Interpretable Machine Learning on Soybean Multi Omics Data Reveals Drought-Driven Shifts of Plant-Microbe Interactions

Hayato Yoshioka[1,2], Pavla Debeljak[2,3], Soizic Prado[2], 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.

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

Conclusion: The application of an interpretable machine learning approach to the plant phenotype prediction framework identifies 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 play a pivotal role 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, 2021), metabolomic (Guo et al., 2023), and microbiome data (Marco, 2022). These inter-omics datasets provide new opportunities to unravel the complex relationships among biological systems. However, integrating and interpreting such diverse data types poses 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). It leverages genome-wide markers to predict phenotypic traits by capturing linear additive genetic contributions. While BLUP can incorporate multiple omics layers simultaneously through multi-kernel approaches, it remains limited to modeling linear relationships among features.

Genome-wide association study (GWAS) has been a classical linear approach for identifying individual SNP markers associated with phenotypic variation (Atwell et al., 2010; Ott et al., 2015). However, GWAS does not integrate multiple omics layers and typically analyzes genome-wide markers in isolation.

In contrast, machine learning approaches such as Random Forest (RF) (Breiman, 2001) are capable of modeling complex nonlinear interactions among multi-omics features. RF is an ensemble method that constructs a collection of decision trees and aggregates their predictions, providing both high predictive performance and interpretable feature importance scores for biological features (Acharjee et al., 2016; Pfeifer, 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 RF model can effectively capture such complexity and can outperform traditional linear models, such as BLUP, in predictive tasks when nonlinearity exists in multi-omics interactions (Yoshioka et al., 2025).

SHAP (SHapley Additive exPlanations) 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 insight into 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 for flowering time (Wang, 2024). In addition, SHAP has been applied to 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 with the aim of identifying interacting biomarkers that bridge plant-microbe interactions. The application of SHAP-based interpretability to uncover such cross-domain interactions in plant-microbe multi-omics datasets remains unexplored, making this approach a novel contribution to the field.

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

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

Materials and Methods

Soybean Multi-Omics Data

We analyzed multi-omics datasets collected from a common panel of $N = 198$ soybean accessions (Figure 1). 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.
- **Rhizosphere metabolome**: a 198×263 matrix of metabolome features extracted from the soybean rhizosphere, with each column representing a metabolome feature detected by tandem mass spectrometry.
- **Rhizosphere microbiota profiles (microbiome)**: a $198 \times 1{,}111$ matrix based on 16S rRNA sequencing of the soybean rhizosphere, where each column represents the relative abundance of a microbial taxon.
- **Plant Phenotype**: a 198×9 matrix including biomass-related traits such as shoot and leaf fresh/dry weights.

Genotypic data were obtained from the global soybean minicore collection (Kajiya-Kanegae, 2021). Microbiome and meta-metabolome data were collected from the same field plots as the phenotypic data to ensure comparability across omics layers. All 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 (Yoshioka et al., 2025). To reduce the computational burden in SHAP analysis of the RF model, SNPs were pruned based on linkage disequilibrium (LD). A representative subset of 3,078 SNP markers (LD threshold = 0.001) was selected from an initial set of 425,858 SNPs (LD threshold = 0.95). Despite the relatively small size of the subset of 3,078 SNPs, they provided approximate coverage of the entire genome (Figure 1S). All features were scaled prior to modeling. See the supplementary data for details on the data used for the analysis.

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).

We have feature matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ and response matrix $\mathbf{Y} \in \mathbb{R}^{n \times q}$. Each omics layer was used to construct a corresponding kernel matrix $\mathbf{K}_l$, capturing pairwise similarities among samples (Hamazaki and Iwata, 2020). The univariate mixed model was performed for each single 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 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 each layer. All coefficient vectors were concatenated across 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 normalized so that:

$$\sum_{k=1}^{p} \mathbf{B}_{\mathbf{k},\mathbf{j}} = 1, \quad \text{for all } j = 1, \ldots, q.$$

To examine the structure of 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, and $\mathbf{D}$ is the diagonal matrix of singular values. 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 same feature matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ and response matrix $\mathbf{Y} \in \mathbb{R}^{n\times q}$ were used.

