Electronic Supplementary Material (ESI) for Digital Discovery.

Supporting Information: Biomass to energy: A machine learning model for optimum gasification pathways

María Victoria Gil,[*,†] Kevin Maik Jablonka,[‡] Susana García,[¶] Covadonga Pevida,[†] and Berend Smit[*,‡]

†*Instituto de Ciencia y Tecnología del Carbono (INCAR), CSIC, Francisco Pintado Fe 26, 33011 Oviedo, Spain*

‡*Laboratory of Molecular Simulation (LSMO), Institut des Sciences et Ingénierie Chimiques (ISIC), Valais, École Polytechnique Fédérale de Lausanne (EPFL), Rue de l'Industrie 17, Sion CH-1951, Switzerland*

¶*The Research Center for Carbon Solutions (RCCS), School of Engineering and Physical Sciences, Heriot-Watt University, EH14 4AS Edinburgh, United Kingdom*

E-mail: victoria.gil@incar.csic.es; berend.smit@epfl.ch

Contents

Section 1. Experimental dataset: Gasification

Experimental results from the steam-oxygen gasification of biomasses with different origins have been used in this work: pine sawdust, chestnut sawdust, torrefied pine sawdust, torrefied chestnut sawdust, almond shells, cacao shells, grape pomace, olive stones, pine kernel shells, and pine cone leaves. The dataset includes the experimental results published in a previous work by our research team.[S1]

The experimental procedure and the plant where the experiments were performed have been described in detail elsewhere.[S1] Briefly, the gasification plant (PID Eng&Tech) used experimentally consists of an atmospheric pressure bubbling fluidized bed reactor. The plant comprises a double-hopper feeding system, a pre-heated gas inlet, a gasification reactor with freeboard, an outlet gas double cyclone cleaning system, a tar trap, a condenser, and control and measuring systems. The gasification reactor is a SS310 cylinder with a height of 1 m and an inner diameter of 77 mm, which ends in a 59.2 mm long and 133 mm diameter freeboard. Both the reactor and the freeboard are surrounded by three independent electrical furnaces that allow the temperature of the reactor wall to increase up to 950 °C. The bed material used in the experiments was 1 kg of coal ash (212 µm to 710 µm). The fuel feeding system includes a main storage hopper (10 L), which is communicated by an auger (Ø = 20 mm) with a smaller secondary one. A water-cooled secondary auger (Ø = 20 mm) directly introduces the fuel sample into the fluidized bed reactor at approximately 6 cm above the bottom of the reactor. Gases are fed into the main hopper by Bronkhorst High-Tech mass flow controllers. A Wilson 307 piston pump feeds in the required mass flow of liquid water to the reactor. Fly ashes and solid particles are retained in a cyclone and in a 100 µm pore diameter ceramic filter. The outlet gas is directed to a heat exchanger where water and heavy tars condense. The lighter tars are captured by a cold-trap and the cleaned gases are sent to the gas analyzers.

Gasification experiments were carried out at different temperatures (T), steam-to-air (SA) ratios, stoichiometric ratios (SR), and steam-to-biomass ratios (SBR). SR is defined as the experimental oxygen-to-fuel weight ratio divided by the oxygen-to-fuel weight ratio for stoichiometric combustion. SA, SR, and SBR are calculated by Eqs. (1), (2), and (3), respectively[S2–S4]:

$$\text{SA} = \frac{\text{Total water supplied (mL/min)}}{\text{Total air supplied (mL/min)}} \tag{1}$$

$$\text{SR} = \frac{(\text{Oxygen/fuel})_{\text{experimental}}}{(\text{Oxygen/fuel})_{\text{stoichiometric}}} \tag{2}$$

$$\text{SBR} = \frac{\text{Total water supplied (g/min)}}{\text{Fuel supplied (g/min)}} \tag{3}$$

The composition of the product gas after gasification was calculated on a nitrogen-free and dry basis. The gas flow rates of the species generated during the experiments were then calculated by means of a nitrogen balance (used as an internal standard). From the results of the gaseous emissions, the gas yield (GAS)[S5], the higher heating value (HHV) of the gas obtained (HHVgas)[S6], and the energy yield (E_{yield}) were determined as defined by Eqs. (4), (5), and (6), respectively:

$$\text{GAS (Nm}^3\text{ gas/kg biom)} = \frac{\text{Q}_{\text{outlet-gas}}}{\text{m}_{\text{biomass}}} \tag{4}$$

$$\text{HHVgas (MJ/Nm}^3) = (11.76 \cdot x_{\text{CO}} + 11.882 \cdot x_{\text{H}_2} + 37.024 \cdot x_{\text{CH}_4}) \cdot 10/1000 \tag{5}$$

$$\text{E}_{\text{yield}}\text{ (MJ/kg biom)} = \text{GAS} \cdot \text{HHVgas} \tag{6}$$

where $Q_{\text{outlet-gas}}$ is the volumetric flow of the outlet gas (Nm^3/h), obtained by applying a balance to the inert gas (N_2), m_{biomass} is the biomass inlet mass flow on a dry basis (kg/h), and x_i (vol%) represents the volumetric percentage of each component in the dry product gas.

Section 2. Exploratory data analysis

Section 2.1. Data dictionary

- **`BIOMASS`**: type of biomass (AS=almond shells, CHE=chestnut sawdust, CHET=torrefied chestnut sawdust, CS=cacao shells, GP=grape pomace, OS=olive stones, PCL=pine cone leafs, PIN=pine sawdust, PINT=torrefied pine sawdust, PKS=pine kernel shells)
- **`SA`**: steam-to-air ratio used in the gasification experiments (vol/vol)
- **`SBR`**: steam-to-biomass ratio used in the gasification experiments (wt/wt)
- **`HHVbiom (MJ/kg)`**: higher heating value (HHV) of the biomass (MJ/kg)
- **`T (K)`**: gasification temperature (K)
- **`SR`**: stoichiometric ratio, i.e., experimental oxygen-to-fuel weight ratio divided by the oxygen-to-fuel weight ratio for stoichiometric combustion
- **`C (%)`**: carbon content of the biomass (wt%) on a dry basis
- **`N (%)`**: nitrogen content of the biomass (wt%) on a dry basis
- **`H (%)`**: hydrogen content of the biomass (wt%) on a dry basis
- **`S (%)`**: sulfur content of the biomass (wt%) on a dry basis
- **`O (%)`**: oxygen content of the biomass (wt%) on a dry basis
- **`ash (%)`**: ash content of the biomass (wt%) on a dry basis
- **`VM (%)`**: volatile matter content of the biomass (wt%) on a dry basis
- **`FC (%)`**: fixed carbon content of the biomass (wt%) on a dry basis
- **`MC (%)`**: moisture content of the biomass (wt%)

- **`volCO2 (%)`**: CO_2 concentration in the gasification gas (vol%)
- **`volCO (%)`**: CO concentration in the gasification gas (vol%)
- **`volCH4 (%)`**: CH_4 concentration in the gasification gas (vol%)
- **`volH2 (%)`**: H_2 concentration in the gasification gas (vol%)
- **`volCOMB (%)`**: concentration of combustible gas (H_2+CO+CH_4) in the gasification gas (vol%)
- **`GAS (m³/kg biom)`**: gas production from the gasification process in m^3/kg biom

To develop the model we used the following input features: SA, SBR, HHVbiom (MJ/kg), T (K), SR, C (%), H (%), O (%), ash (%), VM (%), FC (%) and MC (%). Biomass N and S contents were not included as inputs of the model because it is not expected that they have an effect on the gasification gases studied. The outputs predicted by the model were: volH2 (%), volCO (%), volCOMB (%), and GAS (m^3/kg biom).

Section 2.2. Distributions and correlations

Fig. S1 shows the distributions and correlations between a subset of features. Fig. S2 shows the distributions and correlations between targets. Table S1 shows the descriptive statistics of the gasification data used to build the machine learning model.

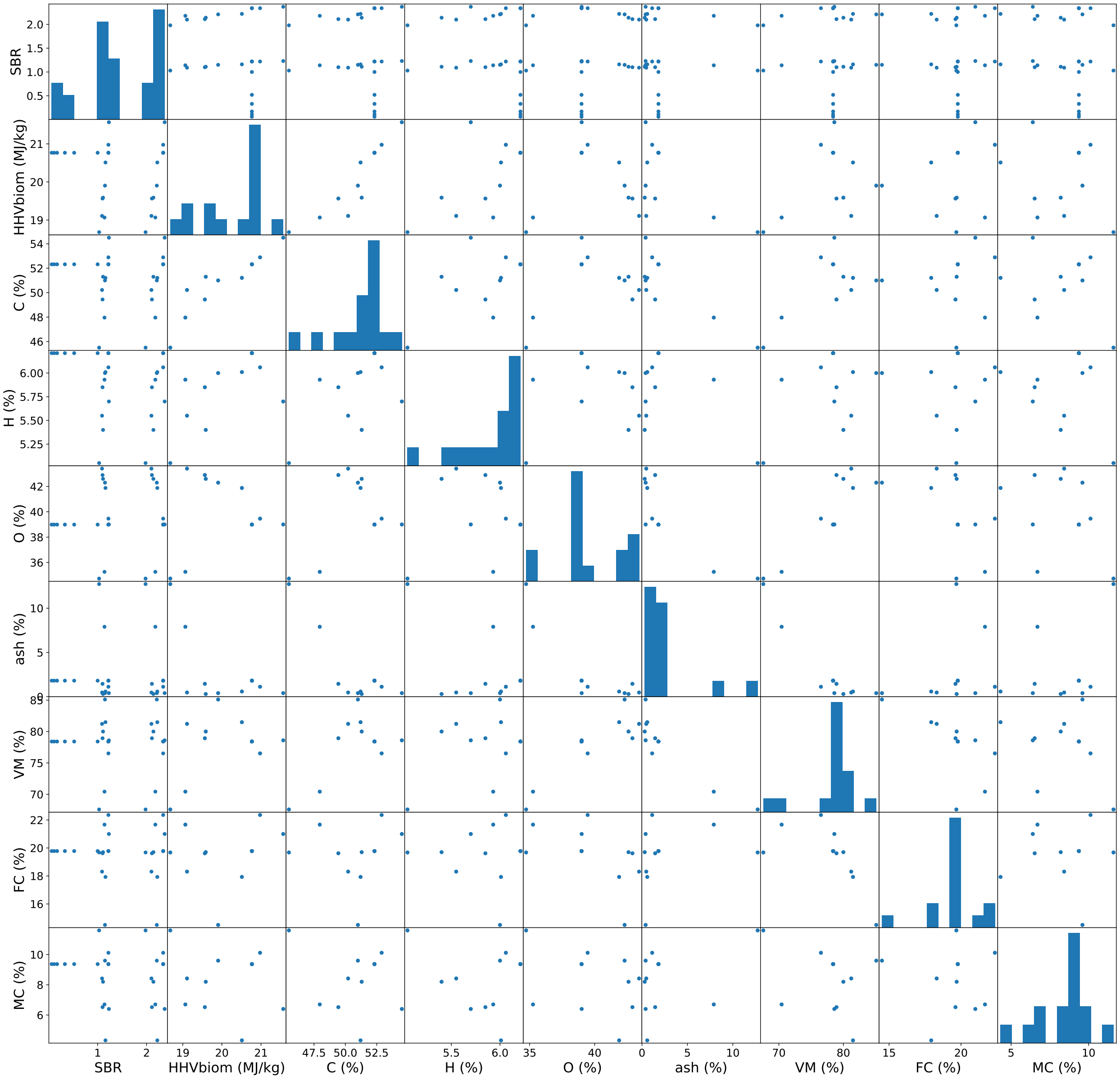


Fig. S1. Distributions and correlations between a subset of features.

