Interpretable multi-omics machine learning reveals drought-driven shifts in plant–microbe interactions

Hayato Yoshioka[1,2], Pavla Debeljak[2,3], Soizic Prado[2], Yushiro Fuji[4], Yasunori Ichihashi[5], Hiroyoshi Iwata[1*]

[1]Department of Agricultural and Environmental Biology, Graduate School of Agricultural and Life Science, The University of Tokyo, Tokyo, Japan.
[2]Muséum National d'Histoire Naturelle, Unité Molécules de Communication et Adaptation des Micro-Organismes, UMR 7245, CNRS, Sorbonne Université, Paris, France.
[3]SupBiotech, Villejuif, France.
[4]Center for Sustainable Resource Science, RIKEN, Yokohama, Japan.
[5]Center for Sustainable Resource Science, RIKEN, Tsukuba, Japan.

*Corresponding author(s). E-mail(s): hiroiwata@g.ecc.u-tokyo.ac.jp;

Abstract

Background: Plant-microbe interactions in the rhizosphere are central to plant growth, nutrient acquisition, and stress resilience. Although multi-omics approaches enable comprehensive profiling of different biological layers, integrating these data to understand the mechanisms underlying plant-microbe symbiosis, particularly under drought stress, remains a challenge.
Results: Genomic, metabolomic, and microbiome data from 198 soybean accessions grown under both control and drought conditions were integrated to identify environment-specific predictive features of the plant phenotypes. We compared best linear unbiased prediction (BLUP), genome-wide association study (GWAS), and a nonlinear machine learning model to evaluate their ability to detect informative features. The machine learning models provided flexible variable selection and outperformed linear models in capturing nonlinear dependencies. Model interpretation using SHapley Additive exPlanations (SHAP) indicated that the isoflavone derivative, daidzin, and the drought-tolerant *Candidatus Nitrosocosmicus*, were major contributors to phenotypic variation, specifically under drought stress. SHAP-based interaction networks indicated cross-omics links, including connections between daidzin, gamma-aminobutyric acid (GABA), and *Paenibacillus*.

Conclusion: The proposed interpretable machine learning approach for plant phenotype prediction identified multi-omics biomarkers and interactions, providing insights into plant adaptation to drought stress through environment-dependent rhizosphere networks and symbiotic associations.

Background

Interactions between plants and microbes are pivotal in regulating plant growth (Ulrich et al., 2019; Xiong et al., 2020). In particular, rhizosphere microorganisms are crucial for plant health as they enhance nutrient uptake and strengthen plant resilience, especially under abiotic stresses, such as drought (Kristin and Miranda, 2013; Alemu et al., 2024). Therefore, understanding the rhizosphere microbiome is essential for deciphering the mechanisms underlying plant-microorganism symbiosis, particularly those related to drought tolerance.

Recent advances in multi-omics technologies have enabled the concurrent profiling of genomic (Kajiya-Kanegae et al., 2021), metabolomic (Guo et al., 2023), and microbiome data (Marco et al., 2022). These inter-omics datasets provide new opportunities to unravel the complex relationships among various biological systems. However, the integration and interpretation of such diverse data types pose significant computational and statistical challenges.

Best Linear Unbiased Prediction (BLUP) is a widely used linear mixed model in plant and animal breeding (Meuwissen et al., 2001; VanRaden, 2008). It leverages genome-wide markers to predict phenotypic traits by capturing linear additive genetic contributions. While BLUP can incorporate multiple omics layers through multi-kernel approaches, it remains limited to modeling linear relationships among features.

Genome-wide association studies (GWAS) are popular for identifying individual SNP markers associated with phenotypic variation (Atwell et al., 2010; Ott et al., 2015). However, GWAS typically analyzes genome-wide markers in isolation. When evaluating the significance of the interactions between genome-wide markers and other omics variables, the number of possible combinations and multiple testing burden may increase exponentially, making such analyses infeasible.

In contrast, machine learning approaches, including Random Forest (RF; Breiman 2001), are capable of nonlinear modeling and feature selection in multi-omics contexts. RF constructs a collection of decision trees and aggregates their predictions, providing both high predictive performance and feature importance scores for biological features (Acharjee et al., 2016; Pfeifer et al., 2022). Nonlinear dependencies across omics layers have been observed in integrative studies (Zhao et al., 2022; Shen et al., 2024). Recently, our group demonstrated that the RF model can effectively capture such complexity and outperform traditional linear models, such as BLUP, in predictive tasks when nonlinearity exists in multi-omics interactions (Yoshioka et al., 2025).

SHapley Additive exPlanations (SHAP) is a game-theoretic approach that quantifies the marginal contribution of each feature by comparing model predictions across all possible subsets of input features (Lundberg and Lee, 2017). Unlike traditional

feature importance scores, SHAP values are both model-specific and globally consistent, offering insights into the interactions among features and their influence on predictions (Lundberg et al., 2019).

Recent studies have employed RF and SHAP to interpret phenotype prediction models, including those used for flowering time (Wang et al., 2024). In addition, SHAP has been used with microbiome data to identify features distinguishing between drought and control conditions (Hagen et al., 2024).

However, to date, no study has systematically integrated metabolomic, microbiomic, and genomic data to identify interacting biomarkers that bridge plant–microbe interactions. Moreover, the application of SHAP-based interpretability to uncover such cross-domain interactions in plant–microbe multi-omics datasets remains unexplored.

Due to the high-dimensional nature of the interaction analysis, we employed XGBoost (Chen and Guestrin, 2016), an ensemble tree-based gradient boosting method. Unlike RF, which aggregates independently grown trees, the gradient boosting model constructs trees sequentially by optimising a loss function, enabling efficient computation in high-dimensional settings.

In this study, we specifically focused on drought-induced shifts in plant-microbe interactions. By comparing plants grown under drought and control conditions, we aimed to capture how the plant-microbe-metabolome network collectively responds to water deficit stress.

Our analysis was conducted in three steps. First, we compared the BLUP and RF models to assess their respective abilities and identify biologically meaningful features from multi-omics datasets. Second, we employed SHAP to elucidate condition-specific features and cross-omics interactions that drive phenotypic variation. Finally, through this integrative multi-omics framework, we identify drought-responsive biomarkers and uncover condition-specific regulatory mechanisms that underpin adaptive responses in plant-soil microbiome systems.

Methods

Soybean Multi-Omics Data

We analysed multi-omics datasets collected from a common panel of $N = 198$ soybean accessions (Fig. 1). All omics and phenotype measurements were obtained from field trials conducted in 2019 under two watering conditions: watering (control) and non-watering (drought) to assess the effects of drought stress (Toda et al., 2022; Sakurai et al., 2022). The datasets included the following:

- **Whole-genome SNP genotypes (genome)**: a $198 \times 425{,}858$ matrix, where each column corresponds to a single nucleotide polymorphism (SNP) marker in the soybean genome (Kajiya-Kanegae et al., 2021).
- **Rhizosphere metabolome**: a 198×263 matrix of metabolome features obtained from rhizosphere samples. Each column represents the normalized peak area of a distinct metabolome feature detected by tandem mass spectrometry (Dang et al., 2025).

- **Rhizosphere microbiota profiles (microbiome)**: a $198 \times 16{,}457$ matrix generated by 16S rRNA gene amplicon sequencing of DNA extracted from rhizosphere samples. Each column represents the relative abundance of a bacterial taxon (Dang et al., 2025).
- **Plant phenotype**: a 198×9 matrix. These traits include biomass-related features such as shoot and leaf fresh/dry weights, growth stage, plant height, main stem length, number of nodes, and number of branches.

The soybean accessions used in this study were selected from the Global Soybean Minicore Collection (Kajiya-Kanegae et al., 2021). Furthermore, all rhizosphere microbiota profiles (microbiome) and rhizosphere metabolome data were collected from the same field plot as the phenotypic measurements. This design ensured consistency across samples and enabled integrative modeling across multiple omics layers.

