

Predicting Dry-Basis Carbon Content in Raw Plant Biomass from Ash and Volatile Matter

Shuanzhu Sun,^{a,*} Yiqing Wang,^b Jiewen Zhu,^a Xiao Zhang,^b Chunlei Zhou,^a and Chenchen Yang^a

Carbon content is needed for fuel characterization and carbon accounting, but measuring every incoming biomass lot by ultimate analysis is not always practical. This study examined whether routine proximate data could provide a preliminary value that remains easy to audit. The source table contained 1,689 solid-fuel records. Filters for feedstock type, completeness, physical limits, proximate closure, and duplication reduced this to 643 samples of untreated wood, grass or plant, organic residue, husk or shell, and straw. A category-stratified test set was reserved before comparing ridge regression, four nonlinear regressors, and three published correlations by nested cross-validation. Ridge based on ash and volatile matter returned the lowest mean root mean square error (RMSE; 2.883 wt.% C). For the 129 held-out samples, R^2 was 0.707, RMSE was 2.412 wt.% C, and mean absolute error (MAE) was 1.767 wt.% C. Its lead over the alternatives was small. The nominal 90% conformal interval covered 95.3% of test samples, although its mean width was 11.841 wt.% C; leave-one-category-out RMSE reached 4.621 wt.% C. The equation is limited to preliminary screening and quality assurance. Certified elemental analysis is still required for high-stakes decisions.

DOI: 10.15376/biores.21.4.11400-11413

Keywords: Lignocellulosic biomass; Elemental carbon; Proximate analysis; Ridge regression; Nested cross-validation; Conformal prediction; Fuel screening

Contact information: *a: Jiangsu Frontier Electric Technology Co., Ltd., No. 58 Suyuan Avenue, Jiangning District, Nanjing, Jiangsu, China; b: State Grid Jiangsu Electric Power Co., Ltd. Xuzhou Power Supply Company, No. 20 Jiefang North Road, Gulou District, Xuzhou, Jiangsu, China;*

** Corresponding author: 15905166613@139.com*

INTRODUCTION

Biomass sold under one feedstock name can still vary markedly from lot to lot. Untreated wood, straw, grasses, husks, and crop residues differ in moisture, ash, volatile matter (VM), fixed carbon (FC), and elemental composition. Nurek *et al.* (2019), for example, reported distinct compositional groups for forest residues and power-plant wood chips. These differences matter during handling and combustion, and they enter the carbon balance used for reporting (Faaij 2006). For an operator receiving changing biomass lots, a single initial characterization is insufficient.

Fuel laboratories routinely obtain proximate values for moisture, ash, VM, and FC. Carbon, hydrogen, nitrogen, sulfur, and often oxygen come from ultimate analysis, which needs dedicated elemental instrumentation and tighter analytical control (Racero-Galaraga *et al.* 2024). Proximate analysis was not treated as a replacement for that measurement. The practical question was narrower: can the results of proximate analysis help select samples for confirmation or reveal a gross inconsistency while the carbon result is pending?

Several equations already connect proximate and ultimate composition, although their calibration domains differ. Parikh *et al.* (2007) derived linear relations from 200 biomass records and tested them on another 50. Shen *et al.* (2010) included ash explicitly, whereas Nhuchhen (2016) covered both raw and torrefied biomass. More recent studies have used nonlinear least-squares correlations (Mohammed *et al.* 2022), neuro-fuzzy systems (Olatunji *et al.* 2022), and stepwise linear correlations (Camacho Ardila *et al.* 2024). Other work has considered neural networks, linear regression (Lawal *et al.* 2021), ensemble learning, and interpretable machine learning (Timilsina *et al.* 2024). Neural-network and hybrid models have also been applied to related biomass properties (Matin *et al.* 2024; Wang and Zhao 2025). These methods broaden the available choices, but greater model complexity does not by itself improve either independently validated accuracy or auditability.

Four points shaped our comparison. First, dry-basis ash, VM, and FC are approximately closed, so an unconstrained model using all three repeats one degree of freedom. Second, hyperparameter tuning has to remain separate from performance estimation; otherwise the reported error can be optimistic (Varma and Simon 2006; Cawley and Talbot 2010). Third, a conventional random holdout says little about a feedstock category absent from training. Finally, a point estimate gives no uncertainty for the individual lot being screened. For this quality-assurance application, auditability means that an analyst can trace the retained measurements, reproduce the split and preprocessing, recalculate a prediction from fixed coefficients, and see when an uncertainty or domain check triggers elemental confirmation. Auditability is not a substitute for accuracy: a more complex model would be preferable if it delivered a material and independently validated gain. Accordingly, model families were first compared on identical folds; parsimony and traceability were considered only after predictive performance had been established.

The study set a deliberately narrow objective: determine how much predictive information is available in two independent coordinates of routine proximate composition for raw plant biomass. To our knowledge, this is the first biomass carbon-prediction study to combine an explicitly non-redundant proximate representation with a test set locked before model comparison, nested hyperparameter selection, distribution-free prediction intervals, and leave-one-category-out validation. This combination allows average accuracy, individual-sample uncertainty, and transfer to an unrepresented feedstock category to be assessed within one auditable workflow. The output was intended as a screening equation with a traceable calculation, not as an alternative to certified elemental analysis.

EXPERIMENTAL

Public Data and Screening

The analysis started from the public supplementary dataset of Timilsina *et al.* (2024): 1,689 records compiled from Phyllis2 (TNO 2026). The table contains fuel category, proximate composition, elemental composition, and higher heating value, but no moisture field. The values were treated as dry basis. Among retained complete records, median ash + VM + FC was 100.00 wt.%, and median ash + C + H + O + N + S was 99.98 wt.%; both closure results support that interpretation. The second sum was not always close to 100%. It was kept as a diagnostic rather than a filter, since the original laboratory and basis metadata were unavailable.

