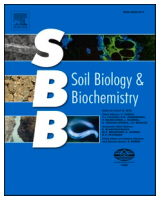
Soil Biology and Biochemistry 223 (2026) 110322



Contents lists available at ScienceDirect

Soil Biology and Biochemistry

journal homepage: www.elsevier.com/locate/soilbio



Beyond SHAP: Exposing the instability of supervised feature importance and the necessity of external validation in soil microbiome research☆

ARTICLE INFO

Keywords:
SHAP interpretability
Feature importance validation
Bradford hill criteria
Leave-top-feature-out method
Soil microbiome modeling

ABSTRACT

Soil Biology and Biochemistry has published six articles employing SHAP for feature importance assessment despite the absence of ground truth validation, exemplified by Du et al. (2026). This study empirically demonstrates that supervised models (Random Forest, XGBoost), with and without SHAP, exhibit substantial instability and inconsistency in feature ranking orders, whereas unsupervised methods (Gaussian Mixture Model, Highly Variable Gene Selection) and Spearman correlation achieve perfect ranking consistency and stability—despite all seven algorithms showing statistically comparable predictive accuracy (0.502–0.511). This dissociation between predictive accuracy and interpretive stability reveals that SHAP-derived rankings constitute model-specific artifacts rather than reliable ecological insights. We introduce a leave-top-feature-out (LTFO) framework grounded in the extended Bradford Hill criteria, incorporating feature consistency and dose-response relationships, as a necessary external validation approach for feature importance claims in soil microbiome and broader ecological modeling research.

Researchers often fail to acknowledge the fundamental principles underlying supervised models, neglecting critical assumptions and overlooking the absence of ground truth in feature assessment calculations (Domingo, 2026; Qian et al., 2026). Furthermore, researchers frequently do not distinguish between methodological types—parametric versus semiparametric versus nonparametric, and linear versus nonlinear. When linear methods are applied to nonlinear data, or when parametric or semiparametric methods are applied to nonparametric data, the resulting outcomes, including metric scores, can be substantially distorted. This mismatch leads to erroneous interpretations and conclusions, as it violates the underlying assumptions of the model when applied against the nonlinear, nonparametric nature of real-world data—unless the linear, parametric nature of the data can first be established. This paper focuses specifically on these issues within supervised models.

Schiller et al. (2025) reviewed 575 explainable artificial intelligence applications in environmental and Earth system sciences, finding that SHAP/Shapley approaches were the most frequently used method. However, only 7 studies (1.2%) explicitly identified trustworthiness as a central objective, leading the authors to emphasize that explainability does not automatically confer trustworthiness. Similarly, Stock et al. (2023) noted that although machine learning has become a popular tool in ecology, many scientific applications suffer from data leakage and other common pitfalls, which can produce misleading results. They argue that discipline-specific model information sheets must be developed to support more rigorous model evaluation.

Soil Biology and Biochemistry has published 6 articles employing SHAP (SHapley Additive exPlanations) to assess feature importance, despite the complete absence of ground truth in feature assessment—a field-wide limitation that undermines the interpretability of these models. A representative case among the 6 SHAP-based studies is that of Du et al. (2026), in which they did not acknowledge that feature importance derived from supervised models, and SHAP's outputs in particular, rely solely on a given supervised model and therefore constitute unvalidated and unreliable estimates of feature importance. This is because supervised models inherently possess two distinct types of accuracy: target prediction accuracy and feature importance reliability. While target prediction accuracy can be rigorously validated against ground truth labels, feature importance lacks any corresponding ground truth for validation. In other words, high target prediction accuracy does not guarantee reliable feature importance.

This asymmetry between the two types of accuracy has a critical consequence: feature importance derived from supervised models reflects each variable's contribution to the model's prediction, not its true ecological or biological association with the outcome, nor a reliable measure of feature importance in any general sense. In other words, a high SHAP value indicates that a feature strongly influenced a particular model's output given the specific data, algorithm, and hyperparameters used—but it does not confirm that the feature is genuinely or consistently important in the underlying biological system. Because no independent, model-free benchmark exists against which SHAP rankings can be checked, apparent feature importance may be an artifact of model architecture, training data idiosyncrasies, multicollinearity among features, or random variation in model fitting, rather than a stable or reproducible signal. This complete absence of ground truth in feature assessment means that SHAP outputs, however mathematically elegant,

☆ **According to ScholarGPS**, Yoshiyasu Takefuji holds notable global rankings in several fields. He ranks 25th out of 1,287,415 scholars in life sciences, 22nd out of 805,705 in COVID-19, and 1st out of 109,919 in environmental sciences.

https://doi.org/10.1016/j.soilbio.2026.110322
Received 1 September 2026; Received in revised form 29 September 2026; Accepted 1 October 2026
Available online 3 October 2026
0038-0717/© 2026 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

cannot be validated internally and therefore require external, independent verification before being treated as evidence of true feature importance. Without such validation, SHAP-derived rankings remain model-specific artifacts rather than generalizable ecological or biological insights.