To reduce the computational burden in the SHAP analysis of the RF model, SNPs were pruned based on linkage disequilibrium (LD). A representative subset of 3,078 SNPs (LD threshold = 0.001) was selected from an initial set of 425,858 SNPs (LD threshold = 0.95). Despite the relatively small size compared with the full data, the 3,078-SNP subset provided approximate coverage of the entire genome loci (Fig. B1). Additionally, the 3,078-SNP subset achieved predictive performance comparable to that of larger sets containing approximately 10k to 34k SNPs in phenotypic predictions (Fig. B2). The objective was to evaluate variables rather than optimise predictive performance. Thus, considering interpretability, computational limits in RF, and the need to balance the number of variables with those in the microbiome and metabolome datasets, we selected the 3,078-SNP set for subsequent RF analysis.

In the field trial, both microbiome and metabolome datasets were obtained from the same plots used for the phenotypic measurements. Note that these datasets were also recently utilized in Dang et al. (2025). Although they were collected from the same field, in the present study, the microbiome data were newly analyzed and integrated according to our experimental design, while the metabolome data were derived from those previously processed in Dang et al. (2025).

The observed phenotypic values and omics features were filtered to remove missing values. After filtering, the dataset sizes (accessions $\times$ features) were as follows. For the control condition: metabolome (182×255), microbiome (182×5424), and phenotype (182×9). For the drought condition: metabolome (179×253), microbiome (179×4771), and phenotype (179×9). All datasets were subsequently scaled prior to modeling. Details of the omics data processing for integration are provided in Appendix A.

See the supplementary information in Appendix A for details on data collection and experimental conditions. The soybean raw genome sequences are available in the DDBJ Sequence Read Archive (DRA) under the BioProject accession number PRJDB7281, as previously processed and reported by Kajiya-Kanegae et al. (2021). The microbiome raw read sequences were originally processed and are available in the DDBJ DRA under the BioProject accession number PRJDB37881. The metabolome raw data were previously processed and analyzed by Dang et al. (2025), and are available in the RIKEN DropMet database (http://prime.psc.riken.jp/menta.cgi/prime/drop_index; ID: DM0071). The plant phenotype data were originally processed and analyzed in

this study. The microbiome, metabolome, and phenotype data are also provided as supplementary materials.

Linear Mixed Model Analysis (BLUP)

To estimate the contributions of individual features to phenotypic variation, we implemented a linear mixed model framework based on BLUP (Meuwissen et al., 2001; VanRaden, 2008).

We have a feature matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ and a response matrix $\mathbf{Y} \in \mathbb{R}^{n \times q}$. Each omics layer was used to construct a corresponding kernel matrix $\mathbf{K}_l$, which captures the pairwise similarities among samples. The univariate mixed model was performed for each phenotype $j = 1, \ldots, q$ as follows:

$$\mathbf{y}_j = \mathbf{1}\mu + \sum_{l=1}^{L} \mathbf{Z}_l \mathbf{u}_{l,j} + \boldsymbol{\epsilon}, \quad \mathrm{Var}[\mathbf{y}_j] = \sum_{l=1}^{L} \mathbf{Z}_l \mathbf{K}_l \mathbf{Z}_l' \sigma_l^2 + \mathbf{I}\sigma_e^2.$$

The multi-kernel linear mixed effects model was computed using the R package `RAINBOWR` (version 0.1.35) (Hamazaki and Iwata, 2020). The model was applied independently to each phenotype $j = 1, \ldots, q$ to obtain trait-specific random effects $\mathbf{u}_{l,j}$ for each omics layer l. Assuming that each kernel matrix was based on a linear inner product, we reconstructed the feature-level coefficients for each trait and omics layer using:

$$\mathbf{K}_l = \frac{1}{p_l} \mathbf{X}_l \mathbf{X}_l', \quad \mathbf{b}_{l,j} = \mathbf{X}_l' \mathbf{K}_l^{-1} \mathbf{u}_{l,j},$$

where $\mathbf{X}_l \in \mathbb{R}^{n \times p_l}$ is each feature matrix (genome, metabolome, or microbiome), and $\mathbf{b}_{l,j} \in \mathbb{R}^{p_l \times 1}$ is the resulting coefficient vector.

These feature matrices were used to construct the corresponding kernel matrices $\mathbf{K}_l$ for every layer. All coefficient vectors were concatenated across the traits and layers to form a global coefficient matrix:

$$\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_q] \in \mathbb{R}^{p \times q}, \quad \mathbf{b}_j = \begin{bmatrix} \mathbf{b}_{1j} \\ \vdots \\ \mathbf{b}_{Lj} \end{bmatrix} \in \mathbb{R}^{p \times 1}.$$

To enable comparison across traits, each column of $\mathbf{B}$ was standardised to have a mean 0 and a variance 1:

$$\sum_{i=1}^{p} \mathbf{B}_{i,j} = 0, \quad \frac{1}{p} \sum_{i=1}^{p} \mathbf{B}_{i,j}^2 = 1, \quad j = 1, \ldots, q.$$

To examine the structure of the feature-trait relationships, we applied singular value decomposition (SVD) to the coefficient matrix:

$$\mathbf{B} = \mathbf{U}\mathbf{D}\mathbf{V}',$$

where $\mathbf{U}$ and $\mathbf{V}$ are the left and right singular vectors, respectively, and $\mathbf{D}$ is the diagonal matrix of singular values. The phenotypes and features were projected into the latent space using $\mathbf{U}\sqrt{\mathbf{D}}$ and $\mathbf{V}\sqrt{\mathbf{D}}$, respectively.

Nonlinear Feature Importance Using RF

We also implemented a nonlinear regression framework using RF to estimate feature contributions. The feature matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ was constructed by concatenating multi-omics layers, while the response matrix $\mathbf{Y} \in \mathbb{R}^{n \times q}$ represents the biomass-related phenotypic traits. Details of the data preprocessing are provided in Appendix A.

For each phenotype $j = 1, \ldots, q$, we trained an independent RF model using the R package `ranger` (version 0.16.0) (Wright and Ziegler, 2017). Hyperparameter tuning was performed using a grid search. See Appendix C and Fig. B3 for details. We confirmed that the predictive performance of the RF model fitted to the original data exceeded the null distribution of predictive performance obtained by refitting the model after randomly permuting the response vector (Fig. B4).

$$\mathbf{y}_j = f_j(\mathbf{X}) + \boldsymbol{\epsilon}_j,$$

where f_j denotes the RF model trained for phenotype j. After training, the feature importance scores were extracted. Let $\mathbf{I} \in \mathbb{R}^{p \times q}$ denote the resulting matrix of feature contributions, where $\mathbf{I}_{i,j}$ indicates the contribution of feature i to phenotype j.

The importance values from the RF were all positive. For the convenience of variable selection, each column of $\mathbf{I} \in \mathbb{R}^{p \times q}$ is normalised such that the total sums to 1 for each phenotype:

$$\sum_{i=1}^{p} \mathbf{I}_{i,j} = 1, \quad j = 1, \ldots, q.$$

Analogous to the coefficient matrix $\mathbf{B}$ derived from BLUP, the normalised importance matrix $\mathbf{I} \in \mathbb{R}^{p \times q}$ from the RF models can be treated as a feature-by-trait matrix:

$$\mathbf{B}^{(\mathrm{RF})} = \mathbf{I} \in \mathbb{R}^{p \times q}.$$

Here, each column of $\mathbf{B}^{(\mathrm{RF})}$ represents the relative contribution of each feature to the prediction of a specific phenotype, aggregated across the trees in the forest (Fig. 1). While the entries of $\mathbf{B}^{(\mathrm{RF})}$ do not correspond to regression coefficients in a strict parametric sense, this matrix serves as a comparable construct for downstream analyses, such as dimensionality reduction or visualisation of feature-trait associations. Following the BLUP, we can perform the SVD for $\mathbf{B}^{(\mathrm{RF})}$ to visualise the structure of feature-trait relationships.

Leaf dry weight, a commonly used indicator of plant biomass that reflects the size of the primary photosynthetic organ, showed representative predictability and consistent associations with omics features across all three omics layers (Fig. 2). Therefore, downstream GWAS and SHAP analyses were conducted using leaf dry weight as a representative phenotype for interpretability and computational feasibility.

Genome-wide association study (GWAS)

GWAS is a linear approach for identifying SNP markers associated with phenotypic variation (Atwell et al., 2010; Ott et al., 2015). Here, we employed GWAS in a conventional setting to evaluate the baseline performance and provide a comparison with SNP selection results from RF importance.