The screening rules were fixed before model fitting. Five categories of unconverted plant feedstock were retained: untreated wood, grass or plant, organic residue, husk or shell, and straw. Processed fuels and chars were outside the study because conversion changes the relation between proximate fractions and carbon. Sludge, plastics, paper, animal wastes, and mixed municipal wastes were also excluded. Each retained row required non-negative ash, VM, FC, and elemental carbon, with carbon strictly between 0 and 100 wt.%. Dry-basis proximate closure had to fall within 99 to 101 wt.%. This ± 1 wt.% interval was used as an internal-consistency check, not as an interlaboratory repeatability specification. When category, ash, VM, FC, and carbon were identical, only the first source row was retained. Carbon was required as the supervised response for model development and error calculation, but it was never supplied as a predictor. At application, the final equation uses only ash and VM. Rows without measured carbon cannot support supervised accuracy assessment and therefore were excluded from model development; residual values and target-dependent performance played no part in deletion decisions.

Of the 1,689 source records, 1,187 belonged to the target categories, 650 had all core variables, and 646 passed the physical and closure checks. Removing three duplicate vectors left 643 samples: 284 untreated wood, 121 grass or plant, 93 organic residue, 75 husk or shell, and 70 straw. The supplementary files retain the source-row audit, processed table, locked split, model outputs, and scripts so that this sequence can be checked.

Basis Definition and Predictors

All mass fractions used by the models were on a dry basis. This basis was maintained throughout; mixing dry-basis and as-received values would introduce a moisture-driven relation. The elemental-carbon variable is denoted by C_d , and the proximate fractions obey Eq. 1, as follows,

$$Ash_d + VM_d + FC_d \approx 100 \quad (1)$$

where Ash_d is the dry-basis ash content, VM_d is the dry-basis volatile matter content, and FC_d is the dry-basis fixed carbon content, all expressed as mass percentages of the dry fuel (wt.%). Because FC is calculated by difference, using ash, VM, and FC together duplicates one degree of freedom. Primary models therefore used ash and VM as independent coordinates. FC was retained for description and for published equations. When moisture, Mar , is measured independently for the same representative sample, Eq. 2 converts the predicted dry-basis carbon value to an as-received value, as follows,

$$C_{ar} = C_d (1 - Mar/100) \quad (2)$$

where C_{ar} is the as-received carbon content, C_d is the dry-basis carbon content, and Mar is the as-received moisture content, all in wt.%. Equation 2 simply changes the reporting basis. It is not another fitted relation: the as-received carbon fraction equals the dry-basis fraction multiplied by the dry-matter fraction.

Candidate Models and Published Correlations

Five regressors were compared. Ridge regression (Hoerl and Kennard 1970) served as the linear model. The nonlinear set comprised radial-basis support vector regression (SVR), random forest (RF) (Breiman 2001), extremely randomized trees (Geurts *et al.* 2006), and histogram gradient boosting. Ridge and SVR were fitted after standardizing ash and VM; the tree models received the original units. The search grids were compact: ridge penalty 0.01, 0.1, 1, 10, or 100; SVR penalty 10 or 100, kernel

coefficient ‘scale’ or 0.1, and insensitive loss 0.1 or 0.5; RF and extra trees with 100 trees, unrestricted or 12 maximum depth, and 1 or 5 minimum samples per leaf; and boosting with 150 iterations, learning rate 0.03 or 0.08, 7 or 15 leaf nodes, and L2 penalty 1. With 514 development samples, a broader search that would add selection variance was avoided. For the ridge sensitivity check, ash only, VM only, FC only, and ash + VM were evaluated under the same splits.

Three published equations were applied without refitting their coefficients. This tests how the published forms transfer to the present biomass domain:

$$C_d = 0.637 FC_d + 0.455 VM_d, \quad (3)$$

$$C_d = 0.635 FC_d + 0.460 VM_d - 0.095 Ash_d, \quad (4)$$

$$C_d = -35.9972 + 0.7698 VM_d + 1.3269 FC_d + 0.3250 Ash_d \quad (5)$$

where C_d is the dry-basis elemental carbon content, and FC_d , VM_d , and Ash_d are the dry-basis fixed carbon, volatile matter, and ash contents, respectively, all in wt.%. Equations 3, 4, and 5 are from Parikh *et al.* (2007), Shen *et al.* (2010), and Nhuchhen (2016), respectively.

Leakage-Controlled Validation

Calculations were run in Python 3.9.6 (Python Software Foundation, Beaverton, OR, USA), NumPy 2.0.2, pandas 2.3.3, SciPy, and scikit-learn 1.6.1 (Pedregosa *et al.* 2011). The seed was fixed at 20260711. Before comparing models, a category-stratified 20% split produced 514 development and 129 test samples. All five feedstock categories occurred in both subsets. The saved assignments were not changed after benchmarking began.

Only the development set was used for model selection. Five shuffled, category-stratified outer folds estimated performance. Within each outer training set, three similarly stratified inner folds selected the hyperparameters with the lowest RMSE. Preprocessing was refitted in every training fold, and the published equations were evaluated on exactly the same outer validation folds. This nesting kept tuning results separate from the performance estimate (Varma and Simon 2006; Cawley and Talbot 2010).

The coefficient of determination (R^2), RMSE, and mean absolute error (MAE) were,

$$R^2 = 1 - [\Sigma(y_i - \hat{y}_i)^2 / \Sigma(y_i - \bar{y})^2], \quad (6)$$

$$RMSE = [\frac{1}{n} \Sigma(y_i - \hat{y}_i)^2]^{1/2}, \quad (7)$$

$$MAE = \frac{1}{n} \Sigma|y_i - \hat{y}_i| \quad (8)$$

where y_i is the measured dry-basis carbon content of sample i , $\hat{y}_i$ is the corresponding predicted value, $\bar{y}$ is the mean of the measured values (all in wt.%), and n is the number of samples.

