An Integrated Statistical and Machine Learning Pipeline (FAME) for Factorial Plant Experiments

Mohammad B. Bagherieh-Najjar

m.b.bagherieh@gmail.com

Golestan University

Raheleh Daqeliyeh

Golestan University

Hamidreza Sadeghipour

Golestan University

Method Article

Keywords: analytical pipeline, machine learning, SHAP, Random Forest, ANOVA, factorial experiment, plant stress, Streamlit

Posted Date: June 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9869936/v1

Additional Declarations: No competing interests reported.

Abstract

Background

Factorial experiments are fundamental to plant stress physiology, yet their analysis typically relies on fragmented workflows spanning multiple software packages. Analysis of variance (ANOVA) identifies significant effects but does not quantify relative factor importance or reveal multivariate trait relationships. Machine learning methods offer complementary insights, but no integrated, reproducible pipeline combining classical statistics with explainable machine learning has been formalized for factorial plant stress experiments.

Results

We present a comprehensive, open-source Python analytical pipeline with a Streamlit web interface that integrates assumption diagnostics, N-way ANOVA with partial eta-squared effect sizes, Kruskal-Wallis non-parametric validation, post-hoc mean comparisons with compact letter display via R's agricolae package, SHAP and permutation feature importance, cross-validated model comparison (Random Forest, Linear Regression, Gaussian Process Regression), bootstrap confidence intervals for standardized regression coefficients, Gaussian Process Regression response surfaces, Multi-Output Random Forest residual correlation analysis, and Principal Component Analysis. All figures are automatically generated at 600 DPI in both PNG and PDF formats with configurable single/two-column A4 layouts. A comprehensive text report is produced for manuscript preparation. The pipeline is validated on three independent factorial datasets: (i) a 2-factor chitosan × water deficit experiment in Viola spp. (n = 36, 7 responses), (ii) a 3-factor irrigation × genotype × silicon experiment in *Vicia faba* (n = 120, 12 responses; originally published in Neyestani et al., 2025), and (iii) a 4-factor mineral nutrition experiment on shikonin production in *Onosma dichroantha* callus (n = 105; originally published in Ghazagh et al., 2023). Across all datasets, the pipeline reproduced published ANOVA results while providing additional insights: quantitative factor importance partitioning via SHAP, identification of context-dependent interactions invisible to main-effects analysis, detection of severe Gaussian Process overfitting (CV R^2 as low as − 56.8) despite excellent training fit, and multivariate trait coupling via residual correlation analysis. Random Forest was the most robust algorithm across all datasets.

Conclusions

The pipeline provides a standardized, reproducible framework that extracts mechanistic insights beyond conventional ANOVA from factorial plant stress experiments. The code which is available from GitHub is adaptable to any factorial design with 2–5 factors and unlimited response variables.

Background

Factorial experiments—in which two or more treatment factors are varied simultaneously—constitute the backbone of experimental plant biology, enabling detection of interactions invisible to single-factor designs [1]. In plant stress physiology, factorial designs are essential because plants integrate signals from multiple coincident environmental variables and ameliorative treatments [2, 3]. Despite sophisticated experimental designs, analytical methods have remained largely static for decades, centered on ANOVA followed by post-hoc mean comparisons.

ANOVA has fundamental limitations. It reports statistical significance but does not quantify the relative magnitude of factor importance; a factor with a small but consistent effect may yield a low p-value without being the dominant driver of phenotypic variation [4]. ANOVA treats each response variable in isolation, obscuring multivariate relationships and trade-offs [5]. Post-hoc letter display algorithms are complex to implement correctly, and errors in letter assignment are documented in published literature [6, 7].

The past decade has witnessed extensive machine learning (ML) applications in plant science, spanning phenotyping [8, 9], genomic prediction [10], stress classification [11], and yield forecasting [12]. However, adoption in factorial experiments remains limited because: (i) sample sizes are modest (n = 24–120), raising overfitting concerns for flexible methods [13]; (ii) many ML methods are "black boxes" without mechanistic interpretability [14]; and (iii) no standardized workflow integrates ML with traditional ANOVA.

Recent advances in explainable artificial intelligence address the interpretability challenge. SHAP (SHapley Additive exPlanations) values decompose predictions into additive feature contributions with strong theoretical guarantees [15, 16]. Permutation importance provides independent non-parametric validation [17]. Together with standardized regression coefficients and bootstrap confidence intervals, these methods enable quantitative factor partitioning beyond ANOVA.

The analytical workflow for factorial experiments currently involves fragmented tools: statistical packages for ANOVA, separate scripts for machine learning, and manual figure preparation. This is error-prone, difficult to reproduce, and inhibits methodological triangulation. We present an integrated pipeline that unifies the entire workflow within a single executable script, accessible through a web interface requiring no programming knowledge.

