

# Interpretable machine learning reveals biochar-specific hydraulic predictors of plant microbial fuel cell performance under unsaturated soil conditions

Sibin Gan<sup>1</sup>

[gansibin@mails.gdut.edu.cn](mailto:gansibin@mails.gdut.edu.cn)

Guangdong University of Technology

Qinyue Yao<sup>2</sup>

---

## Research Article

**Keywords:** Plant microbial fuel cell, Biochar, Unsaturated soil, Matric suction, Interpretable machine learning, Bioelectrochemical engineering

**Posted Date:** August 12th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-10583647/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

# Interpretable machine learning reveals biochar-specific hydraulic predictors of plant microbial fuel cell performance under unsaturated soil conditions

Sibin Gan<sup>1\*</sup>, Qinyue Yao<sup>2</sup>

<sup>1</sup> School of Civil and Transportation Engineering, Guangdong University of Technology, Guangzhou, Guangdong, China

<sup>2</sup> Department of Civil and Intelligent Construction Engineering, Shantou University, Shantou, Guangdong, China

\* Corresponding author: Sibin Gan (gansibin@mails.gdut.edu.cn)

## Abstract

Plant microbial fuel cells (PMFCs) operating in unsaturated soil are affected by dynamic interactions among water status, environmental conditions and biochar amendment, yet the relative predictive contributions of these factors remain unclear. This study analysed a time-resolved dataset generated from a controlled PMFC experiment comprising unamended control soil and soils amended with apple wood, corn straw or reed straw biochar. Relative humidity, temperature, volumetric water content, log-transformed matric suction and treatment indicators were used as predictors, while biochar properties and 16S rRNA sequencing data were retained for treatment-level interpretation. Seven regression approaches, comprising a mean baseline and six machine-learning models, were evaluated using within-treatment blocked five-fold cross-validation to limit temporal leakage, with all preprocessing and model selection conducted within the training folds. Among four electrical outputs, electrical potential and power density were the most predictable. Random forest achieved  $R^2$  values of 0.605 and 0.577 for these targets, respectively. Feature-set ablation showed that hydraulic–environmental variables and treatment indicators provided complementary predictive information. Permutation importance identified the reed straw biochar treatment and matric suction as the strongest treatment-level and continuous predictors, respectively. Microbial community patterns provided independent biological context for treatment-specific responses but were not used as row-level predictors. These findings show that treatment-aware, leakage-controlled machine learning can clarify biochar-specific hydraulic predictors and support material selection and operational assessment within comparable unsaturated-soil PMFC systems.

## Keywords

Plant microbial fuel cell; Biochar; Unsaturated soil; Matric suction; Interpretable machine learning; Bioelectrochemical engineering

## 1. Introduction

Plant microbial fuel cells (PMFCs) recover bioelectricity from living plant–soil systems through the coupling of photosynthesis, rhizodeposition, microbial metabolism and extracellular electron transfer [1–3]. Their low-input operation, in situ applicability and

ecological compatibility make them promising for constructed wetlands, environmental remediation, green infrastructure and self-powered environmental sensing [1–5]. However, soil-based PMFCs do not necessarily operate under continuously saturated conditions. Progressive drying and water redistribution alter pore-water connectivity, ion transport, oxygen diffusion, microbial activity and electrode–soil contact, resulting in substantial fluctuations in electrical output [6–8]. Understanding and predicting these responses is therefore important for process monitoring and operational assessment of soil-based bioelectrochemical systems under unsaturated conditions.

Biochar amendment can modify the hydraulic, electrochemical and biological environment of PMFCs [9]. Its pore structure, specific surface area, surface functional groups, cation-exchange capacity and water-retention characteristics may influence moisture distribution, ion migration, soil resistance, electrode contact and microbial habitat [10–13]. These effects depend strongly on biochar feedstock and production conditions [10–12]. Greater porosity or surface area may improve water retention and microbial attachment, but does not necessarily increase electrical output when ion transport or interfacial electron transfer remains limiting. Plant identity and microbial community composition further contribute to PMFC performance, although enrichment of individual taxa does not necessarily correspond to increased bioelectricity generation [14]. The relevant question is therefore not simply whether biochar improves PMFC output, but how biochar-specific treatment backgrounds modify the relationship between hydraulic state and electrical performance.

Conventional statistical analyses can identify treatment differences and pairwise associations, but may not adequately represent the nonlinear and interdependent structure of unsaturated

PMFC data. Electrical output may respond jointly to volumetric water content, matric suction, temperature, relative humidity and biochar treatment, while several predictors can contain overlapping information. Machine-learning methods provide a means of modelling such nonlinear relationships and quantifying the relative predictive value of different variables, and have increasingly been applied to microbial fuel cells, biochar properties and environmental remediation processes [15 – 18]. Nevertheless, repeated observations collected during continuous drying are temporally dependent. Randomly distributing adjacent observations between training and test sets can therefore inflate performance estimates through temporal leakage [19]. Mathematically related electrical variables may also introduce target leakage [20], whereas treatment-level microbial measurements cannot validly be replicated across individual monitoring times. Reliable analysis consequently requires validation strategies that preserve temporal structure, separate prediction targets and respect the hierarchy of the experimental data.

Accordingly, this study applied a leakage-controlled and interpretable machine-learning framework to time-resolved data generated from a controlled PMFC experiment comprising unamended soil and soils amended with apple wood, corn straw or reed straw biochar. Seven regression algorithms were evaluated using within-treatment blocked five-fold cross-validation. The study aimed to: (1) compare the predictability of electrical potential, power density, current and resistance during progressive soil drying; (2) quantify the complementary predictive contributions of hydraulic–environmental variables and biochar treatment indicators; and (3) identify the dominant predictors and assess whether treatment-level biochar properties and microbial community patterns provided plausible context for the model-derived treatment

differences. Because the 16S rRNA data were not temporally aligned with the electrical observations, they were retained exclusively for treatment-level interpretation. The resulting framework was intended to support process understanding, biochar selection and operational assessment within comparable unsaturated-soil PMFC systems rather than to establish a universally transferable prediction model.

## 2. Materials and Methods

### 2.1 PMFC system and biochar treatments

The soil used in the plant microbial fuel cell (PMFC) experiment was collected from Sangpu Mountain, Shantou University, Guangdong Province, China. It contained 35.41% sand and 64.59% silt and was classified as sandy lean clay (CL). The liquid limit, plastic limit, specific gravity and pH of the soil were 28.3%, 12.8%, 2.66 and 6.55, respectively. PMFC units were constructed in cylindrical pots measuring 100 mm in internal diameter and 150 mm in height and planted with *Hydrocotyle vulgaris*. A circular copper plate with a diameter of 50 mm was buried at a depth of approximately 80 mm and used as the anode. A copper rod measuring 5 mm in diameter and 50 mm in length was positioned near the soil surface and used as the cathode. Four treatments were established, as illustrated in **Figure 1**: unamended control soil (CG), apple wood biochar amendment (AWB), corn straw biochar amendment (CSB) and reed straw biochar amendment (RSB). Each biochar-amended unit contained 60 g of biochar and 1150 g of soil, corresponding to approximately 5% of the total dry solid mass, whereas each control unit contained 1210 g of soil without biochar. Five replicate PMFC units were prepared for each treatment.

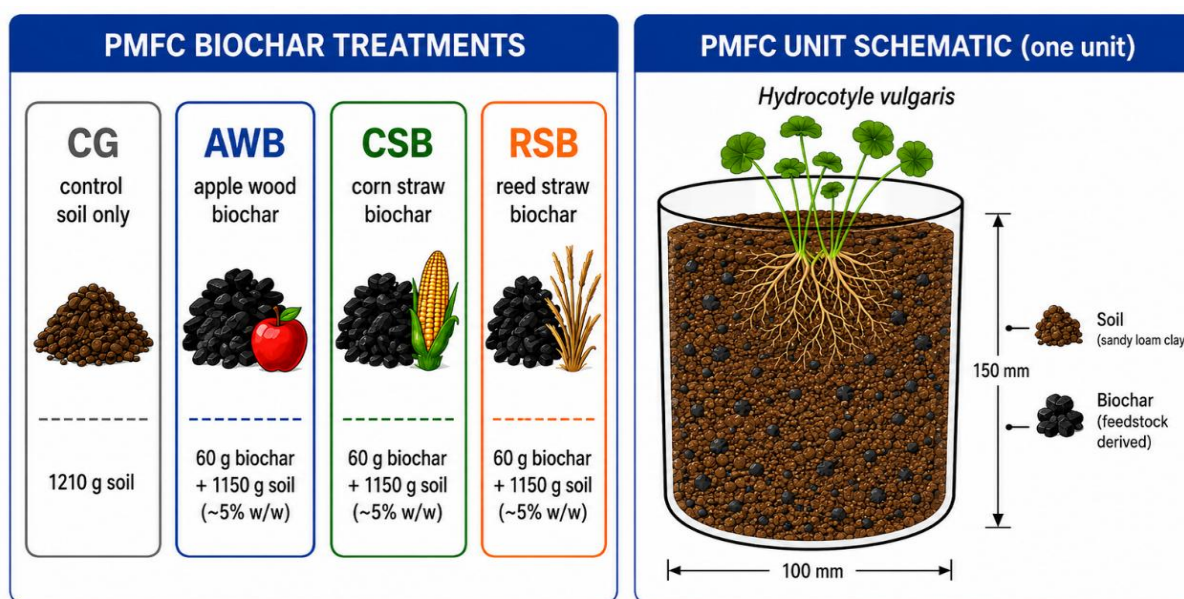


**Figure 1.** Biochar treatments and schematic configuration of the plant microbial fuel cell units. CG, unamended control soil; AWB, apple wood biochar; CSB, corn straw biochar; RSB, reed straw biochar.

The three biochars were produced under oxygen-limited conditions at 600 °C for 180 min. Their physicochemical properties were characterised using scanning electron microscopy, X-ray diffraction, Fourier-transform infrared spectroscopy, elemental analysis and Brunauer–Emmett–Teller surface-area analysis following previously reported procedures [7]. The specific surface areas of AWB, CSB and RSB were 8.001, 22.407 and 29.198 m<sup>2</sup> g<sup>-1</sup>, respectively, and their corresponding cation-exchange capacities were 17.6, 14.7 and 20.3 cmol kg<sup>-1</sup>.

## 2.2 Drying protocol and electro-hydraulic monitoring

Before the drying experiment, all plant microbial fuel cell (PMFC) units were cultivated for four weeks under sufficient irrigation and illumination. Electrical potential, current and volumetric water content were monitored until electrical output became relatively stable. The PMFC units were subsequently saturated, after which irrigation was discontinued and the soil was allowed to dry progressively through evaporation and plant water uptake. The experiment was conducted at a room temperature of 21 ± 3 °C, a relative humidity of 53 ± 5% and a mean

solar radiation of  $13 \pm 2 \text{ MJ m}^{-2} \text{ d}^{-1}$ . The experimental sequence and monitoring configuration are summarised in **Figure 2**.