GWAS was performed using a linear mixed model (LMM) implemented in the R package `gaston`(Perdry and Dandine-Roulland, 2023) (version1.6), incorporating three principal components (PCs) derived from the eigen decomposition of the GRM to account for population structure.

The full set of SNPs (425,858 markers, LD threshold = 0.95) was filtered using a minor allele frequency (MAF) threshold of 0.05 and then used in the GWAS model. After modeling, SNPs with a significance threshold of $p < 1 \times 10^{-5}$ were retained for further analysis.

SHAP: Individual-Features Attribution

To investigate the individual contributions of multi-omics features to phenotype prediction (hereafter referred to as individual feature contributions), we applied SHAP to the trained RF models. SHAP is a game-theoretic approach that quantifies the marginal contribution of each feature by comparing the model predictions across all possible subsets of the input features.

Formally, the SHAP value for feature i is defined as:

$$\phi_i(f, x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \left[f_x(S \cup \{i\}) - f_x(S) \right]. \tag{1}$$

where N is the set of all input features, S is a subset that excludes feature i, and f is the prediction model. The $f_x(S)$ denotes the model prediction for input instance x when only the features in S are considered. The left-hand side of equation (1) is a combination weight designed to give equal consideration to all possible feature orders. The right-hand side represents the marginal contribution to the prediction when feature i is added to subset S. By combining these components, feature importance can be assigned fairly, making SHAP particularly suitable for elucidating the nonlinear effects of features in multi-omics data.

SHAP values were computed for each sample using the `iml` R package (version 0.11.4) (Molnar et al., 2018) and then averaged to obtain global feature importance scores.

After RF modeling under drought and control conditions separately, omics features were ranked based on their total variable importance scores, calculated for each feature i as follows:

$$\theta_i = \sum_{j=1}^{q} \mathbf{B}^{(\mathrm{RF})}{}_{i,j}$$

were retained, where $\mathbf{B}^{(\mathrm{RF})}{}_{i,j}$ denotes the contribution of omics feature i to phenotype j.

Here, we selected features with a total RF importance score of $\theta_i > 0.01$ under either drought or control conditions, yielding a total of 159 features for evaluation. This threshold was chosen to retain features that exhibited non-negligible importance, given that the distribution of the feature importance scores was zero-inflated (Fig. B5). From a practical perspective, applying such a cutoff provides a reasonable preprocessing step for SHAP analysis because it reduces the computational burden while preserving the major informative signals.

Using this subset of 159 features, the SHAP values were calculated to identify environment-dependent differences in the relative contribution of individual features. For each condition, the absolute SHAP values were first calculated for each feature across all samples and then averaged. The ratio of the mean absolute SHAP values (drought divided by control) was calculated to highlight the condition-specific features of the model.

SHAP: Interaction-Features Attribution

Condition Specific Interactions

Although SHAP individual-feature attribution values quantify each omics feature, they do not capture interaction effects. To address this, we applied SHAP interaction analysis to characterise the pairwise relationships among the features (Molnar, 2025).

SHAP interaction values decompose the prediction into individual effects and pairwise interactions.

$$\phi_{i,j}(f,x) = \sum_{S \subseteq N \setminus \{i,j\}} \frac{|S|!\,(N-|S|-2)!}{2(N-1)!}\,\delta_{ij}(S), \tag{2}$$

where $i \neq j$ and

$$\delta_{ij}(S) = f_x(S \cup \{i,j\}) - f_x(S \cup \{i\}) - f_x(S \cup \{j\}) + f_x(S). \tag{3}$$

The weighting factor in equation (2) ensures the symmetry consideration of all feature orderings, whereas the term $\delta_{ij}(S)$ isolates the second-order marginal contribution in equation (3).

Although some features exhibited only minor individual effects on the phenotype, they might contribute to phenotypic variation through cooperative or context-dependent interactions with other omics layers. To capture such synergy-driven patterns, we adopted a broader feature set ($\theta_i > 0.001$ in both conditions), which facilitated the identification of interaction effects that were not reflected in the individual feature importance. The distribution of feature importance values is shown in Fig. B5.

To evaluate the sensitivity of the inferred interaction network to the RF importance threshold θ, we conducted a robustness analysis by varying θ from 0.001 to 0.01. SHAP interaction scores were first computed under the most permissive setting ($\theta = 0.001$). For each stricter threshold, nodes retained according to RF importance were identified for each condition, and the interaction network was restricted to edges whose two incident nodes remained selected.

This yielded a total of 2,729 features for the evaluation. Based on these features, we constructed omics feature interaction networks, where nodes represent individual features and edges denote the strength of interaction between feature pairs.

In high-dimensional SHAP interaction analysis, the number of feature pairs increases exponentially with the number of input variables. Due to this computational burden, the interaction values were computed using the `shap.TreeExplainer` from the `shap` Python package applied to an XGBoost-based gradient boosting model (Chen and Guestrin, 2016), as RF-based implementations were computationally prohibitive in this setting. Since both RF and gradient boosting are tree-based ensemble methods, the resulting SHAP interaction values captured the same class of nonlinear feature interactions underlying the nonlinear prediction models.

Hyperparameter tuning was conducted so that the same set of tuned hyperparameters could be applied to both the control and drought conditions, ensuring comparability between environmental settings. See Appendix C and Fig. B3 for implementation details.

SHAP interaction values were calculated for each sample and averaged to obtain the global 2,729 $\times$ 2,729 symmetrical interaction matrices for each condition. To highlight the environment-specific interactions, we computed a differential SHAP interaction matrix as:

$$\Delta\mathbf{\Phi} = \mathbf{\Phi}_{\text{drought}} - \mathbf{\Phi}_{\text{control}},$$

where $\mathbf{\Phi}$ denotes a SHAP-based pairwise interaction matrix. Positive values of $\Delta\mathbf{\Phi}$ indicated stronger interactions under drought stress, whereas negative values indicated interactions that were more pronounced under control conditions.

The two-step SHAP analysis—first focusing on individual feature importance, then exploring pairwise interaction features—provides a comprehensive and interpretable view of how genomic, metabolomic, and microbial features jointly influence plant phenotypes under contrasting environmental conditions.

Null-Distribution-Based Significance Assessment

To assess the statistical significance of each differential interaction, a permutation-based null distribution was constructed. For each condition, the response vector was randomly permuted. The XGBoost model was then refitted to the permuted data, and SHAP interaction matrices were computed. For each permutation replicate $b \in \{1, \ldots, 100\}$, a null differential interaction matrix was calculated as:

$$\Delta\mathbf{\Phi}^{(b)} = \mathbf{\Phi}^{(b)}_{\text{drought}} - \mathbf{\Phi}^{(b)}_{\text{control}},$$

representing the distribution of interaction differences expected under the null hypothesis of no association between predictors and observed values (phenotypes).

All off-diagonal (upper-triangular) entries from the $B = 100$ null differential interaction matrices were concatenated to form a pooled null distribution, from which the mean μ_0 and standard deviation σ_0 were estimated. The observed differential

interaction values were converted into Z-scores as:

$$Z_{ij} = \frac{\Delta\Phi_{ij} - \mu_0}{\sigma_0},$$

Here, edges with large absolute Z-scores were interpreted as highly environment-specific SHAP interactions.

Edgewise empirical p-values were computed for each feature pair (i, j) using two-sided tests based on the null replicates (Phipson and Smyth, 2016).

$$p_{ij} = \frac{1 + \sum_{b=1}^{B} \mathbb{I}\left(\left|\Delta\Phi_{ij}^{(b)}\right| \geq |\Delta\Phi_{ij}|\right)}{B+1},$$

To examine the robustness of the inferred interaction network with respect to the choice of the Z-score threshold, we systematically constructed networks using a range of alternative cutoffs ($|Z| \geq 10, 20, 30, 40, 50$, and 60). For each threshold, the corresponding two-sided empirical p-values were jointly evaluated to ensure consistency between effect size and statistical extremeness.