We validate the pipeline on three independent datasets spanning different experimental designs (2-, 3-, and 4-factor), species (Viola spp., *Vicia faba*, *Onosma dichroantha*), and response types (growth, pigments, enzymes, secondary metabolites), demonstrating generalizability and revealing insights beyond the original publications.

Implementation

Pipeline Architecture

The pipeline is implemented in Python (~ 1,800 lines, analyzer_core.py) with wrapper scripts for 2–5 factor designs and a Streamlit web interface. The workflow proceeds through eleven sequential stages:

Stage 1: Data ingestion. CSV upload with automatic column matching, missing value detection (NA, na, N/A), and categorical/continuous factor detection.

Stage 2: Assumption diagnostics. Shapiro-Wilk normality test, Levene's test for homogeneity of variances, multi-method outlier detection (overall IQR, modified Z-score).

Stage 3: Pearson correlations. Factor–response linear correlations with significance testing.

Stage 4: N-way ANOVA. Type II sums of squares with partial eta-squared (η^2) effect sizes. Continuous factors are binned into user-specified categories; categorical factors are detected automatically. All two-way interactions are included for ≥ 2 factor designs.

Stage 5: Kruskal-Wallis validation. Rank-based non-parametric alternative reported alongside parametric results with agreement assessment.

Stage 6: Post-hoc comparisons. Compact letter display via R's agricolae package (Tukey's HSD or Fisher's LSD) called as a subprocess. This ensures correct letter assignment using the gold-standard implementation [6]. A Python fallback is provided if R is unavailable.

Stage 7: SHAP feature importance. Random Forest regression (200 estimators, max_depth=None, min_samples_leaf = 2) with TreeExplainer SHAP values normalized to percentage importance. Beeswarm plots visualize directional impacts.

Stage 8: Permutation validation. Twenty-repeat permutation feature importance with standard deviation.

Stage 9: Model comparison. Five-fold cross-validated R^2 for Random Forest, Linear Regression, and Gaussian Process Regression (RBF kernel + WhiteNoise), with best model identification.

Stage 10: Advanced analyses. Standardized β coefficients with 95% bootstrap confidence intervals (1,000 iterations); Multi-Output Random Forest (MORF) with residual correlation matrix; PCA with biplot (treatment groups + response loadings); GPR interaction response surfaces.

Stage 11: Report generation. Comprehensive text report with all tables, statistical outputs, and interpretive summaries.

All analyses use a fixed random seed (42) for reproducibility. Parameters are configurable via the sidebar: ANOVA binning, cross-validation folds, bootstrap iterations, Random Forest estimators, and per-figure layout selection.

Figure Generation

Eight figures are automatically generated at 600 DPI in PNG and PDF formats: (1) SHAP importance rankings, (2) SHAP beeswarm plots, (3) GPR contour surfaces, (4) MORF residual correlation matrix, (5) standardized β coefficients with bootstrap CI, (6) model performance comparison, (7) PCA biplot, and (S1) permutation importance. Font sizes (8–10 pt) are optimized for A4 readability. Figure titles are offset vertically for cropping. Layout width (~ 85 mm single-column, ~ 170 mm two-column) is independently configurable per figure.

Dependencies

Python 3.8+, streamlit, numpy, pandas, matplotlib, scipy, statsmodels, scikit-learn, shap. R (≥ 4.0) with agricolae is optional for post-hoc tests.

Results

Demonstration 1: Two-Factor Chitosan × Water Deficit in Viola spp.

The Viola experiment (n = 36) investigated chitosan foliar application (0, 1, 1.5, 2% w/v) under water deficit (100, 75, 50% FC) with seven responses. The pipeline's diagnostics confirmed that four responses met normality assumptions, with three non-normal distributions identified (total chlorophyll, chla_b_ratio, CAT, POD; Shapiro-Wilk $p < 0.05$) and all meeting homogeneity of variance (Levene's $p > 0.05$). Three CAT outliers were detected at the 50% FC × 2% chitosan treatment—the highest-stress, highest-elicitor combination—and were retained as genuine biological responses.

Functional partitioning of trait regulation. The pipeline's triangulation of ANOVA, SHAP, and permutation importance revealed a striking functional architecture (Figure.1 and Table 1).

Chitosan dominated growth traits (SHAP: fresh weight 83.5%, dry weight 80.4%; permutation: 81.6%, 76.0%), while water deficit dominated photosynthetic pigments (SHAP: chla_b_ratio 92.2%, total chlorophyll 91.3%). Catalase exhibited near-equal dual regulation (SHAP: water deficit 50.9%, chitosan 49.1%; permutation ranked chitosan first at 52.0%). This four-method convergence—spanning ANOVA p-values, η^2, SHAP, and permutation importance—constitutes unusually robust triangulation.

