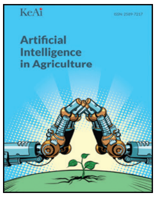
Contents lists available at ScienceDirect

Artificial Intelligence in Agriculture

journal homepage: http://www.keaipublishing.com/en/journals/artificial-intelligence-in-agriculture/



A novel elemental composition based prediction model for biochar aromaticity derived from machine learning

Hongliang Cao [a,b,*], Yaime Jefferson Milan [b], Sohrab Haghighi Mood [b], Michael Ayiania [b], Shu Zhang [c], Xuzhong Gong [d], Electo Eduardo Silva Lora [e], Qiaoxia Yuan [a], Manuel Garcia-Perez [b]

[a] Key Laboratory of Agricultural Equipment in Mid-lower Yangtze River, College of Engineering, Huazhong Agricultural University, No. 1, Shizishan Street, Hongshan District, Wuhan 430070, PR China
[b] Department of Biological Systems Engineering, Washington State University, Pullman, WA 99164, USA
[c] College of Materials Science and Engineering, Nanjing Forestry University, Nanjing 210037, PR China
[d] Key Laboratory of Green Process and Engineering, National Engineering Laboratory for Hydrometallurgical Cleaner Production Technology, Institute of Process Engineering, Chinese Academy of Sciences, Beijing 100190, PR China
[e] Excellence Group in Thermal Power and Distributed Generation, Federal University of Itajubá, Brazil

ARTICLE INFO

Article history:
Received 14 April 2021
Received in revised form 26 June 2021
Accepted 26 June 2021
Available online 7 July 2021

Keywords:
Biochar
C aromaticity
Prediction model
Machine learning
Genetic programming

ABSTRACT

The measurement of aromaticity in biochars is generally conducted using solid state ^{13}C nuclear magnetic resonance spectroscopy, which is expensive, time-consuming, and only accessible in a small number of research-intensive universities. Mathematical modelling could be a viable alternative to predict biochar aromaticity from other much easier accessible parameters (e.g. elemental composition). In this research, Genetic Programming (GP), an advanced machine learning method, is used to develop new prediction models. In order to identify and evaluate the performance of prediction models, an experimental data set with 98 biochar samples collected from the literature was utilized. Due to the benefits of the intelligence iteration and learning of GP algorithm, a kind of underlying exponential relationship between the elemental compositions and the aromaticity of biochars is disclosed clearly. The exponential relationship is clearer and simpler than the polynomial mapping relationships implicated by Maroto-Valer, Mazumdar, and Mazumdar-Wang models. In this case, a novel exponential model is proposed for the prediction of biochar aromaticity. The proposed exponential model appears better prediction accuracy and generalization ability than existing polynomial models during the statistical parameter evaluation.

1. Introduction

Biochar is the product of biomass thermochemical conversion in an oxygen limited or depleted environment (Woolf et al., 2010; Yuan et al., 2017). Biochar has received attention recently as a carbon-negative product and as an effective means to improve soil fertility, as well as other ecosystem gains such as carbon sequestration to mitigate climate change (Cao et al., 2014; Cao et al., 2020; Chen et al., 2021; Czech et al., 2021). The aromaticity is one of the most important property of biochars as it can improve the stability of biochars and affect the soil environment. The biochars, which possess greater proportion of aromatic C, offer greater chemical recalcitrance and resistance to biological degradation (Smith et al., 2017; Wiedemeier et al., 2015a). The fraction of carbons present in aromatic rings is called aromaticity (Wiedemeier et al., 2015a). Moreover, Han et al. (2014) explored a positive relationship between the aromaticity of biochar C and their capacity for organic pollutants adsorption. Singh et al. (2012) discovered that there is a definite link between the amount of CO_2 generated during a long-term incubation of biochars and their initial proportion fraction of aromatic C. Obviously, quantification of the aromaticity of biochar C is highly significant for their further utilization and could be the basis for business models.

To quantitatively characterize aromatic C of biochars, a wide variety of chemical and physical methods have been used, such as Solid state ^{13}C nuclear magnetic resonance (^{13}C NMR) spectroscopy (McBeath et al., 2011; McBeath et al., 2014), Near-edge X-ray absorption fine structure spectroscopy (NEXAFS) (Keiluweit et al., 2010), Benzene polycarboxylic acid (BPCA) analysis (Wiedemeier et al., 2013), lipid analysis (Wiedemeier et al., 2015b). However, these techniques are expensive and time-consuming, and only available in research-intensive universities (Baccile et al., 2014; Wiedemeier et al., 2015a). Consequently, simple and robust alternatives have been sought to evaluate the aromaticity of biochars.

[*] Corresponding author at: Key Laboratory of Agricultural Equipment in Mid-lower Yangtze River, College of Engineering, Huazhong Agricultural University, No. 1, Shizishan Street, Hongshan District, Wuhan 430070, PR China.
E-mail address: hongliangcao@mail.hzau.edu.cn (H. Cao).

https://doi.org/10.1016/j.aiia.2021.06.002

Mathematical modelling can be an economically-viable and efficient alternative for evaluating the degree of aromaticity, since it is possible to accurately predict the aromaticity of biochar C by using some basic feature parameters of biochars. Elemental composition data for any biochars are easily obtained and assessed. Due to this reason, Maroto-Valer et al. (1998a) developed a linear mathematical model for the prediction of the C aromaticity of bituminous coal by using atomic H/C ratio. However, the linear Maroto-Valer model has an application limit that H/C only ranges from 0.5 to 0.8. Therefore, Mazumdar (1999) developed a more accurate prediction model, derived from polyaromatic hydrocarbons (PAH, only including C and H) with a revised densimetric method, for the C aromaticity by elemental compositions. The model has a better prediction capacity than the Maroto-Valer model, mainly because former one considered the structure information of C-atom and H-atom (Mazumdar, 1999). However, when the Mazumdar model is used for the prediction of the C aromaticity of biochars, its prediction capacity is dramatically reduced. Wang et al. (2013) discovered that because Mazumdar's model only considered C- and H-atoms, the model failed when applied to biochar. Biochar is a heterogenous material and does not only involve C-atom and H-atom, it includes other heteroatoms, such as O, N and S. In this case, Wang et al. (2013) further modified the Mazumdar model and obtained a specialized prediction model for biochar aromaticity, which takes into account the influence of heteroatoms of H, O, and N. Unfortunately, the generalization abilities of the modified model is also limited, when the H/C ratio of biochars is more than 1.0. Therefore, developing new prediction model for biochar aromaticity with higher generalization abilities is greatly necessary.