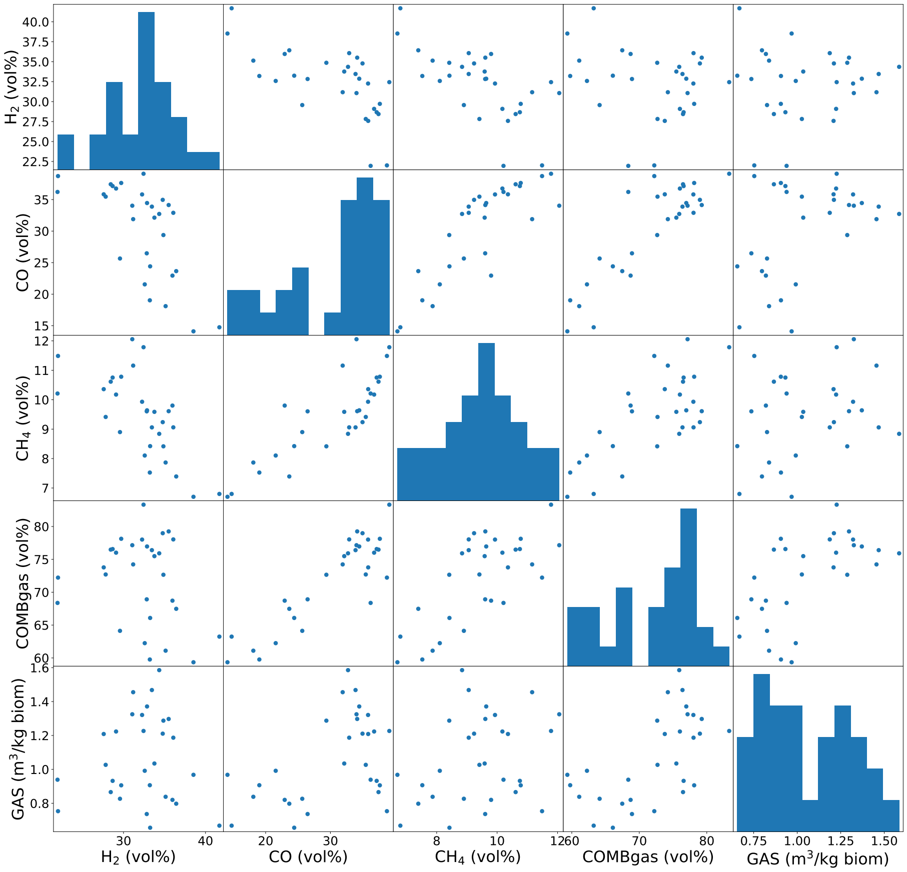


Fig. S2. Distributions and correlations between targets.

Table S1. Descriptive statistics of the gasification data.

Parameter	mean	median	minimum	maximum
T (K)	1153	1173	973	1173
SA	1.96	2.33	0.11	2.33
SR	0.19	0.19	0.13	0.25
SBR	1.43	1.22	0.06	2.37
C (%)	51.20	52.32	45.50	54.50
H (%)	5.92	6.01	5.05	6.21
O (%)	39.70	38.99	34.73	43.41
ash (%)	2.41	1.81	0.30	12.73
VM (%)	78.02	78.41	67.59	85.10
FC (%)	19.56	19.78	14.50	22.36
HHVbiom (MJ/kg)	20.24	20.77	18.68	21.57
MC (%)	8.54	9.37	4.32	11.60
volH2 (%)	32.17	32.85	21.95	41.69
volCO (%)	30.47	33.39	14.11	39.09
volCH4 (%)	9.52	9.61	6.70	12.06
volCOMB (%)	72.16	74.00	59.34	83.32
GAS (m^3/kg biom)	1.07	1.01	0.66	1.59

Section 3. Single-output baseline models

To know the minimum performance that we can expect with our model, before building a Gaussian process regression (GPR) model we evaluated the performance of some simple baseline models, such as the mean dummy regressor and the median dummy regressor. We also evaluated the performance of a more powerful model such as the XGBoost regressor (XGBRegressor, default parameters, version 1.5.1). We used leave-one-out cross-validation (LOOCV) for the training of the baseline models, as we did later for the GPR model. These models are only able to predict single outputs, so independent models were built for each output. The prediction performance of these models is determined by mean of the coefficient of determination, R^2, and the root mean squared error (RMSE) values, which are shown in Tables S2 and S3, respectively.

Table S2. Coefficient of determination, R^2, for the baseline models.

model	H_2 (vol%)	CO (vol%)	COMBgas (vol%)	GAS (m^3/kg biom)
mean dummy regressor	-0.0001	-0.0002	-0.0001	-0.00002
median dummy regressor	0.0360	-0.2120	-0.0813	-0.0903
XGBoost regressor	0.8606	0.8933	0.8826	0.6294

Table S3. Root mean squared error (RMSE) values for the baseline models.

model	H_2 (vol%)	CO (vol%)	COMBgas (vol%)	GAS (m^3/kg biom)
mean dummy regressor	1.0946	1.0655	1.0581	1.0540
median dummy regressor	1.0746	1.1729	1.1002	1.1006
XGBoost regressor	0.4087	0.3480	0.3626	0.6417

Section 4. Model selection and performance assessment

We used Gaussian process regression (GPR) models (with coregionalized kernels) as they are able to capture complex non-linear relationships using only a limited amount of data, besides providing uncertainty estimates. We used the `GPy` Python library[S7] to build and train the Gaussian process regression models. We tried different kernel functions with automatic relevance determination (ARD) in an intrinsic model of coregionalization (ICM).[S8] As we explain in detail below, our selected coregionalized GPR model use as the kernel the sum of the Radial basis function (RBF) kernel and the Linear kernel, with ARD.

A dataset of 30 samples was used to develop the model. All features and outputs were z-score standardized using the mean and standard deviation of the training set (using the scikit-learn Python package[S9]). We evaluated the model performance from three perspectives: predictive performance (using both cross-validation and experimental validation), uncertainty estimation, and randomization test.

Section 4.1. Generalization performance

Since our dataset is small we use leave-one-out cross-validation (LOOCV)[S10] to measure the predictive performance of our model. That is, we train as many models as we have datapoints (N) and then use $N - 1$ points for training and 1 point for testing. We then measure the performance of the model at test points, which shows the ability of the model to generate accurate predictions for unseen data. With the training of many models, we accept a higher computational cost to have a more robust estimate of model performance. For optimization of the kernel hyperparameters we performed 20 random restarts and used hyperparameters corresponding to the best maximum likelihood solution. Based on the findings of Wainer and Cawley[S11] we did not perform nested LOOCV for model selection but directly selected models based on the LOOCV results. We performed 15 replications of each LOOCV run to evaluate the variance in the model performance, and hence we provide all metric results as an average of these repetitions. To evaluate the prediction performance of the models, we used the coefficient of determination, R^2, and the root mean squared error (RMSE).

In addition to LOOCV, to verify the predictive capacity of our model we also performed an experimental validation of the model. For that, we carried out new gasification experiments using two different biomasses that were not used previously for the training and cross-validation of the model: walnut shells (WS) (T=1173 K, SA=2.33, and SR=0.13) and hazelnut shells (HZS) (T=1173 K, SA=2.33, and SR=0.25). We then compared the experimental results of the new gasification experiments with those predicted by the model.

Section 4.2. Uncertainty calibration

To evaluate the uncertainty calibration of our model, we used the UQ360 implementation.[S12] For that, the predictive uncertainty given by our GPR model is expressed as a prediction interval, i.e., a region bounding the mean that we expect an observed value to fall within with a certain probability. For this, we use the metrics Mean Prediction Interval Width (MPIW),[S12] which computes the average width of the 95% prediction interval, and Prediction Interval Coverage Probability (PICP),[S12] which computes the fraction of test data for which the ground truth lies within a 95% prediction interval of the model.

Section 4.3. Model selection

Tables S4 and S5 show the performance metrics for some of the GPR models that we built using different kernel functions.

Table S4. Generalization and calibration metrics for GPR models using different individual kernel functions (all with automatic relevance determination, ARD). Means and standards deviations of 15 runs. R^2: coefficient of determination; RMSE: root mean squared error; PICP: Prediction Interval Coverage Probability[S12]; MPIW: Mean Prediction Interval Width[S12].

kernel	H_2	CO	COMBgas	GAS
	(vol%)	(vol%)	(vol%)	(m^3/kg biom)
Radial Basis Function (RBF)				
R^2	0.8022±0.0051	0.8581±0.0009	0.8402±0.0022	0.7932±0.0024
RMSE	0.4867±0.0062	0.4013±0.0013	0.4230±0.0030	0.4777±0.0085
PICP	0.7244±0.0147	0.5600±0.0181	0.7000±0.0000	0.9000±0.0000
MPIW	1.3187±0.0126	0.7734±0.0141	0.9666±0.0222	1.5850±0.0166
Matérn 3/2				
R^2	0.8542±0.0023	0.9150±0.0008	0.9019±0.0007	0.8333±0.0016
RMSE	0.4179±0.0033	0.3105±0.0014	0.3315±0.0012	0.4304±0.0021
PICP	0.8333±0.0000	0.7689±0.0083	0.7689±0.0083	0.9000±0.0000
MPIW	1.3545±0.0032	0.8826±0.0018	1.0806±0.0033	1.5630±0.0031
Matérn 5/2				
R^2	0.8328±0.0033	0.8827±0.0023	0.8642±0.0016	0.8194±0.0035
RMSE	0.4475±0.0045	0.3648±0.0036	0.3899±0.0023	0.4479±0.0044
PICP	0.7667±0.0000	0.6667±0.0000	0.7667±0.0000	0.9200±0.0163
MPIW	1.3684±0.0085	0.8697±0.0023	1.0781±0.0038	1.6086±0.0066
Rational Quadratic				
R^2	0.7981±0.0015	0.8663±0.0010	0.8473±0.0026	0.8194±0.0027
RMSE	0.4918±0.0018	0.3895±0.0014	0.4135±0.0035	0.4479±0.0033
PICP	0.7444±0.0314	0.5667±0.0000	0.6911±0.0147	0.9000±0.0000
MPIW	1.3469±0.0088	0.8072±0.0063	1.0111±0.0051	1.5967±0.0076
Linear				
R^2	0.8183±0.0000	0.9253±0.0000	0.9087±0.0000	0.8089±0.0000
RMSE	0.4666±0.0000	0.2912±0.0000	0.3198±0.0000	0.4608±0.0000
PICP	0.9333±0.0000	0.9000±0.0000	0.9333±0.0000	0.9667±0.0000
MPIW	1.8188±0.0000	1.0405±0.0000	1.1189±0.0000	1.6967±0.0000

