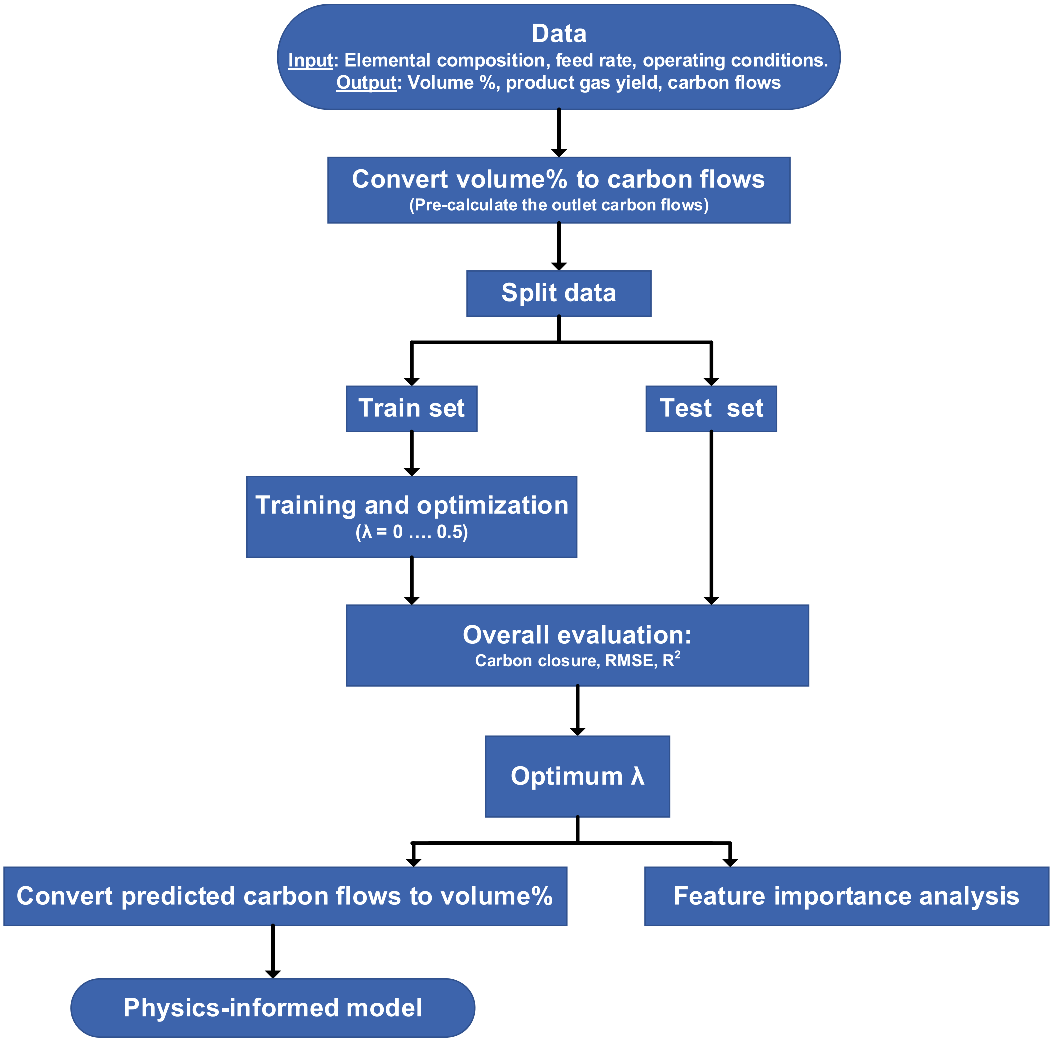
Data
Input: Elemental composition, feed rate, operating conditions.
Output: Volume %, product gas yield, carbon flows
Convert volume% to carbon flows
(Pre-calculate the outlet carbon flows)
Split data
Train set
Test set
Training and optimization
(λ = 0 .... 0.5)
Overall evaluation:
Carbon closure, RMSE, R2
Optimum λ
Convert predicted carbon flows to volume%
Feature importance analysis
Physics-informed model

Fig. 3. Workflow for the development of physics-informed ML model.

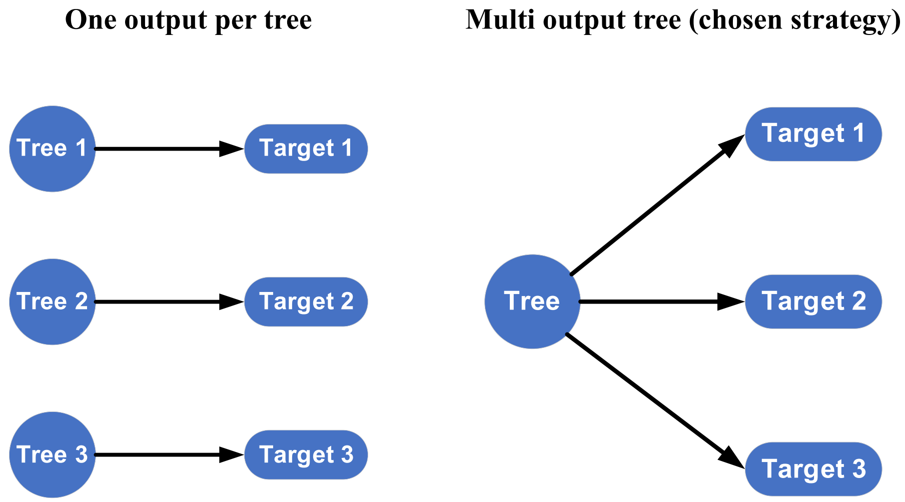

Fig. 4. Comparison of the multi_strategy feature for XGBoost.

whether it has a negative or positive effect on the prediction [15,36,37].

3. Results and discussion

3.1. Data analysis

In Table 1, the statics of the input features and output targets are provided.

For the feedstock composition, it is seen that the minimum values of the ash, oxygen and nitrogen were zero, which are for the virgin plastics pellets (PP or PE) used in some experiments. The use of these pellets also corresponds with the maximum carbon and hydrogen content. The elements with the widest range were carbon and oxygen, while nitrogen had the smallest range. For the process conditions the feed rate ranged from 136 to 519 g/h. The temperature extended from 693 to 856 °C and was more in the range for low temperature gasification and pyrolysis. The StC ratio varied from 0.3 to 1.3 (g/g). The value was never zero, because the moisture content of the feed also contributed to the StC ratio. The value was not too high as too much steam in the system would lead to quenching, making it difficult to reach the desired operating temperature. For the product gas, there were notable variations in the concentrations and flows, which is due to the variation in feedstock composition. The components C_2H_6 and C_3H_6 had the lowest maximum values and the smallest ranges (compared to the other concentrations in vol.% on dry basis (db)). These components are reactive and readily converted to other components especially at higher temperatures. For the thermochemical conversion of biomass-rich feedstocks the concentration of these hydrocarbons are usually low and in the case of C_3H_6 below the detection limit. The syngas components (H_2 and CO) had the highest maximum concentrations, which is mainly related to the thermochemical conversion of biomass-rich feedstocks. In general, when comparing median and mean with one another, it is seen that not all the data was normally distributed. In most cases the data was skewed to the right (positively-skewed, mean > median).

The distribution of the carbon closure of the dataset is shown in Fig. 5. The distribution of the carbon closure was skewed to the right

Table 1
Statical parameter for dataset.

	Unit	Min	Max	Mean	Median
Input features					
Feedstock composition [a]					
Ash (db)	wt.%	0.00	15.4	2.43	1.40
C (db)	wt.%	47.0	85.7	61.3	54.8
H (db)	wt.%	4.30	14.3	8.06	6.00
O (db)	wt.%	0.00	48.8	26.2	32.2
N (db)	wt.%	0.00	1.80	0.85	0.60
Operating conditions					
Feed rate (db)	g/h	136	519	274	215
Temperature	°C	693	856	767	754
StC[b]	g/g	0.30	1.27	0.78	0.78
Output (Target) [c]					
Product gas yield [d]	$Nm^3/kg_{feed(db)}$	0.42	1.25	0.83	0.81
H_2	vol.%	6.14	42.3	27.9	29.2
CO	vol.%	0.28	45.7	21.1	25.2
CO_2	vol.%	0.16	33.8	13.9	16.0
CH_4	vol.%	4.47	33.9	15.8	13.3
C_2H_4	vol.%	2.18	37.8	11.6	5.66
C_2H_6	vol.%	0.00	4.42	1.06	0.73
C_3H_6	vol.%	0.00	10.2	1.49	0.57
Benzene	vol.%	0.43	11.9	3.05	2.54
Char	g_{carbon}/h	1.49	67.1	20.1	13.4
Tar	g_{carbon}/h	2.16	77.6	19.4	14.3
C_2-C_6	g_{carbon}/h	0.00	17.6	3.10	1.66

[a] db – dry basis.
[b] StC includes the moisture content of the feedstock.
[c] All gas concentrations are on a dry, N_2-free, tar-free basis.
[d] Product gas yields refers to dry, N_2-free, tar-free product gas flow divided by feed rate on db basis.