Although the overall qualitative network structure was preserved across the thresholds, a cutoff of $|Z| \geq 30$ was selected for the downstream analyses as a pragmatic compromise between statistical stringency and network interpretability. This threshold corresponds to the extreme tail of the pooled null distribution of SHAP interaction differences (Fig. B6) and effectively filters out interactions that are likely attributable to resampling noise. Under this cutoff, the resulting network comprised 160 connections, representing only 0.004% of all possible feature pairs, yielding a sparse yet interpretable representation of robust environment-specific interactions.

Stability Analysis of Interaction Edges using Bootstrap Resampling

Because SHAP interaction matrices are estimated from finite samples, the individual edge differences in $\mathbf{\Delta\Phi}$ may be sensitive to resampling variability. To assess the robustness of the differential interactions, edge-level stability was quantified using an additional bootstrap procedure applied to the observed data.

Specifically, for each condition, we generated 100 bootstrap replicates by resampling individuals with replacements and refitting the XGBoost model on the resampled pairs $\{(\mathbf{x}_k, y_k)\}$. For each replicate b, we computed the bootstrap differential interaction matrix

$$\Delta\mathbf{\Phi}_{\text{boot}}^{(b)} = \mathbf{\Phi}_{\text{drought,boot}}^{(b)} - \mathbf{\Phi}_{\text{control,boot}}^{(b)}$$

Using these replicates, the complementary stability measures for each edge were defined, (i, j). The presence stability was defined as the number of bootstrap replicates for which a given edge (i, j) is observed as nonzero.

$$n_{ij}^{\text{obs}} = \sum_{b=1}^{B_{\text{boot}}} \mathbb{I}\left(\left|\Delta\Phi_{\text{boot},ij}^{(b)}\right| > \varepsilon\right),$$

where $\varepsilon = 10^{-12}$ is the small numerical tolerance. This measure aims to evaluate the reproducibility of the interaction under resampling, which is particularly relevant for sparse or zero-inflated interaction matrices.

To suppress the spurious edges driven by resampling noise, a stability-based filtering criterion was applied, and edges with $n_{ij}^{\text{obs}} < 2$ were excluded from downstream network analyses. Given the extreme sparsity of SHAP interaction matrices ($\approx 0.015\%$, corresponding to approximately 500 of 3.7 million possible pairs), the appearance of a purely spurious edge in each bootstrap replicate could be treated as a rare event. Under a simple binomial model $X \sim \text{Binomial}(B_{\text{boot}} = 100, p)$ with $p \approx 1.5 \times 10^{-4}$, the probability that the edge observed in at least two bootstrap replicates is approximately 1.1×10^{-4}. See Appendix C.3 for further statistical justification for this conservative threshold. Thus, the stability metric: n_{ij}^{obs} is reported alongside the Z-scores and empirical p-values to facilitate interpretation of the edge-level robustness.

Results

RF and BLUP Models

Fig. 2 presents the SVD of the coefficient matrix of BLUP and the importance matrix of RF for the drought conditions. The results demonstrate that the RF model exhibited a more pronounced and variable pattern of feature importance than the BLUP model. Notably, in BLUP, metabolites are the features most correlated with phenotypes, while other features (such as the genome and microbiome) exhibit a shrinkage pattern, indicating that metabolome features primarily contribute to phenotypic prediction. Under control conditions, a similar trend was observed (Fig. B7).

SHAP analysis of individual-features

SHAP analysis suggests key condition-specific features contributing to phenotypic variations under drought and control conditions. Since metabolomic and microbiomic profiling do not cover the entirety of omics information, the highlighted features should be understood as those selected from the analysed subset described in the Materials and Methods.

Under drought stress, the most highly contributing features were predominantly flavonoids, including isoflavones (e.g. daidzin and its derivatives and genistin) and flavone/flavanone glucosides (e.g. apigenin-7-O-glucoside and naringenin-7-O-glucoside). Flavonoids play a major role in plant responses to drought stress and are associated with plant-microbe interactions (Gargallo-Garriga et al., 2018; He et al., 2022). Additionally, we identified amino acids and their derivatives (asparagine, ornithine, and indole-3-carboxyaldehyde) as major contributing features.

Furthermore, *Candidatus Nitrosocosmicus*, a clade of ammonia-oxidising archaea (AOA), was particularly selected under drought. This clade is notable for its ability to produce manganese catalase (MnKat), which decomposes hydrogen peroxide (H_2O_2), a major reactive oxygen species (ROS) (Sauder et al., 2017). This enzymatic capacity

enables *Ca. Nitrosocosmicus* to survive in oxidative environments, such as drought-stressed soils. Its dominance under ROS stress conditions highlights its potential role in AOA-plant interactions (Lee et al., 2024), consistent with our findings.

In contrast, under control conditions, genetic markers contributed significantly to the observed traits. SHAP re-identified the SNP `Chr06-15056895` as an important feature, which was also detected by GWAS (Fig. B8). This SNP encodes a copper amine oxidase family protein involved in amine metabolism, tissue differentiation, and organ development in plants (Tavladoraki et al., 2016). An SNP `Chr04-22035062`, located near the gene *Glyma.04G140400*, was also selected. This gene encodes an oxidoreductase enzyme that is potentially involved in plant growth (Garrone et al., 2015). Another notable marker, SNP `Chr04-16278976`, is located near *Glyma.04G127100* (Valliyodan et al., 2019). This gene encodes a receptor-like kinase that contains both a protein kinase-like domain (IPR011009) and a Gnk2-homologous domain (IPR002902) and is involved in signal transduction pathways (Shiu and Bleecker, 2001).

In addition to genetic factors, *Duganella* was identified as an important microbial contributor under control conditions in this study. Several species of *Duganella* are known for their plant growth-promoting properties, including the solubilisation of essential nutrients, such as phosphorus, potassium, and zinc (Verma et al., 2014).

Some metabolites, such as gamma-aminobutyric acid (GABA), were selected in both control and drought conditions; however, they were not classified as condition-specific factors because their levels were similarly elevated under both conditions. Further details of each condition are presented in Fig. B9.

SHAP analysis of interaction-features

SHAP interaction values enabled a detailed investigation of multi-omics feature interactions. Condition-specific connectivity patterns were identified by comparing network differences between the drought and control conditions using a Z-score threshold of $|Z| \geq 30$ (Fig. B6). Under this threshold, major features were identified based on their connectivity with other features, including GABA and daidzin, both of which are known to promote plant growth and to influence rhizosphere microbiome composition, particularly under drought stress (Bennett et al., 2004; Xu et al., 2021; Ramesh et al., 2015; Gargallo-Garriga et al., 2018; Zhao et al., 2025). The corresponding networks were obtained by using alternative Z-score thresholds ($|Z| \geq 10, 20, 30, 40, 50, 60$), and are shown in Fig. B10.

Several of these interactions correspond to the known biological relationships. Notably, the model identified a strong interaction between daidzin and the bacterium *Paenibacillus* under drought conditions. This may be supported by a previous study showing that certain strains of *Paenibacillus* produce β-glucosidases that catalyse the hydrolysis of glycosidic bonds in isoflavone glucosides, such as daidzin, thereby converting them into their active aglycone forms, such as daidzein (Park et al., 2013). In addition, the model suggested a biologically plausible interaction between *Paenibacillus* and GABA under drought conditions. This aligns with reports that *Paenibacillus* can express glutamate decarboxylase, facilitating the conversion of glutamate to GABA (Singh et al., 2023).

Under control conditions, a notable interaction was observed involving SNP `Chr04-22035062`, which was also selected in the individual-feature analysis, located downstream of the gene *Glyma.04G140400* (oxidoreductase enzyme), potentially involved in metabolic responses to GABA.

We additionally evaluated the sensitivity of the inferred interaction network to the RF importance threshold θ. The main biological interactions reported in the network remained stable across a wide range of θ values (Table B1). For example, the interaction between *Paenibacillus* and daidzin was retained for $0.001 \leq \theta \leq 0.003$, the interaction between *Paenibacillus* and GABA for $0.001 \leq \theta \leq 0.006$, the interaction between `Chr04_22035062` and GABA for $0.001 \leq \theta \leq 0.01$, and the interaction between *Rhodanobacter* and daidzin for $0.001 \leq \theta \leq 0.005$. These results confirm that the principal interaction structure was robust to moderate changes in the preprocessing threshold (Fig. B11).