Machine learning can effectively perform a modelling task by using artificial intelligence algorithms. The machine learning methods directly "learn" from raw experimental data and develop the functional relationships among the data, even if the physical meaning is unknown or there is no any prior knowledge about the nature of the underlying relationships (Goldberg et al., 2015; Kankar et al., 2011). Because of these advanced features, the intelligence method has been used to predict biochar yield (Cao et al., 2016), to identify inorganic phosphor host (Zhuo et al., 2018), to study nanomaterial toxicity (Winkler et al., 2014), as well as to optimize magnetoelastic Fe–Ga alloy microstructure (Liu et al., 2015). In practice, the adopted intelligence algorithms include Artificial Neural Network (ANN), Support Vector Machine (SVM), Bayesian Network (BN), and Random Forest (RF) (Cao et al., 2016; Jha et al., 2017; Liu et al., 2015; Pan and Pandey, 2016; Winkler et al., 2014; Zhuo et al., 2018), etc. Although these algorithms are strong intelligent modelling methods with a wide applicability, the developed models are "black box" models whose structures and parameters do not supply any insight into the phenomena underlying the data being modeled (Jha et al., 2017; Patil-Shinde et al., 2014). Beside "black box" intelligent algorithms, Genetic Programming (GP) as the advanced subcategory of machine learning methods has the ability to obtain models with clear mathematical expression (analytical models) (Bagheri et al., 2012; Ghugare et al., 2014; Pandey et al., 2015). More importantly, GP is capable of automatically arriving at an optimized mathematic model without making any assumptions regarding the structure and parameters of the developing model. The most attractive feature of GP algorithm is that, depending on the nature of dependencies (whether linear or nonlinear) in the developing data (experimental data), the technique by itself can choose a suitable model that optimally fits the developing data based on Darwinian theory of natural selection (Faris and Sheta, 2013; Pandey et al., 2015; Sharma and Tambe, 2014).

During the modification of the Mazumdar model by considering structure information of heteroatoms (e.g., H, O, and N), there were introduced several ideal assumptions. For instance, one C=O bond was replaced by two C–H bonds and one N atom was replaced by a C atom for the case with high C/N ratios (Wang et al., 2013). Considering ideal assumptions may cause some key model information to be ignored, which might be the reason why the prediction capacity of the modified Mazumdar model is not satisfactory for practical application. Just as discussed above, during the development of GP-based model, the model structure is not specified in advance. Therefore, it is possible to identify a suitable prediction model with higher generalization abilities. Accordingly, the principal objectives of this research are: (1) to develop elemental composition based prediction models for biochar aromaticity with higher generalization abilities, and (2) to discover new mapping relationships and rules between the elemental compositions and C aromaticity of biochars.

2. Methods

2.1. Data collection

To develop robust prediction models for the C aromaticity of biochars, the experimental data including 98 biochar samples derived from ^{13}C NMR spectroscopy were collected from previous literatures (Baldock and Smernik, 2002; Brewer et al., 2011; Cao et al., 2012; Enders et al., 2012; Kaal et al., 2012; Keiluweit et al., 2010; Manna et al., 2020; McBeath et al., 2014; Singh et al., 2012; Wang et al., 2013; Wiedemeier et al., 2015a; Yue et al., 2017), and the detailed data are provided in the supplementary material file (Table S1). The biochars characteristics are diverse because of different feedstock (e.g. 24 kinds of biomass) and different pyrolysis conditions (charring temperature, heating rate, and holding time). In this case, the database covers a wide distribution of C aromaticity and elemental composition. For example, C aromaticity ranges from 0.1000–1.0000, and the ranges of H/C, O/C, N/C atomic ratios are 0.0249–1.8076, 0.0171–0.7166, 0.0006–0.0415, respectively.

In order to develop C aromaticity prediction models, the above collected experimental data were divided into two data sets, e.g. training set and testing set. The training set was employed to identify the model parameters, and the testing set was utilized for the evaluation of the model performance (Cao et al., 2016). To ensure the effectiveness of the developed model, the training and testing sets were chosen randomly. In this research, 60 samples (about 60% of the prepared data) were randomly selected as the training set, and the rest 38 samples were taken as the testing set.

2.2. Modelling methods

2.2.1. Mazumdar-Wang model

Mazumdar (1999) employed a revised densimetric method to accurately predict the C aromaticity of polyaromatic hydrocarbons (PAH, including only C and H):

$$f_a = 1 - \frac{H'}{C'} + \alpha\left(\frac{Mc}{\mathrm{d}} - 5.34\right) \tag{1}$$

where f_a denotes the aromaticity of carbon materials; $\frac{H'}{C'}$ denotes the atomic ratio of H and C; $\frac{Mc}{d}$ denotes the average molar volume of C-atom, decreasing with the increase of the condensation degree of PAH; α denotes a modification coefficient, ranging from 0.115–0.125 for coal samples, increasing with the increase of the C aromaticity. Notice that the constant, 5.34, is the average molar volume of graphite C-atom and is taken as the lower $\frac{Mc}{d}$ limit of C. Moreover, the average molar volume of C-atom, $\frac{Mc}{d}$, can be estimated by a second-order polynomial model of $\frac{H'}{C'}$ (Mazumdar, 1999):

$$\frac{Mc}{d} = 5.34 + 9.15\frac{H'}{C'} - 2.9\left(\frac{H'}{C'}\right)^2 \tag{2}$$

Mazumdar model is proposed considering PAH, which includes only C and H. However, biochar does not only involve C and H, but also O, N and S. These elements can also affect $\frac{Mc}{d}$. Considering this factor, Wang et al. (2013) modified the atomic ratio of $\frac{H'}{C'}$:

$$\frac{H'}{C'} = \frac{\text{H\%}/1 + 2\theta \times \text{O\%}/16}{\text{C\%}/12 + \text{N\%}/14} \tag{3}$$

Where H % /1, O % /16, C % /12, and N % /14 are the molar fractions of corresponding elements divided by their molecular weight; θ is the molar ratio of C=O bond to CO bonds (consisting of aliphatic and O-aryl C, C–O single bond, and C=O double bond). The modified model taken into account the presence of heteratom groups of O and N; One C=O bond is replaced by 2C–H bonds and one N atom is replaced by a C atom. In this research, the modified model is referred as Mazumdar-Wang model.

2.2.2. Genetic programming modelling method

GP simulates Darwin's theory of biological evolution: "survival of the fittest" and "genetic propagation of characteristics"; it is an intelligent evolution strategy which automatically evolves computer programs to solve the task without specifying the structure of the solution in advance (Baumes et al., 2009; Gandomi et al., 2015). Hence, the advantage of GP modelling method is that not only the parameters of models are identified automatically but also it can identify the structure of models.

Herein, we consider a multiple input-single output (MISO) predicting task. Modelling data set is,

$$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_\text{N}, y_\text{N})\}, \tag{4}$$

including N data patterns, where $\boldsymbol{x}_n (n = 1,2,\cdots,N)$ is an M-dimensional input vector ($\boldsymbol{x}_n = [x_{n1}, x_{n2}, \cdots, x_{nM}]^T$), and y_n is the corresponding model output. Using the data set D, the GP can find the exact form and the associated parameters of the unknown MISO model (f). The general form of the model to be solved by GP is given as,

$$y = f(\boldsymbol{x}, \boldsymbol{\omega}) \tag{5}$$

where $\omega = [\omega_1, \omega_2, \cdots, \omega_Q]$ denotes a Q-dimensional parameter vector.