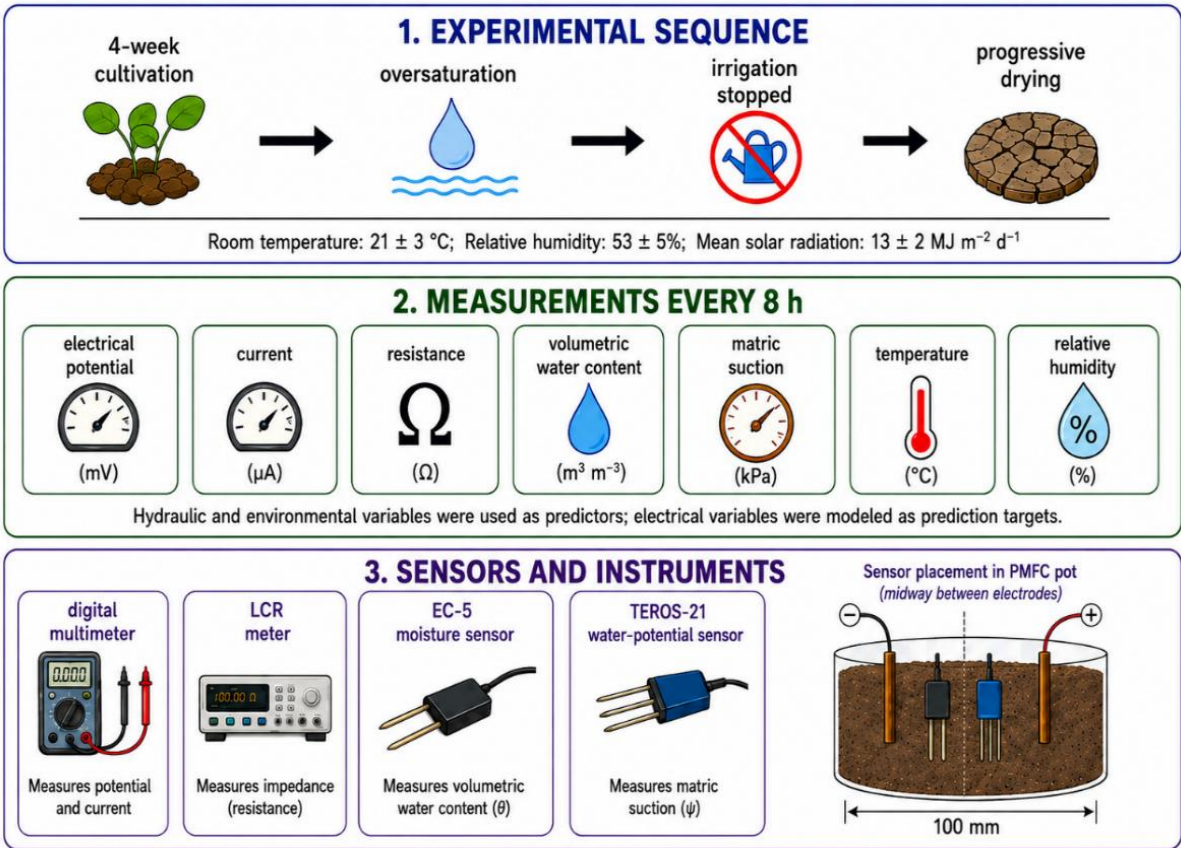


**Figure 2.** Experimental sequence, monitored variables and sensor configuration during progressive drying of the plant microbial fuel cell units. After four weeks of cultivation, the units were saturated, irrigation was discontinued and the soil was allowed to dry progressively. Electrical, hydraulic and environmental variables were recorded at 8 h intervals, with the volumetric-water-content and matric-suction sensors positioned between the anode and cathode.

During progressive drying, electrical, hydraulic and environmental variables were recorded at 8 h intervals. Electrical potential and current were measured using a digital multimeter with resolutions of 0.1 mV and 0.1  $\mu\text{A}$ , respectively. Inter-electrode resistance was measured using a TH2810B<sup>+</sup> LCR meter. Volumetric water content and matric suction were monitored using EC-5 and TEROS-21 sensors, respectively, with both sensors positioned in the soil between the anode and cathode. Ambient temperature and relative humidity were recorded concurrently.

Current density and power density were calculated as:

$$J = I / A \quad (1)$$

$$P = I \times U / A \quad (2)$$

where  $J$  is the current density,  $P$  is the power density,  $I$  is the current,  $U$  is the electrical potential and  $A$  is the electrode area used for normalisation, reported as  $1.96 \times 10^{-5} \text{ m}^2$ .

### 2.3 Microbial community characterisation

After completion of the progressive drying experiment, three root-zone soil samples were collected from each treatment, yielding 12 samples in total. Genomic DNA was extracted using the MagPure Soil DNA KF Kit. The V3–V4 region of the bacterial 16S rRNA gene was amplified using the primer pair 343F/798R, and the purified amplicons were sequenced on an Illumina NovaSeq 6000 platform to generate 250-bp paired-end reads [21]. Raw sequences were trimmed using Cutadapt [22], denoised and resolved using DADA2 [23], and processed within QIIME 2 [24]. Taxonomic assignments were performed against the SILVA 138 database [25]. Bacterial alpha diversity was evaluated using the Chao1 richness estimator [26] and Shannon diversity index [27], and genus-level relative abundances were calculated for subsequent treatment comparisons. Because microbial samples were collected only after drying, they were used for treatment-level interpretation rather than time-resolved prediction.

### 2.4 Dataset construction, targets and predictors

**Table 1** summarises the four experimental treatments and the corresponding time-series and microbial data available for analysis. Three levels of information were distinguished in the analytical framework shown in **Figure 3**. Time-resolved electrical, hydraulic and environmental measurements were used for supervised learning. Biochar physicochemical properties and endpoint 16S rRNA sequencing results were retained separately for treatment-level material and microbial interpretation.

**Table 1.** Experimental treatments and data available for modelling and treatment-level interpretation.

| Treatment | Biochar    | Biochar dose | Time-series data | 16S samples | Use in this study                                 |
|-----------|------------|--------------|------------------|-------------|---------------------------------------------------|
| CG        | None       | 0%           | Available        | 3           | Reference treatment                               |
| AWB       | Apple wood | 5%           | Available        | 3           | Treatment indicator and contextual interpretation |
| CSB       | Corn straw | 5%           | Available        | 3           | Treatment indicator and contextual interpretation |
| RSB       | Reed straw | 5%           | Available        | 3           | Treatment indicator and contextual interpretation |

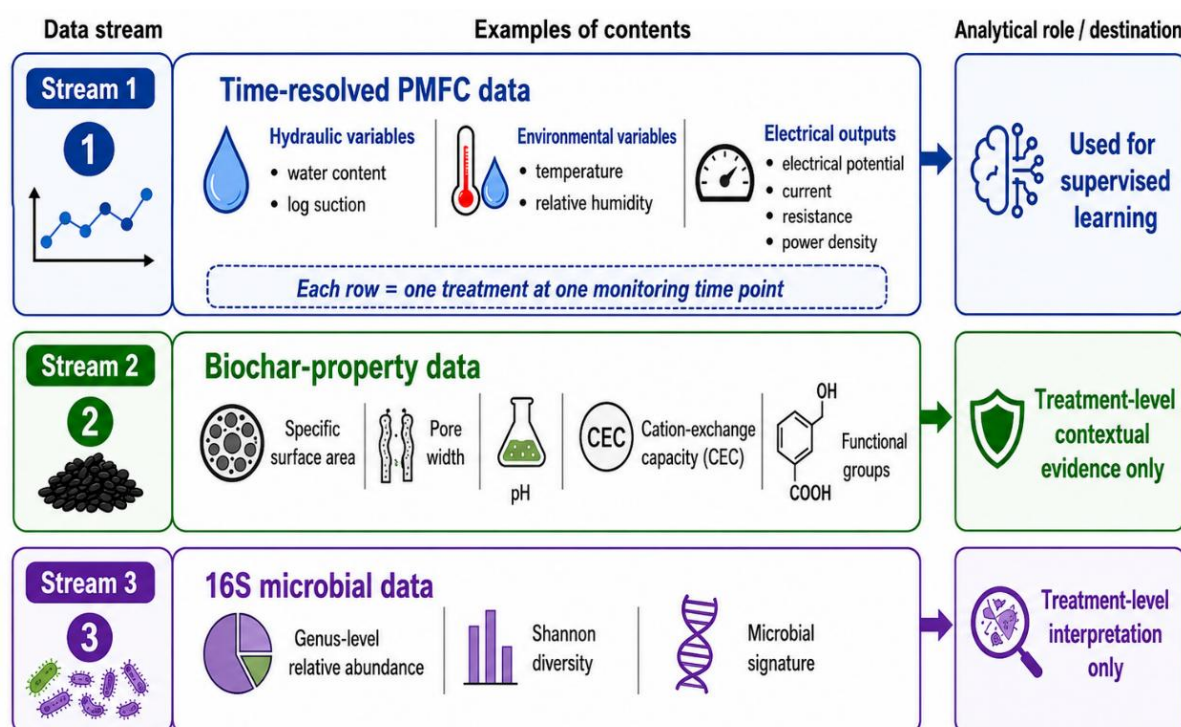


**Figure 3.** Analytical separation of the three data levels used in this study. Time-resolved electrical, hydraulic and environmental observations were used for supervised learning, whereas biochar physicochemical properties and endpoint 16S rRNA profiles were used for treatment-level material and microbial interpretation.

Each record in the modelling dataset corresponded to one treatment at one monitoring time point. Because the records were collected repeatedly during progressive drying, they were

166 treated as temporally ordered observations within each treatment rather than as independent  
167 biological replicates. Calendar dates, elapsed time, row identifiers and source-file information  
168 were excluded from the predictor matrix. Four electrical response variables were modelled  
169 separately: electrical potential, power density, current and resistance. After observations with  
170 missing response values were removed, the effective sample sizes were 345 for electrical  
171 potential, 297 for power density, and 337 each for current and resistance. Electrical potential  
172 and power density were treated as the primary responses because they directly represented  
173 PMFC electrical performance. Current and resistance were retained as auxiliary responses.  
174 Current density was not modelled separately because it was obtained by normalising current to  
175 a constant electrode area. The continuous predictors were relative humidity, temperature,  
176 volumetric water content and matric suction. Matric suction was transformed using the natural  
177 logarithm to reduce the strong right-skewness associated with progressive soil drying.  
178 Treatment identity was represented by three binary variables corresponding to apple wood  
179 biochar (AWB), corn straw biochar (CSB) and reed straw biochar (RSB), with the unamended  
180 control treatment (CG) used as the reference category. To prevent target leakage, no measured  
181 or derived electrical variable was used to predict another electrical response. Model outputs,  
182 residuals, response-derived statistics and feature-importance values were likewise excluded  
183 from the predictors. Biochar properties and microbial-community variables were not merged  
184 with the time-resolved records because they were available only at the treatment level. They  
185 were used only after model fitting to interpret treatment-related patterns, thereby avoiding  
186 artificial replication of treatment-level measurements across the 8 h monitoring records.