Table S5. Generalization and calibration metrics for GPR models using different multiple kernel functions (all with automatic relevance determination, ARD). Means and standards deviations of 15 runs. R^2: coefficient of determination; RMSE: root mean squared error; PICP: Prediction Interval Coverage Probability[S12]; MPIW: Mean Prediction Interval Width[S12].

kernel	H_2	CO	COMBgas	GAS
	(vol%)	(vol%)	(vol%)	(m^3/kg biom)
RBF + Linear				
R^2	0.8924±0.0052	0.9597±0.0012	0.9251±0.0022	0.8187±0.0026
RMSE	0.3589±0.0086	0.2138±0.0031	0.2896±0.0042	0.4488±0.0032
PICP	0.8200±0.0323	0.7667±0.0298	0.7583±0.0179	0.9533±0.0194
MPIW	0.9466±0.0207	0.6557±0.0186	0.9140±0.0169	1.5769±0.0120
Matérn 3/2 + Linear				
R^2	0.8783±0.0036	0.9515±0.0007	0.9169±0.0015	0.8252±0.0024
RMSE	0.3818±0.0056	0.2346±0.0018	0.3050±0.0028	0.4407±0.0031
PICP	0.7511±0.0239	0.7689±0.0147	0.7622±0.0166	0.9400±0.0133
MPIW	0.9384±0.0099	0.6619±0.0057	0.8984±0.0075	1.5378±0.0066
Matérn 5/2 + Linear				
R^2	0.8735±0.0028	0.9523±0.0013	0.9165±0.0018	0.8208±0.0030
RMSE	0.3892±0.0043	0.2327±0.0033	0.3057±0.0032	0.4462±0.0037
PICP	0.7556±0.0157	0.7644±0.0285	0.7556±0.0263	0.9422±0.0147
MPIW	0.9165±0.0134	0.6476±0.0088	0.8914±0.0133	1.5455±0.0069
Rational Quadratic + Linear				
R^2	0.8796±0.0057	0.9581±0.0020	0.9195±0.0017	0.8136±0.0064
RMSE	0.3796±0.0090	0.2181±0.0053	0.3001±0.0031	0.4550±0.0078
PICP	0.8022±0.0333	0.7356±0.0285	0.7333±0.0322	0.9200±0.0237
MPIW	0.8932±0.0194	0.6335±0.0133	0.8839±0.0194	1.5562±0.0122
RBF * Linear				
R^2	0.8703±0.0002	0.9346±0.0003	0.8944±0.0004	0.8093±0.0004
RMSE	0.3941±0.0004	0.2724±0.0005	0.3438±0.0007	0.4603±0.0005
PICP	0.9000±0.0000	1.0000±0.0000	1.0000±0.0000	0.9667±0.0000
MPIW	1.7508±0.0014	1.4649±0.0009	1.7305±0.0008	2.0359±0.0010
RBF + Linear + Matérn 3/2				
R^2	0.9019±0.0034	0.9579±0.0028	0.9150±0.0048	0.8110±0.0104
RMSE	0.3427±0.0059	0.2185±0.0071	0.3083±0.0086	0.4580±0.0125
PICP	0.8267±0.0490	0.7467±0.0317	0.6822±0.0437	0.9222±0.0337
MPIW	0.9171±0.320	0.6378±0.0230	0.8979±0.0349	1.5845±0.0194

According to the metrics shown in Tables S4 and S5, we selected the model that uses as the kernel the sum of the Radial basis function (RBF) kernel and the Linear kernel, with ARD in an intrinsic model of coregionalization (ICM), as the best GPR model. This model shows high R^2 and PICP values, and the lowest RMSE and MPIW values. The parity plots of the selected model predictions of the gasification outputs versus the experimental results (z-score standardized) are shown in Fig. S3. The coefficient of determination, R^2, shows values between 0.82 and 0.98, which indicates a good predictive performance. The low standard deviation of the metrics calculated from the run replications indicates that our model is stable. After selecting the best model, we retrained it on all data to perform an experimental validation of the model, and for prediction purposes on input data collected from the literature.

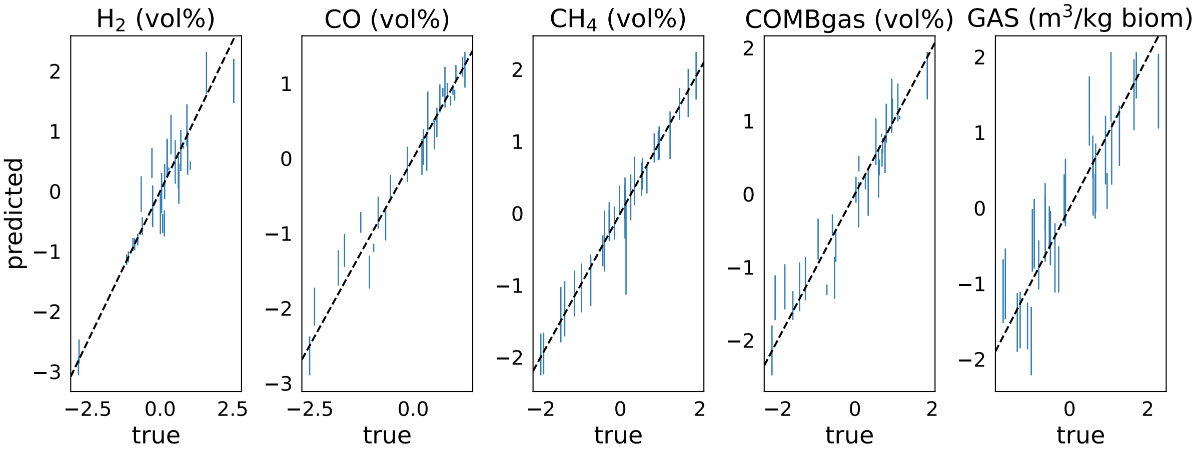


Fig. S3. Leave-one-out cross-validation results (test points). True and predicted values are z-score standardized using the mean and standard deviation of the training set. Error bars show the predicted standard deviations, for CH_4 we compute the error bars considering the error propagation and covariance (see Section 4.5). R^2 (mean and standard deviation of 15 runs): H_2=0.892±0.005, CO=0.960±0.001, CH_4=0.978±0.000, COMBgas=0.925±0.002, GAS=0.819±0.002.

The results of the experimental validation of the model are shown in Fig. S4, where we can compare the experimental and predicted results of novel gasification experiments using two new biomasses not previously used for the training of the model. These results show a good agreement between the experimental results and the corresponding values predicted by the model.

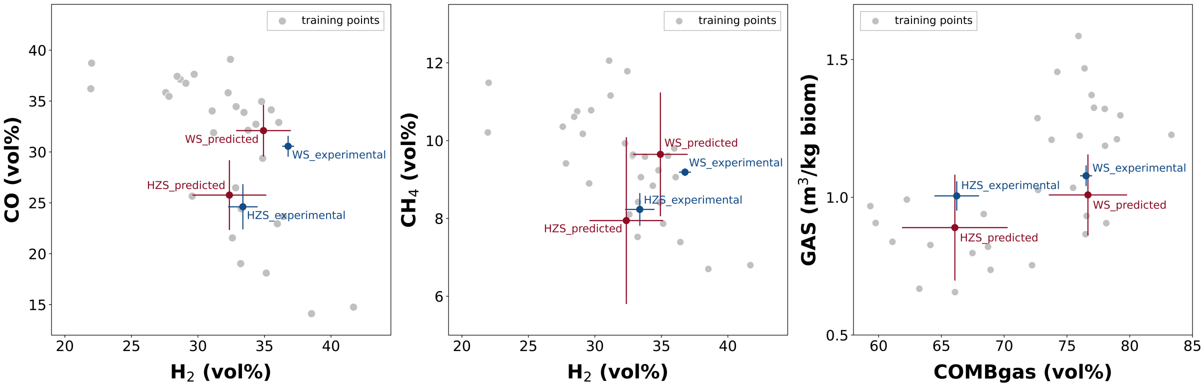


Fig. S4. Results of the experimental validation of the model. For experimental validation we compare new experimental results obtained from the gasification of walnut shells (WS) (T=1173 K, SA=2.33, and SR=0.13) and hazelnut shells (HZS) (T=1173 K, SA=2.33, and SR=0.25) with the values predicted by the model.

Section 4.4. Randomization test: y-scrambling

We used a randomization test, y-scrambling,[S13] to ensure that the model learned something meaningful. To test the null hypothesis that the model uncovered no meaningful pattern in the data, we trained 150 models with scrambled labels. We used leave-one-out cross-validation (LOOCV) to measure the metrics for each of the models. Fig. S5 compares the predicted values with the true values for the test points from the LOOCV results for one of the models trained with randomly shuffled data, showing a very poor prediction performance.

Fig. S6 shows the histograms of the R^2 scores (LOOCV) for the predictions of the 150 different models trained with scrambled labels. The red line indicates the R^2 score obtained for the model trained on the original data, i.e., with unshuffled labels. For all outputs, this value is much better that those obtained for the models trained with shuffled labels. We also calculated an empirical p-value against the null hypothesis that features and targets are independent. The p-value is very low ($p = 0.006$), which indicates a very low probability that the good score obtained with the unshuffled data would be obtained by chance. These results indicate that our model performs poorly on randomly shuffled data, meaning that that the model predictions are not made just by chance, and verifying the robustness of the model.