At the beginning of the GP, a population of candidate solutions is randomly generated to the model identification problem. As seen in Fig. 1(a), a candidate solution is coded with a form of trees (gene expression trees), which forms a candidate model for the model output of y_n when decoded. The tree structure derives from a root node and involves operator nodes and operand nodes. The operator nodes represent mathematical operators, e.g. addition, subtraction, multiplication, division, sine, cosine, exponentiation, etc.; while operand nodes define model inputs ($\boldsymbol{x}$) and parameters (ω). The fitness of each candidate solution for the predicting task is evaluated by its fitness score, which is calculated from fitness functions, such as root mean square error (RMSE). The candidate solutions with high fitness scores are more probably chosen to create a mating pool termed "parents" for the reproduction of new candidate solutions. The new candidate solutions are produced from the parents by evolution operation of crossover and mutation (Fig. 1(b) and (c)). During crossover operation, a pair of parents from the mating pool is selected randomly; and then two new offspring are obtained by tailoring each parent tree at a random point and mutually exchanging the tailored branches between the parents. The crossover operation is addressed with a pre-specified probability value termed "crossover probability". For mutation operation, the changes are applied to the operator and operand nodes of the parent selected randomly to generate a new offspring. The mutation operation is also conducted with a pre-specified probability termed "mutation probability", which is usually of a small probability. In this reproduction and iteration way, the satisfied predicting model is developed automatically. The detailed implementation processes of the algorithm are available in literatures (Faris and Sheta, 2013; Pandey et al., 2015; Patil-Shinde et al., 2014).

2.3. Model performance evaluations

To evaluate the capabilities of the developed models, five statistical parameters (e.g. AAE, MAE, RMSE, R^2, and ρ) were employed. AAE, MAE, RMSE, R^2, and ρ are widely adopted statistical parameters and respectively represent the Average Absolute Error, Maximum Absolute Error, Root Mean Squared Error, coefficient of determination, and correlation coefficient. These coefficients are defined as follows (Cao et al., 2016; Gandomi et al., 2013a):

$$\text{AAE} = \frac{\sum_{i=1}^{N\sum} \left| Y_i^{\text{pred}} - Y_i^{exp} \| \right|}{N}, \tag{6}$$

$$\text{MAE} = max \left| Y_i^{\text{pred}} - Y_i^{exp} \| \right|, \tag{7}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N\sum} (Y_i^{\text{pred}} - Y_i^{exp^2}}{N}}, \tag{8}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N\sum} (Y_i^{exp_i^{\text{pred}\,2}}}{\sum_i^{exp_{ave}^{exp^2}\;{}_{i=1}^{N}} (Y}, \tag{9}$$

$$\rho = \frac{\sum_{i=1}^{N} \left(Y_i^{\text{exp}} - \overline{Y}_{ave}^{\text{exp}}\right)\left(Y_i^{\text{pred}} - \overline{Y}_{ave}^{\text{pred}}\right)}{\sqrt{\sum_{i=1}^{N} \left(Y_i^{\text{exp}} - \overline{Y}_{ave}^{\text{exp}}\right)^2 \sum_{i=1}^{N} \left(Y_i^{\text{pred}} - \overline{Y}_{ave}^{\text{pred}}\right)^2}} \tag{10}$$

Where N is the number of samples, Y_i^{pred} and $Y_i^{exp}$are respectively the predicted and experimental values for the i^{th} output, and $\overline{Y}_{ave}^{\text{pred}}$ and $\overline{Y}_{ave}^{\text{exp}}$ are the average of predicted and experimental values, respectively. Moreover, because correlation coefficient, ρ, will not change significantly via shifting the predicted output of a model equally, and error functions, such as AAE and RMSE, only indicate the error not the correlation. Therefore, a comprehensive performance index (β) is usually used for model performance evaluation (Gandomi et al., 2016):

$$\beta = \frac{\text{RMSE}}{1 + \rho} \tag{11}$$

3. Results and discussion

3.1. Mazumdar-Wang model

Least square method is utilized for the identification of the model parameters (α in Eq. (1) and θ in Eq. (3)), which is implemented handily by using the *lsqcurvefit* function in MATLAB software; the identified parameters of α and θ are equal to 0.0960 and 0, respectively. Fig. 2 shows the performance scattered plots of the model on training and testing data sets, and the corresponding statistical parameters are listed in Table 2.

The tight cloud of points about 45° line demonstrate that the Mazumdar-Wang model has a greatly acceptable predicting capacity for the C aromaticity of biochars, especially due to a high coefficient of determination in testing set ($\text{R}^2 = 0.9130$). However, there is obvious difference between the model parameters obtained in this research ($\alpha = 0.0960$, $\theta = 0$) and those developed by Wang et al. (2013) ($\alpha = 0.110$, $\theta = 0.290$). The primary difference is about the model parameter of molar ratio (θ). Wang et al. (2013) has revealed that molar ratio, θ, is actually not constant, and changes with biochar structure: biochars obtained at low temperature have a