## 2.5 Machine-learning models and preprocessing

Seven regression approaches were evaluated, comprising a mean baseline and six machine-learning models: ridge regression [28], second-order polynomial ridge regression, K-nearest neighbour regression [29], a regression tree, random forest [30] and a lightweight multilayer perceptron. The models were selected to represent progressively different forms of predictive structure. Ridge regression represented a regularised linear model, whereas polynomial ridge regression introduced second-order terms to capture low-order nonlinearities and interactions. K-nearest neighbour regression represented a local, distance-based method. The regression tree and random forest represented single-tree and ensemble tree-based approaches, respectively, while the multilayer perceptron provided a compact neural-network model. The mean baseline predicted the mean response of the corresponding training set and was used to determine whether the machine-learning models provided useful predictive improvement. Missing predictor values were replaced using the median estimated from the corresponding training data. Ridge regression, polynomial ridge regression, K-nearest neighbour regression and the multilayer perceptron were fitted using standardised input features. For polynomial ridge regression, second-order polynomial terms were generated within the modelling pipeline before model fitting. The regression tree and random forest were fitted to median-imputed data without feature standardisation because tree-based partitioning is not affected by differences in predictor scale. The multilayer perceptron contained one hidden layer with eight neurons, used the hyperbolic-tangent activation function and was trained for a maximum of 1200 iterations. Model-specific hyperparameters and their candidate values are summarised in **Table 2**.

209 **Table 2.** Machine-learning models, preprocessing procedures and hyperparameter search spaces.

| Model                          | Model role                                | Preprocessing                                                            | Hyperparameters and candidate values                                                                                                    |
|--------------------------------|-------------------------------------------|--------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------|
| Mean baseline                  | Reference prediction                      | None                                                                     | Training-set mean; no tuned hyperparameters                                                                                             |
| Ridge regression               | Regularised linear model                  | Median imputation and standardisation                                    | Regularisation strength, $\alpha$ : 0.1, 1, 10, 100, 1000                                                                               |
| Polynomial ridge regression    | Low-order nonlinear and interaction model | Median imputation, second-order polynomial expansion and standardisation | Polynomial degree: 2; regularisation strength, $\alpha$ : 0.1, 1, 10, 100, 1000                                                         |
| K-nearest neighbour regression | Local distance-based model                | Median imputation and standardisation                                    | Number of neighbours, k: 3, 5, 9, 15, 21; weights: uniform; distance metric: Euclidean                                                  |
| Regression tree                | Nonlinear single-tree model               | Median imputation; no standardisation                                    | Maximum depth: 3, 5, 7; minimum samples per leaf: 5, 8, 12; minimum samples for splitting: 2                                            |
| Random forest                  | Ensemble tree model                       | Median imputation; no standardisation                                    | Number of trees: 60, 80; maximum depth: 5, 6; minimum samples per leaf: 4, 6; number of features considered at each split: all features |
| Multilayer perceptron          | Neural-network model                      | Median imputation and standardisation                                    | Hidden-layer size: 8; activation: tanh; maximum iterations: 1200; regularisation strength: 0.0001; initial learning rate: 0.001         |

210 All preprocessing steps were fitted exclusively to the training data within each cross-validation  
211 fold. The fitted imputation, standardisation and polynomial-transformation parameters were  
212 subsequently applied to the corresponding validation or test data. No preprocessing parameter  
213 was estimated from the complete dataset before cross-validation. All analyses were performed  
214 in Python version 3.13.0 using scikit-learn version 1.7.2 [31]. NumPy version 2.3.2 and pandas  
215 version 2.3.1 were used for numerical computation and data organisation, respectively.  
216 Stochastic model fitting was controlled using a fixed random seed of 42.

## 217 **2.6 Blocked cross-validation and performance evaluation**

218 Model performance was evaluated using within-treatment blocked five-fold cross-validation.

For each prediction target, observations from the control, apple wood biochar, corn straw biochar and reed straw biochar treatments were first ordered chronologically. The time series of each treatment was then divided into five consecutive blocks while preserving the original observation order. In each outer iteration, the corresponding block from all four treatments was pooled to form the test set, and the remaining four blocks were used for model training. No random shuffling was performed. This procedure ensured that every test set contained observations from all treatments while reducing direct information sharing between adjacent drying records assigned to the training and test sets. Hyperparameter selection was conducted separately within each outer training set using three-fold inner cross-validation, with root mean square error (RMSE) as the optimisation criterion. After the optimal hyperparameter configuration had been identified, the model was refitted using the complete outer training set and evaluated on the held-out temporal block. All preprocessing operations described in Section 2.5 were included within this nested procedure and were estimated using training data only. The overall modelling and evaluation procedure is shown in **Figure 4**. For each candidate model, predictions from the five held-out outer blocks were combined to obtain one pooled out-of-fold prediction for every observation. Model comparison was based primarily on metrics calculated from these pooled predictions rather than on the mean of independently calculated fold-level coefficients.

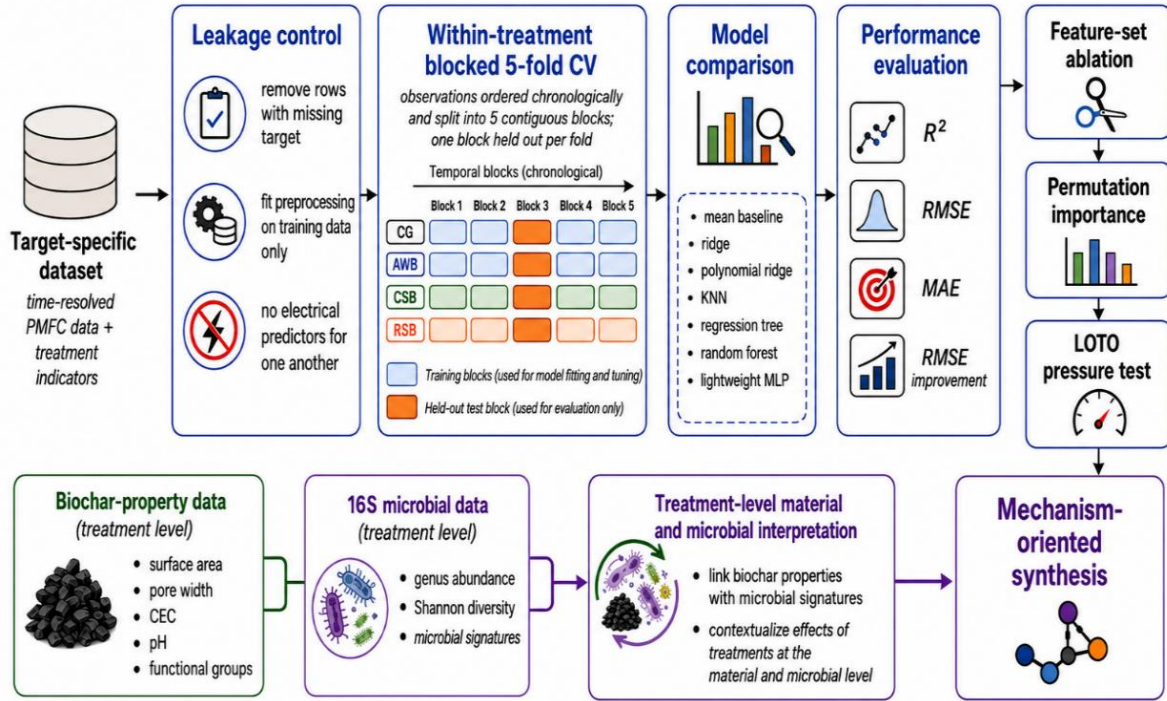


**Figure 4.** Leakage-controlled modelling and treatment-level interpretation strategy. Time-resolved plant microbial fuel cell records were evaluated using within-treatment blocked five-fold cross-validation. Preprocessing and hyperparameter selection were restricted to the training data within each outer fold. Biochar physicochemical properties and 16S rRNA sequencing data were used only for treatment-level material and microbial interpretation.

Predictive performance was quantified using the coefficient of determination ( $R^2$ ), RMSE and mean absolute error (MAE):

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (4)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (5)$$

where  $y_i$  and  $\hat{y}_i$  are the observed and predicted values, respectively,  $\bar{y}$  is the mean observed value and  $n$  is the number of pooled out-of-fold observations. Improvement relative to the mean baseline was calculated as:

$$RMSE\ improvement\ (\%) = \frac{RMSE_{baseline} - RMSE_{model}}{RMSE_{baseline}} \times 100 \quad (6)$$

Positive values therefore indicated lower prediction error than the mean baseline, whereas negative values indicated poorer performance. Subsequent model interpretation focused on electrical potential and power density because these targets showed the strongest overall performance under blocked cross-validation. These estimates represent prediction of held-out temporal blocks within the same experimental system rather than independent external validation.

## 2.7 Feature analysis and transferability assessment

Subsequent analyses focused on the random-forest models for electrical potential and power density, which showed the strongest overall performance under blocked cross-validation. Feature-set ablation was used to assess the relative and complementary predictive information provided by hydraulic–environmental variables and biochar treatment indicators. Five predictor configurations were evaluated: (1) hydraulic–environmental variables only; (2) treatment indicators only; (3) the combined predictor set; (4) the combined set excluding log-transformed matric suction; and (5) the combined set excluding volumetric water content. Each configuration was evaluated using the within-treatment blocked five-fold cross-validation procedure described in Section 2.6, with preprocessing and hyperparameter tuning conducted independently within the corresponding training data. Changes in pooled out-of-fold  $R^2$  and RMSE were used to compare the predictive information retained by each feature set. Predictor dependence was further examined using permutation importance. Within each outer cross-validation fold, one predictor was permuted in the held-out test block while all other predictors were left unchanged. Importance was quantified as the increase in test-set RMSE

relative to the unpermuted predictions. The mean RMSE increase across the five outer folds was used as the overall importance estimate, and the corresponding standard deviation was used to characterise fold-to-fold variability. Because correlated predictors may contain overlapping information, permutation importance can underestimate or distribute importance among related variables [32]. The results were therefore interpreted together with the ablation analysis as indicators of model dependence rather than as measures of independent effects or causality.