Table 1
Multi-method factor importance comparison for Viola spp.

Response	SHAP WD (%)	SHAP CH (%)	Perm WD (%)	Perm CH (%)	β WD	β CH	Dominant
Fresh weight	16.5	83.5	18.4	81.6	−0.019	+ 0.711***	Chitosan
Dry weight	19.6	80.4	24.0	76.0	−0.050	+ 0.663***	Chitosan
chla_b_ratio	92.2	7.8	98.2	1.8	+ 0.493**	−0.048	Water deficit
Total Chl	91.3	8.7	97.8	2.2	−0.944***	+ 0.080	Water deficit
CAT	50.9	49.1	48.0	52.0	+ 0.483**	+ 0.557***	Equal
POD	74.5	25.5	86.7	13.3	+ 0.904***	+ 0.288***	Water deficit
Protein	67.3	32.7	71.0	29.0	+ 0.748***	+ 0.425***	Water deficit

Model performance and overfitting detection. Five-fold cross-validation (Table 2 and Fig. 2) revealed excellent RF performance for POD (R^2 = 0.898 ± 0.063) and good performance for total chlorophyll (R^2 = 0.782 ± 0.377). Critically, GPR produced severely negative CV R^2 for CAT (− 9.281), POD (− 5.471), and protein (− 56.759) despite training R^2 > 0.99 for all traits—catastrophic overfitting invisible without cross-validation.

Table 2
Cross-validated model performance for Viola spp. (5-fold CV R^2 ± SD).

Response	RF	LR	GPR	Best
Fresh weight	0.446 ± 0.177	0.400 ± 0.154	0.115 ± 0.133	RF
Dry weight	0.285 ± 0.274	0.292 ± 0.319	0.063 ± 0.212	LR
chla_b_ratio	0.571 ± 0.312	−0.026 ± 0.483	0.667 ± 0.228	GPR
Total Chl	0.782 ± 0.377	0.184 ± 1.420	0.991 ± 0.004	GPR
CAT	0.534 ± 0.230	0.367 ± 0.308	−9.281 ± 3.164	RF
POD	0.898 ± 0.063	0.855 ± 0.040	−5.471 ± 2.217	RF
Protein	0.635 ± 0.240	0.382 ± 0.464	−56.759 ± 21.872	RF

Multivariate relationships. MORF residual correlation analysis identified two strong couplings beyond treatment effects: fresh weight–dry weight (r = 0.635) and total chlorophyll–POD (r = 0.537). PCA

revealed that the first two components explained 78.7% of variance, with PC1 (49.2%) reflecting antioxidant and protein responses to water deficit and PC2 (29.5%) capturing growth variation driven by chitosan (Fig. 3).

Demonstration 2: Three-Factor Silicon × Irrigation × Genotype in Vicia faba

This dataset (n = 120, 12 responses) was originally analyzed using conventional ANOVA by Neyestani et al. [18], who reported that water deficit reduced growth and RWC while increasing chlorophyll, carotenoids, and GPX activity, and that silicon effects were "minimal and genotype-dependent."

Reanalysis with the pipeline. The pipeline reproduced all published significant effects while providing additional quantitative insights (Supplementary Table S1). SHAP analysis confirmed Irrigation as the dominant factor (62.1–87.7% across responses), with Genotype contributing 5.1–27.8% and Silicon 7.2–15.2%. Permutation importance independently validated these rankings.

Context-dependent silicon interactions. While silicon main effects were non-significant for 8 of 12 responses—consistent with the original conclusion of "minimal effects"—the pipeline identified 8 significant silicon-containing interactions across 5 responses that were not highlighted in the original ANOVA: Irrigation × Silicon was significant for hs ($p = 0.032$, $\eta^2 = 0.063$), fr ($p = 0.003$, $\eta^2 = 0.103$), ft ($p = 0.043$, $\eta^2 = 0.058$), dr ($p = 0.005$, $\eta^2 = 0.096$), dt ($p = 0.045$, $\eta^2 = 0.057$), and chlt ($p = 0.005$, $\eta^2 = 0.095$). Genotype × Silicon was significant for fr ($p = 0.003$, $\eta^2 = 0.141$), dr ($p = 0.019$, $\eta^2 = 0.105$), dt ($p = 0.002$, $\eta^2 = 0.147$), and chlt ($p = 0.034$, $\eta^2 = 0.093$). These context-dependent interactions—invisible to main-effects-only interpretation—indicate that silicon's efficacy depends on both genotype and water availability.