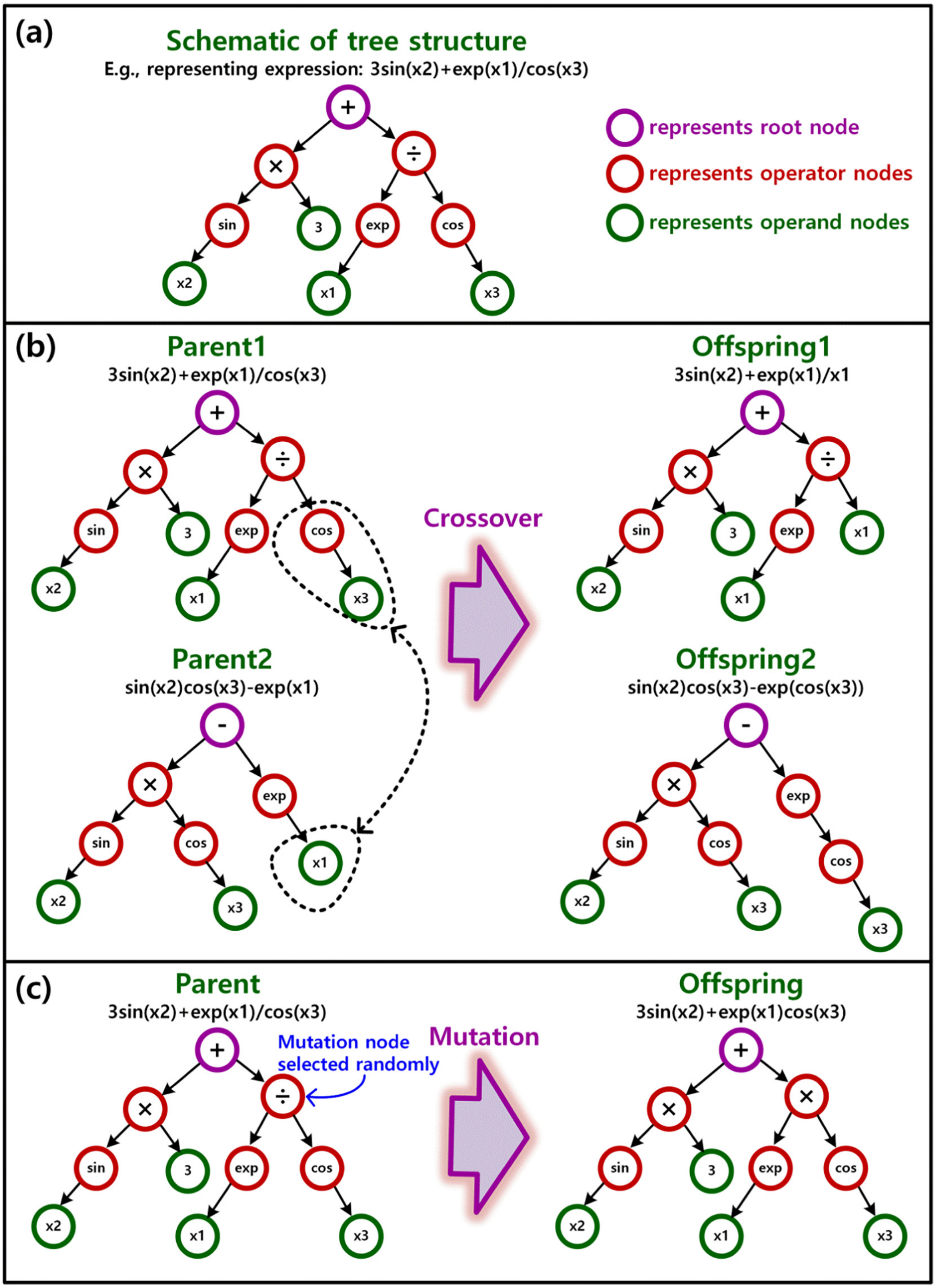


Fig. 1. Schematic of genetic programming: (a) basic tree structure, (b) two offspring trees created by crossover operation, and (c) one offspring tree created by mutation operation.

larger fraction of organic O existing as C–O (accounting for almost 100% of total CO bonds when $\theta = 0$), while those obtained at high temperature have a higher fraction of organic O as C=O (up to $\theta = 1$). In this research, there is a quite number of low temperature biochar samples and the atomic ratio of H/C reaches a maximal value of 1.8076, but Wang et al. (2013) just considered biochars with H/C less than 1. Therefore, we have utilized a wider sample range which could be the reason for the difference of model parameters.

3.2. GP-based models

For the identification of GP-based C aromaticity models, H/C, O/C, N/C are considered as the input parameters. During the training of the model, the population size and the maximum number of the generation were set as 300 and 1000, respectively. To select the parent genes from the pool of available solutions, the tournament selection strategy was adopted (Pandey et al., 2015). The tournament size is set to 4. The detailed parameters of the GP algorithm are given in Table 1, which have

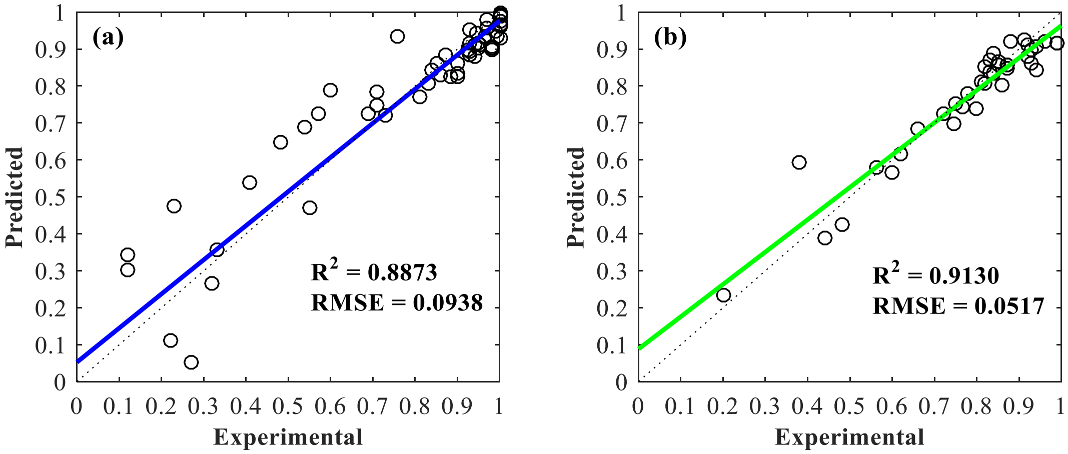


Fig. 2. Performance scattered plots Mazumdar-Wang model with α =0.0960 and θ =0: (a) on training data set and (b) on testing data set.

Table 1
Parameters used for the GP algorithm.

Parameters of the GP algorithm	Parameter settings
Population size	300
Number of generation	1000
Tournament size	4
Maximum depth of tree	3
Crossover probability	0.85
Mutation probability	0.1
Reproduction probability	0.05
Selection method	Plain lexicographic tournament selection
Termination criteria	1000 generation or fitness value less than 1.0×10^{-4}, whichever is earlier
Mathematical operations	{+, −, ×, ÷, cos, sin, tan, exp., log, ()^}

been examined until there was no longer significant improvement in the performance of the GP-based models. Three alternative models (expression trees) were determined by considering their corresponding fitness scores and model complexities (see Fig. 3).

As shown in the Fig. 3, the proposed GP-based C aromaticity prediction models are created by the arrangement of operators, variables, and constants. we can find that those models have same root node, subtraction operation, and then the root node connects different subprograms (genes) to form the final prediction models. Any individual aspect of the problem to be modeled is accentuated by each of the subprograms so that a meaningful overall solution is identified (Faris and Sheta, 2013; Sharma and Tambe, 2014). In this way, the important information underlying in the problem could be achieved via each of the evolved subprograms.

Fig. 4 shows the comparison of the experimental against predicted C aromaticity values from the three prediction models, and the corresponding performance evaluation parameters are given in Table 2. It is interesting to notice that all three GP-based models have better prediction performances than the Mazumdar-Wang model, due to lower AAE, MAE, RMSE, and comprehensive performance index (β), as well as higher R^2 and correlation coefficient (ρ). GP-I model predicts the C aromaticity using a polynomial combination of H/C and O/C; GP-II and GP-III models indicate that there are exponential relationships between the C aromaticity and the elemental atom ratios. Moreover, GP-III model believes N/C has influence for the aromaticity prediction, and considers its effect in the model.

For further verification of the GP-based models, box-and-whisker diagrams of the prediction errors in testing set were conducted, as shown in Fig. 5. Box-and-whisker diagram is a standardized way of displaying the distribution of data based on a five number summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”), which