Transferability across treatment backgrounds was examined using leave-one-treatment-out analysis. In each iteration, all observations from one treatment were reserved for testing, while the remaining three treatments were used for model training. The combined predictor set and the hydraulic–environmental-only predictor set were evaluated separately. All data preprocessing and hyperparameter tuning were performed using the training treatments alone. The leave-one-treatment-out analysis was treated as a pressure test of model sensitivity to treatment background rather than as external validation for previously unseen biochars. In the combined predictor set, the binary indicator corresponding to the held-out treatment had no positive examples in the training data; consequently, the model could not learn the treatment-specific effect of that category. The analysis therefore assessed the extent to which prediction depended on treatment information already represented during training, rather than the ability to predict the performance of arbitrary new biochar materials.

Biochar physicochemical properties and microbial-community measurements were used only for treatment-level interpretation after model fitting. A selected-genus signature was calculated as the summed relative abundance of seven genera previously identified as differentially abundant using linear discriminant analysis effect size in the source study [21,33]: *Aeromonas*,

*Pseudomonas*, *Bacillus*, *Comamonas*, *Cellvibrio*, *Dechloromonas* and *Hydrogenophaga*. The selected-genus signature, alpha-diversity indices, biochar properties and treatment-level electrical summaries were excluded from model training, feature-set ablation and permutation-importance calculations. These variables were used solely to contextualise treatment-related predictive patterns and were not interpreted as direct evidence of causal effects attributable to individual microbial taxa or biochar properties.

### 3. Results

#### 3.1 Comparative performance across electrical targets

**Figure 5** compares the predictive performance of the seven regression approaches across the four electrical targets, while **Table 3** summarises the best-performing model and associated evaluation metrics for each target. Electrical potential and power density exhibited markedly greater predictability than current and resistance, as evidenced by their higher pooled out-of-fold  $R^2$  values and larger reductions in root mean square error (RMSE) relative to the mean baseline.

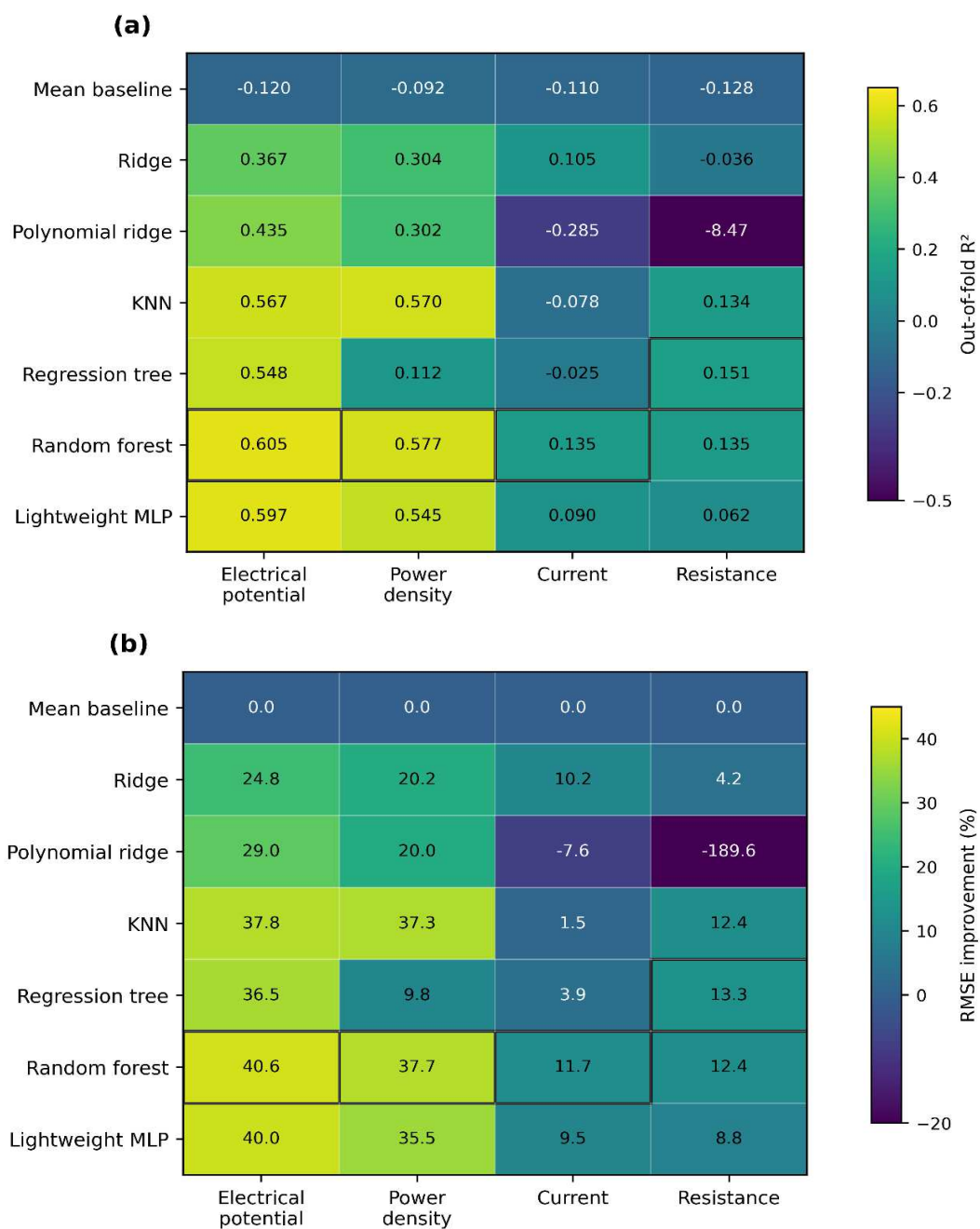


**Figure 5.** Comparison of seven regression approaches for predicting plant microbial fuel cell electrical outputs under within-treatment blocked five-fold cross-validation. (a) Pooled out-of-fold  $R^2$ . (b) Reduction in RMSE relative to the mean baseline. Cell labels show the original values, and black outlines identify the best-performing approach for each target. Colour scales were truncated at  $-0.5$  in panel (a) and  $-20\%$  in panel (b) to improve visual comparability.

**Table 3.** Best-performing models for the four plant microbial fuel cell electrical targets under within-treatment blocked five-fold cross-validation.

| Prediction target    | Effective n | Best model      | $R^2$ | RMSE                       | MAE                        | RMSE reduction relative to baseline (%) |
|----------------------|-------------|-----------------|-------|----------------------------|----------------------------|-----------------------------------------|
| Electrical potential | 345         | Random forest   | 0.605 | 74.663 mV                  | 58.223 mV                  | 40.6                                    |
| Power density        | 297         | Random forest   | 0.577 | 278.681 mW·m <sup>-2</sup> | 171.719 mW·m <sup>-2</sup> | 37.7                                    |
| Current              | 337         | Random forest   | 0.135 | 18.443 µA                  | 10.991 µA                  | 11.7                                    |
| Resistance           | 337         | Regression tree | 0.151 | 7675.812 Ω                 | 2148.978 Ω                 | 13.3                                    |

*Note:* Performance metrics were calculated from pooled out-of-fold predictions. The best-performing model for each target was identified primarily according to the highest  $R^2$  and lowest RMSE. RMSE and MAE are reported in the original units of each target.

Among the candidate approaches, random forest yielded the strongest overall performance for electrical potential, attaining an  $R^2$  of 0.605, an RMSE of 74.663 mV and a mean absolute error (MAE) of 58.223 mV. This corresponded to a 40.6% reduction in RMSE relative to the mean baseline. The lightweight multilayer perceptron and K-nearest neighbour regression produced slightly lower  $R^2$  values of 0.597 and 0.567, respectively. Random forest also achieved the highest  $R^2$  and the lowest RMSE for power-density prediction, with an  $R^2$  of 0.577, an RMSE of 278.681 mW·m<sup>-2</sup> and an MAE of 171.719 mW·m<sup>-2</sup>. The corresponding RMSE reduction relative to the mean baseline was 37.7%. K-nearest neighbour regression showed comparable performance, with an  $R^2$  of 0.570 and a slightly lower MAE; however, random forest was retained as the best-performing model because it provided the more favourable combination of  $R^2$  and RMSE. In contrast, current and resistance displayed limited predictability. Random forest produced the best current predictions, with an  $R^2$  of 0.135, an RMSE of 18.443 µA and an MAE of 10.991 µA. Resistance was best predicted by the regression tree, which yielded an  $R^2$  of 0.151, an RMSE of 7675.812 Ω and an MAE of 2148.978 Ω. The corresponding RMSE

reductions relative to the mean baseline were 11.7% for current and 13.3% for resistance. Given their consistently higher predictive performance, electrical potential and power density were selected as the primary targets for subsequent out-of-fold diagnostic analysis, feature-set ablation and permutation-importance assessment. Random forest was used for these analyses because it achieved the strongest overall performance for both targets.

### **3.2 Out-of-fold prediction characteristics**

**Figure 6** presents the pooled out-of-fold predictions generated by the random-forest models for electrical potential and power density. Predicted electrical potentials showed comparatively close agreement with the corresponding observations, with most data distributed around the 1:1 line. With residuals defined as the observed value minus the predicted value, the mean residual was  $-1.977$  mV, indicating little overall directional bias. Power-density predictions exhibited greater dispersion around the 1:1 line and yielded a mean residual of  $-34.045$  mW·m<sup>-2</sup>.