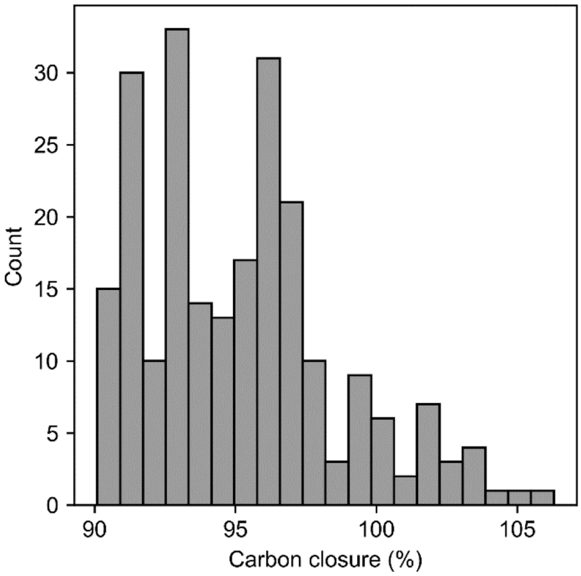

Fig. 5. Carbon closure distribution of the dataset (90 – 95: 51.5 %; 95–100: 38.5 %; 105–100: 9.1 %; >105: 0.9 %).

(positively skewed), and the majority of the carbon closures were below 100 %. More than half of the carbon closures were below 95 %, with only 10 % of the data having a carbon closure greater than 100 %. This indicates that in most cases the carbon exiting the system was underestimated, which is due to experimental error or a minimal amount of missing carbon. The values were however still within the experimental error range (see Section 2.1).

3.2. Data-driven models

After the evaluation of the data statistics, different ML models (only data-driven) were trained and compared. The comparison of the models in terms of RMSE and R^2 is given in Fig. 6.

From Fig. 6a it is seen that the tree-based ML models performed better than the SVR model by having lower RMSE values for the test set. For the tree-based models, the ensemble methods RF and XGBoost performed better than the DT model, by have lower RMSE scores for both the training and the test sets. XGBoost had the lowest RMSE value for the training (0.37) and test (1.12) set. For the R^2 values, all models had values close to one for both the training and test sets. The XGBoost model also had the highest R^2 value (0.955) for the test data, followed by the RF model (0.954), while for the training data the DT model had the highest R^2 (0.994) followed by XGBoost (0.991).

As stated in Section 2.3.1, ensemble methods, compared to tree-methods, are more robust due to their built-in regularization terms and are able to handle complex interactions among variables, which is the case for the complex thermochemical conversion process where the variables are interrelated. Furthermore, due to the algorithm design of the XGBoost, it generally has a high predictive accuracy, compared to other tree-based models, which was also seen for this study. The SVR model is also able to handle complex interactions between variables (see Section 2.3.1), however for this case its predictive accuracy was less compared to XGBoost.

Overall, the SVR model demonstrated the lowest performance, whereas the XGBoost model exhibited the highest performance, followed by the RF model, in terms of prediction accuracy. The superior performance of the XGBoost model (and gradient boosting regression (GBR) ML models in general) in predicting the product yields and gas composition, has been shown in past studies for gasification as well as for various other thermochemical processes such as hydrothermal liquefaction and carbonization, and pyrolysis [6,16,17,29,30,33,37,38]. For gasification specifically, Yang et al. [6] indicated that the GBR model had the highest prediction accuracy (in comparison to SVR and RF) in terms of predicting the syngas yield and composition from the gasification of municipal solids waste. Li et al. [29] compared the performance of NN, SVR, RF and GBR for predicting the product yields and syngas composition of biomass waste gasification and indicated that GBR showed superior prediction accuracy compared to other models. Lastly, Xue et al. [17] evaluated five different ML models including DT and RF, and indicated that XGBoost was the best performing model for predicting the syngas properties of biomass gasification with steam. Based on the results in Fig. 6 as well as the results of previous studies, the XGBoost model was selected for further development of the physics-informed ML model.

3.3. Physics-informed ML model

For the physics-informed ML model, multiple models were trained and optimized (hyperparameter tuning) based on the selected λ, which assigns the weight given to physics-informed part of the ML model. λ was varied from 0 to 0.5 in steps of 0.1 and the RMSE and R^2 values were calculated to evaluate the model's predictive accuracy. To evaluate the improvement of the carbon balance closure, the following approach was employed. The total number of data points with a carbon balance closure between 95 and 105 % was calculated for each model based on the predictions of the test set and compared to the total number of experimental data points in the test set for the same carbon closure interval. The ratio between the two was then estimated as shown in the Eq. (9).

$$Ratio \ \pm 5\% = \frac{N_{model,\ test}(95\% \ \leq Carbon\ closure \ \leq 105\%)}{N_{exp,test}\ (95\% \ \leq Carbon\ closure \ \leq 105\%)} \qquad (9)$$

The ratio for carbon closure between 99 and 101 % was also estimated and used for the comparison. In Fig. 7, the comparison for the different λ values is shown.

The results show that as the contribution of physics increased, the predictive capability of the model decreased due the RMSE increasing and R^2 decreasing. The increase of the RMSE was minimal going from a pure data-driven model to a model with 10 % physics contribution, however, as the physics contribution increased ($\lambda > 0.2$), the increase became more notable. For $\lambda > 0.5$, the predictive accuracy rapidly decreased (RMSE > 4.5 & $R^2 < 0.80$), and for this reason higher λ values were not evaluated. The carbon closure improved with the physics contribution. The improvement was indicated by the increase in ratios of predicted carbon closure to the experimental carbon closure (Eq. (9)) for the chosen intervals (100 ± 5 % and 100 ± 1 % respectively). For the data-driven results, the closure remained the same as for the experimental set for Ratio ± 5 % (equal to 1.0) and slightly decreased for Ratio ± 1 % (equal to 0.33). The ratio of carbon closure between 95 and 105 % (Ratio ± 5 %) increased for λ of 0.1, while for higher λ values the ratio

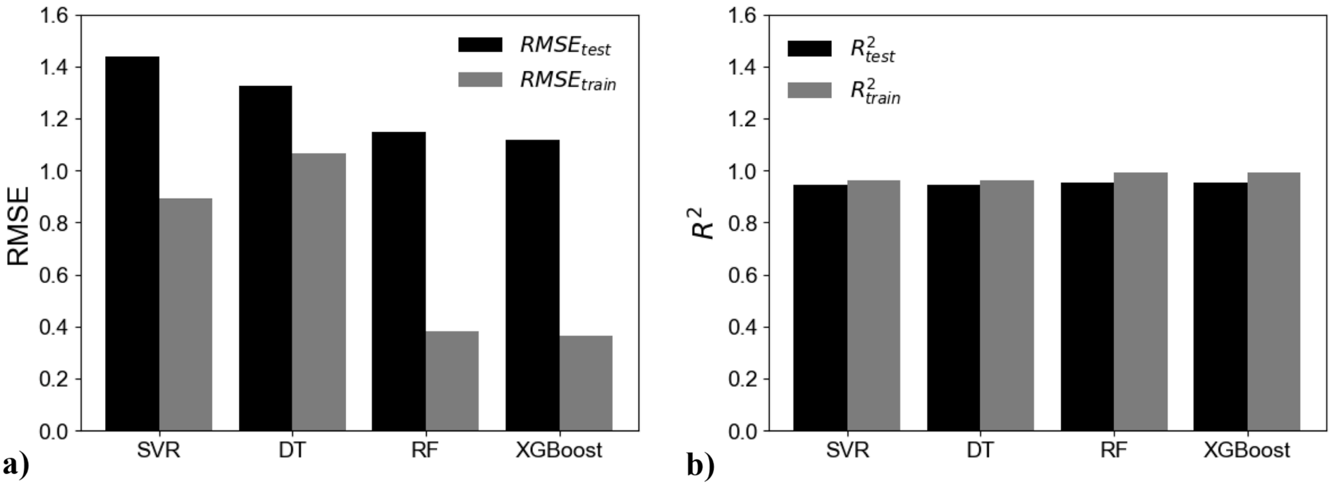

Fig. 6. Comparison of the performance of different data-driven ML a) RMSE b) R^2.