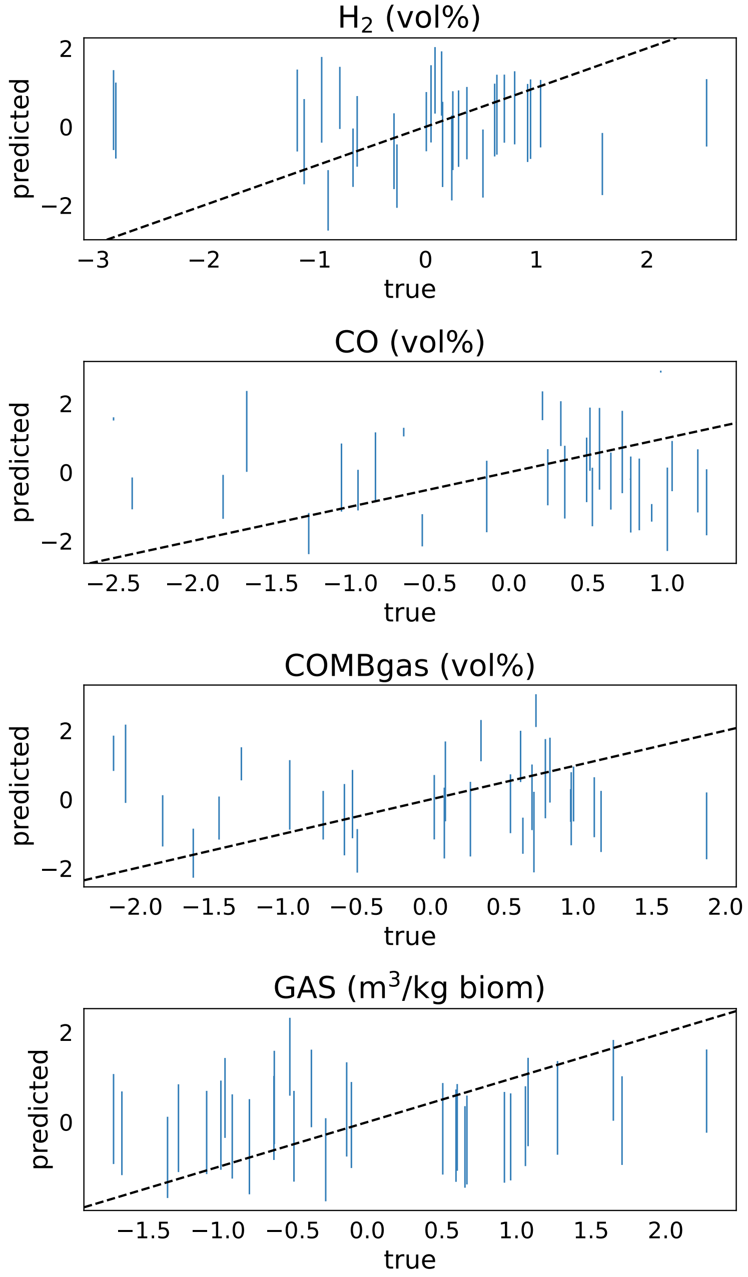


Fig. S5. Example of parity plots for one model trained with randomly shuffled data. Leave-one-out cross-validation results (test points). Errorbars show the predicted standard deviations. R^2: CO=-1.0131, H_2=-0.3707, COMBgas=-0.8129, GAS=-0.0342.

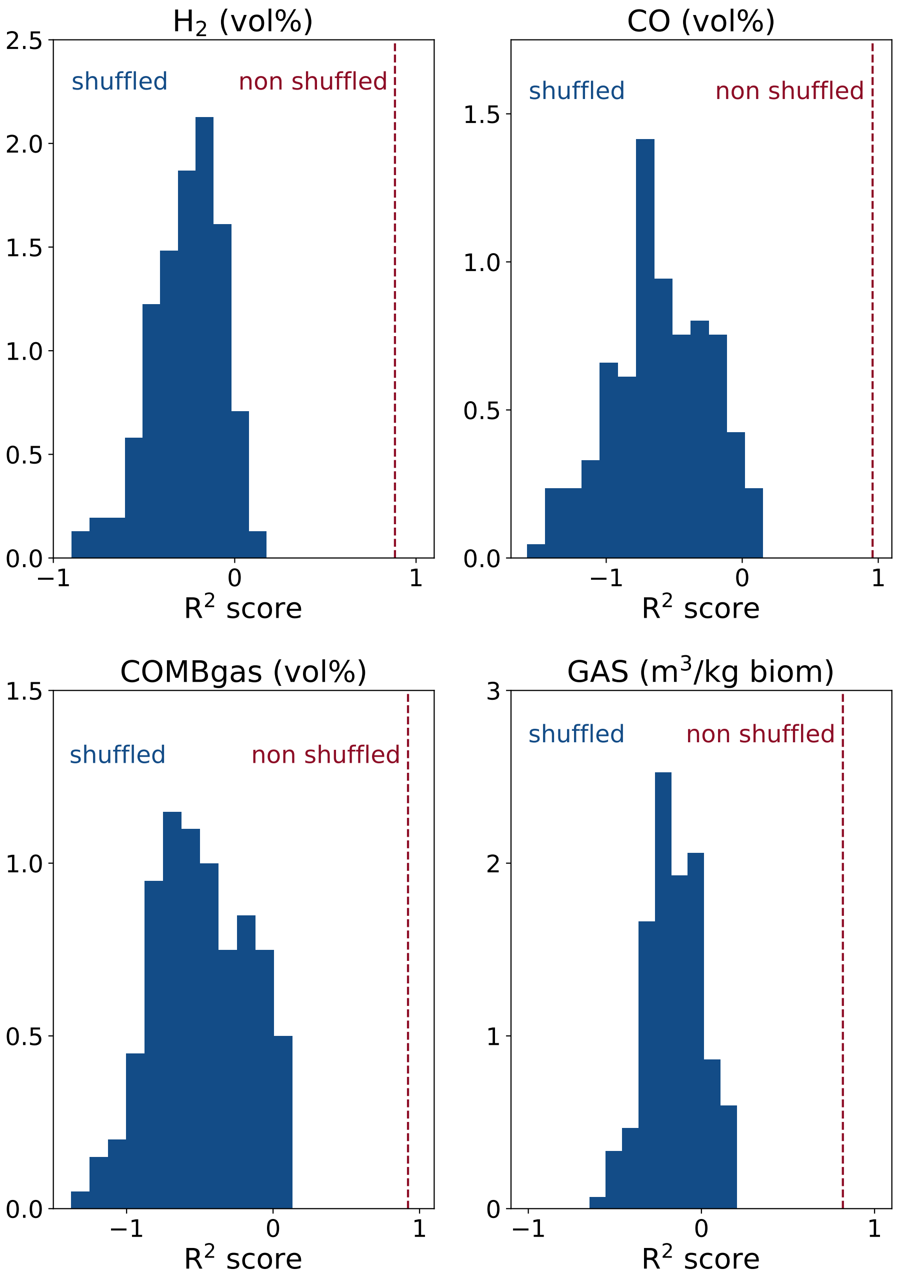


Fig. S6. Results of the randomization test: y-scrambling. Histograms of R^2 scores (LOOCV) for predictions of the models trained with shuffled labels. Results from 150 repetitions (permutation test p-value = 0.006 for all outputs[S14,S15]). Red line indicates the R^2 score obtained for the model trained on the original data, i.e., with unshuffled labels.

Section 4.5. Predicting methane volume fraction: Error propagation

The methane volume fraction is one of the outcomes of the biomass gasification process. However, for this variable, the noise in the data is relatively large, which makes it more difficult to build a sufficiently reliable predictive model based on these data. Therefore, we predicted the CO, H_2 and total combustible gas (COMBgas) concentrations (COMBgas is the sum of H_2, CO, and CH_4), which are variables that can be predicted more accurately with our model. Then, we estimated the methane volume fraction from those outputs.

For the CH_4 concentration, as it is estimated from other outputs, we compute the error bars (or uncertainty) considering the error propagation. The associated covariance term is also taken into account since we consider that the measurements are dependent (and they have non-zero covariances). The parity plot of the CH_4 concentration predictions versus the experimental results is shown in Fig. S3. The value of the coefficient of determination, R^2, indicates a good prediction performance for this output.

Section 5. Feature importance analysis

Section 5.1. Methods

To determine the feature importance we used two approaches: the SHapley Additive exPlanations (SHAP) technique[S16] marginalized over the full dataset to calculate SHAP values, and the partial dependence plots[S17] in which the plotted features are marginalized out over the distribution of all features.

To perform the SHAP analysis for the CH_4 concentration, we built a custom function for the SHAP kernel explainer that accounts for the estimation of this output from the model predicted targets.

We used partial dependence plots[S17] to show the marginal effect that the features have on the predicted outcomes of our machine learning model. The partial dependence function for regression is defined by Eq. (7).

$$\hat{f}_S(x_S) = E_{X_C}\left[\hat{f}(x_S, X_C)\right] = \int \hat{f}(x_S, X_C) d\mathbb{P}(X_C) \tag{7}$$

The x_S are the features for which the partial dependence function is plotted and X_C are the other features used in the machine learning model $\hat{f}$.

Section 5.2. SHAP importance values

We determined the SHAP values for the features that characterize the biomass and the gasification operating process (Fig. S7). The x-axis gives the SHAP value, i.e., the impact of this feature with respect to the baseline (the average gasification products values), and the y-axis displays the most relevant features that impact the output the most. Each point on these graphs corresponds to one SHAP value for a prediction and feature. The color coding gives the value of this feature (red high, blue low). If a feature is irrelevant, all dots, irrespective of color, will be on the baseline. A positive SHAP value with a red color means that a positive value of this feature leads to a higher output value than the average. The global importance of the features (across all biomass) is therefore given by the width of the spread of the SHAP values. In Fig. S7, we show the SHAP summary plots for the H_2, CO, CH_4, and combustible gas (COMBgas) concentrations in the gasification outlet gas, and for the gas yield (GAS).

Focusing on the gasification operating parameters, our model identifies the temperature as one of the most important features for our outputs. Fig. S7a shows that the most important feature shown by the SHAP plots for the H_2 concentration prediction is the gasification temperature, and we can see that high-temperature values have positive importance values, which means that the gasification temperature has a positive influence on the prediction of the H_2 content in the outlet gas. We can also see that a very low gasification temperature (dark blue color points) has a very relevant negative effect on the hydrogen production. We can explain this because higher temperatures favor the endothermic water gas and steam reforming reactions according to Le Chatelier's principle, favoring the conversion of formed methane and char produced during the gasification process.[S18] A similar effect of the gasification temperature is found for the gas yield (Fig. S7e), also highlighting the negative effect of low gasification temperatures on the gas production.[S19] Higher temperatures can favor the production of gas during the biomass devolatilization, as well as promote cracking reactions of secondary hydrocarbons, tars, and char, together with steam reforming and gasification reactions that increase the gas production.[S20,S21] Although of slightly lower importance, the gasification temperature has also a positive influence on the production of combustible gas (Fig. S7d), undoubtedly influenced by the H_2 concentration.