Each model used the same outer folds, so fold RMSE values were compared with two-sided paired Wilcoxon signed-rank tests (Wilcoxon 1945). There were only five pairs; these tests are reported as descriptive. Once the model family and features were fixed, five-fold validation on all development samples selected the final hyperparameter. It was evaluated once on the 129-sample test set. Confidence intervals came from 2,000 paired bootstrap resamples that kept each measured-predicted pair intact.

For split-conformal prediction, the development data were divided by category stratification into proper training and 25% calibration subsets (Angelopoulos and Bates

2023). Ridge tuning used proper training only. Absolute calibration residuals set the finite-sample 90% quantile, after which the interval was evaluated on the untouched test set. Test-set permutation importance used the RMSE increase over 100 repetitions (Fisher *et al.* 2019). For the category-transfer test, one category was excluded at a time, tuned ridge by four-fold validation on the remaining categories, and predicted the excluded category.

RESULTS AND DISCUSSION

Dataset Characteristics

Dry-basis carbon ranged from 19.70 to 62.17 wt.%, with a mean of 46.95 wt.% (Table 1). Untreated wood made up 44.2% of the records and combined the lowest median ash with the highest median carbon; aggregate errors thus reflect the available category balance. After screening, the proximate sum ranged from 99.39 to 100.01 wt.%. For the 529 rows with complete carbon, hydrogen, oxygen, nitrogen, and sulfur values, the ash-plus-elements sum had a median of 99.98 wt.% but ranged from 94.14 to 109.94 wt.%. Those departures were left in the data. Source-specific analytical metadata were not available, and cleaning on a target-linked sum could hide genuine heterogeneity.

Table 1. Descriptive Statistics of the Retained Dataset (n = 643)

Variable (wt.%, dry basis)	Mean	SD	Minimum	Median	Maximum
Ash	5.99	8.03	0.00	3.30	64.00
Volatile matter	76.99	8.22	29.30	78.63	94.73
Fixed carbon	17.02	5.04	0.00	16.82	44.90
Elemental carbon	46.95	5.12	19.70	48.00	62.17

Figure 1 reports the screening losses and the retained category distributions. Figure 2 gives the signed Spearman matrix. Its interpretation must account for both mass-fraction dilution and compositional closure. Carbon correlated with ash at $\rho = -0.709$ because, on a dry whole-fuel basis, increasing noncombustible inorganic mass directly reduces the mass fraction available for carbon-bearing organic matter; ash-explicit correlations have reported the same direction (Shen *et al.* 2010; Camacho Ardila *et al.* 2024). Carbon correlated with VM at $\rho = 0.494$, while ash and VM correlated at -0.732 . FC showed the weakest marginal association with elemental carbon. This is plausible because FC is an operational residue after devolatilization and is calculated by difference; it is not identical to elemental carbon, which is partitioned between volatile products and the char-forming fraction. Mohammed *et al.* (2022) likewise reported no clear simple C-FC trend in its source data. These correlations describe association and do not establish a chemical causal effect.

The category boxplots in Fig. 1c mark 49 observations beyond 1.5 times the interquartile range: 40 on the low-carbon side and nine on the high-carbon side. Of the 40 low-side points, 34 contained at least 15 wt.% ash; their mean ash content was 24.86 wt.%, compared with 4.78 wt.% for observations not marked as boxplot outliers. This pattern supports mineral dilution as the main measurable explanation for many low-side points.

All 49 observations passed the prespecified physical and closure checks and were retained. The public file contains only broad category labels and no subtype information, so proposed explanations such as bark-rich wood or soil-contaminated grass cannot be verified from these data.

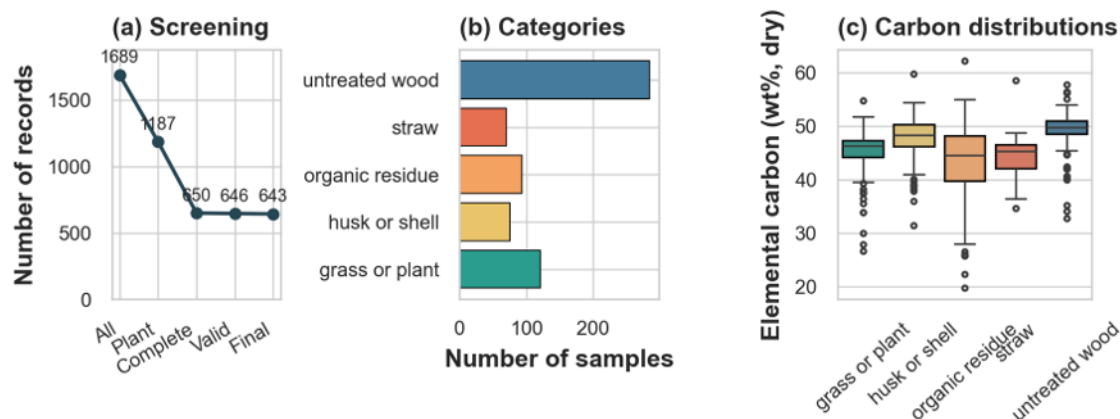


Fig. 1. Sample screening, retained categories, and dry-basis carbon distributions; circles beyond the boxplot whiskers in panel (c) denote observations outside 1.5 times the interquartile range and were retained