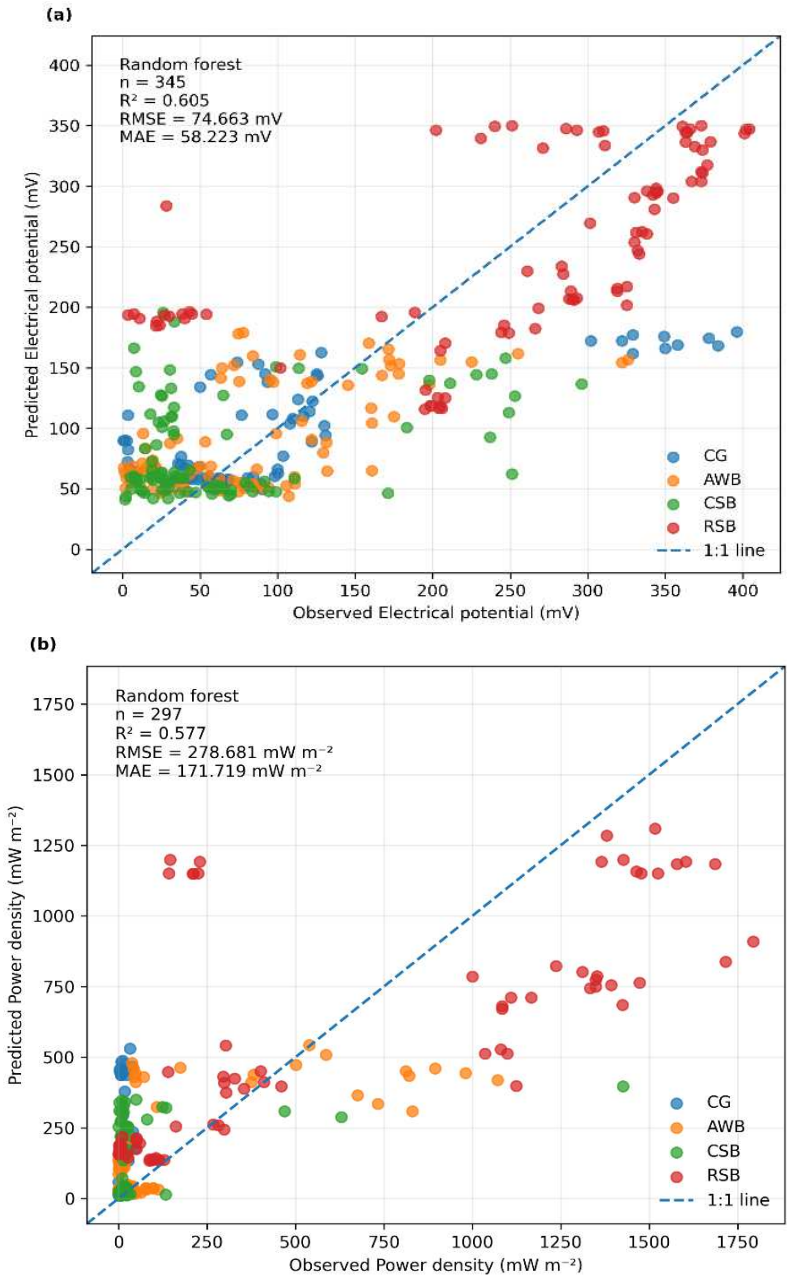


**Figure 6.** Pooled out-of-fold observed and predicted values generated by the random-forest models under within-treatment blocked five-fold cross-validation. (a) Electrical potential and (b) power density. Points represent individual time-resolved observations and are distinguished by treatment; the dashed line denotes perfect agreement between observed and predicted values.

Prediction characteristics varied among the four treatments. The random-forest models reproduced the generally higher electrical outputs observed for the reed straw biochar treatment, but the predicted values were compressed towards the central response range. The highest electrical-potential observations were underestimated, and this compression was more

pronounced for power density. Peak power-density values in the reed straw biochar treatment tended to be underestimated, whereas low values in the control and corn straw biochar treatments tended to be overestimated. Prediction errors were therefore more evident at the upper and lower ends of the response distributions than within the intermediate range. This pattern was particularly pronounced for power density, for which the highest-output reed straw biochar observations showed the largest departures from the 1:1 relationship.

### **3.3 Contributions of hydraulic–environmental and treatment predictors**

**Figure 7** compares the performance of the random-forest models across the five predictor configurations. For electrical potential, the combined predictor set yielded the highest pooled out-of-fold  $R^2$  of 0.605. The hydraulic–environmental-only and treatment-only configurations achieved  $R^2$  values of 0.279 and 0.214, respectively, with RMSE values 35.0% and 41.0% higher than that of the combined predictor set. Excluding log-transformed matric suction reduced  $R^2$  to 0.546 and increased RMSE by 7.1%, whereas excluding volumetric water content reduced  $R^2$  to 0.542 and increased RMSE by 7.7%.

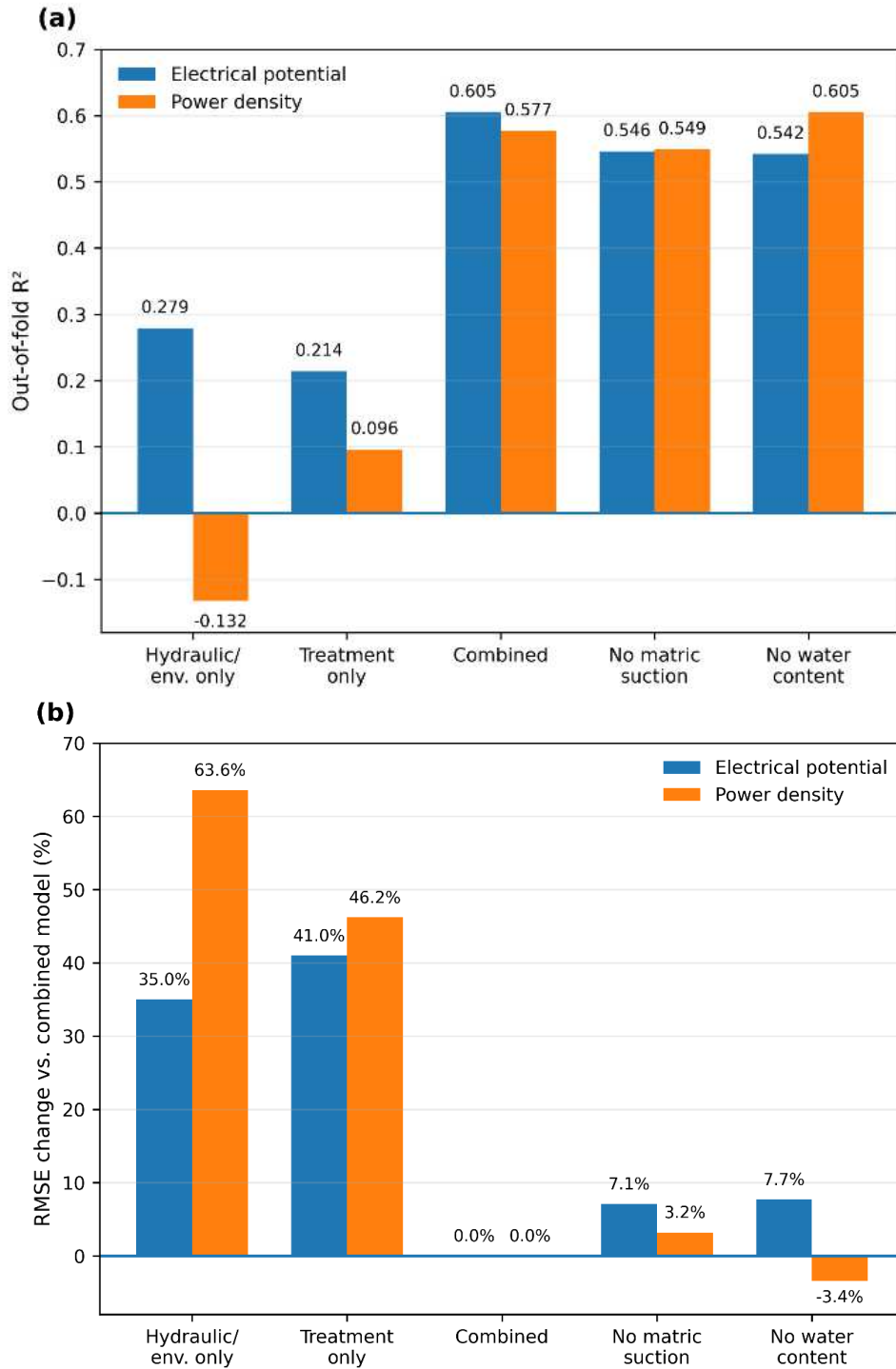


**Figure 7.** Feature-set ablation results for random-forest prediction of electrical potential and power density under within-treatment blocked five-fold cross-validation. (a) Pooled out-of-fold  $R^2$  values for the five predictor configurations. (b) Percentage change in RMSE relative to the complete combined predictor set. Positive values denote higher prediction error, whereas negative values denote lower prediction error.

A similar contrast among feature configurations was observed for power density. The combined

predictor set achieved a pooled out-of-fold  $R^2$  of 0.577. In comparison, the hydraulic–environmental-only configuration produced an  $R^2$  of  $-0.132$ , while the treatment-only configuration yielded an  $R^2$  of 0.096. Their RMSE values were 63.6% and 46.2% higher, respectively, than that of the combined predictor set. Excluding log-transformed matric suction reduced  $R^2$  to 0.549 and increased RMSE by 3.2%. By contrast, excluding volumetric water content increased  $R^2$  to 0.605 and reduced RMSE by 3.4% relative to the complete combined predictor set. cross both prediction targets, configurations containing hydraulic–environmental variables together with treatment indicators consistently outperformed configurations based on either predictor group alone. The effects of excluding individual water-status variables differed between electrical potential and power density.

### 3.4 Permutation importance of individual predictors

**Figure 8** summarises the permutation importance of the individual predictors for the random-forest models of electrical potential and power density. For electrical potential, the reed straw biochar indicator showed the highest importance, with its permutation increasing the held-out RMSE by  $34.30 \pm 19.22$  mV. The corn straw biochar indicator ranked second, followed closely by log-transformed matric suction, for which permutation increased RMSE by  $7.16 \pm 10.32$  mV. Volumetric water content showed a smaller contribution, while temperature, relative humidity and the apple wood biochar indicator occupied the lower positions in the importance ranking.