GPR overfitting recurrence. As in the Viola dataset, GPR produced severely negative CV R^2 for hs (−15.008) and gp (− 6.418) despite excellent training fit, establishing this as a systematic limitation of kernel-based methods at moderate sample sizes rather than an artifact of a single dataset.

Demonstration 3: Four-Factor Mineral Effects on Shikonin accumulation in Onosma dichroantha

This dataset (n = 105) originally analyzed by Ghazagh et al. [19] using statistical modeling and multivariate regression investigated Cu, NO3, Ca, and Cd effects on shikonin content in callus culture.

Factor importance and interaction dominance. The pipeline revealed a notable pattern: no individual main effect was significant for Cu, NO3, or Ca ($p > 0.05$), while Cd showed a moderate main effect ($p = 0.045$, $\eta^2 = 0.084$). However, five of six two-way interactions were highly significant ($p < 0.01$): Cu × NO3 ($\eta^2 = 0.239$), Cu × Cd ($\eta^2 = 0.167$), NO3 × Ca ($\eta^2 = 0.243$), NO3 × Cd ($\eta^2 = 0.242$), and Ca × Cd ($\eta^2 = 0.326$, classified as synergistic). SHAP analysis ranked NO3 as the most important factor (47.4%), followed by

Cu (19.9%), Ca (16.9%), and Cd (15.7%). Permutation importance confirmed this ranking (NO3: 40.4%, Ca: 23.8%, Cd: 20.7%, Cu: 15.1%). Pearson correlations supported the SHAP ranking: NO3 showed a moderate negative correlation with shikonin ($r = -0.607$, $p < 0.001$), while Cu showed a weak negative correlation ($r = -0.198$, $p = 0.043$).

Predictive modeling. Random Forest achieved good cross-validated performance ($R^2 = 0.696 \pm 0.042$), substantially outperforming Linear Regression ($R^2 = 0.265 \pm 0.221$). GPR again overfit severely (CV $R^2 = -12.580$), consistent with the pattern observed across all three datasets. This recurrence strongly supports the recommendation that GPR should only be used for visualization of interaction surfaces (where training fit is the relevant metric) and not for predictive modeling at sample sizes typical of factorial experiments ($n < 200$), unless independent validation data are available.

Cross-Dataset Patterns

Three patterns emerged consistently across all datasets: (i) Random Forest was the most robust algorithm, selected as best model for 15 of 20 response variables across all experiments; (ii) GPR overfit severely for at least one response in every dataset, producing negative CV R^2 despite excellent training fit—a pattern that would remain invisible without the pipeline's automatic cross-validation reporting; and (iii) SHAP and permutation importance converged on the same factor rankings in all cases, validating SHAP as a reliable method for factorial plant stress data.

Discussion

A Unified Analytical Framework

The primary contribution of this work is methodological: an integrated, reproducible pipeline that formalizes the analysis of factorial plant stress experiments. While individual components are well-established, their integration into a single executable workflow with a graphical interface, automatic publication-ready figure generation, and comprehensive text report represents a novel contribution to the plant science toolkit.

Reproducibility. The pipeline encodes the entire workflow in a single script with explicit parameters and a fixed random seed, enabling exact reproduction of all results from raw data. This addresses the fragmentation and undocumented parameter choices that contribute to the reproducibility challenge in computational plant science [20].

Methodological triangulation. The application of ANOVA, η^2, SHAP, permutation importance, and standardized β coefficients to the same question—"which factor matters most?"—provides built-in validation. Convergence across methods substantially increases confidence; divergence flags areas for investigation. Across all three datasets, convergence was the rule rather than the exception.

Overfitting detection. The pipeline's reporting of both training and cross-validated performance prevents misinterpretation of overfit models. The recurrence of GPR overfitting across three independent datasets—spanning different species, stress types, and response categories—establishes this as a systematic limitation warranting caution in small-sample plant ML applications. This finding echoes recent calls for mandatory cross-validation reporting in plant science ML studies [5, 9].

Correct post-hoc letter display. By calling R's agricolae via subprocess, the pipeline ensures correct compact letter display—a non-trivial algorithmic problem [6] where several widely-used packages have produced incorrect assignments.

Biological Insights Across Datasets

Beyond methodological validation, the pipeline revealed biologically meaningful patterns. The functional partitioning in Viola—chitosan dominating growth, water deficit dominating pigments, catalase as a convergent node—demonstrates that elicitor and stress signals can operate through largely independent regulatory networks. The context-dependent silicon interactions in *Vicia faba*, invisible to main-effects-only ANOVA, suggest that silicon's efficacy in drought mitigation is more nuanced than previously reported [18] and depends critically on genotype × environment combinations. The interaction-dominated regulation of shikonin in Onosma—where no single factor was strongly predictive but factor combinations were—illustrates a general principle applicable to secondary metabolite optimization in medicinal plant biotechnology [19].