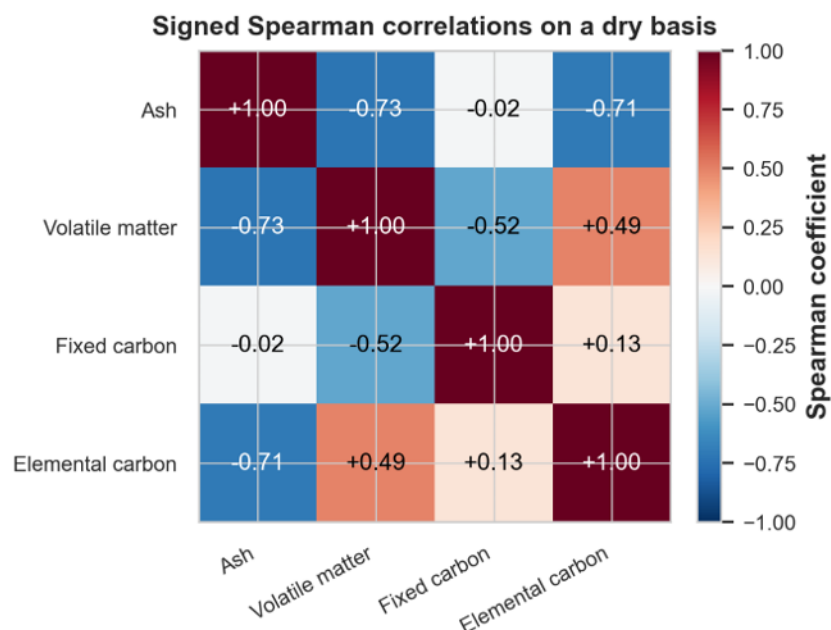


Fig. 2. Signed Spearman correlations among dry-basis proximate fractions and elemental carbon; shading denotes correlation magnitude

Model Comparison and Sensitivity

Ridge had the lowest outer-fold RMSE, 2.883 ± 0.153 wt.% C, and mean R^2 of 0.694 ± 0.044 (Table 2; Fig. 3). The difference from SVR was small, and SVR also had a slightly lower MAE. RF, extra trees, and boosting returned RMSE values of 3.049, 3.091, and 3.105 wt.% C, respectively. Thus, none of the flexible learners produced a material common-fold improvement over ridge.

Among the three fixed equations, Parikh *et al.* (2007) transferred best at 3.071 wt.% C, only 0.188 wt.% C above ridge, whereas the ash-explicit Shen *et al.* (2010) equation reached 3.150 wt.% C. The raw-and-torrefied equation of Nhuchhen (2016) reached 3.773 wt.% C, consistent with the effect of applying coefficients outside their original material domain. Paired Wilcoxon p values were 0.813 for ridge versus SVR and 0.063 to 0.188 for the remaining comparisons. With only five fold pairs, these tests are descriptive; the results identify a practical choice among near-peers rather than a universally superior algorithm.

Table 2. Category-Stratified Nested Validation on the Development Set (n = 514); Mean $\pm$ SD over Five Outer Folds

Model	R ²	RMSE (wt.% C)	MAE (wt.% C)
Ridge	0.694 $\pm$ 0.044	2.883 $\pm$ 0.153	2.049 $\pm$ 0.131
Support vector regression	0.690 $\pm$ 0.048	2.900 $\pm$ 0.147	1.986 $\pm$ 0.139
Random forest	0.653 $\pm$ 0.077	3.049 $\pm$ 0.128	2.172 $\pm$ 0.130
Parikh <i>et al.</i> (2007)	0.653 $\pm$ 0.056	3.071 $\pm$ 0.271	2.337 $\pm$ 0.189
Extra trees	0.649 $\pm$ 0.035	3.091 $\pm$ 0.120	2.162 $\pm$ 0.119
Gradient boosting	0.643 $\pm$ 0.060	3.105 $\pm$ 0.148	2.209 $\pm$ 0.074
Shen <i>et al.</i> (2010)	0.635 $\pm$ 0.054	3.150 $\pm$ 0.234	2.379 $\pm$ 0.210
Nhuchhen (2016)	0.470 $\pm$ 0.111	3.773 $\pm$ 0.165	2.756 $\pm$ 0.186

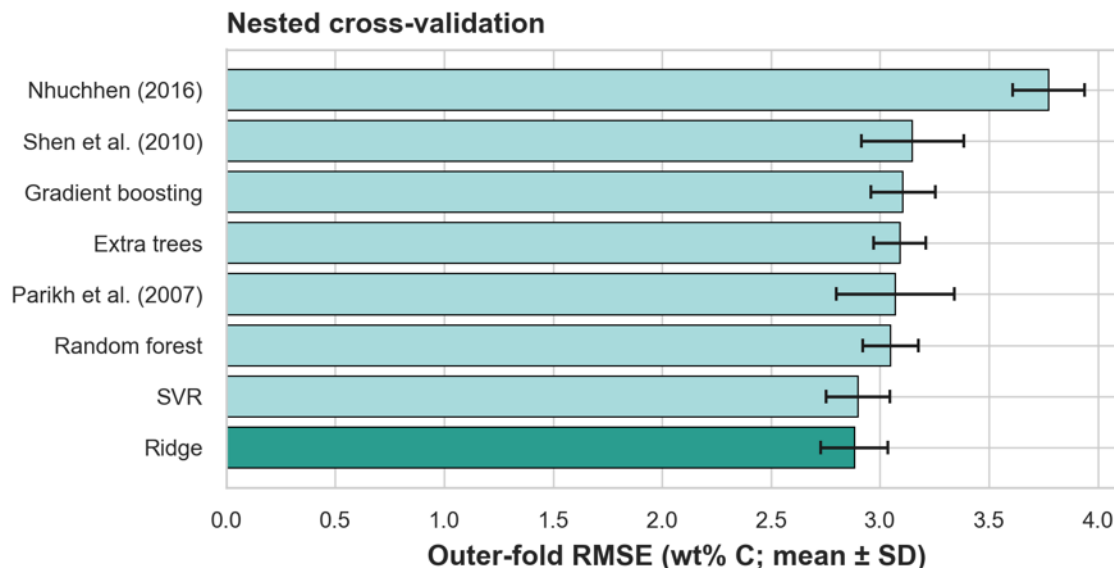


Fig. 3. Outer-fold root mean square error; bars show the mean and error bars show the standard deviation over five folds

Recent studies were also reviewed to place the common-fold comparison in Table 2 in a current context. Because the studies used different datasets, compositional domains, preprocessing, and validation designs, their published metrics were not ranked directly against the present nested-validation results.