To assess the robustness of the detected interactions, their stability across 100 bootstrap replicates was evaluated (Tables B2 and B3). Among bootstrap replicates, several edges showed moderate stability. The edge between SNP `Chr04-22035062` and GABA exhibited relatively higher stability (observed in 14 bootstrap replicates). The interactions between *Paenibacillus* and daidzin (6 replicates) and between *Paenibacillus* and GABA (3 replicates) appeared across resampling iterations. Similarly, an interaction between *Rhodanobacter* and daidzin was detected in 6 replicates, suggesting a reproducible association between the taxon and isoflavone-related metabolites. However, the magnitude of SHAP interaction differences (Drought − Control) was not necessarily proportional to the replicative stability (Fig. B12).

Discussion

The SVD of the feature-trait matrices for BLUP and RF showed a more dispersed and interpretable structure for the RF-based models, particularly under drought conditions (Fig. 2 and B7). In contrast, BLUP showed more shrinking effects, which is typical when the explanatory matrix contains many variables. This suggests that a nonlinear model may be better suited for feature selection in complex biological systems, where additive linear assumptions do not hold.

The overlap between SNPs identified by GWAS and those selected by SHAP highlights the potential synergy between traditional and machine learning-based approaches (Fig. B8). For instance, SNP `Chr06-15056895`, located near the gene *Glyma.06G178400*, was detected by both SHAP and GWAS, reinforcing its putative role in growth-related traits. Such convergence enhances confidence in candidate biomarkers.

Our findings suggest a shift in the dominant omics layers under drought stress, highlighting the enhanced roles of metabolomic and microbial features (Fig. 3). Notably, the strong contribution of flavonoids, including daidzin, supports previous evidence that soybean roots adjust flavonoid exudation to shape the rhizosphere community in response to drought stress (Gargallo-Garriga et al., 2018; He et al., 2022). The identification of amino acids and their derivatives is also consistent with previous findings, as amino acid concentrations increase under drought conditions (Postles

et al., 2016). Additionally, the selection of *Candidatus Nitrosocosmicus*, a major contributor to phenotypic variation under drought stress, further validates the SHAP model's identification of key features in the microbial community (Sauder et al., 2017; Lee et al., 2024).

Interaction analysis provided further plausible plant-metabolic-microbial links (Fig. 4). For example, daidzin, which was selected as an individual feature, was also selected as a feature of the pairwise interactions. Especially, the co-selection of daidzin and *Paenibacillus* aligns with reports that *Paenibacillus* strains produce β-glucosidases to hydrolyse isoflavone glucosides such as daidzin (Park et al., 2013). This enzymatic synergy may be related to ROS management. Additionally, the inferred interaction between *Paenibacillus* and GABA agrees with evidence that this bacterium encodes glutamate decarboxylase (Singh et al., 2023), which could support plant stress resilience. Interestingly, although GABA was not selected as a condition-specific feature based on individual SHAP contributions, network-level analysis revealed a strong association with drought-specific activity. This highlights the advantage of network approaches in capturing complex indirect relationships that are not apparent from individual feature contributions alone.

In addition to these known interactions, the model also suggests other potential connections between plant secondary metabolites and the microbiome that have not yet been directly reported. For example, daidzin is associated with bacteria such as *Rhodanobacter* (Singh et al., 2023), which is known to promote plant growth. GABA was also connected with *Devosia*, a plant growth–promoting genus (Zhang et al., 2024), and *Aquisphaera*, which is frequently detected in the plant rhizosphere (Thompson et al., 2025). These findings highlight the potential to elucidate previously unknown plant-microbe interactions, in which plants may use flavonoids and other metabolome features to modulate the composition of rhizosphere bacteria (Hassan and Mathesius, 2012; Gargallo-Garriga et al., 2018; He et al., 2022).

Interestingly, our analysis showed contrasting patterns of feature associations under drought and control conditions. Under drought stress, metabolomic features were particularly prominent in the modeling, suggesting scenarios in which multiple metabolites are simultaneously produced by plants to cope with drought stress. For instance, the correlation between ornithine and asparagine may reflect amino acid accumulation patterns that influence plant stress responses (Postles et al., 2016). This highlights the importance of metabolite co-regulation and network-level interactions during abiotic stress.

In contrast, under control (non-drought) conditions, genetic determinants were more strongly associated with trait expression, reflecting the predominance of intrinsic growth programs and host-regulated physiological processes in the absence of stress, that is, genetic effects outweighing environmental influences. This is supported by the identification of genotypic markers near genes involved in signal transduction, amine metabolism, and development.

Interaction edges were observed in only a limited number of bootstrap replicates; however, this pattern is a natural consequence of the extremely high-dimensional and sparse interaction spaces in which only 0.015% of all possible feature pairs exhibited non-zero SHAP interactions. A simple binomial argument indicates that observing the

same interaction in as few as 2–4 of 100 bootstrap replicates is highly unlikely under a null model driven purely by sparsity. Given the empirical nonzero rate of SHAP interactions ($p \approx 1.5 \times 10^{-4}$), the probability of observing an edge at least twice, three times, or four times by chance alone is of the order of 10^{-4}, 10^{-7}, and 10^{-9}, respectively. Thus, even modest levels of replicative stability can provide non-trivial statistical support in this setting (Appendix C.3).

Some interactions exhibit large point-estimate differences but limited bootstrap recurrence, whereas others show moderate differences with higher stability (Fig. B12). This pattern may partly arise from the severe multicollinearity inherent in high-dimensional data such as those analyzed in this study, which complicates feature selection. In such settings, the model may alternate between correlated features across bootstrap samples, leading to reduced apparent stability. This issue can potentially be mitigated by incorporating stronger regularization or dimension-reduction approaches.

In the present study, we intentionally prioritized the detection and biological interpretation of interaction effects, using stability as a supportive criterion. Alternative analytical designs are also feasible. For example, greater emphasis could be placed on stability by adjusting the model complexity, regularization strength, or low-dimensional feature selection thresholds to enhance the collective recurrence of interactions. Such strategies would likely enable analyses to focus more explicitly on reproducibility.

Overall, the integrated framework enables scalable feature selection across multiple omics layers, including plant genomic data, and highlights potentially meaningful biological interactions. Despite this choice, there were some limitations. Although the proposed framework identifies synergistic associations across omics layers, it does not establish causal relationships between features. Therefore, the resulting interaction network reflects model-based associated contributions rather than direct mechanistic links, and should be interpreted as a hypothesis-generating representation to guide targeted validation.

Accordingly, further experimental studies are required to validate and extend our findings. Causal biological mechanisms can be explored through targeted experimental designs, such as quantitative trait locus (QTL) mapping and gene expression analyses, which will help refine candidate loci or genes and elucidate the underlying gene regulatory networks.

Moreover, we plan to expand this framework to incorporate larger untargeted metabolomics and microbiome datasets spanning a broader range of organisms, including fungi. Such extensions will facilitate investigations of broader ecological and ecosystem-level processes and provide deeper insights into the complex dynamics that shape these systems.

Conclusions

This study demonstrates the utility of comparing linear and nonlinear modeling approaches to validate feature selection in multi-omics datasets. By comparing BLUP and GWAS with a nonlinear machine learning model across genomic, metabolomic,

and microbiomic layers, we confirmed that the RF model captured additional nonlinear interactions, avoiding the shrinkage effects observed in linear methods. The use of SHAP further ensured that these nonlinear patterns were biologically interpretable and not model artefacts. We identified environment-specific biomarkers linked to soybean biomass under drought and control conditions, providing a basis for predictive breeding and microbiome-informed strategies to improve stress resilience. Future studies incorporating temporal sampling and experimental validation will help translate these multi-omics insights into practical strategies for sustainable crop production in diverse environments.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We are grateful to the following researchers for their contributions to soybean data collection: Masami Hirai Yokota, Megumi Narukawa, Erika Usui, Kie Kumaishi, Takumi Satou, Shungo Kobori, Yutaka Yamada, Akito Kaga, Yusuke Toda, Kengo Sakurai, Hideki Takanashi, Yoshihiro Ohmori, Motoyuki Ishimori, Toru Fujiwara, Hirokazu Takahashi, Mikio Nakazono, Mai Tsuda, Yuji Yamasaki, Hisashi Tsujimoto, Izumi Higashida, and all the technical staff at the Arid Land Research Center, Tottori University. We appreciate the advice of Dr. Komlan Avia.