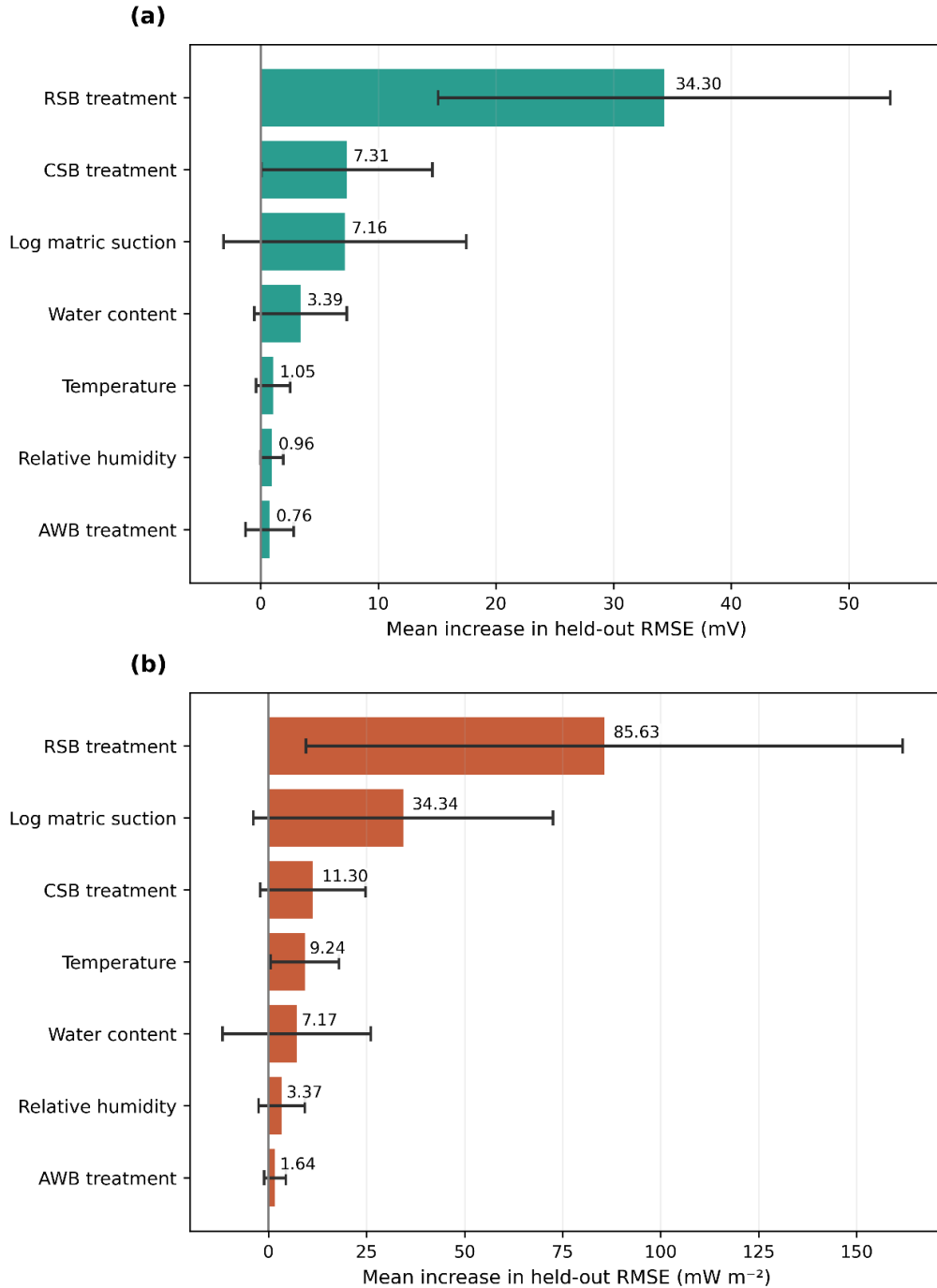


**Figure 8.** Permutation importance of individual predictors in the random-forest models for (a) electrical potential and (b) power density. Importance was quantified as the increase in held-out root mean square error after permutation of each predictor. Bars represent the mean increase across the five outer folds, and error bars denote the corresponding standard deviation.

The reed straw biochar indicator was also the highest-ranked predictor for power density, producing an RMSE increase of  $85.63 \pm 76.08 \text{ mW} \cdot \text{m}^{-2}$  after permutation. Log-transformed

matric suction ranked second, with an RMSE increase of  $34.34 \pm 38.20 \text{ mW} \cdot \text{m}^{-2}$  followed by the corn straw biochar indicator, temperature and volumetric water content. Relative humidity and the apple wood biochar indicator showed comparatively low permutation importance. Thus, the reed straw biochar indicator ranked first for both electrical targets, while log-transformed matric suction was the highest-ranked continuous predictor. The relatively large standard deviations, particularly for power density, indicated substantial fold-to-fold variability in the estimated importance values.

### **3.5 Treatment-level biochar and microbial patterns**

**Figure 9** summarises the treatment-level selected-genus signature, Shannon diversity and maximum power density. The selected-genus signature differed among the biochar treatments, with the highest mean value recorded for apple wood biochar at  $54.07 \pm 7.76\%$ , followed by corn straw biochar at  $37.41 \pm 15.98\%$ . Reed straw biochar exhibited the lowest value among the three biochar treatments, at  $12.39 \pm 5.67\%$ .

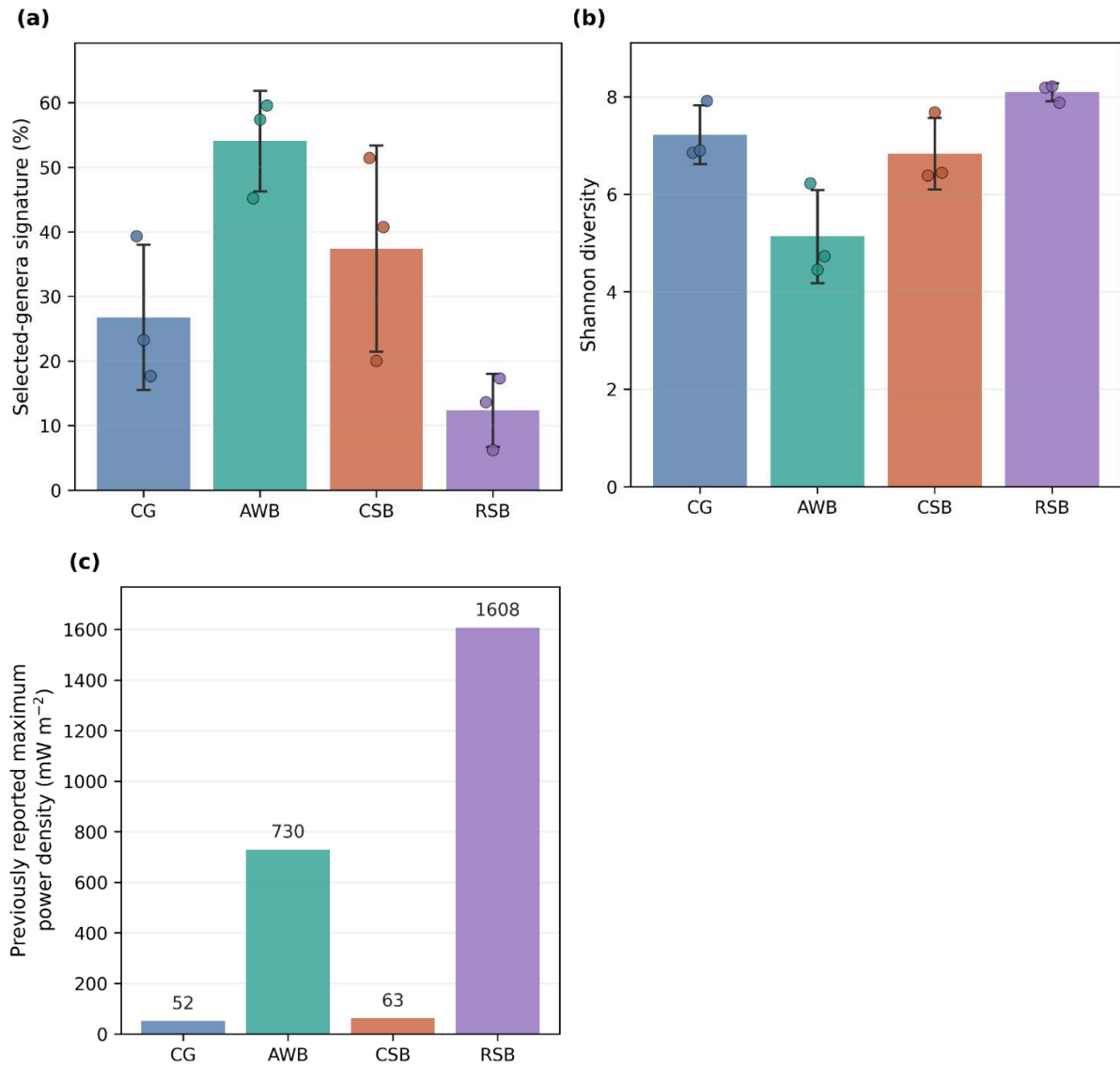


**Figure 9.** Treatment-level microbial signatures, diversity and previously reported PMFC output across biochar treatments. (a) Selected-genus signature calculated from seven LEfSe-supported genera. (b) Shannon diversity. Bars show mean  $\pm$  SD with individual samples. (c) Previously reported maximum power density, included only for treatment-level interpretation.

Shannon diversity also varied among treatments. Reed straw biochar showed the highest mean Shannon index, reaching  $8.10 \pm 0.19$ . The treatment ranking based on Shannon diversity therefore differed from that based on the selected-genus signature. For treatment-level comparison, the previously reported maximum power densities were  $1608 \text{ mW} \cdot \text{m}^{-2}$  for reed straw biochar,  $730 \text{ mW} \cdot \text{m}^{-2}$  for apple wood biochar,  $63 \text{ mW} \cdot \text{m}^{-2}$  for corn straw biochar and  $52 \text{ mW} \cdot \text{m}^{-2}$  for the unamended control. Accordingly, maximum power density followed the order

RSB >AWB >CSB  $\approx$ CG, whereas the selected-genus signature was highest in AWB and lowest in RSB.

### **3.6 Cross-treatment transferability and temporal robustness**

Leave-one-treatment-out evaluation resulted in substantially weaker predictive performance than within-treatment blocked cross-validation, as shown in **Figure 10** and **Table 4**. For electrical potential, the best-performing non-baseline model was K-nearest neighbour regression using the combined predictor set, which achieved a pooled  $R^2$  of 0.186, an RMSE of 107.191 mV and an MAE of 76.908 mV. This corresponded to a 21.9% reduction in RMSE relative to the leave-one-treatment-out mean baseline. When only hydraulic–environmental predictors were used, ridge regression yielded a negative  $R^2$  of  $-0.167$  and an RMSE of 128.299 mV.