Table 3. Contextual Comparison with Recent Biomass Elemental-Composition Studies

Study	Method and Inputs	Reported Carbon Performance	Interpretive Boundary
Mohammed <i>et al.</i> (2022)	Nonlinear least squares; VM and FC; 450 training plus 50 validation records	AAPE 2.12%; R^2 0.9993	Restricted domain ($60 < VM < 90$; $10 < FC < 30$); different validation
Olatunji <i>et al.</i> (2022)	FCM-ANFIS; ash, VM, and FC	RMSE 3.894 wt.% C; MAD 2.442 wt.% C; CC 0.898	Different dataset and validation design; not a common-fold result
Camacho Ardila <i>et al.</i> (2024)	Stepwise linear correlation; proximate inputs	RMSE 2.65 wt.% C; AAE 3.86%; ABE -1.10%	Different dataset and metrics; not a common-fold result

Note: AAPE, average absolute percentage error; AAE, average absolute error; ABE, average bias error; CC, correlation coefficient; FCM-ANFIS, fuzzy c-means adaptive neuro-fuzzy inference system. Published metrics are reproduced for context and are not directly rank-comparable with Table 2.

Adding VM to ash changed development RMSE from 2.923 to 2.883 wt.% C (Table 4). On the locked test set, the ordering reversed: ash alone gave 2.360 wt.% C, compared with 2.412 wt.% C for ash + VM; VM and FC alone were clearly weaker. The two-variable model was expected to keep its advantage. It did not. Ash + VM was kept as the primary model because that choice had been made before the test set was opened. The reversal indicates that VM adds little information that transfers reliably once ash is included. Table 4 also provides the requested selective predictor-exclusion tests: ash-only, VM-only, and FC-only models were evaluated under the same saved splits. Carbon was excluded from every predictor set and served only as the response used to fit and evaluate each supervised model.

Table 4. Ridge Feature Sensitivity under Identical Nested Splits and on the Locked Test Set

Feature Set	Development RMSE	Development MAE	Test R^2	Test RMSE	Test MAE
Ash + VM	2.883 $\pm$ 0.153	2.049 $\pm$ 0.131	0.707	2.412	1.767
Ash only	2.923 $\pm$ 0.199	2.049 $\pm$ 0.161	0.720	2.360	1.721
VM only	4.164 $\pm$ 0.417	2.808 $\pm$ 0.283	0.495	3.168	2.288
FC only	4.995 $\pm$ 0.313	3.680 $\pm$ 0.344	0.061	4.320	3.329

Note: Error values are wt.% C.

Independent Test and Final Equation

The locked test set was not used for model or feature choice. Tuning on all development samples gave a ridge penalty of 1.0. Back-transformation to the original mass-percentage scale gave the following equation; coefficients are shown to three significant figures to reflect the precision justified for heterogeneous natural materials. Full-precision coefficients remain in the accompanying code for exact reproducibility.

$$C_d = 58.3 - 0.617 \text{ Ash}_d - 0.101 \text{ VM}_d \quad (9)$$

where C_d is the predicted dry-basis carbon content, and Ash_d and VM_d are the dry-basis ash

and volatile matter contents, respectively, all in wt.%. Using $VM_d \approx 100 - Ash_d - FC_d$, Eq. 9 can be rearranged as follows,

$$C_d \approx 48.2 - 0.516 Ash_d + 0.101 FC_d \quad (10)$$

where C_d is the predicted dry-basis carbon content, and Ash_d and FC_d are the dry-basis ash and fixed carbon contents, respectively, all in wt.%.

For the 129 test samples, R^2 was 0.707, RMSE was 2.412 wt.% C, and MAE was 1.767 wt.% C (Table 5; Fig. 4). Bootstrap intervals were tighter for RMSE and MAE than for R^2 , which is more sensitive to the composition and variance of a finite test set. The two lowest-carbon test samples contained 26.00 and 26.60 wt.% C and 36.50 and 45.80 wt.% ash, respectively; their predictions were 30.48 and 24.64 wt.% C. Their low carbon values are therefore consistent with strong mineral dilution. The largest absolute residuals were not confined to the lowest-carbon edge: six of the eight largest occurred within the broad organic-residue category. The public table lacks material subtype, supplier, and laboratory metadata, so a more specific feedstock-level attribution is not supportable. All of these samples were retained, and the residual tails in Fig. 5 show where one equation fails to represent the full feedstock heterogeneity. Equation 9 was used for all reported predictions.

Table 5. Performance on the Untouched Category-Stratified Test Set (n = 129)

Metric	Estimate	Bootstrap 95% Confidence Interval
R^2	0.707	0.455-0.824
RMSE (wt.% C)	2.412	2.007-2.830
MAE (wt.% C)	1.767	1.505-2.056