Funding

This work was supported by JSPS KAKENHI (JP23KJ0506, JP22K21352), JST CREST (JPMJCR16O2), JST ALCA-Next (JPMJAN23D1), and JST-Mirai (JPMJMI120C7).

Data availability

The soybean genome data were previously processed by Kajiya-Kanegae et al. (2021), and are available in the DDBJ Sequence Read Archive (DRA) under the BioProject accession number PRJDB7281.

The metabolome raw data were previously processed by Dang et al. (2025), available on the RIKEN DropMet website (http://prime.psc.riken.jp/menta.cgi/prime/drop_index; ID: DM0071). The data are also provided as supplementary materials.

The microbiome raw read sequences are available in the DDBJ DRA under the BioProject accession number PRJDB37881 and were originally generated and published in this study. The data are also provided as supplementary materials.

The plant phenotype data were originally processed and analyzed in this study, and the data are provided as supplementary materials.

All codes and data are available at GitHub (https://github.com/Yoska393/ShapPlantMicro).

Contributions

H.Y., P.D., S.P., and H.I. conceived the study. Y.F. and Y.I. curated the data. H.Y. analyzed the data. H.Y., P.D., S.P., and H.I. wrote and reviewed the manuscript.

References

Acharjee, A., Kloosterman, B., Visser, R. G., and Maliepaard, C. (2016). Integration of multi-omics data for prediction of phenotypic traits using random forest. *BMC Bioinformatics*, 17:363–373.

Alemu, A., Åstrand, J., Montesinos-Lopez, O. A., y Sanchez, J. I., Fernandez-Gonzalez, J., Tadesse, W., Vetukuri, R. R., Carlsson, A. S., Ceplitis, A., Crossa, J., et al. (2024). Genomic selection in plant breeding: Key factors shaping two decades of progress. *Molecular Plant*, 17(4):552–578.

Atwell, S., Huang, Y. S., Vilhjálmsson, B. J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A. M., Hu, T. T., Jiang, R., Muliyati, N. W., Zhang, X., Amer, M. A., Baxter, I., Brachi, B., Chory, J., Dean, C., Debieu, M., de Meaux, J., Ecker, J. R., Faure, N., Kniskern, J. M., Jones, J. D. G., Michael, T., Nemri, A., Roux, F., Salt, D. E., Tang, C., Todesco, M., Traw, M. B., Weigel, D., Marjoram, P., Borevitz, J. O., Bergelson, J., and Nordborg, M. (2010). Genome-wide association study of 107 phenotypes in arabidopsis thaliana inbred lines. *Nature*, 465(7298):627–631.

Bennett, J. O., Yu, O., Heatherly, L. G., and Krishnan, H. B. (2004). Accumulation of genistein and daidzein, soybean isoflavones implicated in promoting human health, is significantly elevated by irrigation. *Journal of Agricultural and Food Chemistry*, 52(25):7574–7579.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, San Francisco, CA, USA. ACM.

Dang, T., Fuji, Y., Kumaishi, K., Usui, E., Kobori, S., Sato, T., Narukawa, M., Toda, Y., Sakurai, K., Yamasaki, Y., Tsujimoto, H., Hirai, M. Y., Ichihashi, Y., and Iwata, H. (2025). I-svvs: integrative stochastic variational variable selection to explore joint patterns of multi-omics microbiome data. *Briefings in Bioinformatics*, 26(3):bbaf132.

Gargallo-Garriga, A., Preece, C., Sardans, J., Oravec, M., Urban, O., and Peñuelas, J. (2018). Root exudate metabolomes change under drought and show limited capacity for recovery. *Scientific Reports*, 8(1):12696.

Garrone, A., Archipowa, N., Zipfel, P. F., Hermann, G., and Dietzek, B. (2015). Plant protochlorophyllide oxidoreductases a and b: Catalytic efficiency and initial reaction steps *. *Journal of Biological Chemistry*, 290(47):28530–28539.

Guo, X., Sarup, P., Jahoor, A., Jensen, J., and Christensen, O. F. (2023). Metabolomic-genomic prediction can improve prediction accuracy of breeding values for malting quality traits in barley. *Genetics Selection Evolution*, 55(1):61.

Hagen, M., Dass, R., Westhues, C., Blom, J., Schultheiss, S. J., and Patz, S. (2024). Interpretable machine learning decodes soil microbiome's response to drought stress. *Environmental Microbiome*, 19(1):35.

Hamazaki, K. and Iwata, H. (2020). Rainbow: Haplotype-based genome-wide association study using a novel snp-set method. *PLoS Computational Biology*, 16(2):1–17.

Hassan, S. and Mathesius, U. (2012). The role of flavonoids in root–rhizosphere signalling: opportunities and challenges for improving plant–microbe interactions. *Journal of Experimental Botany*, 63(9):3429–3444.

He, D., Singh, S. K., Peng, L., Kaushal, R., Vílchez, J. I., Shao, C., Wu, X., Zheng, S., Morcillo, R. J. L., Paré, P. W., and Zhang, H. (2022). Flavonoid-attracted aeromonas sp. from the arabidopsis root microbiome enhances plant dehydration resistance. *The ISME Journal*, 16(11):2622–2632.

Kajiya-Kanegae, H., Nagasaki, H., Kaga, A., Hirano, K., Ogiso-Tanaka, E., Matsuoka, M., Ishimori, M., Ishimoto, M., Hashiguchi, M., Tanaka, H., et al. (2021). Whole-genome sequence diversity and association analysis of 198 soybean accessions in mini-core collections. *DNA Research*, 28(1):dsaa032.

Kristin, A. and Miranda, H. (2013). The root microbiota—a fingerprint in the soil? *Plant and Soil*, 370(1–2):671–686.

Lee, U.-J., Gwak, J.-H., Choi, S., Jung, M.-Y., Lee, T. K., Ryu, H., Awala, S. I., Wanek, W., Wagner, M., Quan, Z.-X., and Rhee, S.-K. (2024). “¡i¿ca¡/i¿. nitrosocosmicus” members are the dominant archaea associated with plant rhizospheres. *mSphere*, 9(12):e00821–24.

Lundberg, S. M., Erion, G. G., and Lee, S.-I. (2019). Consistent individualized feature attribution for tree ensembles.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Marco, S., Loredana, M., Riccardo, V., Raffaella, B., Walter, C., and Luca, N. (2022). Microbe-assisted crop improvement: a sustainable weapon to restore holobiont functionality and resilience. *Horticulture research*, 9:uhac160.

Meuwissen, T., Hayes, B., and Goddard, M. (2001). Prediction of the total genetic value using genome-wide dense marker maps. *Genetics*.

Molnar, C. (2025). *Interpretable Machine Learning*. 3 edition.

Molnar, C., Bischl, B., and Casalicchio, G. (2018). iml: An r package for interpretable machine learning. *JOSS*, 3(26):786.

Ott, J., Wang, J., and Leal, S. M. (2015). Genetic linkage analysis in the age of whole-genome sequencing. *Nature Reviews Genetics*, 16(5):275–284.

Park, D.-J., Lee, Y.-S., and Choi, Y.-L. (2013). Characterization of a cold-active beta-glucosidase from paenibacillus xylanilyticus kj-03 capable of hydrolyzing isoflavones daidzin and genistin. *The Protein Journal*, 32(7):579–584.

Perdry, H. and Dandine-Roulland, C. (2023). *gaston: Genetic Data Handling (QC, GRM, LD, PCA) and Linear Mixed Models*. R package version 1.6.

Pfeifer, B., Baniecki, H., Saranti, A., Biecek, P., and Holzinger, A. (2022). Multi-omics disease module detection with an explainable greedy decision forest. *Scientific reports*, 12(1):16857.

Phipson, B. and Smyth, G. K. (2016). Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *arXiv preprint arXiv:1603.05766*.

Postles, J., Curtis, T. Y., Powers, S. J., Elmore, J. S., Mottram, D. S., and Halford, N. G. (2016). Changes in free amino acid concentration in rye grain in response to nitrogen and sulfur availability, and expression analysis of genes involved in asparagine metabolism. *Frontiers in Plant Science*, Volume 7 - 2016.