Comparison with Existing Tools

Existing plant science analysis workflows typically involve SPSS, SAS, or R for ANOVA, separate Python/R scripts for ML, and manual figure preparation. Specialized tools such as ImageJ/Fiji [21] and PlantCV [22] address image-based phenotyping; AlphaFold [23] and related tools address protein structure; and platforms like CyVerse [24] provide cloud-based analysis environments. However, none provides an integrated ANOVA-to-ML pipeline specifically designed for factorial experiments with automatic publication-ready output. The pipeline fills this niche, complementing rather than competing with existing specialized tools.

Limitations and Future Development

The pipeline supports completely randomized and randomized complete block designs; extension to split-plot and other designs would broaden applicability. ML methods are limited to RF, LR, and GPR; incorporation of gradient boosting (XGBoost, LightGBM) and regularized regression would expand options. A command-line interface would benefit batch processing. The PCA biplot and MORF correlation matrix require ≥ 2 responses; a univariate-only mode could be added for single-response experiments.

Conclusions

We present an integrated, open-source Python pipeline for statistical and machine learning analysis of factorial plant stress experiments. Validated on three independent datasets spanning 2–4 factors, 1–12 responses, and different species and stress types, the pipeline consistently: (i) reproduced published ANOVA results, (ii) provided additional quantitative insights through SHAP importance partitioning and permutation validation, (iii) detected and reported GPR overfitting invisible without cross-validation, and (iv) identified multivariate trait relationships through MORF residual analysis and PCA. The pipeline is accessible through a web interface, generates publication-ready figures in multiple formats, and produces a comprehensive text report. The complete code is provided as supplementary material and is adaptable to any factorial design with 2–5 factors.

Methods

Pipeline Implementation

The pipeline is implemented in Python 3.8 + with Streamlit for the web interface. Complete source code is provided in Supplementary File S1. The Streamlit application is launched by executing streamlit run Home.py in the terminal. Wrapper scripts for 2-, 3-, 4-, and 5-factor analyses are provided in the pages/ directory.

Demonstration Datasets

Dataset 1 (Viola spp.): Uniform viola plants at the 6–8 leaf stage were subjected to water deficit (100, 75, 50% FC) and chitosan foliar application (0, 1, 1.5, 2% w/v) in a completely randomized factorial design with three replications (n = 36). Seven responses were measured: fresh weight, dry weight, chlorophyll a/b ratio, total chlorophyll, catalase, peroxidase, and total soluble protein. Full experimental details are provided in Supplementary Methods.

Dataset 2 (*Vicia faba*)

Published by Neyestani et al. [18]. Three *Vicia faba* genotypes (G20, G61, G62) were subjected to two irrigation regimes (Ctrl, Drought) and three silicon treatments (Ctrl, N-Si, Si) with five replications (n = 120). Twelve responses were measured.

Dataset 3 (*Onosma dichroantha*)

Published by Ghazagh et al. [19]. Callus cultures were exposed to combinations of Cu, NO3, Ca, and Cd at five levels each in a Response Surface Design (RSM) with four replications of cube points (n = 105). Shikonin content was the response variable.

Analysis Parameters

Default parameters were: ANOVA categories = 3 (Dataset 1), 3 (Dataset 2), 5 (Dataset 3); cross-validation folds = 5; bootstrap iterations = 1,000; Random Forest estimators = 200; random seed = 42. Post-hoc comparisons used Tukey's HSD for Datasets 1 and 3, and Fisher's LSD for Dataset 2 (consistent with the original publication).

Abbreviations

ANOVA
Analysis of variance
CAT
Catalase
CD
Critical difference
CV
Cross-validation
FC
Field capacity
GPR
Gaussian Process Regression
GPX
Guaiacol peroxidase
HSD
Honestly Significant Difference
LR
Linear Regression
MDA
Malondialdehyde
ML
Machine learning
MORF
Multi-Output Random Forest
PCA
Principal Component Analysis
POD
Peroxidase
RF
Random Forest
RWC

Relative water content
SHAP
SHapley Additive exPlanations

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets analyzed in this study are included in the supplementary information files. The complete Python code for the pipeline is freely released under an MIT open-source, available from https://github.com/mbbagherieh/FAME-Factorial-Analysis-ML-Engine- upon publication; and is available from the corresponding author upon request.

Competing interests

The authors declare no competing interests.

Funding

This research was supported by Golestan University through a MSc research grant to RD.

Authors' contributions

MBBN: Conceptualization, supervision, Python code development, data analysis, writing – review and editing. RD: Investigation and data curation. HS: Methodology, validation, review and editing. All authors read and approved the final manuscript.