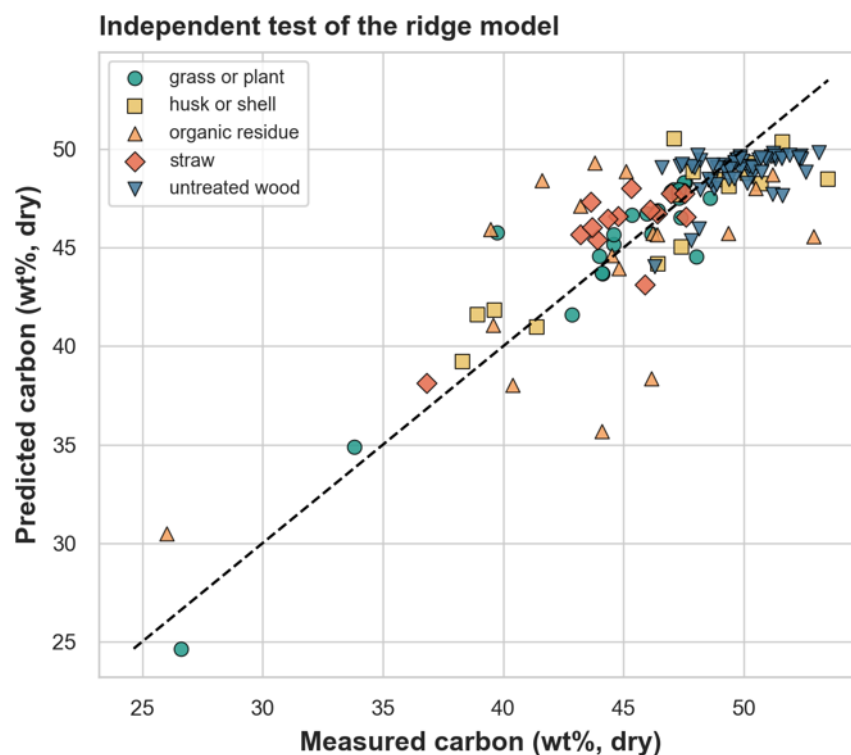


Fig. 4. Measured and ridge-predicted dry-basis carbon on the untouched test set

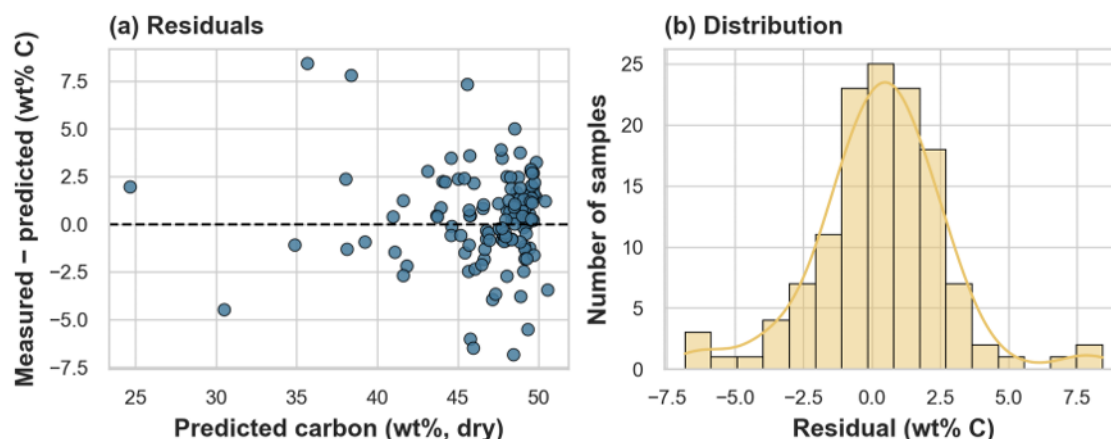


Fig. 5. Independent-test residuals against predictions and residual distribution

Uncertainty and Category Transfer

Permuting ash raised test RMSE by 4.244 ± 0.312 wt.% C; the increase for VM was 0.128 ± 0.055 wt.% C (Fig. 6a). Ash thus carried most of the stable predictive information available to this fitted ridge model. The predictors form a closed composition, so this result should not be read as evidence that VM is chemically irrelevant.

The nominal 90% conformal interval had a half-width of 5.920 wt.% C. It covered 123 of 129 test samples (95.3%), and its mean total width was 11.841 wt.% C. Coverage was conservative, but the interval was broad. It is useful as a warning during screening, not as support for a high-stakes decision from one prediction.

When each category was excluded from training in turn, RMSE ranged from 2.009 wt.% C for grass or plant to 4.621 wt.% C for organic residue (Table 6; Fig. 6b). Straw and untreated wood had R^2 values of 0.229 and 0.263 despite moderate absolute errors; their restricted carbon ranges partly explain this result. The organic-residue RMSE was nearly twice the locked mixed-category test RMSE. This category-level transfer loss is compatible with the concentration of large independent-test residuals in the broad organic-residue label. A random split can therefore look reassuring even when transfer to an unfamiliar category is weak, a limitation that cannot be inferred from the high within-domain R^2 values reported by some recent correlations (Mohammed *et al.* 2022).

Table 6. Ridge Performance When Each Category Was Completely Excluded from Training

Held-Out Category	n	R^2	RMSE (wt.% C)	MAE (wt.% C)
Grass or plant	121	0.786	2.009	1.416
Husk or shell	75	0.646	2.898	2.182
Organic residue	93	0.628	4.621	3.832
Straw	70	0.229	3.318	2.347
Untreated wood	284	0.263	2.450	1.938

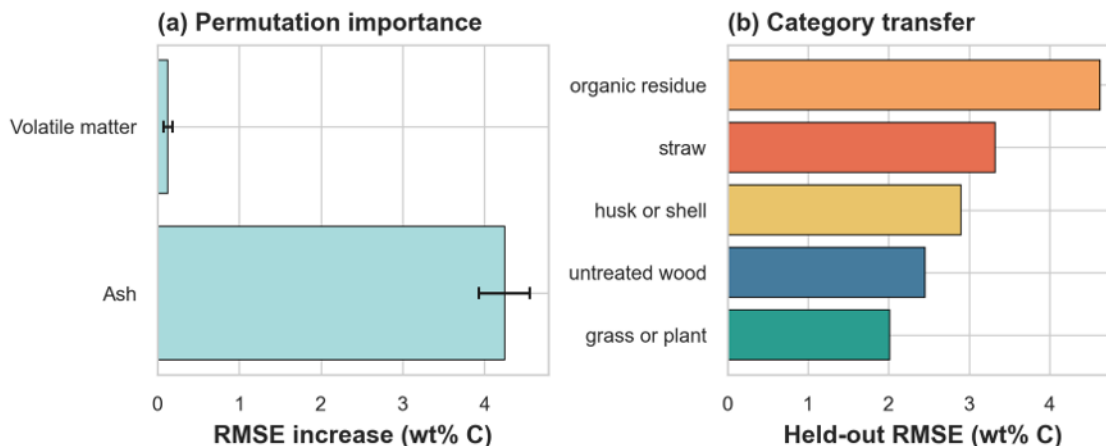


Fig. 6. Model diagnostics: (a) independent-test permutation importance and (b) error for a completely unseen biomass category