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 by grid search. See Appendix C and Figure 2S for details.

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

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

Following BLUP, each column of $\mathbf{I}$ was normalized such that:

$$\sum_{k=1}^{p} \mathbf{I_{k,j}} = 1, \quad \text{for all } j = 1, \ldots, q.$$

Analogous to the coefficient matrix $\mathbf{B}$ derived from BLUP, the normalized 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 trees in the forest (Figure 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 visualization of feature-trait associations. Following the BLUP, we can perform the SVD for $\mathbf{B}^{(\mathrm{RF})}$ to visualize the structure of feature-trait relationships.

Genome-wide association study (GWAS)

GWAS is a classical linear approach for identifying SNP markers associated with phenotypic variation (Atwell et al., 2010; Ott et al., 2015). Although GWAS does not integrate multi-omics layers, it provides 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) (version 1.6), incorporating three principal components (PCs) derived from the eigen decomposition of the genomic relationship matrix (GRM) to account for population structure.

The full set of SNPs (425,858 markers, LD threshold = 0.95) was filtered by 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.

SHAP: individual-features Attribution

To investigate the individual contributions of multi-omics features to phenotype prediction (hereafter referred to as individual-features contributions), we first applied SHapley Additive exPlanations (SHAP) to the trained RF models. 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. This enables an interpretable assessment of nonlinear feature importance.

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

$$\phi_k(f, x) = \sum_{S \subseteq N \setminus \{k\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left[f_{S \cup \{k\}}(x_{S \cup \{k\}}) - f_S(x_S) \right], \tag{1}$$

where N is the set of all input features, S is a subset that excludes feature k, and f is the prediction function. This formulation ensures that each feature's contribution is assessed fairly across all possible contexts.

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. For interpretability and computational feasibility, we selected leaf dry weight as a representative phenotype and focused on RF-selected feature set.

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

$$i_k = \sum_{j=1}^{t} \mathbf{B}^{(\mathrm{RF})}{}_{k,j}$$

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

Here, the features with a total RF importance score $i_k > 0.01$ under either drought or control conditions. This yielded a total of 159 features for evaluations.

Using this subset, 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, then averaged. The ratio of the mean absolute SHAP values (drought divided by control).

SHAP: Interaction-Features Attribution

While SHAP individual-features attribution values quantify each omics feature, they do not capture interaction effects. To address this, we further applied SHAP interaction analysis to characterize pairwise relationships among features across omics layers.

SHAP interaction values decompose the prediction into individual effects and pairwise interactions. For this analysis, we constructed feature interaction networks where nodes represent individual-features and edges denote the strength of interaction between feature pairs. A larger feature set was used (features with $i_k > 0.001$) under either drought or control conditions, to capture a broader view of potential interactions. This yielded a total of 2,729 features for evaluations.

Interaction values were computed using `shap.TreeExplainer` from the `shap` Python package, applied to an `xgboost`-based gradient boosting model (Chen and Guestrin, 2016). See also Appendix C and Figure 2S for implementation and hyperparameter tuning details. SHAP interaction values were calculated for each sample and averaged to obtain a global $2,729 \times 2,729$ symmetric interaction matrix for each condition. To enhance interpretability, values were transformed using $\log_{10}(x+1)/2$, and negative values were set to zero.

To highlight environment-specific interactions, we computed a differential SHAP interaction matrix:

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

where $\mathbf{S}$ denotes the SHAP-based pairwise interaction matrix. Positive values in $\Delta\mathbf{S}$ indicate interactions that are stronger under drought stress, while negative values reflect interactions that are more pronounced under control conditions. To visualize the most substantial differences, we selected the top 100 interactions based on the absolute values of $\Delta\mathbf{S}$, representing approximately the top 0.001% of all possible interactions.

Together, 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.