Acknowledgments

The authors thank Golestan University for research facilities. We acknowledge the use of DeepSeek AI (DeepSeek) for consultation on Python code debugging and editorial assistance. The authors retain full responsibility for the scientific content.

References

1. Montgomery DC. Design and Analysis of Experiments. 10th ed. Wiley; 2019.
2. Mittler R, Zandalinas SI, Fichman Y, Van Breusegem F. Reactive oxygen species signalling in plant stress responses. Nat Rev Mol Cell Biol. 2022;23(10):663–79.
3. Zandalinas SI, Mittler R. Plant responses to multifactorial stress combination. New Phytol. 2022;234(4):1161–7.
4. Nakagawa S, Lagisz M, O'Dea RE, Rutkowska J, Yang Y, Noble DWA, Senior AM. The orchard plot: cultivating a forest plot for use in ecology, evolution, and beyond. Res Synth Methods. 2021;12(1):4–12.
5. Newman SJ, Furbank RT. Explainable machine learning models for crop improvement. J Exp Bot. 2021;72(10):3446–60.
6. Piepho HP. An algorithm for a letter-based representation of all-pairwise comparisons. J Comput Graph Stat. 2004;13(2):456–66.
7. Graves S, Piepho HP, Selzer L, Dorai-Raj S, multcompView. Visualizations of Paired Comparisons. R package version 0.1-9; 2024.
8. Gill M, Anderson R, Hu H. Machine learning in plant science: applications in phenotyping, stress detection, and yield prediction. Trends Plant Sci. 2022;27(12):1234–48.
9. Arya S, Sandhu KS, Singh J, Kumar S. Deep learning for plant stress phenotyping: trends and future perspectives. Trends Plant Sci. 2023;28(7):780–94.
10. Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, de los Campos G, et al. Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci. 2017;22(11):961–75.
11. Zhang J, Huang Y, Pu R, Gonzalez-Moreno P, Yuan L, Wu K, Huang W. Monitoring plant diseases and pests through remote sensing technology: a review. Comput Electron Agric. 2022;199:107170.
12. van Klompenburg T, Kassahun A, Catal C. Crop yield prediction using machine learning: a systematic literature review. Comput Electron Agric. 2020;177:105709.
13. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. 2nd ed. Springer; 2009.
14. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206–15.
15. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30:4765–74.
16. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2(1):56–67.
17. Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics. 2010;26(10):1340–7.

18. Neyestani S, Hoseini M, Bagherieh-Najjar MB, Sadeghipour H. Genotype-dependent amelioration of water deficit stress in broad bean by silicon application. Iran J Plant Physiol. 2025;15(2):5123–38.
19. Ghazagh F, Bagherieh-Najjar MB, Nezamdoost T. Unraveling the interaction of copper, cadmium, calcium, and nitrate on phenolics, flavonoids, and shikonin contents of Onosma dichroantha calli by statistical modeling. Environ Sci Pollut Res. 2023;30(25):67234–48.
20. Baker M. 1,500 scientists lift the lid on reproducibility. Nature. 2016;533(7604):452–4.
21. Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, et al. Fiji: an open-source platform for biological-image analysis. Nat Methods. 2012;9(7):676–82.
22. Gehan MA, Fahlgren N, Abbasi A, Berry JC, Callen ST, Chavez L, et al. PlantCV v2: image analysis software for high-throughput plant phenotyping. PeerJ. 2017;5:e4088.
23. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9.
24. Merchant N, Lyons E, Goff S, Vaughn M, Ware D, Micklos D, Antin P. The iPlant collaborative: cyberinfrastructure for enabling data to discovery for the life sciences. PLoS Biol. 2016;14(1):e1002342.