Why an Auditable Model Was Competitive

Initially, it was expected that at least one nonlinear model would improve on the two-parameter ridge equation. None did so consistently on identical outer folds. The data structure offers a plausible explanation: ash, VM, and FC are nearly closed, only two coordinates are independent, and most of the stable signal was associated with ash. Ridge reduced coefficient variance and produced a fixed equation that can be checked line by line. In this application, that traceability is useful because a screening prediction may trigger certified ultimate analysis, flag an inconsistent laboratory result, or be rejected when an uncertainty or domain check fails. This result is specific to the present data and target; it is not a general argument against nonlinear learning. If a neural, hybrid, or ensemble model produced a material gain on a locked external evaluation, that gain could justify the additional complexity. Here, flexible learners could fit local noise and category structure but did not lower common-fold validation error.

Published performance values require careful qualification. Mohammed *et al.* (2022) reported carbon AAPE of 2.12% and R^2 of 0.9993 for a nonlinear model within a restricted VM-FC domain, Olatunji *et al.* (2022) reported carbon RMSE of 3.894 wt.% for a neuro-fuzzy model, and Camacho Ardila *et al.* (2024) reported carbon RMSE of 2.65 wt.% for a stepwise linear correlation. These values were obtained from different datasets, compositional ranges, preprocessing rules, and validation designs, so they cannot be ranked directly against the common-fold results in Table 2. The present work instead asked whether model complexity improved error under identical splits and whether performance persisted on an untouched test set and completely unseen categories. The Parikh equation (Parikh *et al.* 2007) transferred reasonably well, consistent with a stable linear relation. The poorer result for the raw-and-torrefied equation of Nhuchhen (2016) further illustrates the importance of a declared application domain. Accuracy reported for higher heating value or biomass ash is also not directly comparable because the targets differ (Matin *et al.* 2024; Timilsina *et al.* 2024; Wang and Zhao 2025).

Value and Engineering Boundary for Lignocellulosic Feedstocks

While an elemental result is pending, the proximate-based value can support an initial check or flag a large disagreement between the two analyses. If moisture was measured on the same representative sample, Eq. 2 provides the corresponding as-received

value. This remains a laboratory screening calculation. It is not a real-time sensor, since proximate analysis still involves sampling, preparation, and thermal measurement.

A test MAE of 1.77 wt.% C can become material when multiplied by a large fuel mass, and the conformal interval shows substantial uncertainty for an individual sample. Elemental confirmation is appropriate near a decision threshold, outside the training range, or for a new category. Sampling error, moisture heterogeneity, biogenic-fossil allocation, mass flow, and combustion completeness introduce further uncertainties that this model does not address. The equation should not be used for regulatory reporting, fuel settlement, or contractual acceptance.

Limitations and Validation Priority

The public table retains broad category labels and numerical values but omits some original identifiers, material subtypes, laboratories, analytical standards, and sources. Consequently, hypotheses such as bark-rich wood or soil-contaminated grass cannot be verified. Source-aware grouping could not be reconstructed, and records from one original study may occur in both development and test sets. Leave-one-category-out testing examines one form of distribution shift; it does not resolve this source-grouping limitation. Differences among analytical protocols may contribute additional heterogeneity when method metadata are missing (Shen *et al.* 2010; Racero-Galaraga *et al.* 2024). Untreated wood also accounted for 44.2% of the data, and stratification preserved rather than removed that imbalance.

No plant- or supplier-specific samples were used in this study. The next validation should collect paired proximate and ultimate analyses from the same representative samples and record supplier, batch, laboratory, and sampling information. Before any local update, the grouping rule and evaluation protocol should be fixed by source or batch. Calibration is warranted only after the new samples have been checked against the compositional domain of the public data.

CONCLUSIONS

1. Screening reduced 1,689 source records to 643 auditable raw plant-biomass samples. On 129 untouched samples, the ridge equation based on ash and volatile matter (VM) gave $R^2 = 0.707$, RMSE = 2.412 wt.% C, and MAE = 1.767 wt.% C. It provides a first-pass value with a traceable calculation.
2. VM added little information after ash and did not improve the locked test result. The nominal 90% interval was broad, and unseen-category RMSE reached 4.621 wt.% C. These checks are as important as the average test error when judging use of the equation.
3. The method should be used for preliminary screening and quality assurance only. Certified elemental analysis remains necessary for fuel settlement, contractual acceptance, regulatory reporting, and plant-specific carbon accounting.

ACKNOWLEDGMENTS

This work was supported by State Grid Corporation of China (SGCC) (Project No. J2025190). OpenAI Codex (OpenAI, San Francisco, CA, USA) was used to support English-language editing, sentence restructuring, and literature cross-checking. The tool was not used to create experimental data or alter the locked model outputs. The authors retain full responsibility for the manuscript.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability

The source dataset is publicly available through the cited Phyllis2-based compilation. The processed dataset, locked split, screening audit, analysis scripts, and numerical outputs are available from the corresponding author upon reasonable request.