Results

RF and BLUP Models

Figure 2 presents the SVD of the coefficient matrix of BLUP and the importance matrix of RF under drought conditions. The results demonstrate that the RF model shows a more pronounced and variable pattern of feature importance compared to the BLUP model. Notably, in BLUP, metabolites are the features most correlated with phenotypes, while other features (genome, microbiome) exhibit a shrunk pattern, indicating that metabolome features primarily contribute to phenotypic prediction. Under control conditions, a similar trend was observed (Figure 3S).

SHAP analysis of individual-features

SHAP analysis identified key condition-specific features contributing to phenotypic variations under drought and control conditions (Figure 3).

Under drought stress, the most highly contributing features were predominantly flavonoids, including isoflavones (e.g., daidzin and its derivatives, genistin) and flavone/flavanone glucosides (e.g., apigenin-7-O-glucoside, naringenin-7-O-glucoside). Flavonoids are known to play a major role in responding 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 the derivatives (asparagine, ornithine, and indole-3-carboxyaldehyde) as major contributing features.

Furthermore, *Candidatus Nitrosocosmicus*—a clade of ammonia-oxidizing archaea (AOA)—was particularly selected under drought conditions. 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), which is consistent with our findings.

In contrast, under control conditions, genetic markers made a particularly strong contribution to the observed traits. SHAP re-identified the SNP `Chr06-15056895` as an important feature, which was also detected by GWAS analysis (Figure 5S), which encodes a copper amine oxidase family protein involved in amine metabolism, tissue differentiation, and organ development in plants (Tavladoraki et al., 2016). A SNP `Chr04-22035062`, which is located near the gene *Glyma.04G140400* was also selected. This gene encodes an oxidoreductase enzyme, 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. Several species of *Duganella* are known for their plant growth-promoting properties, including the solubilization 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, as their levels were similarly elevated under both conditions. Further details for each condition are presented in Figure 4S.

SHAP analysis of interaction-features

SHAP interaction values enabled a detailed investigation of multi-omics feature interactions. The top 100 condition-specific connectivity patterns were identified by comparing network differences between drought and control conditions (Figure 4). For each condition, we identified major features based on their connections with other features, such as GABA and daidzin, both known for their roles in promoting plant growth, influencing rhizosphere microbiome compositions, 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).

Several of these interactions correspond to known biological relationships. Notably, the model identified a strong interaction between daidzin and the bacterium *Paenibacillus* under drought conditions. This is supported by a previous finding showing that certain strains of *Paenibacillus* produce β-glucosidases, which catalyze 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 into 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 succinic acid and in interactions with microbial taxa such as *Candidatus Berloella*, and *Terrabacter*. See also Figure 4S, Table 1S, 2S, 3S for the details of the results for each condition.

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 (Figure 2, 3S). In contrast, BLUP showed more shrinking effects, which is typical when the explanatory matrix contains a large number of 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 RF highlights the potential synergy between traditional and machine learning-based approaches (Figure 5S). For instance, SNP `Chr06-15056895`, located near the gene *Glyma.06G178400*, was detected by both RF and GWAS, reinforcing its putative role in growth-related traits. Such convergence enhances confidence in candidate biomarkers.

Our findings demonstrate a clear shift in the dominant omics layers under drought stress, highlighting the enhanced roles of metabolomic and microbial features (Figure 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 it is known that amino acid concentrations increase under drought conditions (Postles et al., 2016). Additionally, the selection of *Candidatus Nitrosocosmicus*, known as a major contributor to phenotypic variation specifically 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 gave further insights by uncovering plausible plant-metabolic-microbial links (Figure 4). For example, daidzin, which was selected as an individual