Figures

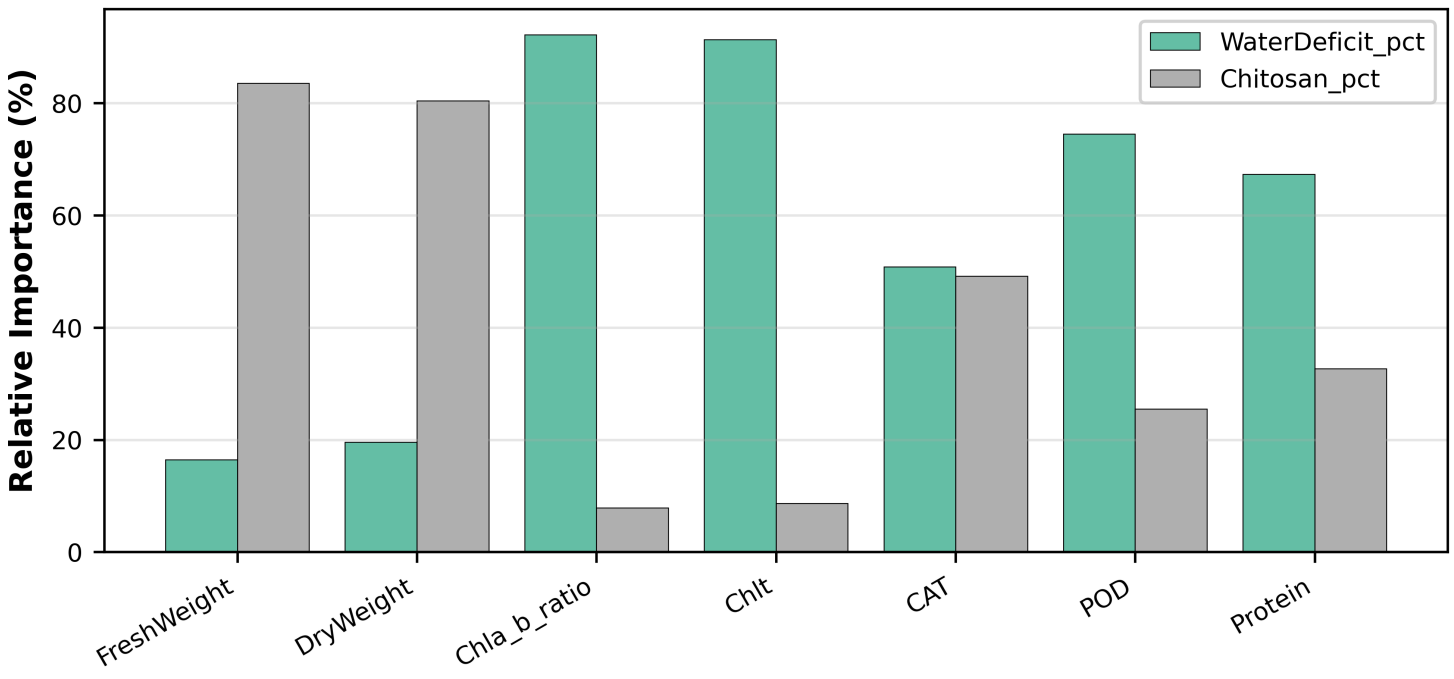

Figure 1

SHAP feature importance for Viola spp. under combined water deficit and chitosan treatment. Bar heights represent the percentage of total predictive importance attributed to each factor.

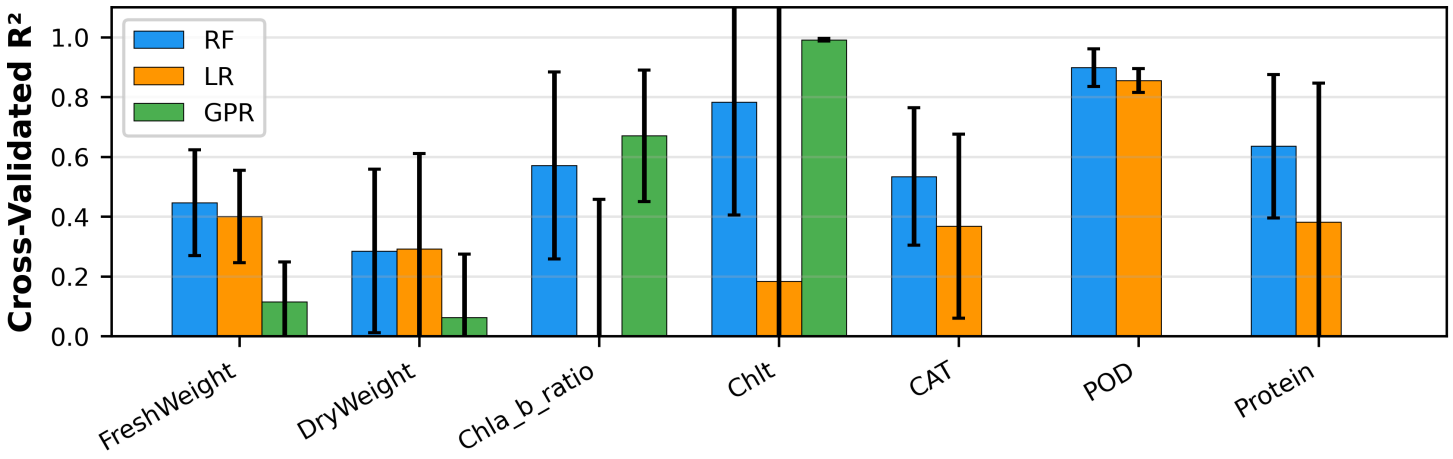

Figure 2

Cross-validated model performance comparison (5-fold CV R^2 ± SD) for Viola spp. Random Forest (blue) is the most robust algorithm. Gaussian Process Regression (green) achieves excellent performance for total chlorophyll (R^2 = 0.991) but severely overfits for CAT, POD, and protein, producing negative CV R^2 values despite perfect training fit.