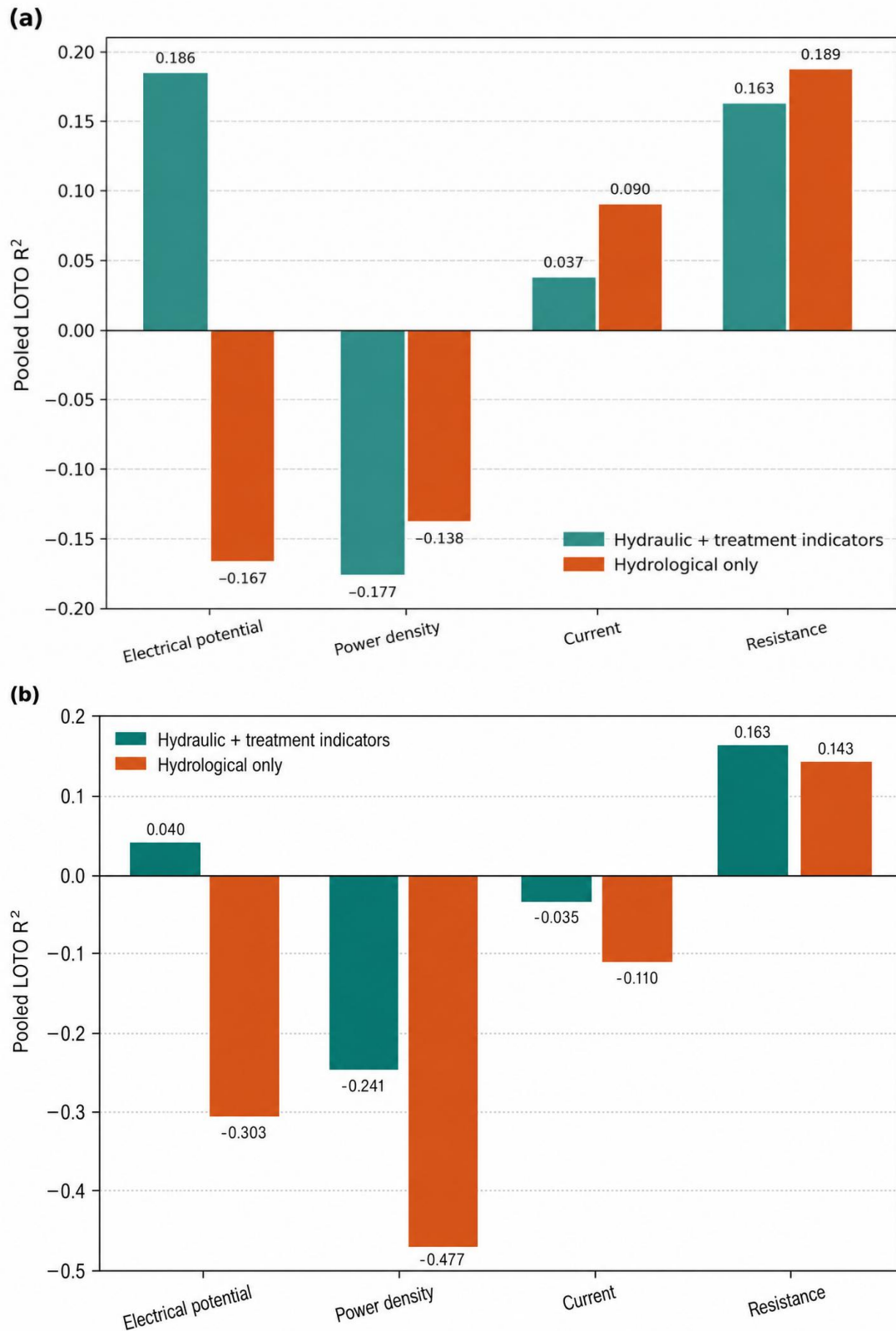


**Figure 10.** Leave-one-treatment-out (LOTO) assessment of cross-treatment predictive performance for four plant microbial fuel cell electrical targets. (a) Pooled LOTO  $R^2$  values of the best-performing non-baseline model under each predictor configuration. (b) Pooled LOTO  $R^2$  values obtained using random-forest models. Predictor configurations comprised hydraulic–environmental variables combined with

treatment indicators and hydraulic–environmental variables alone. Negative  $R^2$  values indicate poorer predictive performance than the corresponding LOTO mean-baseline model.

**Table 4.** Leave-one-treatment-out performance for the primary prediction targets.

| Prediction target    | Predictor configuration                      | Selected model    | Pooled ( $R^2$ ) | RMSE                       | RMSE improvement (%) |
|----------------------|----------------------------------------------|-------------------|------------------|----------------------------|----------------------|
| Electrical potential | Hydraulic–environmental treatment indicators | + KNN             | 0.186            | 107.191 mV                 | 21.9                 |
| Electrical potential | Hydraulic–environmental only                 | Ridge             | −0.167           | 128.299 mV                 | 6.5                  |
| Power density        | Hydraulic–environmental treatment indicators | + Regression tree | −0.177           | 464.894 mW·m <sup>−2</sup> | 3.1                  |
| Power density        | Hydraulic–environmental only                 | Ridge             | −0.138           | 457.088 mW·m <sup>−2</sup> | 4.7                  |

**Note.** Metrics were calculated from pooled leave-one-treatment-out predictions. RMSE improvement was calculated relative to the corresponding mean-baseline model. Negative  $R^2$  values indicate poorer performance than the mean baseline.

Cross-treatment prediction of power density was weaker. The regression tree selected for the combined predictor set produced a pooled  $R^2$  of −0.177, an RMSE of 464.894 mW·m<sup>−2</sup> and an MAE of 235.585 mW·m<sup>−2</sup>. The hydraulic–environmental-only ridge model also produced a negative  $R^2$  of −0.138, with an RMSE of 457.088 mW·m<sup>−2</sup>. Thus, although both models slightly reduced RMSE relative to the corresponding mean baseline, neither explained the between-treatment variation in power density.

## 4. Discussion

### 4.1 Target-dependent predictability and model behaviour

Electrical potential and power density were more predictable than current and resistance, indicating that the four electrical outputs differed in their sensitivity to the available hydraulic and treatment-related predictors. Electrical potential followed a comparatively smooth drying response and retained clear treatment contrasts, whereas current was more susceptible to short-term fluctuations at low signal levels. Resistance may also respond strongly to local variations in electrode–soil contact and conductive pathways that were not represented by the predictor

set. Although power density incorporates variability from both electrical potential and current, its pronounced treatment gradient preserved sufficient structure for prediction.

The generally stronger performance of nonlinear models suggests that PMFC responses were not adequately represented by simple additive relationships. Progressive drying alters pore-water continuity, ionic transport, oxygen availability, microbial activity and electrode contact simultaneously [6–8], producing nonlinear responses and interactions with biochar treatment. Random forest, K-nearest neighbour regression and the lightweight multilayer perceptron were better able to accommodate these relationships than regularised linear models, although their performance does not imply that they reproduced the underlying physical processes completely. Random-forest predictions were compressed towards the central response range, particularly for the highest electrical-potential and power-density observations. This behaviour is consistent with ensemble averaging and the limited representation of extreme high-output observations in the training data. Consequently, peak values were underestimated, while some low-output observations were overestimated.

The  $R^2$  values of 0.605 for electrical potential and 0.577 for power density therefore indicate moderate, rather than high, predictive performance. The reductions in RMSE relative to the mean baseline confirm that the models captured systematic variation beyond average response levels, but substantial point-level variability remained unexplained. The models are consequently more appropriate for identifying dominant response patterns and comparing predictor contributions than for precise reconstruction of individual or extreme PMFC outputs.

## **4.2 Complementary hydraulic and treatment information**

Feature-set ablation showed that hydraulic–environmental variables and treatment indicators

provided complementary predictive information. Hydraulic–environmental variables represented the evolving state of the PMFCs during drying, whereas treatment indicators captured persistent differences among the control and biochar-amended systems. Neither predictor group alone adequately represented both sources of variation, particularly for power density, for which the combined model produced a marked performance improvement. This result suggests that the relationship between hydraulic state and electrical output differed among treatments rather than following a single response function across all PMFCs.

The limited performance loss caused by removing either volumetric water content or matric suction indicates partial redundancy between these water-status descriptors. Volumetric water content measures the amount of water stored in the soil, whereas matric suction reflects its energetic state; nevertheless, the two are linked through the soil-water retention relationship and covary during drying [6,7,12]. Their predictive information may also overlap with temperature, relative humidity and treatment identity. The slight improvement in power-density prediction after removal of volumetric water content therefore indicates limited incremental information beyond that already provided by matric suction and treatment background, rather than physical irrelevance of soil water content.

These ablation results quantify predictive complementarity, not causal contributions. Treatment indicators represent composite differences in biochar properties, hydraulic behaviour, electrode–soil contact and microbial habitat. The analysis therefore cannot attribute their predictive value to any single material characteristic. Directly measured material, hydraulic and interfacial descriptors will be required to separate these effects and improve prediction for previously untested biochars.

### **4.3 Matric suction and biochar-specific treatment effects**

Among the continuous predictors, log-transformed matric suction showed the strongest model dependence for both electrical potential and power density. Unlike volumetric water content, which quantifies soil water storage, matric suction reflects the energetic state of pore water. Increasing suction during drying is associated with drainage of larger pores, thinning of aqueous films and greater tortuosity of ion-transport pathways, which may reduce ionic connectivity and weaken electrode–soil contact. Matric suction therefore appears to integrate several hydraulic constraints relevant to PMFC performance under unsaturated conditions.

Its predictive importance should not, however, be interpreted as evidence of an independent causal effect. Matric suction covaried with water content, drying stage, plant water uptake and treatment-specific retention behaviour, while permutation importance may be redistributed among correlated predictors. Its high ranking therefore indicates that the models relied strongly on the information it carried rather than that suction alone controlled electrical output.

The reed straw biochar indicator ranked first for both primary targets, showing that the RSB treatment contained predictive information not captured by the measured hydraulic and environmental variables. RSB also had the highest specific surface area and cation-exchange capacity and produced the greatest treatment-level electrical output. These properties may have favoured water retention, ion exchange and maintenance of conductive pathways during drying. Nevertheless, the treatment indicator represented the entire RSB experimental background rather than any single material or microbial attribute. Its importance therefore supports a biochar-specific treatment effect but does not identify the responsible property. Future models should replace categorical treatment labels with directly measured descriptors, including pore

structure, electrical conductivity, surface chemistry, soil-water retention parameters and electrode-interface characteristics.

#### **4.4 Microbial patterns as treatment-level contextual evidence**

Treatment-level microbial profiles were not directly aligned with electrical performance. Apple wood biochar exhibited the highest selected-genus signature, yet its maximum power density remained below that of reed straw biochar. Corn straw biochar also showed enrichment of several selected genera, including *Pseudomonas*, while its electrical output remained close to that of the unamended control. Enrichment of the selected taxa alone was therefore insufficient to explain treatment differences in PMFC performance.

Reed straw biochar showed the lowest selected-genus signature but the highest Shannon diversity and maximum power density. This contrast suggests that broader community organisation may have been more relevant to electrical performance than the abundance of a limited set of genera. Greater diversity could support functional complementarity or redundancy under changing soil conditions [34], although this interpretation remains tentative because functional genes, extracellular electron-transfer activity and taxon-specific functions were not measured.

These patterns are consistent with the view that PMFC output emerges from interactions among microbial communities, plant processes and bioelectrochemical conditions rather than from individual taxa alone [14]. The high output associated with reed straw biochar coincided with greater microbial diversity and distinct material and hydraulic characteristics; therefore, the microbial data cannot be separated from other treatment-specific effects.

The microbial evidence was also limited by its temporal resolution. The 16S rRNA profiles

were obtained only after drying, with three biological samples per treatment, whereas electrical and hydraulic variables were recorded every 8 h. The microbial results should therefore be interpreted as treatment-level contextual evidence rather than as predictors of short-term electrical variation or proof of microbial causation.