feature, was also selected as a feature of pairwise interactions. Especially, the co-selection of daidzin and *Paenibacillus* aligns with reports that *Paenibacillus* strains produce β-glucosidases to hydrolyze isoflavone glucosides such as daidzin (Park et al., 2013). This enzymatic synergy may enhance nutrient availability and ROS detoxification under drought. Additionally, the inferred interaction between *Paenibacillus* and GABA is consistent with the bacterium's ability to produce glutamate decarboxylase (Singh et al., 2023), which could further support plant stress resilience. Interestingly, while GABA was not selected as a condition-specific feature based on individual SHAP contributions, network-level analysis revealed its 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 confirming these known interactions, the model also suggests other possible connections between plant secondary metabolites and the microbiome that have not yet been directly reported. For example, daidzin shows an association with bacteria such as *Rhodanobacter* (Singh et al., 2023), known to promote plant growth. GABA is indirectly connected through *Paenibacillus* with a cluster of microbiome taxa, including *Devosia*—reported as a plant growth–promoting genus (Zhang et al., 2024)—as well as *Niastella* and *Aquisphaera*, which are frequently detected in the plant rhizosphere, although their specific roles are not well understood (Akter et al., 2021; Thompson et al., 2025). These findings highlight the potential for elucidating previously unknown plant-microbe interactions, where 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 revealed contrasting patterns of feature associations under drought and control conditions. Under drought stress, metabolome features were particularly prominent in the modeling, suggesting scenarios where 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 known to influence plant stress responses (Postles et al., 2016). This highlights the importance of metabolite co-regulation and network-level interactions under 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—i.e., genetic effects outweighing environmental influences. This is supported by the identification of genotypic markers near genes involved in signal transduction, amine metabolism, and development.

The integrated framework developed in this study enabled feature selection across multiple omics layers, including plant genomes, and successfully uncovered biologically meaningful interactions. In future work, we aim to extend this approach to incorporate larger untargeted metabolomics datasets and microbiome data covering a broader range of organisms, including fungi. Expanding the framework in this way will allow us to address wider ecological and ecosystem-level questions and gain 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 data. By comparing BLUP and GWAS with a nonlinear machine learning model (random forest) across genomic, metabolomic, and microbiomic layers, we confirmed that the RF model captured additional nonlinear interactions, avoiding shrinking effects observed in linear methods. The use of SHAP further ensured that these nonlinear patterns are biologically interpretable and not model artifacts. As a result, 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 across 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: Yushiro Fuji, Masami Hirai Yokota, Yasunori Ichihashi, 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

All codes and data are available on GitHub (https://github.com/Yoska393/ShapPlantMicro). The metabolome data are available from the RIKEN DropMet website (http://prime.psc.riken.jp/menta.cgi/prime/drop_index; ID: DM0071, DM0072).

Contributions

H.Y., P.D., S.P, and H.I. conceived the study. H.Y. analysed 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.

Akter, S., Park, J.-H., Rahman, M. M., and Huq, M. A. (2021). Niastella soli sp. nov., isolated from rhizospheric soil of a persimmon tree. *International Journal of Systematic and Evolutionary Microbiology*, 71(7).

Alemu, A., Åstrand, J., Montesinos-López, O., Sánchez, J., Fernández-Gónzalez, J., Tadesse, W., and Vetukuri, R. e. a. (2024). Genomic selection in plant breeding: key factors shaping two decades of progress. *Molecular Plant*, 17:552–578.

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.

Callahan, B., McMurdie, P., Rosen, M., Han, A., Johnson, A., and Holmes, S. (2016). Dada2: High-resolution sample inference from illumina amplicon data. *Nature Methods*, 13(7):581–583.

Caporaso, J., Lauber, C., Walters, W., Berg-Lyons, D., Lozupone, C., Turnbaugh, P., Fierer, N., and Knight, R. (2011). Global patterns of 16s rrna diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences*, 108(Supplement 1):4516–4522.

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., and Iwata, H. (2023). Integrative framework for stochastic variational variable selection for joint analysis of multi-omics microbiome data. *bioRxiv*. Preprint.

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.

Ichihashi, Y. e. a. (2020). Multiomics analysis on an agroecosystem reveals the significant role of organic nitrogen to increase agricultural crop yield. *Proceedings of the National Academy of Sciences*, 117(25):14552–14560.

Kaga, A. e. a. (2012). Evaluation of soybean germplasm conserved in nias genebank and development of mini core collections. *Breeding Science*, 61(5):566–592.

Kajiya-Kanegae, H. e. a. (2021). Whole-genome sequence diversity and association analysis of 198 soybean accessions in mini-core collections. *DNA Research*, 28(1):032.