Ramesh, S. A., Tyerman, S. D., Xu, B., Bose, J., Kaur, S., Conn, V., Domingos, P., Ullah, S., Wege, S., Shabala, S., Feijó, J., Ryan, P. R., and Gilliham, M. (2015). Gaba signalling modulates plant growth by directly regulating the activity of plant-specific anion transporters. *Nature Communications*, 6(1):7879.

Sakurai, K., Toda, Y., Kajiya-Kanegae, H., Ohmori, Y., Yamasaki, Y., Takahashi, H., Takanashi, H., Tsuda, M., Tsujimoto, H., Kaga, A., et al. (2022). Time-series multispectral imaging in soybean for improving biomass and genomic prediction accuracy. *The Plant Genome*, 15(4):e20244.

Sauder, L. A., Albertsen, M., Engel, K., Schwarz, J., Nielsen, P. H., Wagner, M., and Neufeld, J. D. (2017). Cultivation and characterization of candidatus nitrosocosmicus exaquare, an ammonia-oxidizing archaeon from a municipal wastewater treatment system. *The ISME Journal*, 11(5):1142–1157.

Shen, X., Wang, C., Zhou, X., Zhou, W., Hornburg, D., Wu, S., and Snyder, M. P. (2024). Nonlinear dynamics of multi-omics profiles during human aging. *Nature Aging*, 4(11):1619–1634.

Shiu, S.-H. and Bleecker, A. B. (2001). Plant receptor-like kinase gene family: Diversity, function, and signaling. *Science's STKE*, 2001(113):re22–re22.

Singh, R. P., Kumari, K., Sharma, P. K., et al. (2023). Characterization and in-depth genome analysis of a halotolerant probiotic bacterium paenibacillus sp. s-12, a multifarious bacterium isolated from *rauvolfia serpentina*. *BMC Microbiology*, 23:192.

Tavladoraki, P., Cona, A., and Angelini, R. (2016). Copper-containing amine oxidases and fad-dependent polyamine oxidases are key players in plant tissue differentiation and organ development. *Frontiers in Plant Science*, Volume 7 - 2016.

Thompson, R. M., del Carmen Montero-Calasanz, M., George, D., and Fox, E. M. (2025). From pollution to reforestation: the hidden microbiome of alnus glutinosa nodules over 30 years. *Scientific Reports*, 15(1):23373.

Toda, Y., Sasaki, G., Ohmori, Y., Yamasaki, Y., Takahashi, H., Takanashi, H., Tsuda, M., Kajiya-Kanegae, H., Lopez-Lozano, R., Tsujimoto, H., et al. (2022). Genomic prediction of green fraction dynamics in soybean using unmanned aerial vehicles observations. *Frontiers in Plant Science*, 13:828864.

Ulrich, D. E. M., Sevanto, S., Ryan, M., Albright, M. B. N., Johansen, R. B., and Dunbar, J. M. (2019). Plant-microbe interactions before drought influence plant physiological responses to subsequent severe drought. *Scientific Reports*, 9(1):249.

Valliyodan, B., Cannon, S. B., Bayer, P. E., Shu, S., Brown, A. V., Ren, L., Jenkins, J., Chung, C. Y.-L., Chan, T.-F., Daum, C. G., Plott, C., Hastie, A., Baruch, K., Barry, K. W., Huang, W., Patil, G., Varshney, R. K., Hu, H., Batley, J., Yuan, Y., Song, Q., Stupar, R. M., Goodstein, D. M., Stacey, G., Lam, H.-M., Jackson, S. A., Schmutz, J., Grimwood, J., Edwards, D., and Nguyen, H. T. (2019). Construction and comparison of three reference-quality genome assemblies for soybean. *The Plant Journal*, 100(5):1066–1082.

VanRaden, P. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, 91(11):4414–4423.

Verma, P., Yadav, A., Kazy, S., Saxena, A., and Suman, A. (2014). Evaluating the diversity and phylogeny of plant growth promoting bacteria associated with wheat (triticum aestivum) growing in central zone of india. *Int J Curr Microbiol Appl Sci*, 3:432–447.

Wang, P., Lehti-Shiu, M. D., Lotreck, S., Segura Abá, K., Krysan, P. J., and Shiu, S.-H. (2024). Prediction of plant complex traits via integration of multi-omics data. *Nature Communications*, 15(1):6856.

Wright, M. and Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in c++ and r. *Journal of Statistical Software*, 77.

Xiong, W., Song, Y., Yang, K., Gu, Y., Wei, Z., Kowalchuk, G. A., Xu, Y., Jousset, A., Shen, Q., and Geisen, S. (2020). Rhizosphere protists are key determinants of plant health. *Microbiome*, 8(1):27.

Xu, B., Long, Y., Feng, X., Zhu, X., Sai, N., Chirkova, L., Betts, A., Herrmann, J., Edwards, E. J., Okamoto, M., Hedrich, R., and Gilliham, M. (2021). Gaba signalling modulates stomatal opening to enhance plant water use efficiency and drought resilience. *Nature Communications*, 12(1):1952.

Yoshioka, H., Mary-Huard, T., Aubert, J., Toda, Y., Ohmori, Y., Yamasaki, Y., Tsujimoto, H., Takahashi, H., Nakazono, M., Takanashi, H., Fujiwara, T., Tsuda, M., Kaga, A., Inaba, J., Fuji, Y., Hirai, M. Y., Nose, Y., Kumaishi, K., Usui, E., Kobori, S., Sato, T., Narukawa, M., Ichihashi, Y., and Iwata, H. (2025). Integration of proxy intermediate omics traits into a nonlinear two-step model for accurate phenotypic prediction. *bioRxiv*.

Zhang, C., van der Heijden, M. G. A., Dodds, B. K., Nguyen, T. B., Spooren, J., Valzano-Held, A., Cosme, M., and Berendsen, R. L. (2024). Correction: A tripartite bacterial-fungal-plant symbiosis in the mycorrhiza-shaped microbiome drives plant growth and mycorrhization. *Microbiome*, 12(1):30.

Zhao, A., Li, Q., Meng, P., Liu, P., Wu, S., Lang, Z., Song, Y., and Macho, A. P. (2025). Reduced content of gamma-aminobutyric acid enhances resistance to bacterial wilt disease in tomato. *Plant Biotechnology Journal*, 23(3):792–806.

Zhao, T., Zeng, J., and Cheng, H. (2022). Extend mixed models to multilayer neural networks for genomic prediction including intermediate omics data. *Genetics*, 221(1):iyac034.

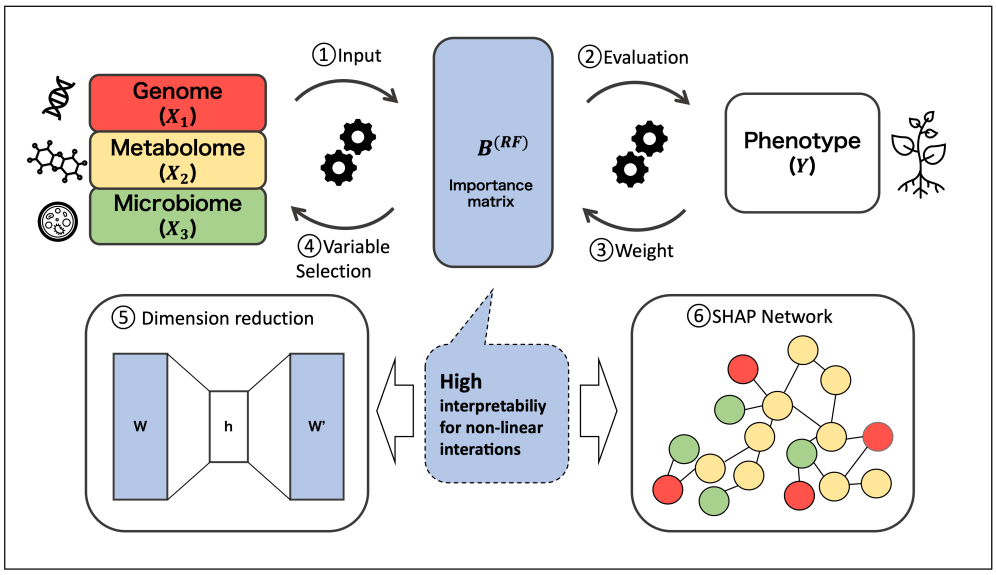

Fig. 1: Concept of our framework. (1) Multi-omics data (Genome, Metabolome, Microbiome) are integrated as a single matrix and introduced into a soybean phenotype prediction model. (2) Phenotype (Y) is predicted using RF. (3) The model is analyzed retrospectively from the output Y to optimize the weights (importance: $B^{(\mathrm{RF})}$) of predictive factors within the omics data. (4) Variable selection is performed based on the nonlinear matrix $B^{(\mathrm{RF})}$, and a new prediction model is built using the selected variables. (5) Relationships among omics layers are interpreted in a low-dimensional latent space. (6) SHAP for network analysis of omics factors.