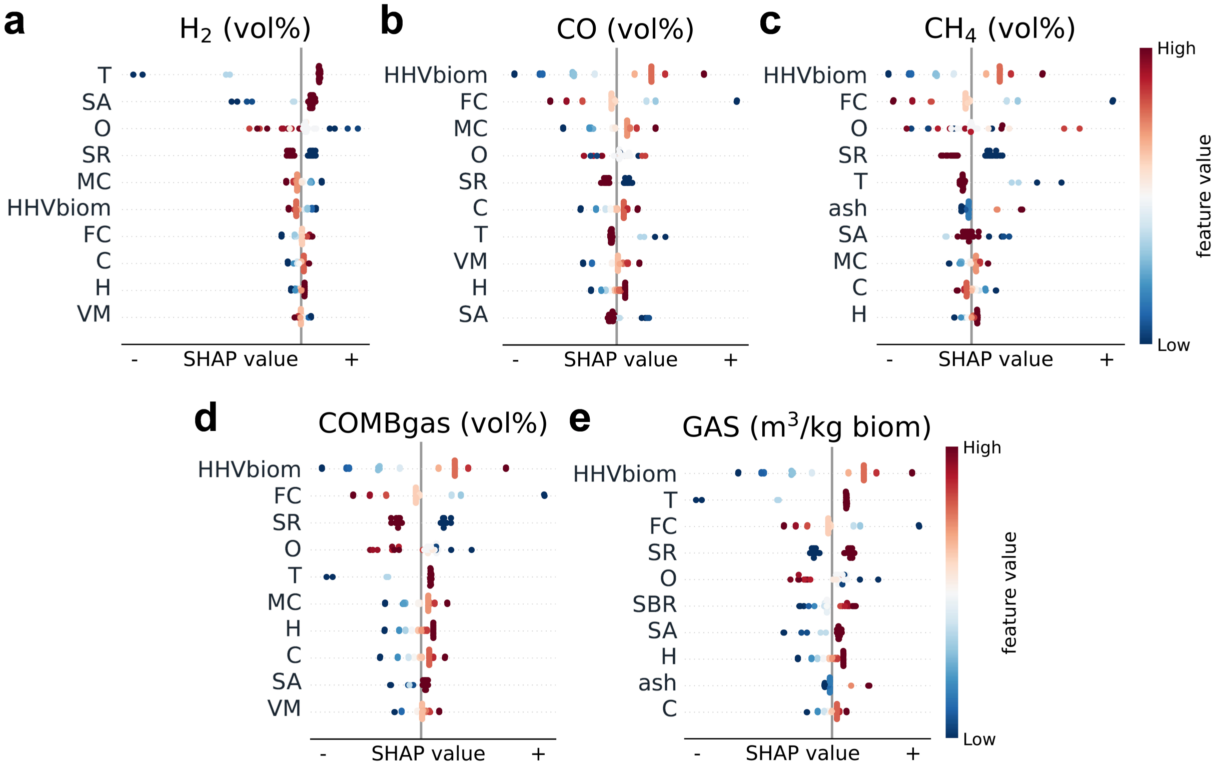


Fig. S7. Impact of the feature values on the gasification predictions. The plots show on the x-axis the SHapley Additive exPlanations (SHAP) values that indicate the impact of a feature on the model output compared to a baseline (vertical lines at $x = 0$). The y-axis shows the most relevant features, that impact the output the most. The color coding for the points indicates the feature values. For the SHAP analysis, we used all the points in our dataset, which were used to build the coregionalized GPR model. T: temperature (K), SA: steam-to-air ratio, SR: stoichiometric ratio, SBR: steam-to-biomass ratio, C: carbon (wt%), H: hydrogen (wt%), O: oxygen (wt%), ash: ash content (wt%), VM: volatile matter (wt%), FC: fixed carbon (wt%), HHVbiom: biomass higher heating value (MJ/kg), MC: moisture content (wt%). **a** H_2 predictions. **b** CO predictions. **c** CH_4 predictions, computed by predicting the combustible volume fraction along with H_2 and CO. **d** Combustible gas (COMBgas) predictions. **e** Gas yield (GAS) predictions.

On the contrary, the gasification temperature has a negative effect on the prediction of the CO (Fig. S7b) and CH_4 (Fig. S7c) concentrations, and higher gasification temperatures decrease their values. The concentration of CO in the produced gas is the result of different competing reactions that occur during the conversion process of the biomass. There are several reactions that can produce CO during the gasification process, such as the partial oxidation, water gas, steam methane reforming, tar reforming, and Boudouard reactions, while it is consumed by the water gas shift (WGS) reaction. A higher production of CO by the partial oxidation reaction is probable at lower temperatures, whereas at higher temperatures a higher CO production by endothermic reactions can shift the WGS equilibrium towards the consumption of CO. Higher temperatures also favor the endothermic steam methane reforming reaction, decreasing the CH_4 content produced during the biomass conversion. We can therefore conclude that higher gasification temperatures favor the H_2 production at the expense of CO and CH_4.

Looking at the other gasification operating parameters, we can see in Fig. S7a that the steam-to-air (SA) ratio has a very high importance for the H_2 concentration prediction. The feature importance analysis indicates that high values of SA have a positive effect on the H_2 content, while negative importance values are shown by low SA ratios. This confirms the importance of performing the gasification process with steam, compared to air, if our objective is a higher production of H_2. Although with lower importance, a positive effect of the SA ratio on the gas yield is also shown, since it favors the overall conversion reactions. However, in the case of the CO and CH_4 concentrations, lower SA ratios have positive importance values and hence higher SA ratios show a negative importance, which indicates a negative influence of this feature on the prediction of the CO and CH_4 contents. This can be explained because higher steam contents in the gasifying agent favor the WGS reaction, producing a higher amount of H_2 at the expense of CO. Likewise, a higher steam content favors the steam methane reforming reaction, which also produces H_2 from CH_4.

Another gasification operating parameter that has some influence on the outputs is the stoichiometric ratio (SR). SR is a measure of the amount of oxygen used in the gasification process compared to the stoichiometric amount of oxygen needed for a complete combustion of the feedstock. From the SHAP plots in Fig. S7, it can be deduced that higher SR values have a negative effect on the prediction of the H_2, CO, CH_4 and combustible gas concentrations. However, a positive effect of SR is seen in the case of the gas yield prediction. When a higher amount of air is used, the combustion reaction is favored and the conversion increased, in turn reducing the partial oxidation reaction that would convert the biomass organic matter to CO. This can explain a lower concentration of combustible gases since they would be converted to CO_2 to a higher extent. On the other hand, the air combustion reactions can avoid an excessive decrease in the reactor temperature by the endothermic reactions, favoring the overall conversion and increasing the gas production, which explains the positive effect detected on the gas yield. Indeed, the main reason for using air in the gasifying atmosphere is to compensate for the high demand for heat by the reforming endothermic reactions. Finally, the steam-to-biomass ratio, SBR, was also a studied gasification operation parameter, which only shows a certain importance on the gas yield (Fig. S7e), probably because the SA ratio and SR values have a higher relative relevance when the gasification process is carried out under a steam-oxygen atmosphere.

The effect of the biomass properties on the process performance has hardly been studied in the literature. From the feature importance analysis, we can see in Fig. S7 that the most important biomass property is the biomass calorific value (HHVbiom). The high importance of HHVbiom to explain the gasification results has not been reported in previous studies in the literature, where the importance of the biomass C content has usually been highlighted.[S22–S24] The SHAP values show that all outputs, except H_2, increase if we increase the value of the biomass calorific value. In the case of H_2, the biomass calorific value has a relatively lower (negative) importance for the prediction of the H_2 concentration, which makes sense as it is favoring the production of the other gases.

Fig. S7 also shows that the fixed carbon (FC) content of the biomass has a substantial importance on the prediction of the gasification outputs. High values of this feature correspond with low SHAP values for the CO, CH_4 and combustible gas concentrations, and also for the gas yield. This indicates that by decreasing the value of the biomass FC content, we increase the predictions of these gasification outputs. From this, a negative influence of this biomass property on the production of combustible gas from the gasification process can be deduced. On the contrary, the biomass FC content shows a positive influence on the H_2 concentration prediction. The biomass FC content is related to the volatile matter (VM) and ash contents. Biomass samples usually have a relatively much higher content of volatile matter than FC and ash. Therefore, we can say that a higher FC value is usually related to a lower value of VM. Although the biomass VM content shows lower importance on the predictions, its favorable effect can also be found from the feature importance analysis in Fig. S7. Accordingly, from these results, we can deduce that the biomass VM content has a positive influence on the CO and combustible gas production, but negative on the H_2 concentration.

The biomass O content also shows a relevant importance on the predictions of the gasification outputs. The most relevant finding shown by the feature importance plots in Fig. S7 is the negative influence of the biomass O content on the prediction of the H_2 and combustible gas concentrations and on the gas yield. A higher value of the O content could be related to lower contents of C and H (majority elements in the biomass elemental composition). For the CO and combustible gas concentrations, and the gas yield, the biomass C and H contents show a significant positive influence on their predictions, i.e., they have an opposite effect than the O content, which means that higher contents of C and H would favor the prediction of these outputs. The importance of the C and H contents is lower in the case of the H_2 concentration, which agrees with the higher importance of the O content for this output, and we can therefore deduce that a low biomass O content could be related with a higher H_2 production.

Finally, the biomass moisture content (MC) shows a positive effect on the predictions of CO, CH_4 and combustible gas concentrations. Higher contents of moisture in biomass can decrease the gasifier temperature by heat consumption, decreasing the conversion of other gases to H_2. This effect is noticeable in this study to a certain extent, although biomass with significantly high moisture contents were not used in this work. If this type of biomass wants to be gasified, a study including biomasses with really high MC contents is recommended to have a better understanding of the effect of the moisture content.

Section 5.3. Partial dependence plots

To determine the feature importance in our study we also used a second approach, i.e., the partial dependence plots between some of the most relevant features, in which the plotted features are marginalized out over the distribution of all features.[S17] Firstly, in Fig. S8a we show the dependence between the two most important gasification operating parameters, the gasification temperature (T) and steam-to-air (SA) ratio. We can clearly see that the gasification temperature has a higher importance for the predictions than the SA ratio, since small changes in the temperature have a strong impact on the outputs, while the effect is lower when the SA ratio changes. In relation to the SA ratio, it should be pointed out that the positive effect of increasing SA on the H_2 prediction is especially relevant at lower values of the steam-to-air ratio, when SA<1, and this effect decreases with the excess steam in the gasifying atmosphere. This agrees with the positive effect of a steam atmosphere on the hydrogen production. Accordingly, the lower effect of higher SA values can also be detected in the other outputs. From this plot, we can also conclude that higher temperatures and higher SA ratios significantly increase the production of H_2, which in turn leads to higher production of combustible gas and gas yield from the biomass gasification process. The importance analysis performed through our model not only tells us the positive or negative influence of the process parameters, but it gives an importance value to each feature, and this allows us to know which of the gasification operating parameters are more relevant for the prediction of each of the process outcomes. In practice, this can help us to prioritize some experimental conditions as a function of the desired results.