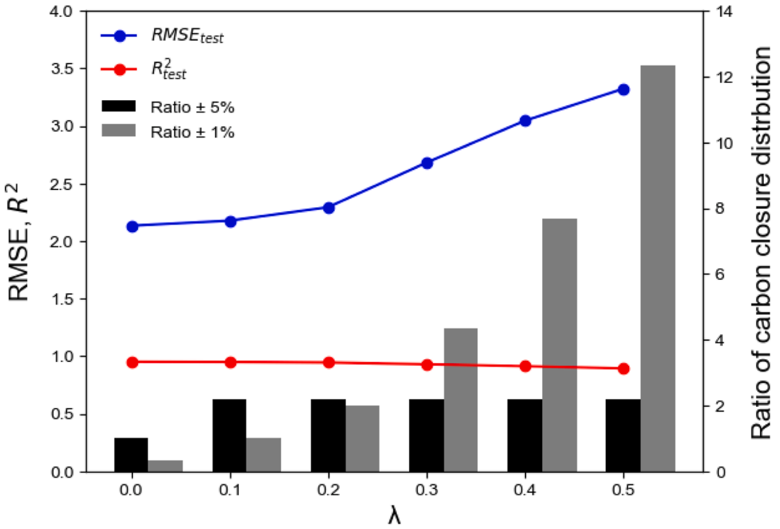

Fig. 7. Comparison of statistical parameters and carbon closure improvement for physics-informed models with different λ-values (Predictions are based on the test set).

remained the same. The ratio of carbon closure between 99 and 101 % increased exponentially with the addition of physics constraint. The trend in the ratios suggest that for a λ ≥ 0.1 the amount of data points with a carbon closure between 95 and 105 % is not increasing, but the distribution is only becoming narrower around a median of one (100 % carbon closure).

Fig. 7 was also drawn for the training set and can be found in the SI. Similar trends were seen for the statistical parameters. For λ > 0.1, the increase in the RMSE was more significant than for the test set predictions, but the absolute values were lower, e.g., for λ of 0.5, $\mathrm{RMSE}_{\mathrm{test}}$ was 3.3 and $\mathrm{RMSE}_{\mathrm{train}}$ was 2.9. For the carbon closure, the Ratio ± 5 % increased up to λ of 0.3 after which the ratio decreases minimally, which is due to the overprediction of carbon exiting the system leading to carbon closures greater than 105 %. This was not seen for the test data where the ratio remained constant, however as will be shown below, the distribution of carbon closure does shift towards higher values. The Ratio ±1 % increased exponentially similar to Fig. 7.

Fig. 8 shows the distributions of the carbon closures based on the predictions from the test set for different λ values as well as the carbon closure of the experimental test set.

It is seen that for the data-driven model and models with a low λ, the carbon exiting the system was underpredicted, leading to a carbon closure lower than 100 %. This was due to the training data (and test data) mainly having data points with a carbon closure less than 100 %, as shown in Fig. 5 (see Section 3.1, right-skewed distribution). As λ increased the distribution of the carbon closure improved, with the ratio of $\mathrm{carbon}_{\mathrm{out}}$ to $\mathrm{carbon}_{\mathrm{in}}$, becoming normally distributed with a median of one. The distribution also became narrower. For a λ of 0.1, all carbon closures were already ≤ 95 %, which was why no significant increase was seen in the Ratio ± 5 % for higher values of λ (see Fig. 7). Fig. 8 shows that already with a 10 % physics-contribution, a significant improvement was made in the distribution of the carbon closure. In Section 3.1 it was shown that for the overall dataset, 51 % of the data points were below 95 % carbon closure (see Fig. 5). It is, however, noted that as λ continued to increase the distribution of the carbon closures shifted more to the left (higher carbon closures). In some cases, the carbon exiting the system was overpredicted, leading to carbon closures greater than 100 %. For a λ of 0.3, 85 % of the carbon closures were between 95 and 100 %, while the 15 % were above 100 % (2.12 % was above 105 %).

The effect of changing the carbon closure threshold for the data selection to 85 and 95 % respectively was also investigated with the results shown in Fig. 9.

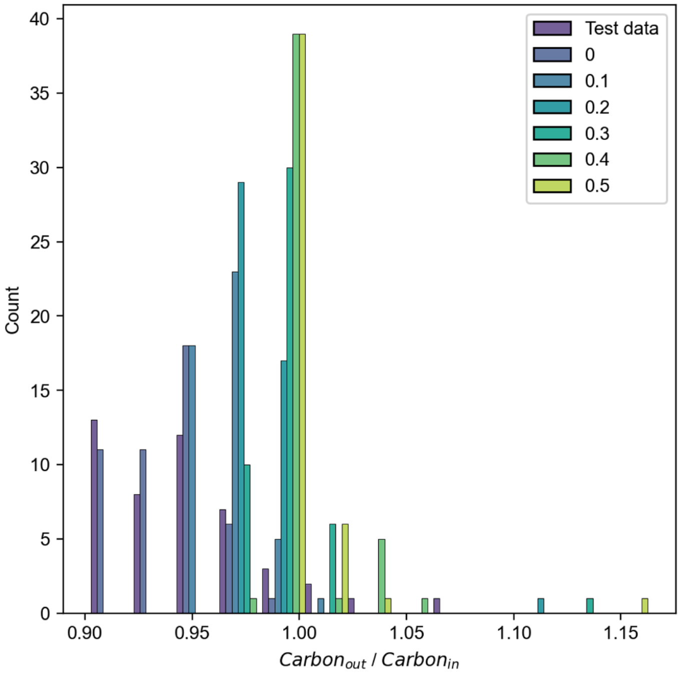

Fig. 8. Comparison of carbon distributions with experimental test data and models with different λ values.

From Fig. 9a & b, it is seen that the trends for both the statistical parameters and carbon closure evaluation were comparable with Fig. 7. For Fig. 9a, the amount of data points in the total dataset increased to 324 by lowering the carbon closure threshold to 85 %. Compared to Fig. 7, the RMSE values were higher for the given λ values, due to the larger changes in concentrations (carbon containing species) and product gas yields for many data points to improve the carbon closure. The R^2 remained similar to that of 90 % carbon closure threshold. The increase in the Ratio ± 5 % followed a similar trend i.e., increasing up to a λ of 0.2 and then remaining constant. The Ratio ± 1 % continued to increase with λ, however when comparing the results with Fig. 7, it is seen that the ratio increase was less than for the dataset with a 90 % carbon closure threshold.

Upon increasing the carbon closure threshold to 95 %, the amount of

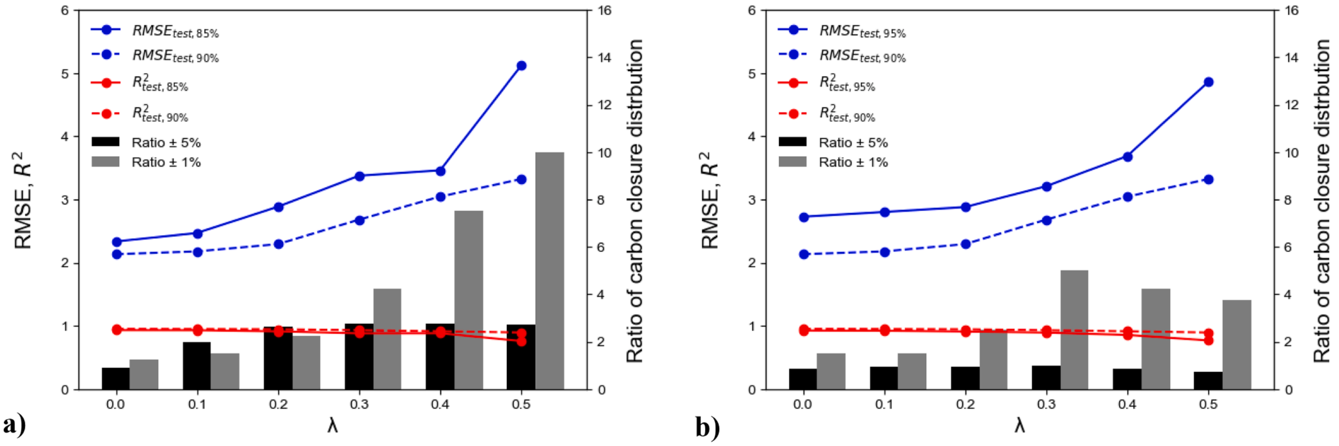

Fig. 9. Comparison of statistical parameters and carbon closure improvement for physics-informed models with different λ-values a) Dataset with carbon closure interval 85 – 115 % b) Dataset with carbon closure limit interval 95 – 105 %.

data points in the complete dataset decreased to 112 and the variability of the data decreased as many feedstocks were removed from the dataset. By comparing the results with Figs. 7 and 9a it is seen that the RMSE was higher and also continued to increase with the increase in λ. For R^2, both the values and trends were similar. The Ratio ± 5 % did not increase with λ values as the threshold for the experimental data was already 95 %. The Ratio ± 1 % increased up to a λ of 0.3 after which it decreased together with the Ratio ± 5 % due to the overprediction of carbon exiting the system, leading the closures to be higher than 101 and 105 %, respectively. Overall, when comparing the results, it is seen that a carbon threshold of 90 % was the optimum selection based on the statistical parameters and carbon closures.