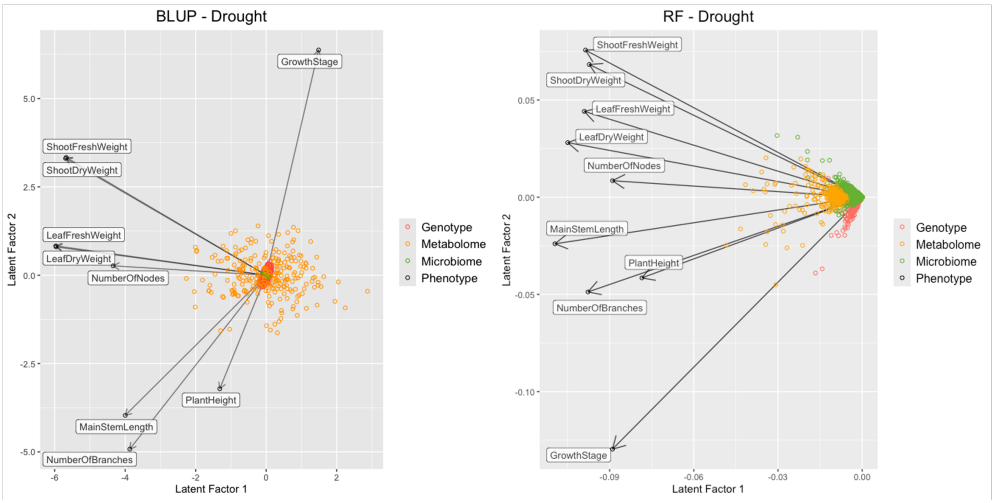

Fig. 2: Feature–trait structures derived from BLUP and RF for predicting biomass-related phenotypes under drought conditions. Feature importance matrices from both models were subjected to singular value decomposition (SVD), and the first two latent factors are visualized. Points represent features and phenotypes projected into the latent space. Colors indicate data types (genotype, metabolome, or microbiome), and phenotype vectors are shown as arrows.

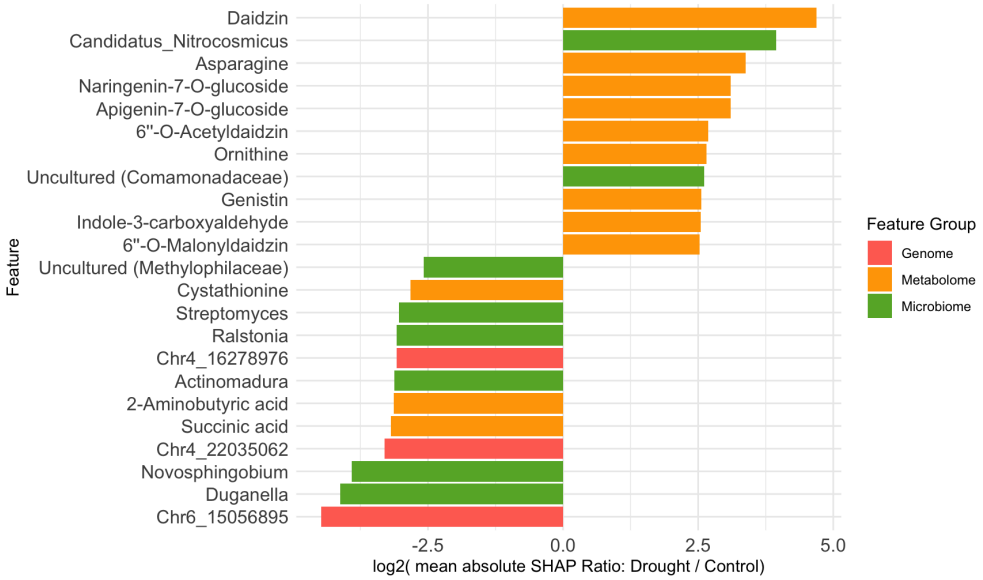

Fig. 3: SHAP-based individual feature importance. For each control and drought condition, absolute SHAP values were computed and their ratios were calculated. The bar plot shows the $\log_2$ ratio of the mean absolute SHAP values (drought divided by control). Features with $\log_2$(mean absolute SHAP ratio) > 2.5 are displayed. Positive values indicate features more specific to drought conditions, whereas negative values indicate features more specific to control conditions.

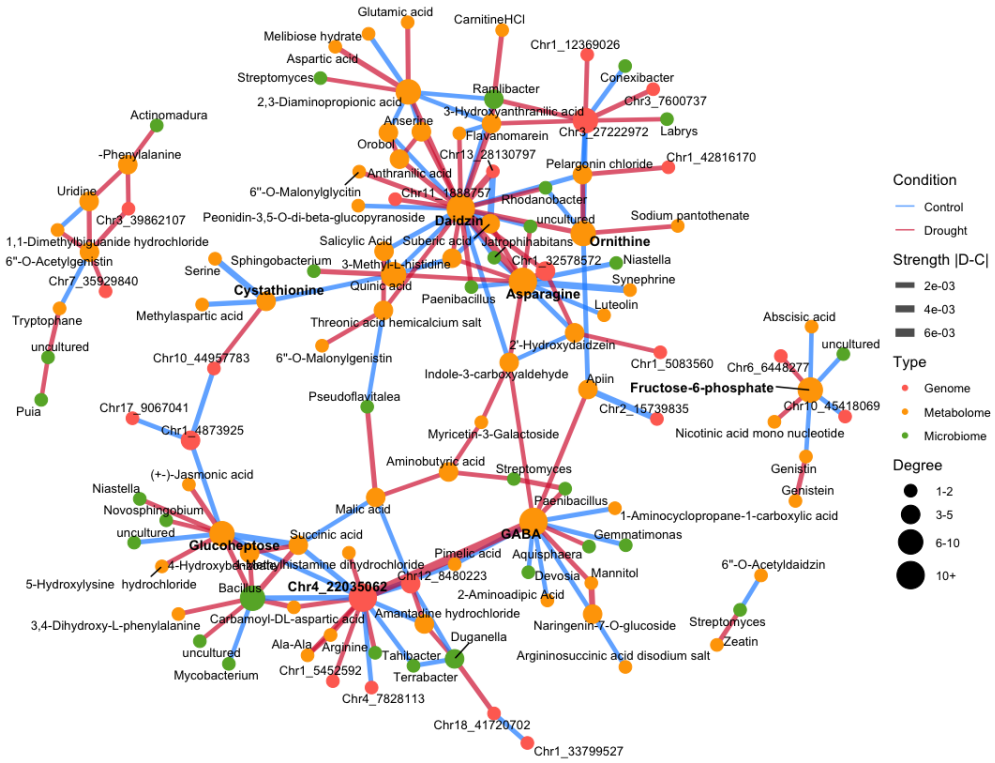

Fig. 4: Condition-specific SHAP interaction difference network. The model was trained separately to predict biomass (leaf dry weight) under drought and control conditions. Environment-specific interactions were identified from the differential SHAP interaction matrix (drought minus control) using Z-scores derived from a pooled null distribution. Only interactions with large absolute Z-scores ($|Z| \geq 30$) were retained. Red edges represent interactions that are stronger under drought conditions, whereas blue edges indicate interactions that are more pronounced under control conditions. Node colors denote omics layers: genome SNP markers (red), microbiome features (green), and metabolome features (orange). Node size reflects the number of connections (degree), edge width corresponds to the magnitude of the interaction difference between drought and control, and highly connected features (hubs) are highlighted in bold font.