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

Kumaishi, K. e. a. (2022). High through-put method of 16s rrna gene sequencing library preparation for plant root microbial community profiling. *Scientific Reports*, 12(1):19289.

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, D. e. a. (2013). Practical innovations for high-throughput amplicon sequencing. *Nature Methods*, 10(10):999–1000.

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. e. a. (2022). Microbe-assisted crop improvement: a sustainable weapon to restore holobiont functionality and resilience. *Horticulture Research*.

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

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. e. a. (2022). Multi-omics disease module detection using an explainable greedy decision forest. *Scientific Reports*, 12:16857.

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.

Quast, C. e. a. (2012). The silva ribosomal rna gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, 41(D1):590–596.

R Core Team (2023). R: Language and environment for statistical computing.

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. e. a. (2022). Time-series multispectral imaging in soybean for improving biomass and genomic prediction accuracy. *Plant Genome*, 15(4):20244.

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.

Sawada, Y. e. a. (2009). Widely targeted metabolomics based on largescale ms/ms data for elucidating metabolite accumulation patterns in plants. *Plant Cell Physiology*, 50(1):37–47.

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. e. a. (2022). Genomic prediction of green fraction dynamics in soybean using unmanned aerial vehicles observations. *Frontiers in Plant Science*, 13.

Toju, H. and Baba, Y. (2018). Dna metabarcoding of spiders, insects, and springtails for exploring potential linkage between above- and below-ground food webs. *Zoological Letters*, 4:1–12.

Tsugawa, H., Kanazawa, M., Ogiwara, A., and Arita, M. (2014). Mrmprobs suite for metabolomics using large-scale mrm assays. *Bioinformatics*, 30(16):2379–2380.

Uchida, K. e. a. (2020). Identification of a unique type of isoflavone o-methyltransferase, gmiomt1, based on multi-omics analysis of soybean under biotic stress. *Plant and Cell Physiology*, 61(11):1974–1985.

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.

Van Rossum, G. and Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.

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. e. a. (2024). Prediction of plant complex traits via integration of multi-omics data. *Nature Communications*, 15:6856.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

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.

Yilmaz, P. e. a. (2014). The silva and "all-species living tree project (ltp)" taxonomic frameworks. *Nucleic Acids Research*, 42(D1):643–648.

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.