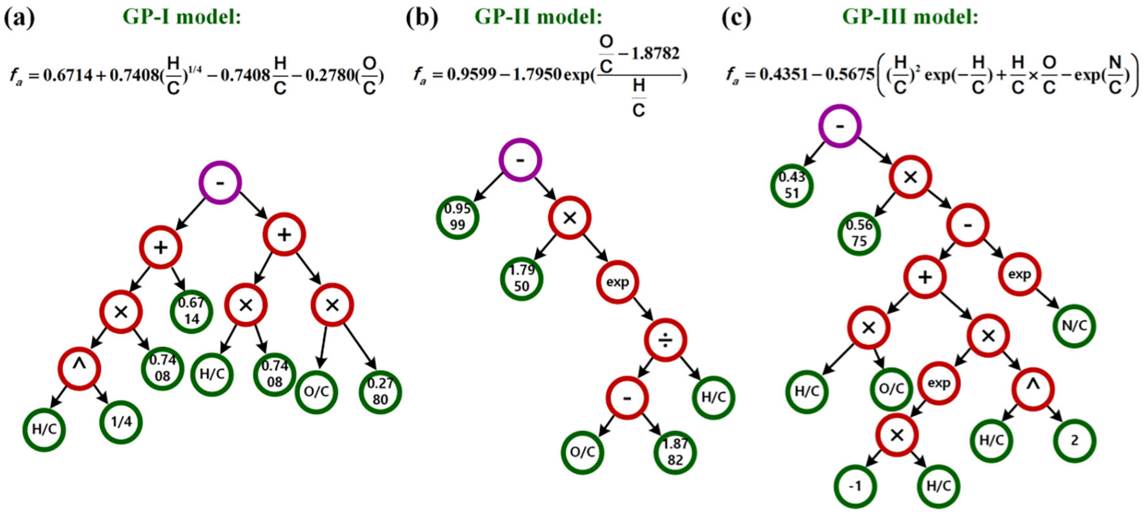

(a) GP-I model:
$f_a = 0.6714 + 0.7408(\frac{H}{C})^{1/4} - 0.7408\frac{H}{C} - 0.2780(\frac{O}{C})$
-
+
+
×
0.6714
×
×
^
0.7408
H/C
0.7408
O/C
0.2780
H/C
1/4
(b) GP-II model:
$f_a = 0.9599 - 1.7950\exp(\frac{\frac{O}{C} - 1.8782}{\frac{H}{C}})$
-
0.9599
×
1.7950
exp
÷
-
H/C
O/C
1.8782
(c) GP-III model:
$f_a = 0.4351 - 0.5675\left((\frac{H}{C})^2\exp(-\frac{H}{C}) + \frac{H}{C}\times\frac{O}{C} - \exp(\frac{N}{C})\right)$
-
0.4351
×
0.5675
-
+
exp
N/C
×
×
H/C
O/C
exp
^
×
H/C
2
-1
H/C

Fig. 3. Three alternative biochar aromaticity prediction models derived from GP algorithms. They are named as GP-I, GP-II, and GP-III, respectively.

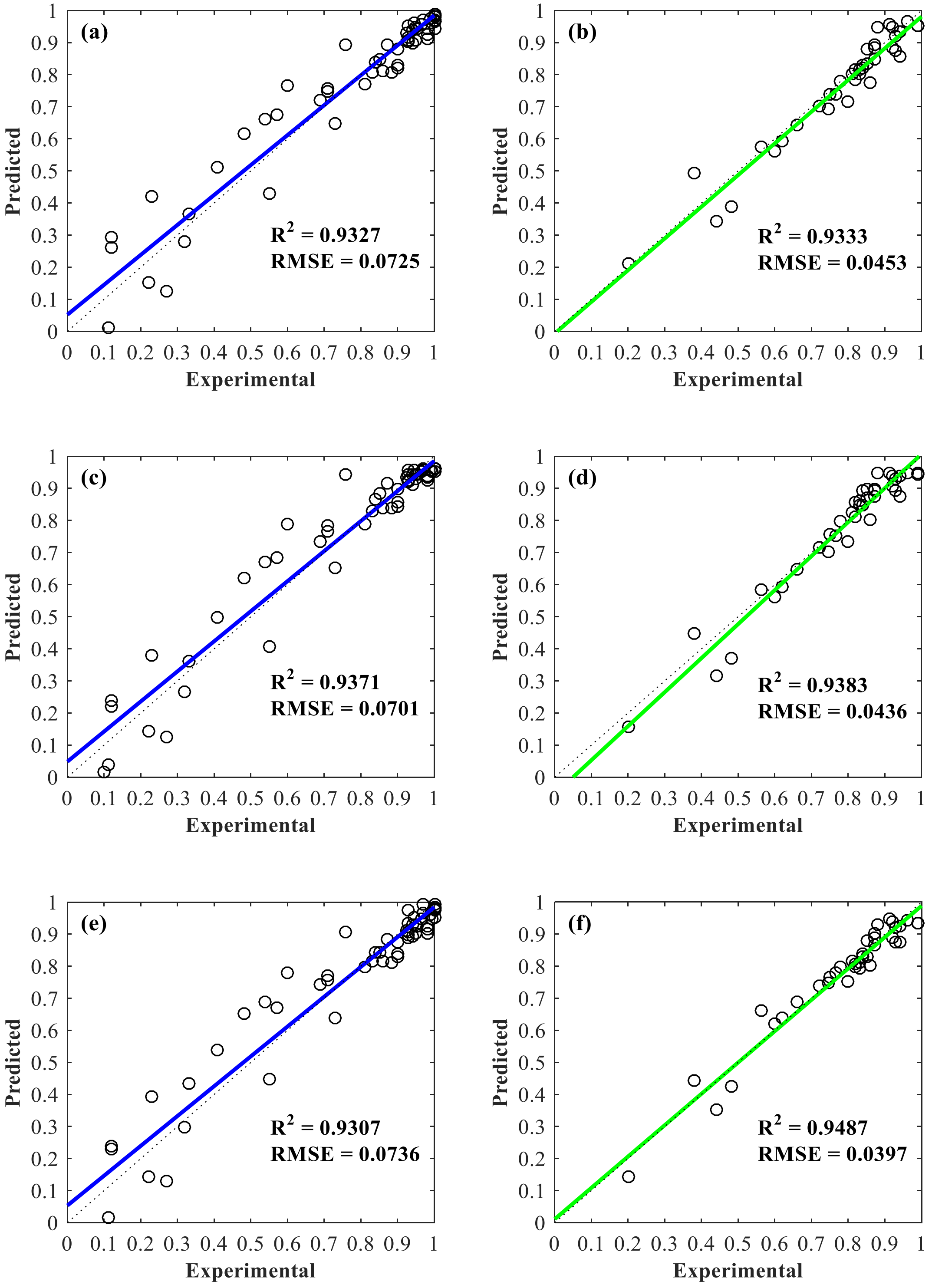


Fig. 4. Experimental versus predicted C aromaticity values using the GP-based prediction models: (a) GP-I model on training data set, (b) GP-I model on testing data set, (c) GP-II model on training data set, (d) GP-II model on testing data set, (e) GP-III model on training data set, and (f) GP-III model on testing data set.

can quantitatively describe features of the prediction errors to analysis the accuracy and reliability of the models. Obviously, GP-II and GP-III models outperform the GP-I and Mazumdar-Wang models. GP-II and GP-III have smaller median values (0.0017 and 0.0075) than GP-I and Mazumdar-Wang models (0.0165 and 0.0113), which means the error centres of former two models are closer to zero. Moreover, there are no outliers for the predictions of GP-II and GP-III models when the error distributions are addressed. Regarding the structure of models, GP-II and GP-III models predict the C aromaticity by exponential functions of elemental atom ratios, but GP-I and Mazumdar-Wang models utilize polynomial combinations. It is obvious that exponential models have better prediction accuracy and generalization capabilities over polynomial models for the prediction of the C aromaticity. These results implicate that there exists an exponential relationship between the

Table 2
Performance evaluation parameters for the three GP-based C aromaticity prediction models.