3.3.1. Model performance for individual outputs and selection of λ

To further assess and compare the predictive capability of data-driven and physics-informed models, the statistical parameters for the individual outputs were compared for different values of λ. The results are reported in the SI. An increase in the RMSE with the increase in λ, is seen for all outputs, while R^2 in turn decreased with λ. The largest increase in RMSE was seen for concentrations of H_2, CO and CO_2 and the char carbon flow, while for the C_2-C_6 carbon flow the RMSE was consistently higher. For the concentration of H_2, CO, CO_2 and also for the tar carbon flow, the largest increase in RMSE was seen for $\lambda > 0.3$ (for the char flow it was $\lambda > 0.2$). The results suggest that λ should not be greater than 0.3 in order to maintain the predictive capabilities of the model. The parity plots (for both test and training set) of the data-driven model and that of the physics-informed model ($\lambda = 0.3$) are presented in Fig. 10.

The RMSE and R^2 for both models are also compared in a separate figure for each output to have a better overview. The figure can be found in the SI. From the results it is seen that for both models (data-driven and physics-informed) the RMSE value for the various outputs was ⟨ 3, while the R^2 was ⟩ 0.6 and for most outputs > 0.9. The highest RMSE (and lowest R^2) value for both models was seen for the C_2-C_6 carbon flow, where the output was scattered and most of the predictions for the training and test data were outside the ± 10 % interval. The deviation was normally distributed, it is thus not a systematic over- or under-prediction of the data. The output that had the lowest RMSE score (≤ 0.06) for both the data-driven and physics-informed model was the product gas yield, which is important as this output parameter influences the predictions of the concentrations for the major gas components (see Fig. 3). The highest R^2 value (1.00) was seen for various outputs for the data-driven model and for the physics-informed model for the tar output.

In general, it is seen that the predictive capability of the model decreased with the contribution of the carbon balance constraint, which is in line with the results of previous sections. The largest change was seen for the RMSE values, while for the R^2 values only slightly decreased with the physics contribution. The majority of the predictions (both test and training) fall within a 10 % deviation from the experimental results. For lower concentrations, some of the data points were above the 10 % deviation (overpredicted). From the Fig. S.3 in the SI, it is seen that the largest changes would be seen for the concentrations of H_2, CO and CO_2 as well as the char carbon flow, however upon comparing the parity plots of these outputs it is seen that the majority of the predictions fell within the ± 10 % deviation, only for the lower concentrations the some values were above the + 10 % threshold.

From the results of the overall comparison, as well as for the individual outputs, it is concluded that for the physics-informed model the optimum value of λ is 0.3. For a value of 0.3, it is seen that the carbon closure improved to 2.2 for Ratio ± 5 % and to 6.7 for Ratio ± 1 %. From the carbon distribution it is seen that the majority of the carbon closures were between 95 and 100 %, which was a significant improvement compared to the experimental dataset. Furthermore, the predictive accuracy was maintained with the model having an overall RMSE of 2.68, which was only 0.55 higher than that of the data-driven model (RMSE = 2.13). The overall R^2 of the data-driven model was 0.95, while that of the physics-informed model was 0.93.

3.3.2. Model interpretability

To have a better understanding of the feature importance and gain meaningful insights from the model results, the SHAP analysis was performed. It should be noted at the time of writing this paper the TreeExplainer did not yet support the vector leaf output of the multi-output feature of XGBoost V2.0 and thus the normal explainer was used for the interpretability analysis. The analysis was performed for both the data-driven and physics-informed model ($\lambda = 0.3$) and compared to evaluate how the feature importance changed with the carbon closure constraint. The analysis was done for the total dataset, five randomly selected datasets (47 datapoints in each set) taken from the main dataset as well as for a dataset containing only biomass and plastics feedstock respectively. The results for the total dataset, one randomly selected dataset, the biomass and plastics are presented in Fig. 11 for the data-driven model. The SHAP results for the other four datasets tested can be found in the SI for both the data-driven and physics-informed model. Note that the outputs presented are in the form of the direct outputs of the model, thus in g/h of carbon flow (except H_2 vol.%), instead of concentrations as presented in the parity plots.

The SHAP plots present the absolute mean feature importance score of all the model inputs. The higher the score the more important the feature and the features are also ranked according to the score. The absolute mean score is sum of the feature importance scores for the

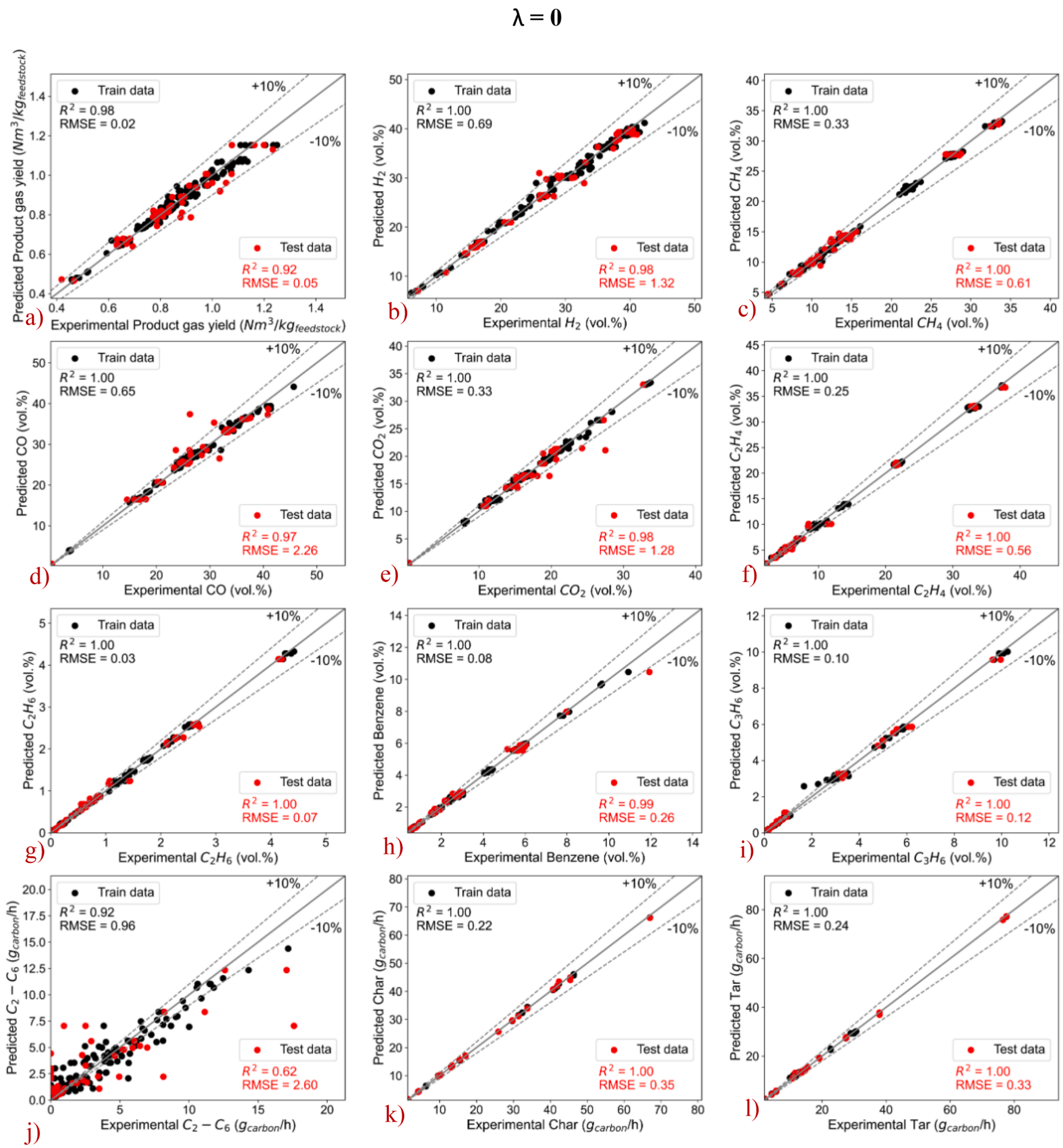

Fig. 10. Parity plot comparison of model outputs a-l) data-driven model m-x) physics-informed model (λ = 0.3).

individual outputs, which is indicated with the different colours (see the legend of Fig. 11a–d). The larger the impact, the larger the mean absolute value. For the data-driven model, the carbon content of the feedstock was the most important feature, followed closely by the feed rate. For the analysis of the whole dataset, four of the five randomized datasets and the biomass dataset, the carbon content had the highest feature importance score, while for set two, the feed rate had the highest feature importance score (see SI). The process conditions namely bed temperature and StC where in most cases the third and fourth most important feature, however, the scores were significantly lower compared to the carbon content and feed rate. The concentrations of the other components in the feedstock appeared to have minimal effect, with the nitrogen content of the feedstock having the lowest feature importance for all the datasets. For the plastics dataset, the feature importance scores differed notably from the other datasets, with the feed rate being the most important feature, followed by the process conditions, StC and bed temperature. The feedstock composition appeared to have a minimal feature importance, with nitrogen again having the lowest score.