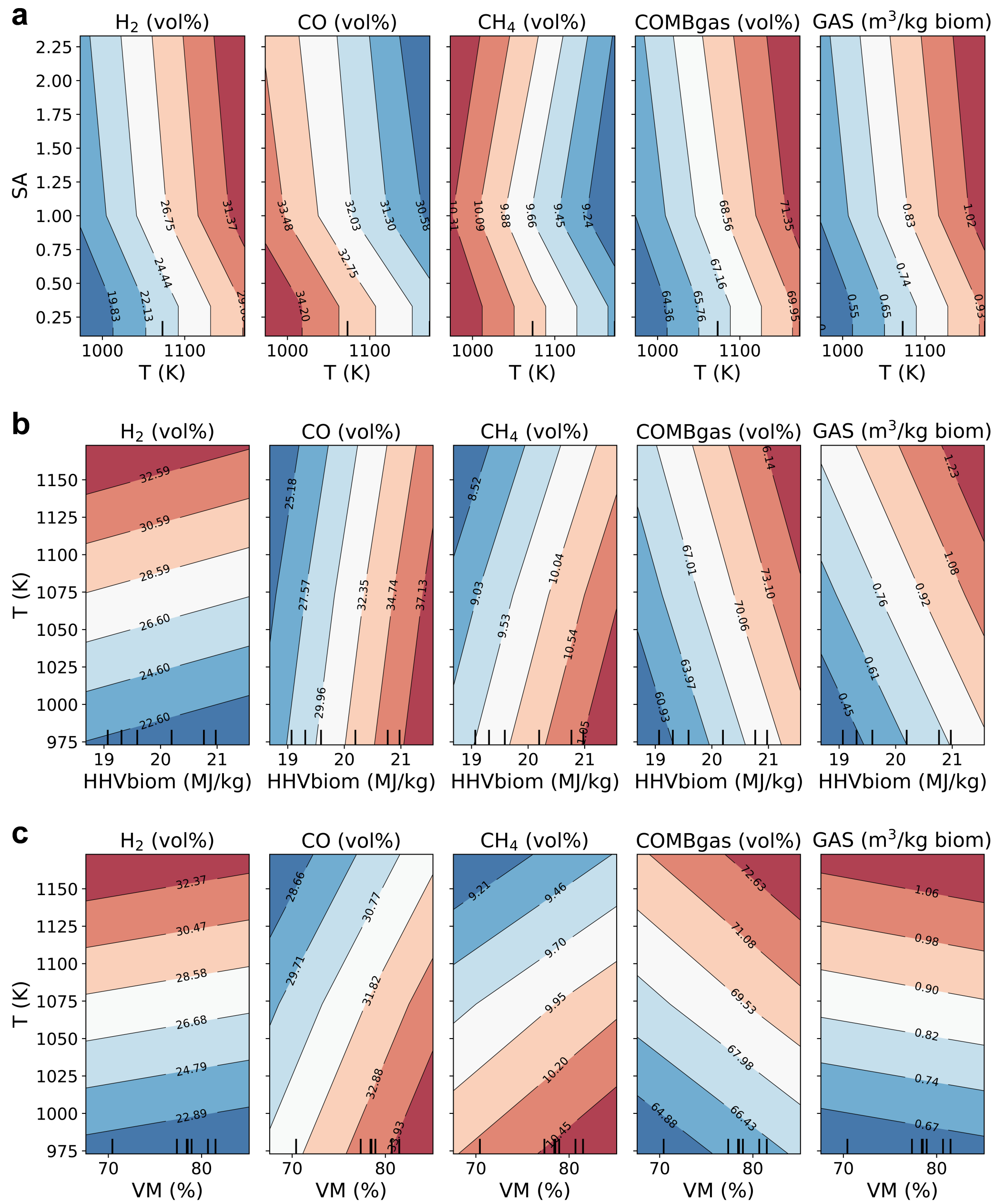


Fig. S8. Partial dependence plots between features. **a** Gasification temperature (T) and steam-to-air (SA) ratio in the gasifying atmosphere. **b** Biomass calorific value (HHVbiom) and gasification temperature (T). **c** Biomass volatile matter (VM) content and gasification temperature (T).

Secondly, given the high importance of the process temperature, it can be interesting to compare the influence on the outputs of some relevant biomass properties with that of the gasification process temperature. For this purpose, we show in Fig. S8b the partial dependence plots for the gasification temperature and the biomass calorific value

(HHVbiom), which, interestingly, was revealed as one of the biomass properties with higher importance on the prediction of the gasification outputs, as mentioned above. We can see that the biomass calorific value is a really relevant parameter for the production of CO, CH_4 and combustible gas, and also for the gas yield. Although high temperatures are unquestionably favorable to obtain high concentrations of combustible gas and high gas yield, the importance of the calorific value of the biomass feedstock cannot be underestimated if we want to obtain a high-energy gasification product. Fig. S8a also shows that the importance of the biomass calorific value is much lower for the prediction of the H_2 concentration, which means that the biomass HHV is not a so relevant parameter for the hydrogen production. From these results, we could deduce that since CO and CH_4 are more immediate products from the biomass conversion, while H_2 is produced from the conversion of the previously produced gases, the influence of a higher energy contribution by the solid biomass is lower for the hydrogen production.

On the other hand, in order to better understand the effect of the VM matter content on the outputs, we show the partial dependence plot for the VM content and the gasification temperature (Fig. S8c). As expected, we can see that the effect of the temperature is clearly higher than that of VM for the prediction of the H_2 concentration, but also for the CH_4 and combustible gas contents, and for the gas yield. However, importantly, the results show that the biomass VM content can have a significant influence on the CO concentration in the outlet gas. The positive effect of VM on these gases indicates a positive relationship between the volatile matter of the biomass and its conversion to CO, and CH_4 at a lower extent, which would be generated during the devolatilization step of the gasification process. On the other hand, the negative effect of VM on the H_2 concentration can be highlighted, indicating that a higher hydrogen production could be favored by biomasses with lower VM contents (and hence higher FC contents).

To summarize, from our feature importance analysis, we want to highlight the competing behavior of the different components of the output gas based on the gasification process parameters and biomass properties. As relevant remarks, we have learned that the production of CO, CH_4 and overall combustible gas is significantly favored by high values of the calorific value of the biomass feedstock, and also that high biomass contents of C, H and VM could promote the combustible gas generation. However, the H_2 production would mainly be favored by high gasification temperatures and steam-to-air ratios in the gasifying atmosphere, and besides, high FC (and hence low biomass VM) together with low O contents appear to be favorable conditions for the hydrogen generation.

Section 6. Selection of key performance indicators (KPIs) for biomass applications

To select the key performance indicator (KPI) parameters that help us to choose the best use for a given biomass, we analyze the experimentally measured gasification outputs, i.e., the CO, H_2, CH_4 and combustible gas (COMBgas) concentrations in the gasification outlet gas, and the gas yield (GAS). In addition, based on those predicted outputs, we also estimate other gasification outcomes that can be relevant to evaluate the performance of the gasification process, i.e., the molar H_2/CO ratio, the gas calorific value (HHVgas) and the gas energy yield (E_{yield}). The estimation of these parameters is described in Section 1.

At first, we can think that the combustible gas production (COMBgas) and the hydrogen production (H_2 (vol%)) could be good indicators of the biomass potential according to the main possible uses of the gasification product gas, i.e., heat/electricity generation, and synthesis of other products such as hydrogen, synthetic natural gas, methanol or hydrocarbon fuels. In Fig. S9 we show the results of the predicted gasification outputs for a sample of all biomasses gathered from the literature. We have not used these biomasses in our training set, so these results are predictions of our model. In this case, we use our model to predict the gasification outputs under fixed process operating conditions (T=1173 K, SA=2.33, and SR=0.25). The values for each gasification output have been scaled between 0 and 1 to be able to use the same color palette for all the output variables and so compare them. In Fig. S9, the biomasses shown in the x-axis are ordered according to the results for the combustible gas concentration (COMBgas), which shows blue colors when it has lower values (left side of the plot) and red colors when COMBgas is higher (right side of the plot). We have chosen this output as a preliminary reference parameter because if we want to maximize the production of energy from the biomass conversion, we would like to maximize the combustible gas concentration from the gasification process. From this plot, we can see that the order of the biomasses according to the CO concentration is quite similar to that established for COMBgas. The order of the biomasses according to the CH_4 concentration does not match exactly that for CO and COMBgas, but a similar color pattern is shown, with lower values on the left side and higher values on the right side. In this case, the order of the individual biomasses is slightly different, since some biomasses have relatively very high values of CH_4 content (darker red colors than the reference pattern), such as eucalyptus sawdust (EUCS), poplar sawdust (POP), beech wood (BEE), or prune pits (PRU).

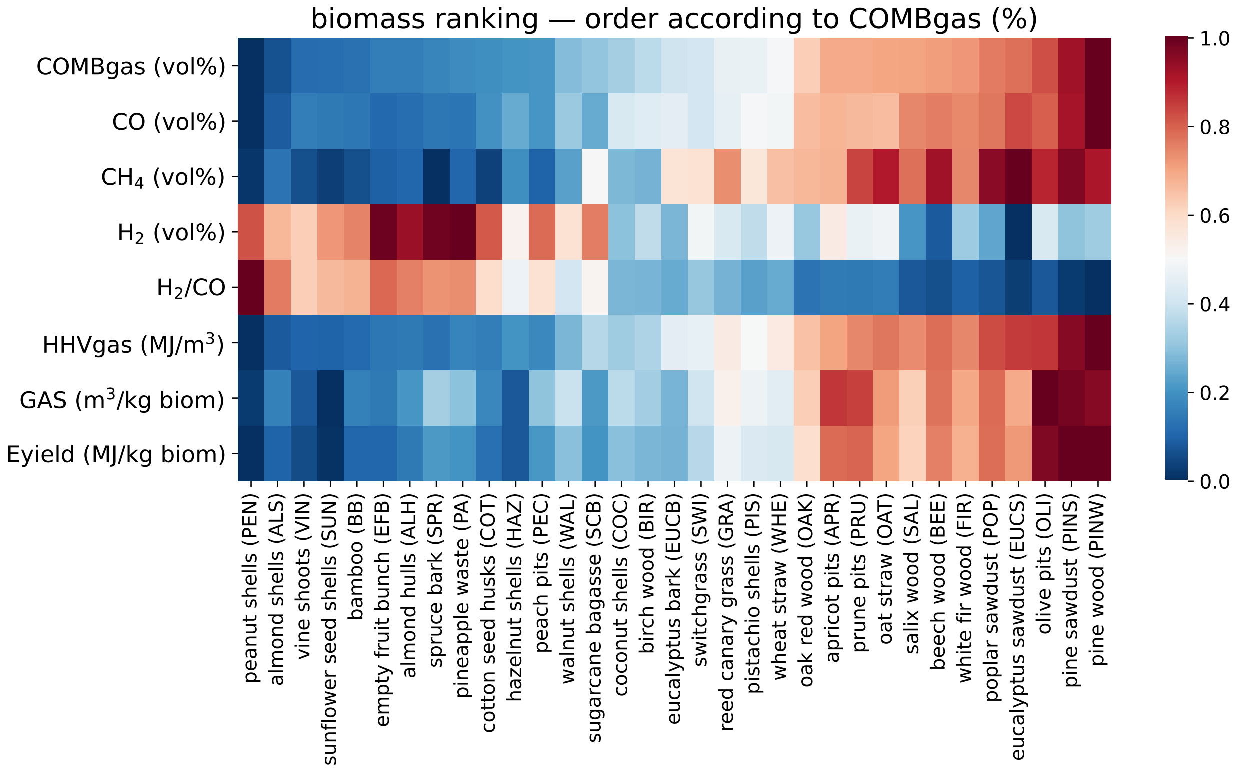


Fig. S9. Biomass ranking plots to select the key performance indicators (KPIs) for biomass applications. All predicted gasification outputs are shown for some biomasses gathered from the literature. The values for each gasification output have been scaled between 0 and 1 for comparison purposes. Biomasses shown in the x-axis are ordered according to the results for the combustible gas concentration (COMBgas). The gasification outputs have been predicted under fixed process operating conditions of T=1173 K, SA=2.33, and SR=0.25.