Data set	Parameters	M-W model	GP-I model	GP-II model	GP-III model
Training ($N = 60$)	Average Absolute Error (AAE)	0.0677	0.0537	0.0529	0.0547
	Maximum Absolute Error (MAE)	0.2155	0.1471	0.1439	0.1412
	Root Mean Squared Error (RMSE)	0.0938	0.0725	0.0701	0.0736
	Coefficient of determination (R^2)	0.8873	0.9327	0.9371	0.9306
	Correlation coefficient (ρ)	0.9429	0.9657	0.9681	0.9647
	Comprehensive performance (β)	0.0483	0.0369	0.0356	0.0375
Testing ($N = 38$)	Average Absolute Error (AAE)	0.0357	0.0348	0.0337	0.0320
	Maximum Absolute Error (MAE)	0.1002	0.0949	0.1240	0.0857
	Root Mean Squared Error (RMSE)	0.0517	0.0453	0.0436	0.0397
	Coefficient of determination (R^2)	0.9129	0.9333	0.9383	0.9487
	Correlation coefficient (ρ)	0.9581	0.9717	0.9756	0.9752
	Comprehensive performance (β)	0.0264	0.0230	0.0220	0.0201

elemental compositions and C aromaticity, which can more accurately characterize the underlying rule and information for the prediction of biochar aromaticity.

3.3. Sensitivity analysis of GP-II and GP-III models

The above statistical results (see Table 2 and Figs. 4 and 5) clearly indicate GP-II and GP-III models have better prediction performances over GP-I and Mazumdar-Wang models. Moreover, the results also show that GP-II and GP-III models possess comparable prediction performances. For instance, during the verification in testing set data, GP-II has smaller median value, but at the same time has larger MAE and RMSE; and the two models have almost equal comprehensive performance index (β). The primary difference of the two models is GP-III model includes the contribution from the input parameter of N/C for the C aromaticity prediction, which results in a model with greater complexity. But GP-II model does not take into account the influence of N/C, and has a simple and clear model structure (see Fig. 3). To consider the influence of different input parameters, we conducted a sensitivity analysis for the two models.

The sensitivity (S_i) of each input parameter is expressed as follow (Gandomi et al., 2013b):

$$S_i = 100 \times \frac{N_i}{\sum_{j=1}^{n} N_j}, \tag{12}$$

$$N_i = fii_{minmax}, \tag{13}$$

in which $f_{\max}(x_i)$ and $f_{\min}(x_i)$ are respectively the maximum and minimum value of the predicted output (i.e. C aromaticity) over the i^{th} input domain, where other input parameters are equal to their mean values. Table 3 presents the results of above analysis for the GP-II and GP-III models. Obviously, the influence of N/C for the C aromaticity prediction is highly small, which can be ignored, because its sensitivity for the outputs of the GP-III model is only 2.23%. Therefore, GP-II model should be the best alternative among these models for the prediction of biochar aromaticity on account of excellent predicting capacity and simple model structure.

Table 3
Sensitivity analysis of different atom ratios for the GP-II and GP-III models.

Models	Sensitivity (%)		
	H/C	O/C	N/C
GP-II	74.25	25.75	–
GP-III	63.53	34.24	2.23

Furthermore, the sensitivity analysis (Table 3) also clearly demonstrates that H/C dominantly determines the prediction of biochar aromaticity because of larger sensitivity weight over O/C. That could be an important reason for the fact that H/C is widely used as an approximating aromatic index (Baldock and Smernik, 2002; Cai et al., 2019; Hammes et al., 2006; Keiluweit et al., 2010; Wiedemeier et al., 2015a). Fig. 6(a) shows 3D distribution of the biochar aromaticity under the prediction of GP-II model, and Fig. 6(b) is the corresponding growing pathway of the biochar aromaticity with the decrease of H/C. As far as we know, as pyrolysis temperature increases, raw biomass experiences charring processes, such as dehydration, depolymerization, volatilization and crystallization, eventually leading to the formation of H and O depleted aromatic C structures (Chen and Yuan, 2011; Fang et al., 2014; Wiedemeier et al., 2015a; Xiao et al., 2016). In this case, the biochar aromaticity increases with decreasing of H/C. In addition, the sensitivity of C aromaticity for O/C reduces with the decrease of H/C (the red shadow area in Fig. 6(b) becomes narrow with the decrease

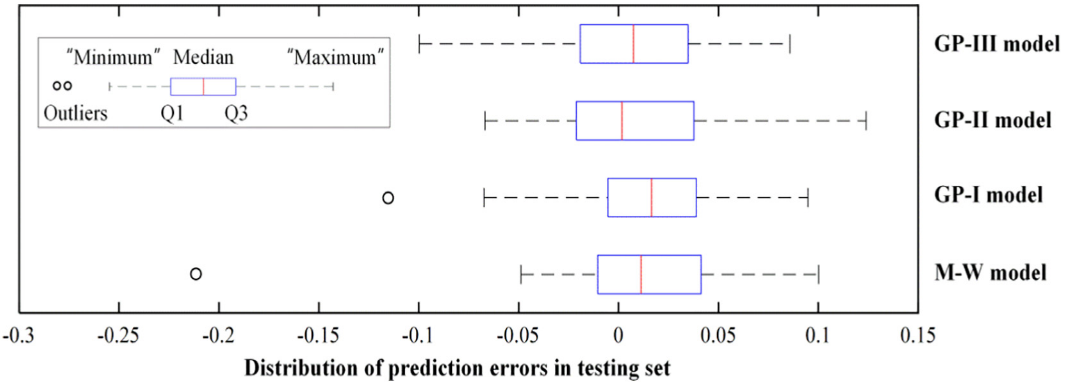


Fig. 5. Box-and-whisker diagrams of the developed predicted models in testing set.

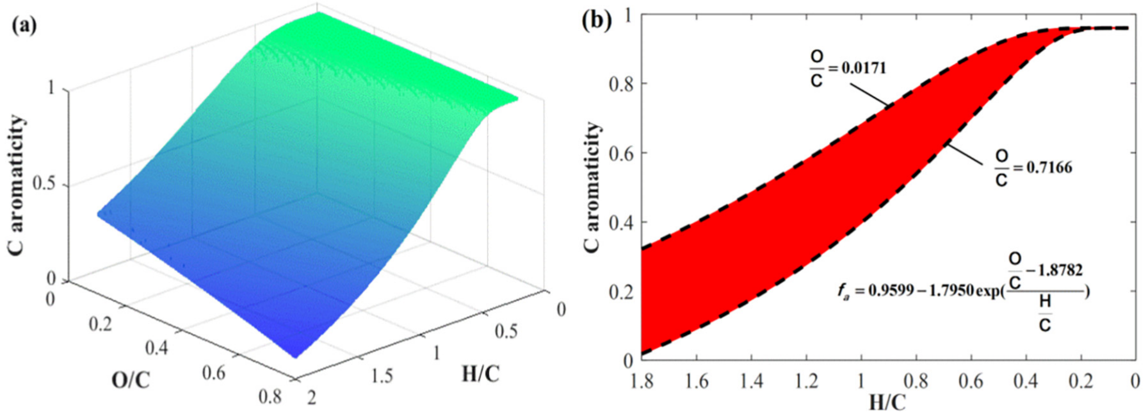


Fig. 6. Predicting distributions of the GP-II model for biochar aromaticity: (a) 3D characteristics under O/C - H/C plane, (b) 2D characteristics with H/C as independent variable, where red shadow area is formed with different O/C and the black dotted lines represent the upper and lower limits of O/C within the collected data set.

of H/C), which primarily is due to the formation of similar molecular structures (tiny aromatic cluster graphene-like structures) at relative high pyrolysis temperature (Fang et al., 2014; Maroto-Valer et al., 1998b; McBeath et al., 2011; Wang et al., 2013). Obviously, the prediction characteristics of GP-II model is greatly consistent with the reported research results. Therefore, the GP-II model is an excellent prediction models for biochar aromaticity with high predicting accuracy and generalization ability.