REFERENCES CITED

- Angelopoulos, A. N., and Bates, S. (2023). “Conformal prediction: A gentle introduction,” *Foundations and Trends in Machine Learning* 16(4), 494-591. <https://doi.org/10.1561/22000000101>
- Breiman, L. (2001). “Random forests,” *Machine Learning* 45(1), 5-32. <https://doi.org/10.1023/A:1010933404324>
- Camacho Ardila, Y., Jaimes Figueroa, J. E., and Wolf Maciel, M. R. (2024). “Mathematical models for predicting the higher heating value and ultimate analysis of biomass,” *Industrial Crops and Products* 208, article 117777. <https://doi.org/10.1016/j.indcrop.2023.117777>
- Cawley, G. C., and Talbot, N. L. C. (2010). “On over-fitting in model selection and subsequent selection bias in performance evaluation,” *Journal of Machine Learning Research* 11, 2079-2107.
- Faaij, A. P. C. (2006). “Modern biomass conversion technologies,” *Mitigation and Adaptation Strategies for Global Change* 11(2), 343-375. <https://doi.org/10.1007/s11027-005-9004-7>
- Fisher, A., Rudin, C., and Dominici, F. (2019). “All models are wrong, but many are useful: Learning a variable’s importance by studying an entire class of prediction models simultaneously,” *Journal of Machine Learning Research* 20(177), 1-81.
- Geurts, P., Ernst, D., and Wehenkel, L. (2006). “Extremely randomized trees,” *Machine Learning* 63(1), 3-42. <https://doi.org/10.1007/s10994-006-6226-1>
- Hoerl, A. E., and Kennard, R. W. (1970). “Ridge regression: Biased estimation for nonorthogonal problems,” *Technometrics* 12(1), 55-67. <https://doi.org/10.1080/00401706.1970.10488634>
- Lawal, A. I., Aladejare, A. E., Onifade, M., Bada, S., and Idris, M. A. (2021). “Predictions of elemental composition of coal and biomass from their proximate analyses using ANFIS, ANN and MLR,” *International Journal of Coal Science & Technology* 8(1), 124-140. <https://doi.org/10.1007/s40789-020-00346-9>
- Matin, A., Špelić, K., Jurišić, V., Matin, B., Grubor, M., Tomić, I., Majdak, T., and Brandić, I. (2024). “The possibility of modelling agricultural biomass ash by neural

- networks concerning proximate analysis inputs,” *Thermal Science* 28(6A), 4771-4780. <https://doi.org/10.2298/TSCI240618238M>
- Mohammed, I. Y., James, D., El-Yakubu, B. J., and Bawa, M. A. (2022). “Proximal-based predictive model for estimating ultanal attributes of lignocellulosic biomass,” *Cleaner Chemical Engineering* 4, article 100071. <https://doi.org/10.1016/j.clce.2022.100071>
- Nhuchhen, D. R. (2016). “Prediction of carbon, hydrogen, and oxygen compositions of raw and torrefied biomass using proximate analysis,” *Fuel* 180, 348-356. <https://doi.org/10.1016/j.fuel.2016.04.058>
- Nurek, T., Gendek, A., and Roman, K. (2019). “Forest residues as a renewable source of energy: Elemental composition and physical properties,” *BioResources* 14(1), 6-20. <https://doi.org/10.15376/biores.14.1.6-20>
- Olatunji, O. O., Adedeji, P. A., Madushele, N., Akinlabi, S., and DiCarlo, A. A. (2022). “Modelling biomass elemental composition: A neurofuzzy approach,” *Procedia Computer Science* 200, 1736-1745. <https://doi.org/10.1016/j.procs.2022.01.374>
- Parikh, J., Channiwala, S. A., and Ghosal, G. K. (2007). “A correlation for calculating elemental composition from proximate analysis of biomass materials,” *Fuel* 86(12-13), 1710-1719. <https://doi.org/10.1016/j.fuel.2006.12.029>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., *et al.* (2011). “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research* 12, 2825-2830.
- Racero-Galaraga, D., Rhenals-Julio, J. D., Sofan-German, S., Mendoza, J. M., and Bula-Silvera, A. (2024). “Proximate analysis in biomass: Standards, applications and key characteristics,” *Results in Chemistry* 12, article 101886. <https://doi.org/10.1016/j.rechem.2024.101886>
- Shen, J., Zhu, S., Liu, X., Zhang, H., and Tan, J. (2010). “The prediction of elemental composition of biomass based on proximate analysis,” *Energy Conversion and Management* 51(5), 983-987. <https://doi.org/10.1016/j.enconman.2009.11.039>
- Timilsina, M. S., Sen, S., Uprety, B., Patel, V. B., Sharma, P., and Sheth, P. N. (2024). “Prediction of HHV of fuel by machine learning algorithm: Interpretability analysis using Shapley Additive Explanations (SHAP),” *Fuel* 357, article 129573. <https://doi.org/10.1016/j.fuel.2023.129573>
- TNO (2026). “Phyllis2: Database for biomass and waste,” (<https://phyllis.nl/>), Accessed 11 July 2026.
- Varma, S., and Simon, R. (2006). “Bias in error estimation when using cross-validation for model selection,” *BMC Bioinformatics* 7, article 91. <https://doi.org/10.1186/1471-2105-7-91>
- Wang, X., and Zhao, S. (2025). “The PSO/GA/ANN modeling and prediction for the higher heating values of solid fuels: The machine learning approach,” *Thermal Science* 29(4A), 2881-2897. <https://doi.org/10.2298/TSCI240922012W>
- Wilcoxon, F. (1945). “Individual comparisons by ranking methods,” *Biometrics Bulletin* 1(6), 80-83. <https://doi.org/10.2307/3001968>

Article submitted: Peer review completed: August 29, 2026; Revised version received: August 30, 2026; Accepted: September 18, 2026; Published: October 9, 2026.
DOI: 10.15376/biores.21.4.11400-11413