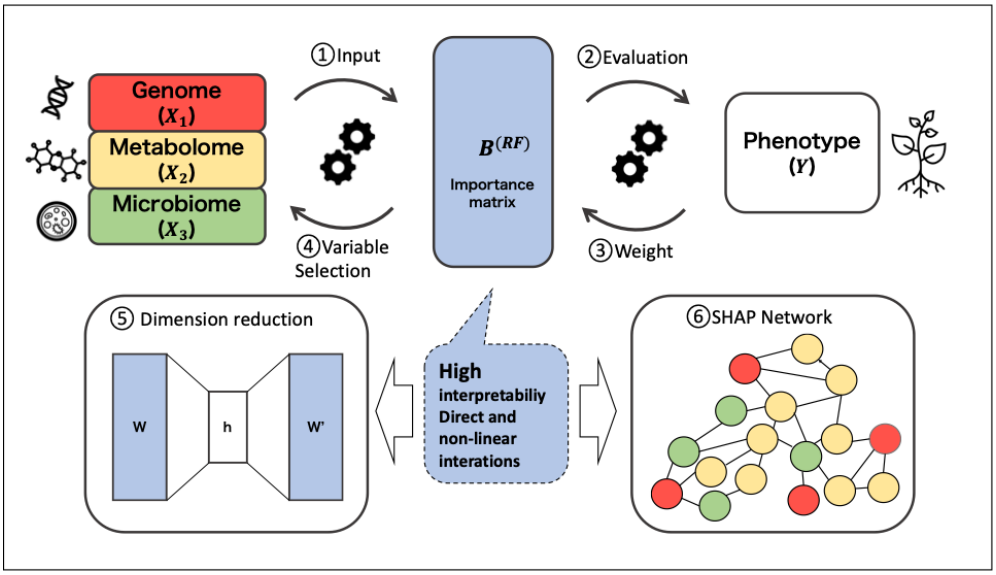

Fig. 1: Concept of our approach. (1) Multi-omics data (Genome, Metabolome, Microbiome) are integrated as a single matrix and introduced into a soybean phenotype prediction model. (2) Phenotype (Y) is predicted using regression (Random Forest), and prediction accuracy is evaluated in comparison with conventional GP models. (3) The model is analyzed retrospectively from the output Y to optimize the weights ($B^{(\mathrm{RF})}$) of predictive factors within the omics data. (4) Variable selection is performed based on the non-linear weight 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 for interpretation. (6) SHAP (SHapley Additive exPlanations) is used for network analysis of omics factors.

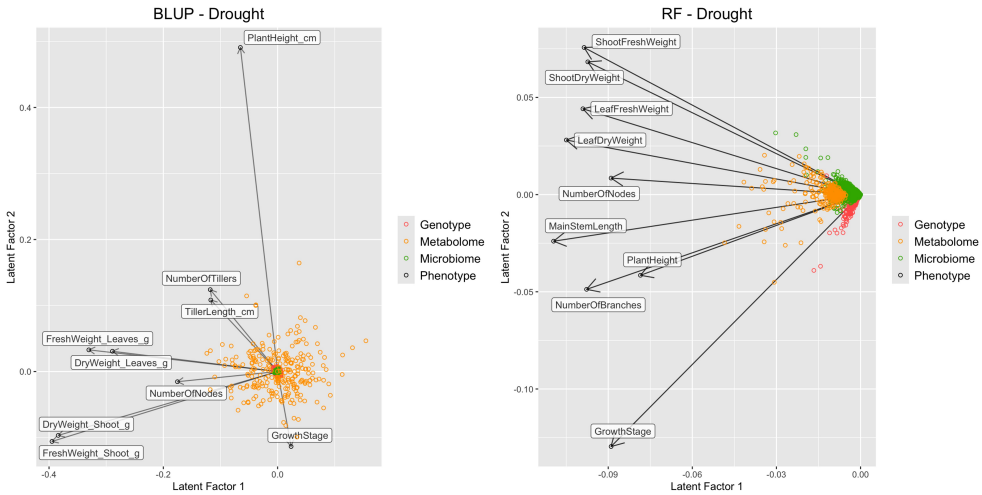

Fig. 2: Feature-trait structures derived from BLUP and RF in predicting biomass related phenotypes. The model was performed under drought conditions. Feature importance matrices from both models were subjected to singular value decomposition (SVD), and the first two latent factors are visualized. Points represent features and phenotypes projected into the latent space. Color indicates data type: genotype, metabolome, or microbiome. Vectors for phenotypes are shown as arrows.