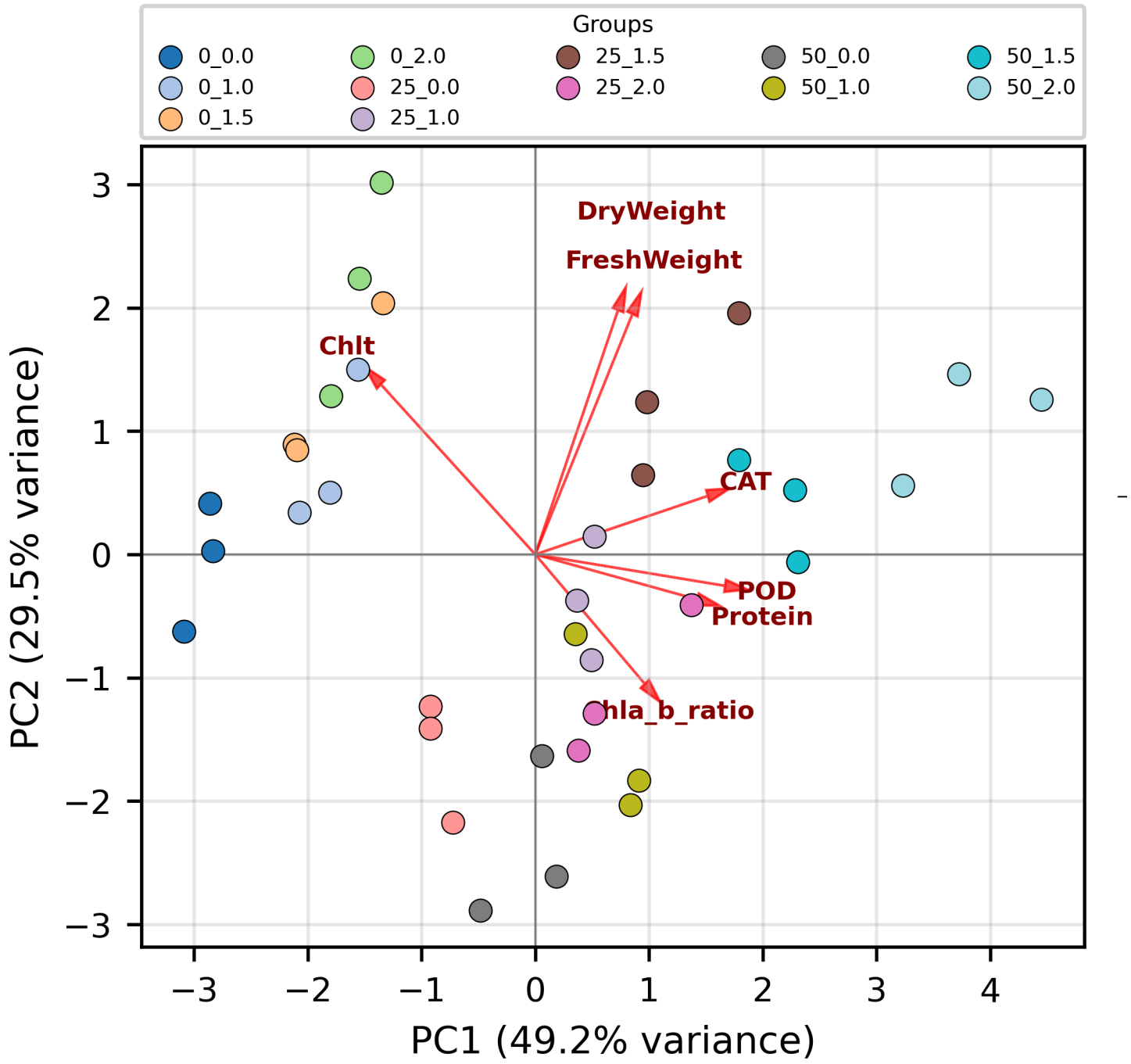

Figure 3

PCA biplot of Viola spp. responses under combined treatments.** PC1 (49.2% variance) separates treatments primarily by water deficit level, reflecting antioxidant and protein responses. PC2 (29.5% variance) separates treatments by chitosan concentration, reflecting growth trait variation. Vectors indicate response variable loadings.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 2factorAnalysisresults.zip
- 3factoranalysisresults.zip
- 4factoranalysisresults.zip