The calorific value of the product gas, HHVgas, follows the same general tendency established by the combustible gas, and the order of the biomasses in this case is quite similar to that for COMBgas. Indeed, the calorific value of the product gas is determined from the concentration of combustible gases in the gasification product. However, some biomasses show a darker red color for HHVgas than for COMBgas, which is due to high CH_4 concentration values (e.g., eucalyptus sawdust (EUCS), poplar sawdust (POP), beech wood (BEE), or prune pits (PRU)). A higher value of the CH_4 content for a given biomass increases the calorific value of the produced gas at a higher extent, since the calorific value of CH_4 is higher than that of the other combustible gases in the blend. This indicates that the CH_4 concentration can have a relevant effect on the energy production.

The gas yield, GAS, also shows the same general color pattern established by the combustible gas, but we can see some differences. GAS is related with the biomass conversion, since it is a measure of the amount of solid fuel that is transformed into gas. We can see that a higher gas production does not always mean a higher gas calorific value. For example, eucalyptus sawdust (EUCS) has lower gas yield but higher gas calorific value, which is associated to a higher CH_4 content, while apricot pits (APR) and prune pits (PRU) show higher gas yields. From these results, we can deduce that if we want to obtain the maximum conversion to energy from the gasification process, we can not only trust on the combustible gas concentration in the outlet gas, or on its heating value, as performance indicator parameters, since the gas yield can also be a relevant parameter that depends on the biomass composition. On the other hand, E_{yield} is a parameter related with both the gas yield and the calorific value of the produced gas, and it represents the amount of energy in the product gas per mass unit of biomass, giving a measure of the conversion efficiency to energy. Fig. S9 shows that the energy yield is very influenced by the gas yield value, which means that a higher gas yield can significantly increase E_{yield}. Therefore, E_{yield} could be the best KPI for the energy production from biomass gasification, since it accounts for the calorific value and the conversion efficiency.

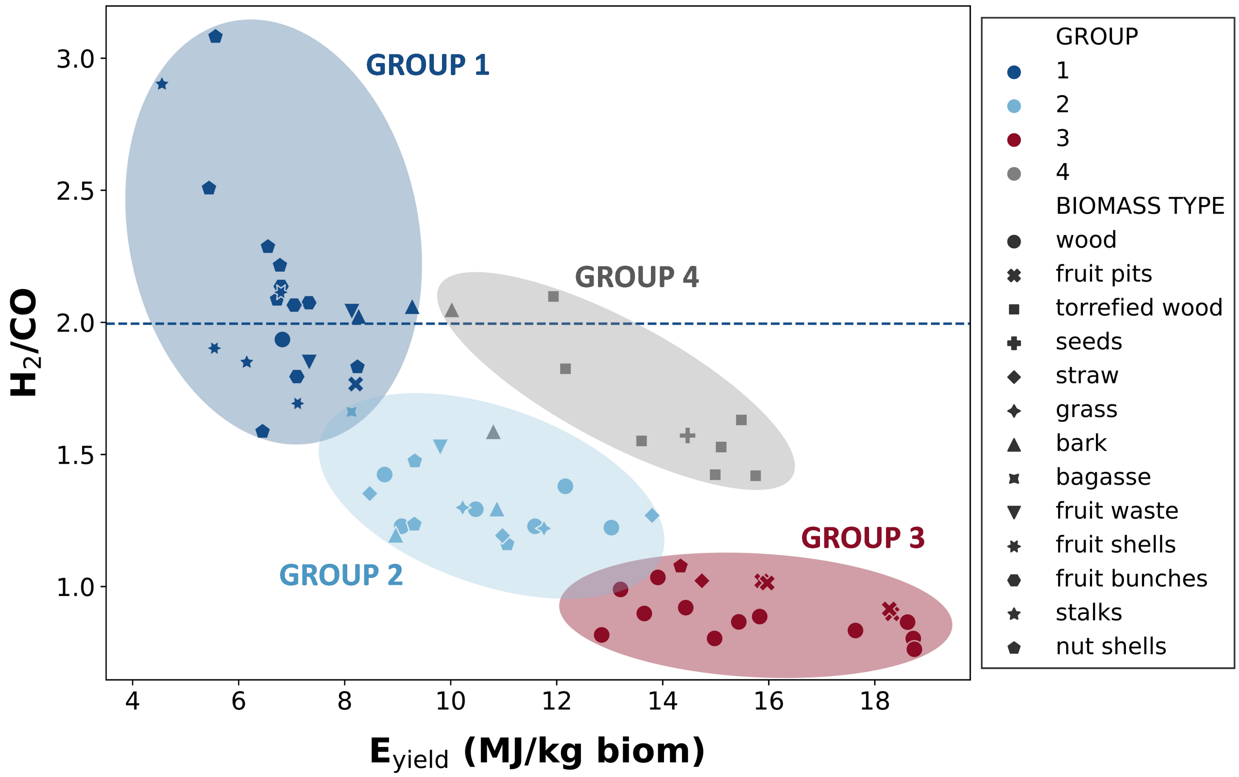
In contrast, Fig. S9 shows that the values of H_2 concentration and H_2/CO ratio follow a different trend. We can say that for these two outputs the distribution of the biomasses follows an inverse order than that established by the combustible gas concentration, showing an opposite color scale, with red colors on the left side of the plot, i.e., higher output values, and blue colors on the right side of the plot, where the biomasses with the lowest values of these variables are located. However, the H_2 concentration values do not match exactly the same order than the H_2/CO ratios. The H_2/CO ratio is usually calculated for synthesis applications of the gasification gas and used as a reference parameter. This means that the H_2/CO ratio would be a more precise KPI than the H_2 concentration for these applications.

If we look at the full plot, we can deduce that biomasses with red colors on the right side of the plot are characterized by high values of CO, CH_4 and combustible gas concentrations, as well as high values of the gas calorific value, gas yield and gas energy yield, i.e., gasification outputs mainly associated with the production of energy from the biomass gasification process. However, biomass with red colors on the left side of the plot produce higher values of H_2 from gasification, and hence higher H_2/CO ratio values. This means that biomasses able to produce higher values of E_{yield} are expected to have a higher potential to be used for applications based on the energy production, while biomasses that give higher values of H_2/CO ratio would have a higher potential to be used for synthesis applications. From this evaluation, E_{yield} and H_2/CO ratio are the parameters that we select as key performance indicators for the exploration of biomass applications.

Section 7. *K*-means clustering

K-means is an unsupervised clustering algorithm[S25] to cluster unlabeled data to a number of groups or clusters. *K*-means clustering provides an unsupervised grouping of samples with similarity. With this analysis, we can find clusters of data points (biomasses) which share similar characteristics, being different from other data points that are in a different cluster. In our work, we performed a cluster analysis to identify the groups of biomasses that share similar gasification outputs, with the objective of finding common biomass properties that can explain the different gasification results in each group.

We applied *k*-means clustering to the gasification outputs, including the H_2, CO, CH_4 and combustible gas (COMBgas) concentrations in the gasification gas, gas yield (GAS), molar H_2/CO ratio, and gas energy yield (E_{yield}). `K-means` scikit-learn Python algorithm was used, with `k-means++` as method for initialization and a number of clusters of 4.[S9] We used several methods to choose the number of clusters and we obtained similar results, selecting four clusters for our analysis. The groups of biomasses that the *k*-means algorithm identified based on the gasification outputs are shown in Fig. S10 as a function of the E_{yield} and H_2/CO ratio values, where we can see the distribution of the different biomass types studied into the different groups. In Fig. S11 we represent the identified clusters as a function of the different biomass properties, showing the distribution of the biomass types studied into the groups.



GROUP 1	GROUP 2	GROUP 3	GROUP 4
cotton stalks	sugarcane bagasse	oat straw	pine bark
vine shoots	corn stover	apricot pits	grape seeds
pineapple waste	barley straw	prune pits	spruce torrefied
orange peel/seeds	wheat straw	olive pits	beech torrefied
empty fruit bunches	switchgrass	palm kernels	poplar torrefied
cotton seed husks	reed canary grass	willow wood	eucaliptus torrefied
almond hulls	walnut shells	salix wood	pine torrefied
peanut shells	coconut shell	oak red wood	
cacao shells	eucalyptus bark	white fir wood	
sunflower seed shells	oak bark	beech wood	
almond shells	demolition wood	pine wood	
hazelnut shells	forest residue wood	eucalyptus sawdust	
spruce bark		poplar sawdust	
bamboo		pine sawdust	

Fig. S10. Biomass groups identified by the *k*-means cluster analysis represented as a function of the E_{yield} and H_2/CO ratio values. *K*-means cluster analysis was based on the gasification outputs. Gasification operating conditions: T=1173 K, SA=2.33, and SR=0.25. The distribution of the different biomass types studied into the different clusters is also shown.