#### **4.5 Transferability limitations and implications for PMFC monitoring and biochar selection**

The reported performance represents within-system temporal prediction rather than broad model generalisability. Although within-treatment blocked cross-validation reduced leakage between adjacent drying observations, all data originated from the same soil, plant species, electrode configuration, experimental batch and four treatment categories. The models therefore predicted held-out drying periods within an already represented PMFC system, not independent experiments or previously untested biochars.

Leave-one-treatment-out evaluation confirmed this limitation. Electrical-potential prediction retained only modest cross-treatment performance, whereas power-density models produced negative pooled  $R^2$  values when an entire treatment was withheld. The learned relationships were therefore strongly dependent on treatment backgrounds represented during training, particularly for power density. The current models are consequently more appropriate for analysing variation within known treatment classes than for extrapolating to new biochar materials.

Limited transferability also reflects the use of categorical treatment indicators, which encoded composite differences in biochar structure, surface chemistry, water retention, conductivity, electrode contact and microbial habitat. Such labels do not describe the properties of an unseen material. Future models should instead incorporate transferable descriptors, including

Brunauer – Emmett – Teller surface area, pore-size distribution, cation-exchange capacity, electrical conductivity, soil-water retention parameters and electrode-interface characteristics. Within these constraints, the framework remains useful for process assessment. The contribution of hydraulic-state information, particularly matric suction, indicates that water-status monitoring may help identify drying-related deterioration in electrical output. Treatment-specific models may also support preliminary biochar screening and comparison of hydraulic conditions associated with performance decline. These applications should, however, be regarded as decision support within comparable biochar–soil systems rather than as universally deployable prediction tools. External validation will require independent experimental batches spanning additional biochar feedstocks, pyrolysis conditions, amendment rates, soils, plants and electrode configurations. Drying – rewetting cycles, synchronised microbial sampling and complementary electrochemical measurements should also be incorporated. Overall, the present analysis identifies a meaningful but system-dependent relationship among hydraulic state, biochar treatment and PMFC performance. Replacing treatment labels with measurable material, hydraulic, electrochemical and microbial descriptors will be essential for improving model transferability and supporting more general PMFC monitoring and material selection.

#### **Author contributions**

Sibin Gan: Conceptualization, Methodology, Software, Formal analysis, Data curation, Visualization, Writing – original draft, and Writing – review and editing. Qinyue Yao: Investigation, Validation, Data curation, Visualization, and Writing – review and editing.

#### **Acknowledgements**

The authors thank the researchers involved in generating the original PMFC experimental and

603 16S rRNA sequencing datasets used in this reanalysis.

604 During the preparation of this manuscript, the authors used OpenAI's ChatGPT to assist with  
605 language editing, manuscript organisation, reference-format checking and conceptual planning  
606 of the graphical abstract. All AI-assisted material was critically reviewed and revised by the  
607 authors, who take full responsibility for the accuracy, originality and integrity of the final  
608 manuscript.

## 609 **Statements and Declarations**

## 610 **Funding**

611 This work was supported by the National Natural Science Foundation of China under Grant No.  
612 52261160382.

## 613 **Competing interests**

614 The authors declare no competing financial or non-financial interests that are directly or  
615 indirectly related to the work submitted for publication.

## 616 **Data availability**

617 The datasets analysed during the current study are available from the corresponding author on  
618 reasonable request.

619

## References

1. Strik DPBTB, Hamelers HVM, Snel JFH, Buisman CJN (2008) Green electricity production with living plants and bacteria in a fuel cell. *Int J Energy Res* 32(9):870–876. <https://doi.org/10.1002/er.1397>
2. Kaku N, Yonezawa N, Kodama Y, Watanabe K (2008) Plant/microbe cooperation for electricity generation in a rice paddy field. *Appl Microbiol Biotechnol* 79(1):43–49. <https://doi.org/10.1007/s00253-008-1410-9>
3. Logan BE, Rabaey K (2012) Conversion of wastes into bioelectricity and chemicals by using microbial electrochemical technologies. *Science* 337(6095):686–690. <https://doi.org/10.1126/science.1217412>
4. Chong PL, Chuah JH, Chow CO, Ng PK (2025) Plant microbial fuel cells: a comprehensive review of influential factors, innovative configurations, diverse applications, persistent challenges, and promising prospects. *Int J Green Energy* 22(3):599–648. <https://doi.org/10.1080/15435075.2024.2421325>
5. Xu C, Zhang J, Wang A, Liu W, Xu Y, Zhang G, Wang S (2025) Impact of sediment microbial fuel cells on the distribution of different forms of phosphorus in lake sediment and water. *Environ Technol* 46(18):3669–3680. <https://doi.org/10.1080/09593330.2025.2474255>
6. Or D, Smets BF, Wraith JM, Dechesne A, Friedman SP (2007) Physical constraints affecting bacterial habitats and activity in unsaturated porous media: a review. *Adv Water Resour* 30(6–7):1505–1527. <https://doi.org/10.1016/j.advwatres.2006.05.025>
7. Gan S, Chen B, Li L, Sushkova S, Garg A (2024) Effect of three different types of biochar on bioelectricity generated from plant microbial fuel cells under unsaturated soil condition. *ACS Appl Bio Mater* 7(10):6554–6567. <https://doi.org/10.1021/acsabm.4c00727>
8. Chen B, Li L, Cai W, Garg A (2024) Bioelectricity generation in plant microbial fuel cells: influence of vegetation density and unsaturated soil properties. *Biomass Bioenergy* 181:107053. <https://doi.org/10.1016/j.biombioe.2024.107053>
9. Zhao X, Xu Y, Yin F, Li Y, Li X, Wei Q (2024) Co-Fe-N@biochar anode for improvement the electricity generation performance of microbial fuel cell. *Environ Technol*

- 45(24):5048–5062. <https://doi.org/10.1080/09593330.2023.2283797>
10. Sohi SP, Krull E, Lopez-Capel E, Bol R (2010) A review of biochar and its use and function in soil. *Adv Agron* 105:47–82. [https://doi.org/10.1016/S0065-2113\(10\)05002-9](https://doi.org/10.1016/S0065-2113(10)05002-9)
  11. Lehmann J, Rillig MC, Thies J, Masiello CA, Hockaday WC, Crowley D (2011) Biochar effects on soil biota: a review. *Soil Biol Biochem* 43(9):1812–1836. <https://doi.org/10.1016/j.soilbio.2011.04.022>
  12. Abel S, Peters A, Trinks S, Schonsky H, Facklam M, Wessolek G (2013) Impact of biochar and hydrochar addition on water retention and water repellency of sandy soil. *Geoderma* 202–203:183–191. <https://doi.org/10.1016/j.geoderma.2013.03.003>
  13. Huggins T, Wang H, Kearns J, Jenkins P, Ren ZJ (2014) Biochar as a sustainable electrode material for electricity production in microbial fuel cells. *Bioresour Technol* 157:114–119. <https://doi.org/10.1016/j.biortech.2014.01.058>
  14. Rusyn I (2021) Role of microbial community and plant species in performance of plant microbial fuel cells. *Renew Sustain Energy Rev* 152:111697. <https://doi.org/10.1016/j.rser.2021.111697>
  15. Tardast A, Rahimnejad M, Najafpour G, Ghoreyshi A, Premier GC, Bakeri G (2014) Use of artificial neural network for the prediction of bioelectricity production in a membrane-less microbial fuel cell. *Fuel* 117:697–703. <https://doi.org/10.1016/j.fuel.2013.09.047>
  16. Zhu X, Wang X, Ok YS (2019) The application of machine learning methods for prediction of metal sorption onto biochars. *J Hazard Mater* 378:120727. <https://doi.org/10.1016/j.jhazmat.2019.06.004>
  17. Yuan X, Suvarna M, Low S, Dissanayake PD, Lee KB, Li J, et al (2021) Applied machine learning for prediction of CO<sub>2</sub> adsorption on biomass waste-derived porous carbons. *Environ Sci Technol* 55(17):11925–11936. <https://doi.org/10.1021/acs.est.1c01849>
  18. Palansooriya KN, Li J, Dissanayake PD, Suvarna M, Li L, Yuan X, et al (2022) Prediction of soil heavy metal immobilization by biochar using machine learning. *Environ Sci Technol* 56(7):4187–4198. <https://doi.org/10.1021/acs.est.1c08302>
  19. Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Aroita G, et al (2017) Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* 40(8):913–929. <https://doi.org/10.1111/ecog.02881>

20. Kaufman S, Rosset S, Perlich C, Stitelman O (2012) Leakage in data mining: formulation, detection, and avoidance. *ACM Trans Knowl Discov Data* 6(4):15. <https://doi.org/10.1145/2382577.2382579>
21. Gan S, Liu W, Chen B, Garg A, Zhussupbekov A (2025) Effects of three different biochars on microbial community and their corresponding bioelectricity generation mechanism in PMFCs under unsaturated conditions. *Environ Qual Manag* 35(1):e70145. <https://doi.org/10.1002/tqem.70145>
22. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* 17(1):10–12. <https://doi.org/10.14806/ej.17.1.200>
23. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP (2016) DADA2: high-resolution sample inference from Illumina amplicon data. *Nat Methods* 13(7):581–583. <https://doi.org/10.1038/nmeth.3869>
24. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* 37(8):852–857. <https://doi.org/10.1038/s41587-019-0209-9>
25. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 41(D1):D590–D596. <https://doi.org/10.1093/nar/gks1219>
26. Chao A (1984) Nonparametric estimation of the number of classes in a population. *Scand J Stat* 11(4):265–270
27. Shannon CE (1948) A mathematical theory of communication. *Bell Syst Tech J* 27(3):379–423. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>
28. Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 12(1):55–67. <https://doi.org/10.1080/00401706.1970.10488634>
29. Cover T, Hart P (1967) Nearest neighbor pattern classification. *IEEE Trans Inf Theory* 13(1):21–27. <https://doi.org/10.1109/TIT.1967.1053964>
30. Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32. <https://doi.org/10.1023/A:1010933404324>
31. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al (2011) Scikit-learn: machine learning in Python. *J Mach Learn Res* 12:2825–2830

- 709 32. Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A (2008) Conditional variable  
710 importance for random forests. *BMC Bioinformatics* 9:307. [https://doi.org/10.1186/1471-](https://doi.org/10.1186/1471-2105-9-307)  
711 2105-9-307
- 712 33. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, et al (2011)  
713 Metagenomic biomarker discovery and explanation. *Genome Biol* 12(6):R60.  
714 <https://doi.org/10.1186/gb-2011-12-6-r60>
- 715 34. Louca S, Polz MF, Mazel F, Albright MBN, Huber JA, O'Connor MI, et al (2018) Function  
716 and functional redundancy in microbial systems. *Nat Ecol Evol* 2:936–943.  
717 <https://doi.org/10.1038/s41559-018-0519-1>