For the datasets where the carbon content had a high feature

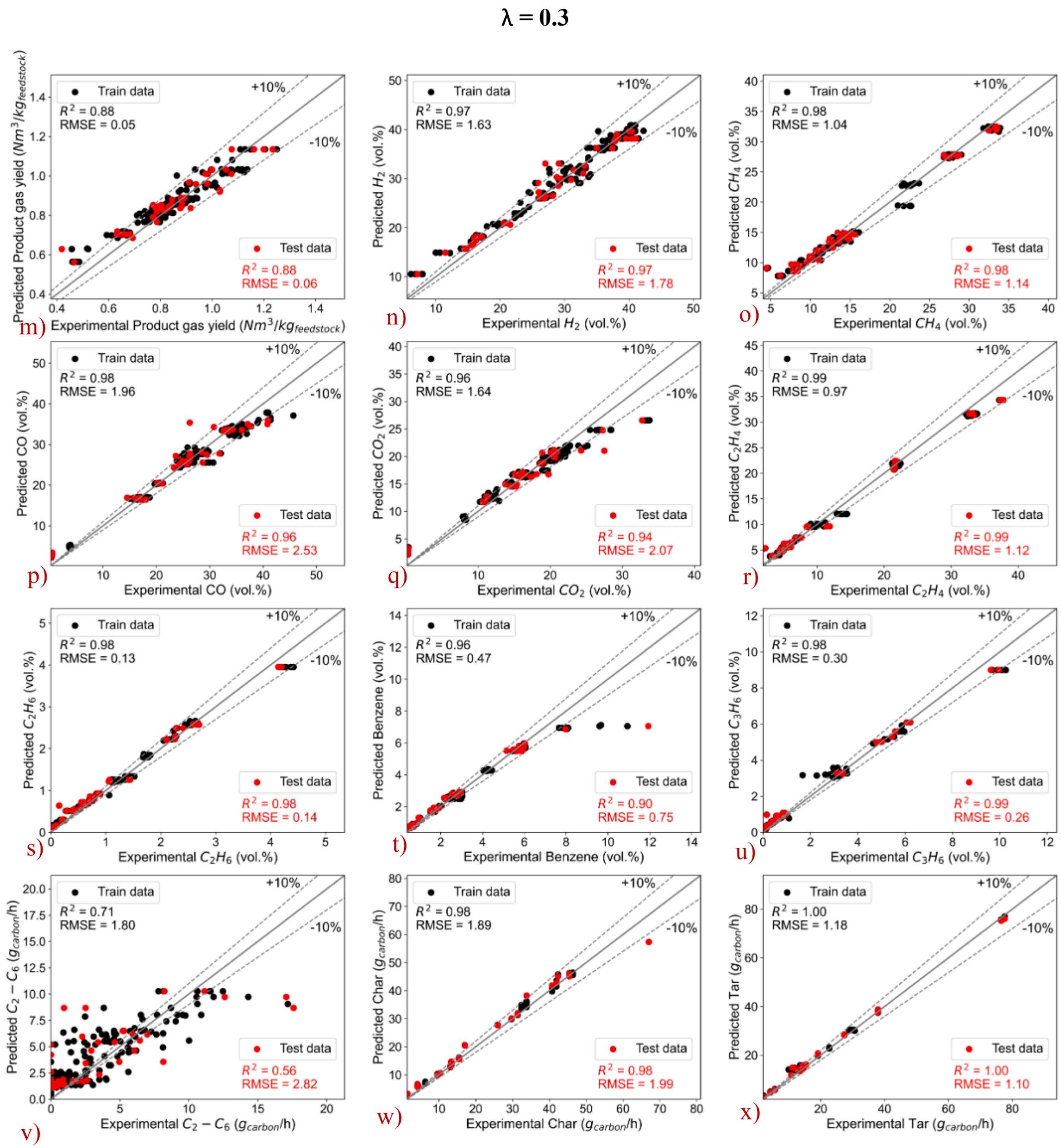

Fig. 10. (*continued*).

importance, the main targets that contributed most to the feature importance score were the H_2 concentration, CO, C_2H_4, benzene and tar carbon flow rates. This indicates that the carbon content had the largest impact on these outputs. For the feed rate feature importance score the CH_4, CO, CO_2, C_2H_4 and char carbon flow rates were the outputs with the highest scores and thus most influenced by the feed rate. Based on the complete dataset, the SHAP analysis was also done for each individual target to have an indication of the influence of each feature on each output. The results of the analysis can be found in the SI (Fig. S.7).

Same as for the global overview the feed rate and carbon content were the two inputs that had the highest feature importance for all outputs. Most outputs increased with the increase in feed rate, which was expected as more products are formed when the feed rate increases. The product gas yield was the only output that decreased with the feed rate, which could be due to the manner in which it was presented, namely the volume of gas divided by the feed rate. Additionally, for the biomass-rich feedstocks, the feed rate was usually higher (average 291 g/h). Compared to plastic gasification, biomass gasification typically results in higher char yields and consequently lower product gas yields. This in turn could cause the model to display the feed rate as having a negative impact on the product gas yield. The carbon content of the feed also had a negative correlation with the product gas yield. This is once

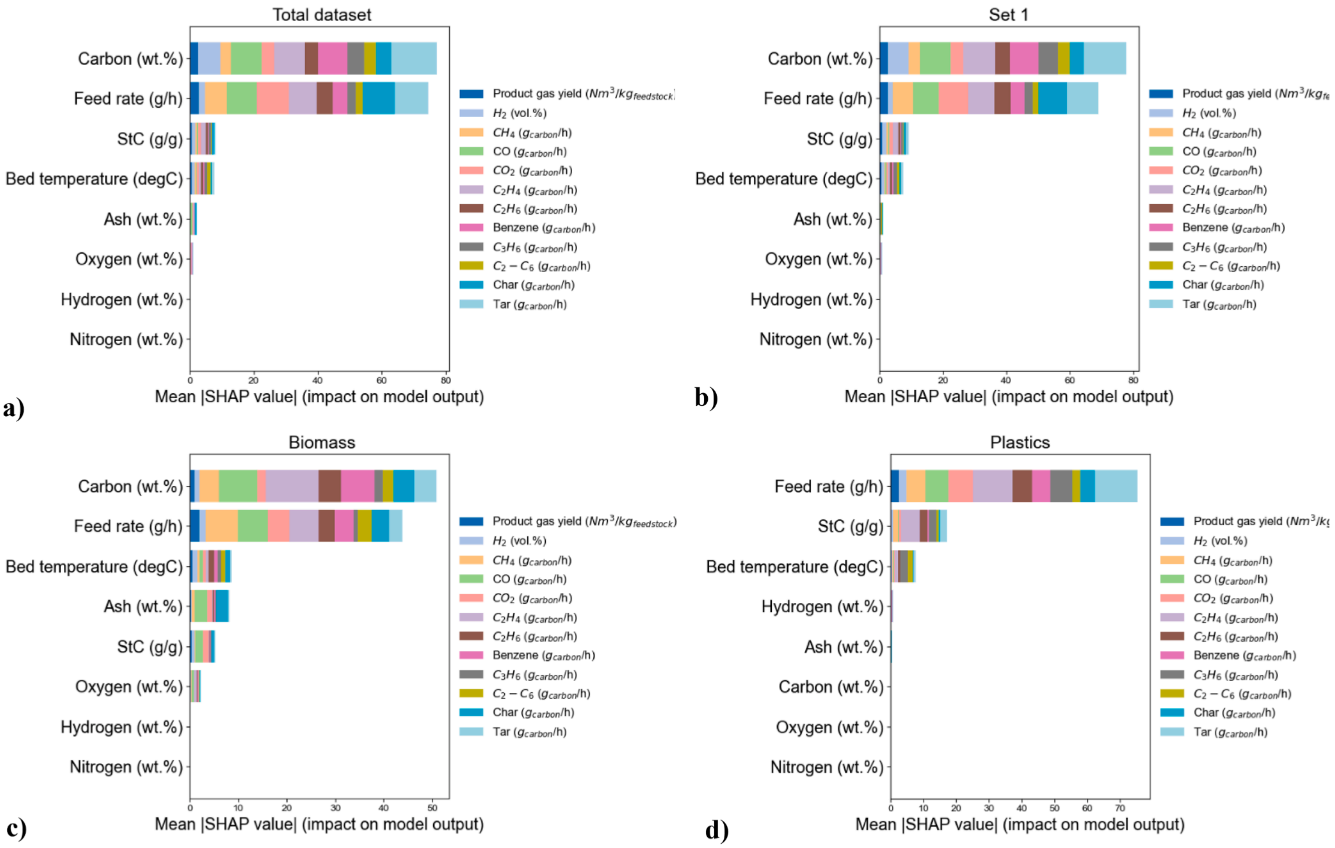

Fig. 11. SHAP results for data-driven model a) complete dataset b) randomly selected dataset 1 c) biomass dataset d) plastics dataset.

more related to the feed rate, meaning that for the plastic-rich feedstocks the feed rate was generally lower (average 207 g/h), but the carbon content of the feed was usually higher.