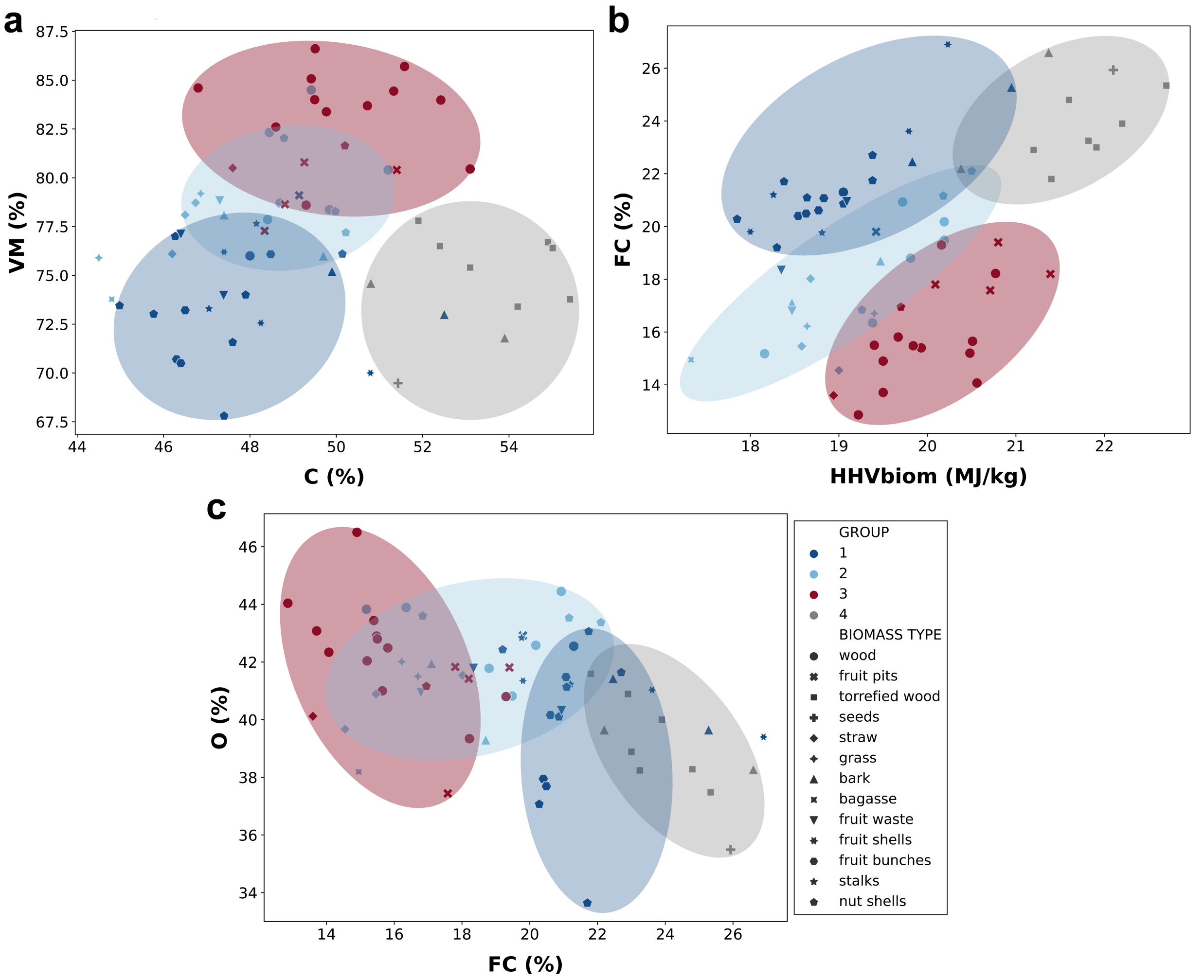


Fig. S11. Biomass groups identified by the *k*-means cluster analysis represented as a function of different biomass properties. *K*-means cluster analysis was based on the gasification outputs. Gasification operating conditions: T=1173 K, SA=2.33, and SR=0.25. The distribution of the different biomass types into the clusters is shown in the legend.

Section 8. Tuning the gasification gas characteristics by changing the process conditions

With our model we can tune the gasification gas characteristics by changing the process conditions. The feature importance analysis shows that the gasification temperature and steam-to-air ratio are the most important process variables. Therefore, we show here how the gasification outputs change if we modify these process parameters. In Fig. S12 we represent the predictions of Eyield and H_2/CO ratio for three gasification temperatures and three SA ratios. The values of these gasification outputs decrease with the decrease in both temperature and SA ratio. For example, in Fig. S12a we can see that at lower gasification temperatures (973 K and 1073 K), we are not able to achieve H_2/CO values around 2 for almost any biomass, and we would need a further WGS reaction if we want to increase the H_2 content in the syngas. The Eyield values are also lower for gasification temperatures below 1173 K, although we can still obtain relatively high values for some biomasses at 1073 K. Fig. S12b shows that the decreasing effect of the SA ratio is lower than that of the temperature, and we can still obtain H_2/CO values close to 2 or relatively high Eyield values for some of the biomasses.

In general, we can conclude that a gasification temperature of 1173 K and a SA ratio of 70/30 are the best-operating conditions if we want to maximize the energy content in the gasification gas for a given biomass. However, these results indicate that by changing the temperature and SA ratio we can obtain a target H_2/CO ratio in the gasification gas. For example, for some biomasses that are able to give high values of the H_2/CO ratio, it would be possible to produce a H_2/CO ratio of two if we use a specific gasification temperature between 1073 K and 1173 K (Fig. S12a). For those biomasses, a H_2/CO ratio of two could also be obtained at 1173 K if we adjust the SA ratio conveniently between 25/75 and 70/30 (Fig. S12b).

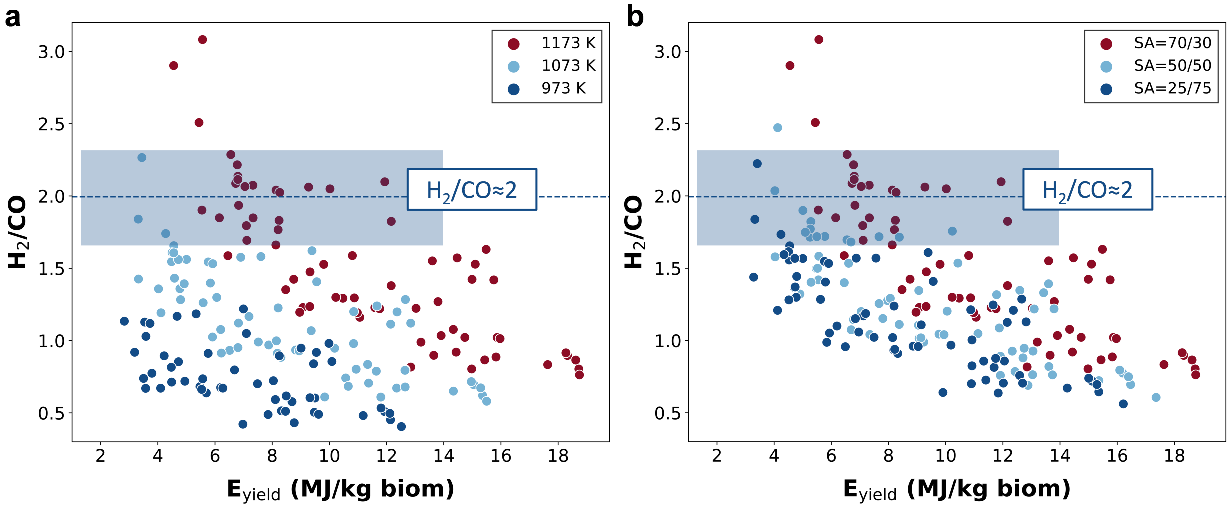


Fig. S12. Predictions of the key performance indicators (KPIs), E_{yield} and H_2/CO ratio, under different process conditions of temperature (T) and steam-to-air (SA) ratio. a Effect of the gasification temperature (T) on the predictions. **b** Effect of the steam-to-air (SA) ratio on the predictions.

References

(S1) González-Vázquez, M.; García, R.; Gil, M.; Pevida, C.; Rubiera, F. Comparison of the gasification performance of multiple biomass types in a bubbling fluidized bed. *Energy Conversion and Management* **2018**, *176*, 309–323.

(S2) Narváez, I.; Orío, A.; Aznar, M. P.; Corella, J. Biomass Gasification with Air in an Atmospheric Bubbling Fluidized Bed. Effect of Six Operational Variables on the Quality of the Produced Raw Gas. *Industrial & Engineering Chemistry Research* **1996**, *35*, 2110–2120.

(S3) Gil, J.; Corella, J.; Aznar, M. P.; Caballero, M. A. Biomass gasification in atmospheric and bubbling fluidized bed: Effect of the type of gasifying agent on the product distribution. *Biomass and Bioenergy* **1999**, *17*, 389–403.

(S4) Hernández, J. J.; Aranda, G.; Barba, J.; Mendoza, J. M. Effect of steam content in the air–steam flow on biomass entrained flow gasification. *Fuel Processing Technology* **2012**, *99*, 43–55.

(S5) Barisano, D.; Canneto, G.; Nanna, F.; Alvino, E.; Pinto, G.; Villone, A.; Carnevale, M.; Valerio, V.; Battafarano, A.; Braccio, G. Steam/oxygen biomass gasification at pilot scale in an internally circulating bubbling fluidized bed reactor. *Fuel Processing Technology* **2016**, *141*, 74–81.

(S6) AENOR, UNE-EN ISO 6976: Gas natural. Cálculo del poder calorífico, densidad, densidad relativa e índice de Wobbe a partir de la composición. 2005.

(S7) GPy, GPy: A Gaussian process framework in python. `http://github.com/SheffieldML/GPy`, since 2012.

(S8) Álvarez, M. A.; Rosasco, L.; Lawrence, N. D. Kernels for Vector-Valued Functions: A Review. 2011; `https://arxiv.org/abs/1106.6251`.

(S9) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(S10) Raschka, S. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. 2018; `http://arxiv.org/abs/1811.12808`.

(S11) Wainer, J.; Cawley, G. Nested cross-validation when selecting classifiers is overzealous for most practical applications. *Expert Systems with Applications* **2021**, *182*, 1–9.

(S12) Ghosh, S.; Liao, Q. V.; Ramamurthy, K. N.; Navratil, J.; Sattigeri, P.; Varshney, K. R.; Zhang, Y. Uncertainty Quantification 360: A Holistic Toolkit for Quantifying and Communicating the Uncertainty of AI. **2021**, 1–7.

(S13) Rücker, C.; Rücker, G.; Meringer, M. Y-randomization and its variants in QSPR/QSAR. *Journal of Chemical Information and Modeling* **2007**, *47*, 2345–2357.

(S14) Good, P. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*; Springer Series in Statistics; Springer New York, NY, 2000.

(S15) Ojala, M.; Garriga, G. C. Permutation tests for studying classifier performance. *Journal of Machine Learning Research* **2010**, *11*, 1833–1863.

(S16) Lundberg, S. M.; Lee, S.-I. In *Advances in Neural Information Processing Systems 30*; Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc., 2017; pp 4765–4774.

(S17) Friedman, J. H. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* **2001**, *29*, 1189–1232.

(S18) Ku, X.; Jin, H.; Lin, J. Comparison of gasification performances between raw and torrefied biomasses in an air-blown fluidized-bed gasifier. *Chemical Engineering Science* **2017**, *168*, 235–249.

(S19) Campoy, M.; Gómez-Barea, A.; Villanueva, A. L.; Ollero, P. Air-Steam Gasification of Biomass in a Fluidized Bed under Simulated Autothermal and Adiabatic Conditions. *Industrial & Engineering Chemistry Research* **2008**, *47*, 5957–5965.

(S20) Gil, J.; Aznar, M. P.; Caballero, M. A.; Francés, E.; Corella, J. Biomass Gasification in Fluidized Bed at Pilot Scale with Steam-Oxygen Mixtures. Product Distribution for Very Different Operating Conditions. *Energy & Fuels* **1997**, *11*, 1109–1118.

(S21) Fremaux, S.; Beheshti, S.-M.; Ghassemi, H.; Shahsavan-Markadeh, R. An experimental study on hydrogen-rich gas production via steam gasification of biomass in a research-scale fluidized bed. *Energy Conversion and Management* **2015**, *91*, 427–432.

(S22) Dellavedova, M.; Derudi, M.; Biesuz, R.; Lunghi, A.; Rota, R. On the gasification of biomass: Data analysis and regressions. *Process Safety and Environment Protection* **2012**, *90*, 246–254.

(S23) Mirmoshtaghi, G.; Skvaril, J.; Campana, P. E.; Li, H.; Thorin, E.; Dahlquist, E. The influence of different parameters on biomass gasification in circulating fluidized bed gasifiers. *Energy Conversion and Management* **2016**, *126*, 110–123.

(S24) Motta, I. L.; Marchesan, A. N.; Maciel Filho, R.; Wolf Maciel, M. R. Correlating biomass properties, gasification performance, and syngas applications of Brazilian feedstocks via simulation and multivariate analysis. *Industrial Crops and Products* **2022**, *181*, 114808.

(S25) Lloyd, S. P. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory* **1982**, *28*, 129–137.