4. Conclusions

The presented results clearly indicate that, compared with polynomial models (e.g. Mazumdar-Wang model), the proposed exponential models (see GP-II and GP-III models) can more accurately characterize the underlying mapping relationship between the elemental compositions and the C aromaticity of biochars. Particularly, GP-II model not only has high predicting accuracy and generalization ability, but also possesses a simple model structure. Furthermore, the research demonstrates that mathematical modelling has high potential utility as a powerful tool for predicting biochar aromaticity using easily accessible feature parameters of biochars, such as elemental compositions, which can greatly save experimental time and cost. Especially, GP intelligence modelling method can accurately identify suitable alternatives with clear mathematic expressions.

Author statement

Hongliang Cao conducted most of the modelling work and writing of this manuscript. Yaime Jefferson Milan, Sohrab Haghighi Mood, Michael Ayiania, Shu Zhang, Xuzhong Gong, and Electo Eduardo Silva Lora collected experimental data, and analyzed characterization data; Qiaoxia Yuan addressed the sensitivity analysis of the models. Sohrab Haghighi Mood, Shu Zhang, and Manuel Garcia-Perez contributed writing of the manuscript. All authors have given approval to the final version of the manuscript.

Declaration of interests

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.

Acknowledgments

This research was financially supported by the National Nature Science Foundation of China (31971807, 51876093), the National Key R&D Program of China (2018YFE0183600), the Fundamental Research Funds for the Central Universities (2662019PY081), the financial support from the China Scholarship Council (201806765007), and the National Research Council of Brazil - CNPq. Moreover, the authors would like to thank Chunjiang Zhao, Editor, Artificial Intelligence in Agriculture, and anonymous reviewers for their help and suggestion.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.aiia.2021.06.002.

References

Baccile, N., Falco, C., Titirici, M., 2014. Characterization of biomass and its derived char using 13C-solid state nuclear magnetic resonance. Green Chem. 16, 4839–4869.

Bagheri, M., Bagheri, M., Gandomi, A.H., Golbraikh, A., 2012. Simple yet accurate prediction method for sublimation enthalpies of organic contaminants using their molecular structure. Thermochim. Acta 543, 96–106.

Baldock, J.A., Smernik, R.J., 2002. Chemical composition and bioavailability of thermally altered Pinus resinosa (Red pine) wood. Org. Geochem. 33, 1093–1109.

Baumes, L.A., Blansché, A., Serna, P., Tchougang, A., Lachiche, N., Collet, P., Corma, A., 2009. Using genetic programming for an advanced performance assessment of industrially relevant heterogeneous catalysts. Mater. Manuf. Process. 24, 282–292.

Brewer, C.E., Unger, R., Schmidt-Rohr, K., Brown, R.C., 2011. Criteria to select biochars for field studies based on biochar chemical properties. Bioenerg. Res. 4, 312–323.

Cai, H., Liu, J., Xie, W., Kuo, J., Buyukada, M., Evrendilek, F., 2019. Pyrolytic kinetics, reaction mechanisms and products of waste tea via TG-FTIR and Py-GC/MS. Energ. Convers. Manag. 184, 436–447.

Cao, X., Pignatello, J.J., Li, Y., Lattao, C., Chappell, M.A., Chen, N., Miller, L.F., Mao, J., 2012. Characterization of wood chars produced at different temperatures using advanced solid-state 13C NMR spectroscopic techniques. Energy Fuel 26, 5983–5991.

Cao, H., Xin, Y., Wang, D., Yuan, Q., 2014. Pyrolysis characteristics of cattle manures using a discrete distributed activation energy model. Bioresour. Technol. 172, 219–225.

Cao, H., Xin, Y., Yuan, Q., 2016. Prediction of biochar yield from cattle manure pyrolysis via least squares support vector machine intelligent approach. Bioresour. Technol. 202, 158–164.

Cao, H., Wu, X., Syed-Hassan, S.S.A., Zhang, S., Mood, S.H., Milan, Y.J., Garcia-Perez, M., 2020. Characteristics and mechanisms of phosphorous adsorption by rape straw-derived biochar functionalized with calcium from eggshell. Bioresour. Technol. 318, 124063.

Chen, B., Yuan, M., 2011. Enhanced sorption of polycyclic aromatic hydrocarbons by soil amended with biochar. J. Soils Sediments 11, 62–71.

Chen, X., Yang, S., Jiang, Z., Ding, J., Sun, X., 2021. Biochar as a tool to reduce environmental impacts of nitrogen loss in water-saving irrigation paddy field. J. Clean. Prod. 290, 125811.

Czech, B., Kończak, M., Rakowska, M., Oleszczuk, P., 2021. Engineered biochars from organic wastes for the adsorption of diclofenac, naproxen and triclosan from water systems. J. Clean. Prod. 288, 125686.

Enders, A., Hanley, K., Whitman, T., Joseph, S., Lehmann, J., 2012. Characterization of biochars to evaluate recalcitrance and agronomic performance. Bioresour. Technol. 114, 644–653.

Fang, Q., Chen, B., Lin, Y., Guan, Y., 2014. Aromatic and hydrophobic surfaces of wood-derived biochar enhance perchlorate adsorption via hydrogen bonding to oxygen-containing organic groups. Environ. Sci. Technol. 48, 279–288.

Faris, H., Sheta, A.F., 2013. Identification of the Tennessee Eastman chemical process reactor using genetic programming. Int. J. Adv. Sci. Technol. 50, 121–140.

Gandomi, A.H., Roke, D.A., Sett, K., 2013a. Genetic programming for moment capacity modeling of ferrocement members. Eng. Struct. 57, 169–176.

Gandomi, A.H., Yun, G.J., Alavi, A.H., 2013b. An evolutionary approach for modeling of shear strength of RC deep beams. Mater. Struct. 46, 2109–2119.

Gandomi, A.H., Alavi, A.H., Ryan, C., 2015. Handbook of Genetic Programming Applications. Springer, New York.

Gandomi, M., Soltanpour, M., Zolfaghari, M.R., Gandomi, A.H., 2016. Prediction of peak ground acceleration of Iran's tectonic regions using a hybrid soft computing technique. Geosci. Front. 7, 75–82.

Ghugare, S.B., Tiwary, S., Tambe, S.S., 2014. Computational intelligence based models for prediction of elemental composition of solid biomass fuels from proximate analysis. Int. J. Syst. Assur. Eng. Manag. 1–14.