Next to the feed rate and carbon content, the bed temperature and StC both had a positive contribution to the product gas yield. For higher temperatures, more gas was typically produced, and the addition of steam promotes the formation of products such as H_2, CO and CO_2 which in turn increased the product gas volume. For the H_2 concertation, the carbon content had the highest feature importance, having a negative influence on the output. Next to the carbon content, the feed rate had the highest importance, although the results were inconclusive on the effect of the feed rate. The StC had the third highest feature importance and promotes the formation of H_2. The increase was due to the steam reacting with either solid carbon to form H_2 and CO (water-gas reaction), or with CO to form H_2 and CO_2 (water-gas shift reaction). Steam also reacts with CH_4 (and other hydrocarbons) to form CO and H_2 (steam reforming) [15,39]. Wang et al. [15] and Xue et al. [17] both showed the positive influence of steam (in the form of steam-to-biomass ratio) on the production of H_2 as well as the gas yield. For this reason, the CO_2 flow rate was also shown to increase with the StC ratio. The influence of the StC ratio for CO shows no clear correlation, most likely due to steam either promoting formation CO (water-gas reaction) or consuming the CO for the water-gas shift reaction. Furthermore, for the CO and CO_2 flow rates, the carbon content was shown to have a negative influence on both outputs, which was due to the correlation with the oxygen content in the feed. A higher carbon content in the feed would mean a lower oxygen content (see Fig. S.1), which would lead to lower CO and CO_2 flows, as the only source of oxygen comes from the feed and some steam (no air in the reactor, indirect gasification).

The hydrocarbons, as well as benzene, and tar flow rates all increased with the carbon content, which was also seen from the correlations in Fig. S.1. As mentioned, plastic-rich feedstocks had a higher carbon content compared to the biomass-rich and in-between feedstocks. The plastic feedstocks for this study mainly consisted of PP, PE and DKR-350 (polyolefin-rich waste) and thus the thermal cracking of these components to the monomers would lead to the production of olefins as well as benzene (due to secondary reactions). Additionally, for gasification of biomass the carbon content has also been shown to contribute to the production of CH_4 [15,17]. For the C_3H_6 carbon flow rate, the bed temperature was shown to be an important feature, having a negative influence. Higher temperatures lead to the conversion of C_3H_6 to smaller molecules such as C_2H_4 and CH_4 and other secondary reactions to form more complex molecules [9]. For the C_2-C_6 carbon flow rate, the bed temperature was once more the third most important input feature, however, the results were inconclusive on the impact that the operating parameters had on the flow.

The char carbon flow appears to decrease with the carbon content (second most importance input feature). Carbon in the feed undergoes thermal cracking and other reactions such as water gas-shift, resulting in the formation of gaseous products such as hydrocarbons, CO and CO_2. Additionally, carbon reacts to form tar (see Fig. S.7 l & x), which consequently reduces the char yield. The decrease in the char yield with carbon content was also seen by Li et al. [29]. Lastly, as previously mentioned, higher carbon content was associated with plastics-rich feedstocks which yield less char compared to gasification of biomass.

The results for the whole dataset, one randomly selected dataset, the biomass and plastics are presented in Fig. 12 for the physics-informed model.

Similar to the data-driven model, the carbon content and the feed rate were the two most important features, for the total dataset as well as the randomly divided datasets and the biomass dataset. The main targets that contributed to the feature importance of the two inputs were also similar to that of the data-driven model. Furthermore, for the total dataset as well as for the five randomly divided datasets (Fig. S.6 in the

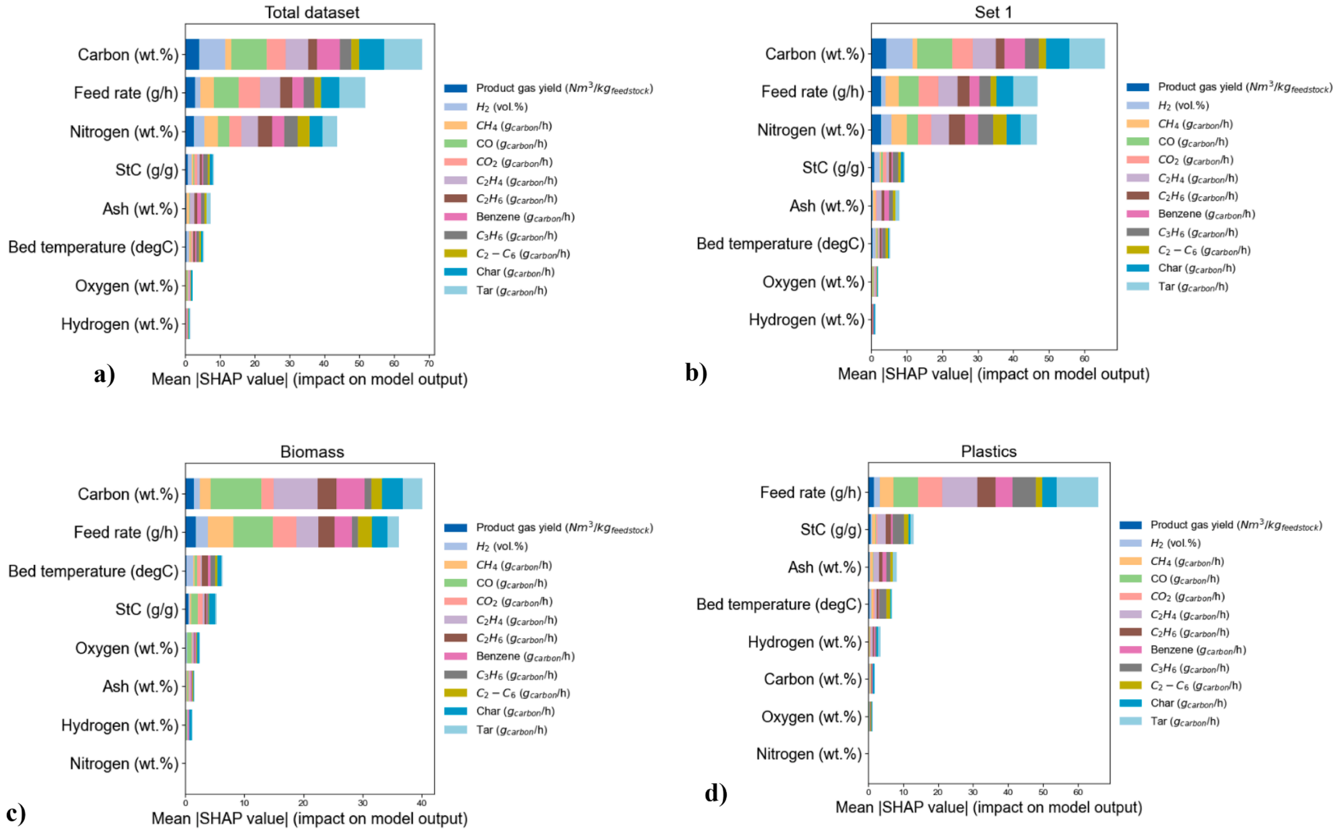

Fig. 12. SHAP results for physics-informed model a) complete dataset b) randomly selected dataset 1 c) biomass dataset d) plastics dataset.

SI) the nitrogen content of the feed had a high feature importance, which was not the case for the data-driven model. For the model training as well as for the SHAP analysis, the datasets used were the same for the data-driven and physics-informed models, thus there was no difference in variability. The increased feature importance was due to the model training and physics. The nitrogen content was not directly used for the physics-contribution (based on the carbon content) and thus indirectly became an important feature. The SHAP analysis was also done on the total dataset for the other values of λ and it was seen that for $\lambda \geq 0.2$, the nitrogen content feature importance does increase but always remains the third most important feature (after feed rate and carbon content). For the biomass and plastics datasets, the variability in the nitrogen content was less than for the whole datasets and randomly divided datasets and for this reason the nitrogen content did not have a high feature importance. The feedstock with the highest nitrogen content was the textile waste (in-between feedstock) which was due to the presence of components such as polyamide (nylon) and polyurethane (used for elastane), which are commonly used in textiles. Furthermore, for the plastics dataset, the feed rate was once more the most important feature, followed by the StC same as for the data-driven model. It was also seen that the ash yield of the feedstock became a more important feature, compared to the data-driven model.

The results of the feature importance on the individual outputs for the physics-informed model are presented in Fig. S.7 m to x. The results for most output features were comparable to that of the data-driven model in terms of both feature importance and correlation with the outputs. For the data-driven models, carbon content and feed rate were the two most important input features while for the physics-informed model, the nitrogen content of the feedstock became important as well. The nitrogen content had a negative influence on all output features. In the study of Yang et al. [6] on municipal solid waste gasification, the nitrogen content was also identified as a significant feature, ranking among the top three. In general, nitrogen exhibited a negative correlation with components such as H_2, CO_2 and CH_4 (mol.%), while showing a positive correlation with the tar yield. Li et al. [29] also concluded that nitrogen content was an important input feature for the syngas (negative correlation) and tar yield (positive correlation). For the H_2 concentration, the negative correlation was related to the nitrogen in feedstock reacting to form ammonia and hydrogen cyanide, which in turn reduced the hydrogen yields in the product gas. Fig. S.1 also indicated a negative linear correlation between the nitrogen content and various features. Notably, strong negative correlations were observed for the model outputs rather than the inputs, with the exception of the hydrogen content in the feed.