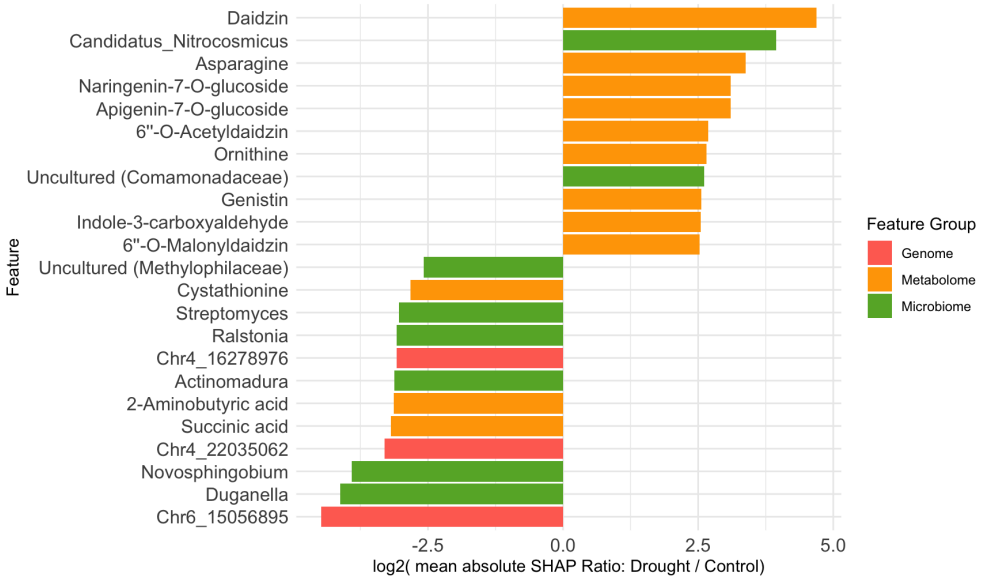

Fig. 3: SHAP individual feature importance. For each control/drought condition, the absolute SHAP values were computed, and their ratio was taken. The bar plot shows the $\log_2$ ratio of mean absolute SHAP values (drought divided by control). Features with $\log_2$(mean absolute SHAP ratio) > 2.5 are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

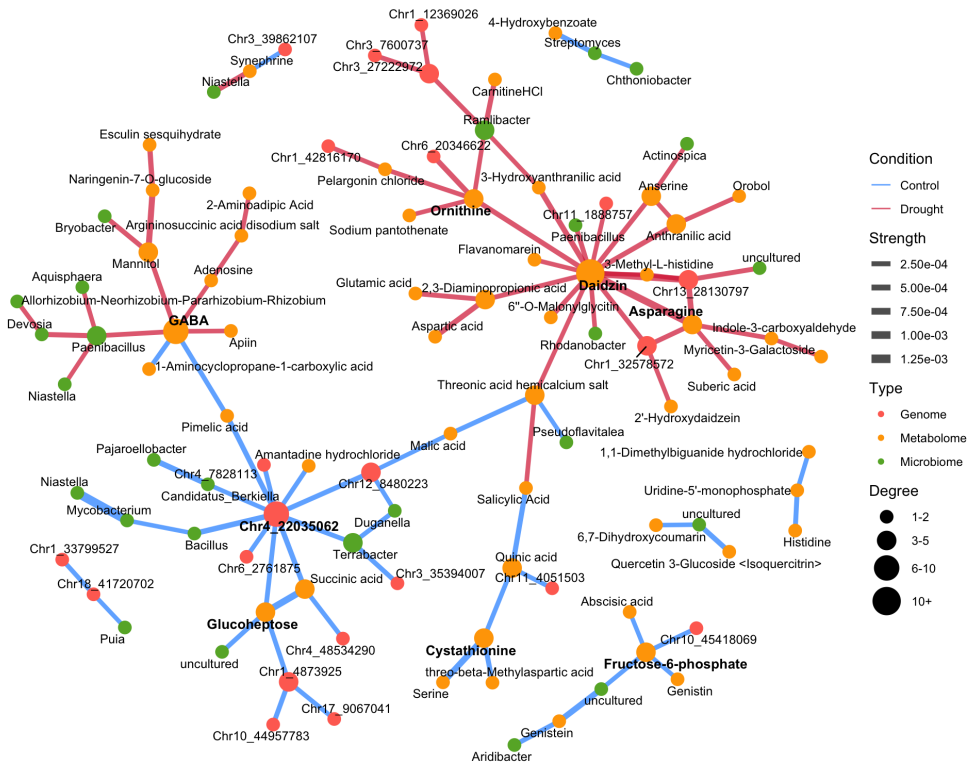

Fig. 4: Condition-Specific SHAP Interaction Difference Network. The model was built to predict biomass (leaf dry weight) under drought and control conditions separately. The top 100 condition-specific interactions were then identified by calculating the difference (drought minus control) between the two networks. Red edges represent interactions that are stronger under drought conditions, while blue edges represent those stronger under control conditions. Node colors indicate omics layers: genome SNP marker (red), microbiome feature (green), metabolome feature (orange). Node size reflects the number of connections, and edge width corresponds to the magnitude of interaction difference between drought and control. The most connected features (hubs) are highlighted in bold font.