Goldberg, E., Scheringer, M., Bucheli, T.D., Hungerbühler, K., 2015. Prediction of nanoparticle transport behavior from physicochemical properties: machine learning provides insights to guide the next generation of transport models. Environ. Sci. Nano 2, 352–360.

Hammes, K., Smernik, R.J., Skjemstad, J.O., Herzog, A., Vogt, U.F., Schmidt, M.W.I., 2006. Synthesis and characterisation of laboratory-charred grass straw (Oryza sativa) and chestnut wood (Castanea sativa) as reference materials for black carbon quantification. Org. Geochem. 37, 1629–1633.

Han, L., Sun, K., Jin, J., Wei, X., Xia, X., Wu, F., Gao, B., Xing, B., 2014. Role of structure and microporosity in phenanthrene sorption by natural and engineered organic matter. Environ. Sci. Technol. 48, 11227–11234.

Jha, S.K., Bilalovic, J., Jha, A., Patel, N., Zhang, H., 2017. Renewable energy: present research and future scope of artificial intelligence. Renew. Sust. Energ. Rev. 77, 297–317.

Kaal, J., Schneider, M.P.W., Schmidt, M.W.I., 2012. Rapid molecular screening of black carbon (biochar) thermosequences obtained from chestnut wood and rice straw: a pyrolysis-GC/MS study. Biomass Bioenergy 45, 115–129.

Kankar, P.K., Sharma, S.C., Harsha, S.P., 2011. Fault diagnosis of ball bearings using machine learning methods. Expert Syst. Appl. 38, 1876–1886.

Keiluweit, M., Nico, P.S., Johnson, M.G., Kleber, M., 2010. Dynamic molecular structure of plant biomass-derived black carbon (biochar). Environ. Sci. Technol. 44, 1247–1253.

Liu, R., Kumar, A., Chen, Z., Agrawal, A., Sundararaghavan, V., Choudhary, A., 2015. A predictive machine learning approach for microstructure optimization and materials design. Sci. Rep-Uk 5, 11551.

Manna, S., Singh, N., Purakayastha, T.J., Berns, A.E., 2020. Effect of deashing on physicochemical properties of wheat and rice straw biochars and potential sorption of pyrazosulfuron-ethyl. Arab. J. Chem. 13, 1247–1258.

Maroto-Valer, M.M., Andrésen, J.M., Snape, C.E., 1998a. Verification of the linear relationship between carbon aromaticities and HC ratios for bituminous coals. Fuel 77, 783–785.

Maroto-Valer, M.M., Andrésen, J.M., Snape, C.E., 1998b. Verification of the linear relationship between carbon aromaticities and HC ratios for bituminous coals. Fuel 77, 783–785.

Mazumdar, B.K., 1999. Molecular structure and molar volume of organic compounds and complexes with special reference to coal. Fuel 78, 1097–1107.

McBeath, A.V., Smernik, R.J., Schneider, M.P.W., Schmidt, M.W.I., Plant, E.L., 2011. Determination of the aromaticity and the degree of aromatic condensation of a thermosequence of wood charcoal using NMR. Org. Geochem. 42, 1194–1202.

McBeath, A.V., Smernik, R.J., Krull, E.S., Lehmann, J., 2014. The influence of feedstock and production temperature on biochar carbon chemistry: a solid-state 13C NMR study. Biomass Bioenergy 60, 121–129.

Pan, I., Pandey, D.S., 2016. Incorporating uncertainty in data driven regression models of fluidized bed gasification: a Bayesian approach. Fuel Process. Technol. 142, 305–314.

Pandey, D.S., Pan, I., Das, S., Leahy, J.J., Kwapinski, W., 2015. Multi-gene genetic programming based predictive models for municipal solid waste gasification in a fluidized bed gasifier. Bioresour. Technol. 179, 524–533.

Patil-Shinde, V., Kulkarni, T., Kulkarni, R., Chavan, P.D., Sharma, T., Sharma, B.K., Tambe, S.S., Kulkarni, B.D., 2014. Artificial intelligence-based modeling of high ash coal gasification in a pilot plant scale fluidized bed gasifier. Ind. Eng. Chem. Res. 53, 18678–18689.

Sharma, S., Tambe, S.S., 2014. Soft-sensor development for biochemical systems using genetic programming. Biochem. Eng. J. 85, 89–100.

Singh, B.P., Cowie, A.L., Smernik, R.J., 2012. Biochar carbon stability in a clayey soil as a function of feedstock and pyrolysis temperature. Environ. Sci. Technol. 46, 11770–11778.

Smith, M.W., Helms, G., McEwen, J., Garcia-Perez, M., 2017. Effect of pyrolysis temperature on aromatic cluster size of cellulose char by quantitative multi cross-polarization 13C NMR with long range dipolar dephasing. Carbon 116, 210–222.

Wang, T., Camps-Arbestain, M., Hedley, M., 2013. Predicting C aromaticity of biochars based on their elemental composition. Org. Geochem. 62, 1–6.

Wiedemeier, D.B., Hilf, M.D., Smittenberg, R.H., Haberle, S.G., Schmidt, M.W.I., 2013. Improved assessment of pyrogenic carbon quantity and quality in environmental samples by high-performance liquid chromatography. J. Chromatogr. A 1304, 246–250.

Wiedemeier, D.B., Abiven, S., Hockaday, W.C., Keiluweit, M., Kleber, M., Masiello, C.A., McBeath, A.V., Nico, P.S., Pyle, L.A., Schneider, M.P.W., Smernik, R.J., Wiesenberg, G.L.B., Schmidt, M.W.I., 2015a. Aromaticity and degree of aromatic condensation of char. Org. Geochem. 78, 135–143.

Wiedemeier, D.B., Brodowski, S., Wiesenberg, G.L.B., 2015b. Pyrogenic molecular markers: linking PAH with BPCA analysis. Chemosphere 119, 432–437.

Winkler, D.A., Burden, F.R., Yan, B., Weissleder, R., Tassa, C., Shaw, S., Epa, V.C., 2014. Modelling and predicting the biological effects of nanomaterials. SAR QSAR Environ. Res. 25, 161–172.

Woolf, D., Amonette, J.E., Street-Perrott, F.A., Lehmann, J., Joseph, S., 2010. Sustainable biochar to mitigate global climate change. Nat. Commun. 1, 1.

Xiao, X., Chen, Z., Chen, B., 2016. H/C atomic ratio as a smart linkage between pyrolytic temperatures, aromatic clusters and sorption properties of biochars derived from diverse precursory materials. Sci. Rep-Uk 6, 22644.

Yuan, X., He, T., Cao, H., Yuan, Q., 2017. Cattle manure pyrolysis process: kinetic and thermodynamic analysis with isoconversional methods. Renew. Energy 107, 489–496.

Yue, Y., Lin, Q., Xu, Y., Li, G., Zhao, X., 2017. Slow pyrolysis as a measure for rapidly treating cow manure and the biochar characteristics. J. Anal. Appl. Pyrolysis 124, 355–361.

Zhuo, Y., Mansouri Tehrani, A., Oliynyk, A.O., Duke, A.C., Brgoch, J., 2018. Identifying an efficient, thermally robust inorganic phosphor host via machine learning. Nat. Commun. 9, 4377.