In general, it is seen that the feedstock properties had a higher feature importance compared to the operating conditions. As the dataset contains a wide variety of feedstocks, the composition in the feedstock also varies greatly. The greater variability in the feedstock properties (compared to operating conditions) could be why stronger correlations were seen. To better highlight the feature importance of the operating conditions, a suggestion is to split the dataset into a biomass-rich and plastics-rich sets and to train two separate ML models. The variation in the feedstock properties would reduce and the effect of operating conditions would become more apparent.

4. Outlook

The physics-informed ML model, developed in this study, enables the prediction of all carbon containing products for the indirect gasification of both biomass and plastic waste feedstocks, while ensuring scientific consistency. The wide range of feedstocks makes the model applicable to various fluidized bed processes, from biomass gasification to produce renewable fuels and chemicals to the chemical recycling of plastic waste. Performing calculations for waste valorization (through chemical recycling) helps to promote a circular economy, to reduce landfilling and emissions. Predicting the complete carbon distribution enables the tracking of both desired products and unwanted carbon byproducts (such as tar, CO_2 etc.) and thus process optimization can be done by both increasing product yields and minimizing unwanted byproducts.

On a potential industrial scale, the versatile model can be used for evaluating the product slate of new feedstocks under various conditions, assisting with the optimization of the process for the desired yields. Using the model as a screening tool, reduces cost and time associated with extensive feedstock characterization and experimental work. Furthermore, the model can be integrated with the process control system for real-time adjustments based on variations in feedstock composition. This in turn assist with making the process more economically viable and has environmental benefits by improving feedstock utilization and conversion, increasing product yields and reducing operation costs through process optimization.

Currently the model is limited to one technology namely bubbling fluidized beds on lab-scale and only considers indirect gasification. The model can be broadened to include direct gasification data, various bed materials, and parameters such as the equivalence ratio and calorific values of the feedstock (for energy balance calculations). Currently, the model considers the elemental composition as the only input for the feedstock characteristics, however, some plastics feedstocks are similar in chemical composition (e.g. PE and PP) but result in different product distributions. Incorporating polymer composition as input would expand the model applicability and assist with understanding the influence of polymer composition on the product distribution. On a physics basis, the mass balance of other components such as hydrogen and oxygen could be added together with the energy balance. Including more physical contributions would reduce the data demand for training the models.

To further broaden the industrial application of the model, other technologies such as dual fluidized beds, internal circulating fluidized bed reactors should be considered as well as different scales of technologies. This would, however, require incorporating literature data as complete experimental datasets are not readily available for each technology at different scales. Test campaigns are also time consuming and costly. Incorporating literature data is challenging due to limited availability of complete datasets necessary to calculate the carbon balance. Expanding the model (both for different technologies and scales), while still incorporating the physical mechanism (or adding more constraints) would require a combination of literature data and experimental data (for extensive campaigns to gather complete datasets). Alternatively, simulated data could also be used for training and evaluation.

5. Conclusions

A novel physics-informed ML model to predict the yields and composition for indirect gasification/thermochemical conversion of various feedstocks (biomass to plastics) was developed. As opposed to prior developed data-driven models, which only predicted the main products without scientific constraints, this physics-informed ML model predicted the entire carbon product slate, while ensuring scientific consistency in the form of carbon closure over the system. For the study, four data-driven models were trained and evaluated to select the ML model for the development of the physics-informed model. It was concluded that tree-based models were more suited for modelling the complex thermochemical process. XGBoost was the best model overall (based on R^2 and RMSE statistical parameters) and selected for further development. From the physics-informed model, it was concluded that already with a 10 % physics contribution, the carbon closure improved significantly compared to the experimental values. The λ could not be too high as this leads to an overprediction of the carbon outlet (carbon closures greater than one) and decrease in the predictive accuracy of the model. A 30 % physics contribution (70 % data-driven contribution) was the optimal value for the model. At 30 %, the predicative capabilities of model with regards to the overall model and well as of the individual outputs was maintained and carbon closure notably improved. From the feature importance analysis, it was concluded that the most important input features were related to feedstock properties, rather than the operating conditions (apart from feed rate). The feature importance results were comparable for both the data-driven and physics-informed models in terms of absolute feature importance and the correlations with the outputs.

The developed physics-informed model is first step towards improving data-driven ML models for application of waste gasification/ thermal cracking. The model covers a broad range of feedstocks, making it versatile for applications to different processes from biomass conversion for the production of renewable fuels and chemicals to the chemical recycling of plastic waste. Due to the model predicting the entire carbon product slate (wanted and unwanted products), while ensuring scientific consistency, the model offers economic and environmental benefits for optimizing process conditions to both maximize product yields and minimizing waste streams. Currently, the model is limited to indirect gasification in fluidized bed reactors, but can be broadened to include more technologies (and scales) and have more inputs such as bed material type, equivalence ratio, calorific values of feedstocks and the polymer/biomass composition of the waste feedstocks. For the physics contribution, the model can be extended to include the hydrogen and oxygen balance as well as the energy balance. Including these contributions would assist with further ensuring scientific consistency and would reduce the data demand for training the models.

Nomenclature

C_i	Carbon flow rate of component i (g/h)
g_i	Gradient
h_i	Hessian
I_i	Total set of leaf nodes
L_i	Loss function
N	Number of points
R^2	Coefficient of determination
T	Number of leave nodes of the tree
y_i	Experimental value
$\hat{y}_i$	Predicted value
β	Penalty term
γ	Minimal loss reduction required for splitting a new leaf
λ	Weight of physics contribution to ML model
Ω	Regularization term
ω	Weight values for regularization

Abbreviations

ANN	Artificial neural networks
CFD	Computational fluid dynamics
db	Dry basis
DT	Decision tree
GBR	Gradient boosting regression
GC	Gas chromatographer
GC-FID	GC with a flame ionisation detector
ID	Internal diameter
ML	Machine learning
MSE	Mean squared error
NN	Neural networks
PE	Polyethylene
PINN	Physics-informed neural networks
PP	Polypropylene
RDF	Refuse derived fuel
RF	Random forest
RMSE	Root-mean-square-error
SHAP	Shapley Additive Explanations
SI	Supporting information
StC	Steam-to-carbon ratio
SVM	Support vector machine
SVR	Support vector regression
XGBoost	Extreme gradient boosting

Funding information

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The author would like to acknowledge the Netherlands Ministry of Economic Affairs as part of TNO's biomass research program for providing funding for the publication fees.

CRediT authorship contribution statement

Surika van Wyk: Writing – review & editing, Writing – original draft, Validation, Methodology, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

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

Acknowledgements

The author would like to thank Carlos Mourao Vilela and Berend Vreugdenhil for their review of the article and their valuable insights. Special thanks to Pejman Shoeibi Omrani for reviewing the machine learning section, providing highly useful advice, and assisting with the explainability of the models.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ceja.2024.100699.

Data availability

Data will be made available on request.

References

[1] S.K. Sansaniwal, M.A. Rosen, S.K. Tyagi, Global challenges in the sustainable development of biomass gasification: an overview, Renew. Sustain. Energy Rev. 80 (2017) 23–43, https://doi.org/10.1016/j.rser.2017.05.215.

[2] S.S. Siwal, Q. Zhang, N. Devi, A.K. Saini, V. Saini, B. Pareek, S. Gaidukovs, V. K. Thakur, Recovery processes of sustainable energy using different biomass and wastes, Renew. Sustain. Energy Rev. 150 (2021) 111483, https://doi.org/10.1016/j.rser.2021.111483.

[3] G. Lopez, M. Artetxe, M. Amutio, J. Bilbao, M. Olazar, Thermochemical routes for the valorization of waste polyolefinic plastics to produce fuels and chemicals. A review, Renew. Sustain. Energy Rev. 73 (2017) 346–368, https://doi.org/10.1016/j.rser.2017.01.142.

[4] A.D. Korberg, B.V. Mathiesen, L.R. Clausen, I.R. Skov, The role of biomass gasification in low-carbon energy and transport systems, Smart Energy 1 (2021) 100006, https://doi.org/10.1016/j.segy.2021.100006.

[5] J.-P. Lange, S.R.A. Kersten, S.D. Meester, K. Ragaert, Plastic recycling stripped naked – from circular product to circular industry with recycling cascade, (2024).

[6] Y. Yang, H. Shahbeik, A. Shafizadeh, S. Rafiee, A. Hafezi, X. Du, J. Pan, M. Tabatabaei, M. Aghbashlo, Predicting municipal solid waste gasification using machine learning: a step toward sustainable regional planning, Energy 278 (2023) 127881, https://doi.org/10.1016/j.energy.2023.127881.

[7] S. Ascher, I. Watson, S. You, Machine learning methods for modelling the gasification and pyrolysis of biomass and waste, Renew. Sustain. Energy Rev. 155 (2022) 111902, https://doi.org/10.1016/j.rser.2021.111902.

[8] A. Kushwah, Modelling approaches for biomass gasifiers: a comprehensive overview, Sci. Total Environ. (2022).

[9] O. Dogu, M. Pelucchi, R. Van De Vijver, P.H.M. Van Steenberge, D.R. D'hooge, A. Cuoci, M. Mehl, A. Frassoldati, T. Faravelli, K.M. Van Geem, The chemistry of chemical recycling of solid plastic waste via pyrolysis and gasification: state-of-the-art, challenges, and future directions, Prog. Energy Combust. Sci. 84 (2021) 100901, https://doi.org/10.1016/j.pecs.2020.100901.

[10] V. Marcantonio, L. Di Paola, M. De Falco, M. Capocelli, Modeling of biomass gasification: from thermodynamics to process simulations, Energies 16 (2023) 7042, https://doi.org/10.3390/en16207042.

[11] D. Baruah, D.C. Baruah, Modeling of biomass gasification: a review, Renew. Sustain. Energy Rev. 39 (2014) 806–815, https://doi.org/10.1016/j.rser.2014.07.129.

[12] S. Safarian, R. Unnþórsson, C. Richter, A review of biomass gasification modelling, Renew. Sustain. Energy Rev. 110 (2019) 378–391, https://doi.org/10.1016/j.rser.2019.05.003.

[13] D. Serrano, D. Castelló, Tar prediction in bubbling fluidized bed gasification through artificial neural networks, Chem. Eng. J. 402 (2020) 126229, https://doi.org/10.1016/j.cej.2020.126229.

[14] Y. Cheng, E. Ekici, G. Yildiz, Y. Yang, B. Coward, J. Wang, Applied machine learning for prediction of waste plastic pyrolysis towards valuable fuel and chemicals production, J. Anal. Appl. Pyrolysis 169 (2023) 105857, https://doi.org/10.1016/j.jaap.2023.105857.

[15] Z. Wang, L. Mu, H. Miao, Y. Shang, H. Yin, M. Dong, An innovative application of machine learning in prediction of the syngas properties of biomass chemical looping gasification based on extra trees regression algorithm, Energy 275 (2023) 127438, https://doi.org/10.1016/j.energy.2023.127438.

[16] M.V. Gil, K.M. Jablonka, S. Garcia, C. Pevida, B. Smit, Biomass to energy: a machine learning model for optimum gasification pathways, Digit. Discov. 2 (2023) 929–940, https://doi.org/10.1039/D3DD00079F.

[17] P. Xue, T. Chen, X. Huang, Q. Hu, J. Hu, H. Zhang, H. Yang, H. Chen, Prediction of syngas properties of biomass steam gasification in fluidized bed based on machine learning method, Int. J. Hydrog. Energy 49 (2024) 356–370, https://doi.org/10.1016/j.ijhydene.2023.08.259.

[18] T. Bikmukhametov, J. Jäschke, Combining machine learning and process engineering physics towards enhanced accuracy and explainability of data-driven models, Comput. Chem. Eng. 138 (2020) 106834, https://doi.org/10.1016/j.compchemeng.2020.106834.

[19] S. Ren, S. Wu, Q. Weng, Physics-informed machine learning methods for biomass gasification modeling by considering monotonic relationships, Bioresour. Technol. 369 (2023) 128472, https://doi.org/10.1016/j.biortech.2022.128472.

[20] K.M. Shaw, P.E. Poh, Y.K. Ho, Z.Y. Chen, I.M.L. Chew, Modeling the anaerobic digestion of palm oil mill effluent via physics-informed deep learning, Chem. Eng. J. 485 (2024) 149826, https://doi.org/10.1016/j.cej.2024.149826.

[21] Y. Dong, L. Song, Q. Zhao, Z. Ding, L. Qiu, C. Lu, G. Chen, A physics-guided eXtreme gradient boosting model for predicting the initial productivity of oil wells, Geoenergy Sci. Eng. 231 (2023) 212402, https://doi.org/10.1016/j.geoen.2023.212402.

[22] B. Zhu, S. Ren, Q. Weng, F. Si, A physics-informed neural network that considers monotonic relationships for predicting NO emissions from coal-fired boilers, Fuel 364 (2024) 131026, https://doi.org/10.1016/j.fuel.2024.131026.

[23] W.L. Van de Kamp, P.J. De Wild, H.A.M. Knoef, J.P.A. Neeft, J.H.A. Kiel, Tar measurement in biomass gasification, standardisation and supporting R&D, TNO (past ECN), 2006. https://publications.tno.nl/publication/34628627/3oZ90f/c06046.pdf (accessed March 6, 2024).

[24] G. Katsaros, D.S. Pandey, A. Horvat, G.A. Almansa, L.E. Fryda, J.J. Leahy, S. A. Tassou, Gasification of poultry litter in a lab-scale bubbling fluidised bed reactor: impact of process parameters on gasifier performance and special focus on tar evolution, Waste Manag. 100 (2019) 336–345, https://doi.org/10.1016/j.wasman.2019.09.014.

[25] G. Katsaros, D.S. Pandey, A. Horvat, G. Aranda Almansa, L.E. Fryda, J.J. Leahy, S. A. Tassou, Experimental investigation of poultry litter gasification and co-gasification with beech wood in a bubbling fluidised bed reactor–effect of equivalence ratio on process performance and tar evolution, Fuel 262 (2020) 116660, https://doi.org/10.1016/j.fuel.2019.116660.

[26] H.L. Zhu, Y.S. Zhang, M. Materazzi, G. Aranda, D.J.L. Brett, P.R. Shearing, G. Manos, Co-gasification of beech-wood and polyethylene in a fluidized-bed reactor, Fuel Process. Technol. 190 (2019) 29–37, https://doi.org/10.1016/j.fuproc.2019.03.010.

[27] BRISK2 | EUHorizon2020, (2023). https://brisk2.eu/(accessed June 3, 2023).

[28] S. Van Wyk, H.W.J.P. Neomagus, J.R. Bunt, R.C. Everson, Coal reactivity and selection for solid-based pre-reduction of sponge iron, Int. J. Coal Prep. Util. 40 (2020) 233–246, https://doi.org/10.1080/19392699.2017.1384729.

[29] J. Li, L. Li, Y.W. Tong, X. Wang, Understanding and optimizing the gasification of biomass waste with machine learning, Green Chem. Eng. 4 (2023) 123–133, https://doi.org/10.1016/j.gce.2022.05.006.

[30] S. Liu, Y. Yang, L. Yu, F. Zhu, Y. Cao, X. Liu, A. Yao, Y. Cao, Predicting gas production by supercritical water gasification of coal using machine learning, Fuel 329 (2022) 125478, https://doi.org/10.1016/j.fuel.2022.125478.

[31] E.E. Ozbas, D. Aksu, A. Ongen, M.A. Aydin, H.K. Ozcan, Hydrogen production via biomass gasification, and modeling by supervised machine learning algorithms, Int. J. Hydrog. Energy 44 (2019) 17260–17268, https://doi.org/10.1016/j.ijhydene.2019.02.108.

[32] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco California USA, ACM, 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785.

[33] T. Katongtung, T. Onsree, N. Tippayawong, Machine learning prediction of biocrude yields and higher heating values from hydrothermal liquefaction of wet biomass and wastes, Bioresour. Technol. 344 (2022) 126278, https://doi.org/10.1016/j.biortech.2021.126278.

[34] Optuna: A hyperparameter optimization framework–optuna 3.6.1 documentation, (2023). https://optuna.readthedocs.io/en/stable/(accessed June 3, 2023).

[35] XGBoost Developers, XGBoost release 2.1.0-dev, 2023. https://readthedocs.org/projects/xgboost/downloads/pdf/latest/(accessed December 24, 2023).

[36] S. Ascher, X. Wang, I. Watson, W. Sloan, S. You, Interpretable machine learning to model biomass and waste gasification, Bioresour. Technol. 364 (2022) 128062, https://doi.org/10.1016/j.biortech.2022.128062.

[37] Q. Liu, G. Zhang, J. Yu, G. Kong, T. Cao, G. Ji, X. Zhang, L. Han, Machine learning-aided hydrothermal carbonization of biomass for coal-like hydrochar production: parameters optimization and experimental verification, Bioresour. Technol. 393 (2024) 130073, https://doi.org/10.1016/j.biortech.2023.130073.

[38] A. Alabdrabalnabi, R. Gautam, S.M. Sarathy, Machine learning to predict biochar and bio-oil yields from co-pyrolysis of biomass and plastics, Fuel 328 (2022) 125303, https://doi.org/10.1016/j.fuel.2022.125303.

[39] P. Parthasarathy, K.S. Narayanan, Hydrogen production from steam gasification of biomass: influence of process parameters on hydrogen yield–a review, Renew. Energy 66 (2014) 570–579, https://doi.org/10.1016/j.renene.2013.12.025.