This paper introduces the necessary external feature validation framework, such as a leave-top-feature-out (LTFO) approach, grounded in the extended Bradford Hill criteria with two mandatory components: feature consistency and dose-response relationships to calculate reliable feature importance (Hill, 1965; Thapa et al., 2026; Loughman et al., 2026; Loos et al., 2026; Zhao et al., 2026; Pengo et al., 2026). The LTFO operates as follows: select the top n features from the full set (set1), remove the highest-ranked feature from the full set to create a reduced dataset, re-fit and re-select the top n-1 features from the reduced dataset (set2), and compare the two ordered sets of feature ranking orders. Eliminating the highest-ranked feature potentially creates the strongest impact on feature ranking orders, while removing the lowest-ranked feature generates the weakest impact.

While this paper examines a single dataset, LTFO has been validated successfully across a wide range of domains and disciplines employing diverse datasets to demonstrate its robustness and generalization (Takefuji, 2025a, 2025b; 2026a; 2026b; 2026c; 2026d; 2026e; 2026f). While existing ablation or perturbation studies, such as Jaccard-based and other methods, discard critical feature ranking order information by treating features as unordered sets, the LTFO examines ordered sets of feature rankings, preserving ranking information to assess feature consistency and dose-response relationships.

This paper empirically demonstrates these pitfalls regarding the use of supervised models with and without SHAP for feature assessment across multiple algorithms such as supervised models (Random Forest: RF, RF-SHAP, XGBoost: XGB, XGB-SHAP), unsupervised models (Gaussian Mixture Model: GMM, Highly Variable Gene Selection: HVGS) and non-target-prediction nonparametric statistical method like Spearman's correlation. Bacteria_Abundance serves as a target with 100000 samples and 11 features. For purposes of reproducibility and transparency, Python code, cvshap2.py was publicly available on GitHub (2026).

Table 1 summarizes the 5-fold cross-validation accuracy and corresponding feature ranking orders (Top-6 and Top-5) across seven algorithms, comprising supervised classifiers with and without SHAP (RF, XGB, RF-SHAP, XGB-SHAP), unsupervised methods (HVGS, GMM), and a correlation-based approach (Spearman).

Cross-validation accuracy remained tightly clustered across all seven methods, ranging narrowly from 0.502 (XGB-SHAP) to 0.511 (XGB and Spearman), with a spread of only 0.009. This near-equivalence in predictive performance indicates that no single algorithmic family — whether supervised, unsupervised, or correlation-based — offers a meaningful advantage in classification accuracy over the others, suggesting that the underlying signal-to-noise structure of the dataset imposes a comparable ceiling on predictive performance regardless of the modeling approach employed.

Despite comparable accuracy, the supervised approaches exhibited pronounced instability in feature ranking order, both in the presence and absence of SHAP-based interpretation. Among the Top-6 features, RF, XGB, RF-SHAP, and XGB-SHAP each produced markedly different rankings and, in several cases, different feature compositions altogether. Notably, Fungi_Abundance (%) occupied the top rank in RF, RF-SHAP, and XGB-SHAP, yet was ranked fifth in XGB — a striking reordering given otherwise similar accuracy. Similarly, CO2_Emission (μg/g/day) ranked first in XGB but only fourth or lower in the SHAP-based variants. This inconsistency persisted, and in some cases intensified, when transitioning from Top-6 to Top-5 rankings: for instance, Fungi_Abundance (%) dropped out of the Top-5 entirely in RF, RF-SHAP, and XGB-SHAP despite ranking first in the Top-6 for three of these four methods, while CO2_Emission (μg/g/day) shifted from first (Top-6, XGB) to entirely absent (Top-5, XGB). The application of SHAP did not resolve this instability; rather, RF-SHAP and XGB-SHAP diverged from their non-SHAP counterparts (RF and XGB, respectively) in both feature composition and rank order, indicating that SHAP-based attribution introduces its own source of ranking variability rather than stabilizing the rankings derived from the underlying supervised model.

Table 1
Cross-validation accuracy and feature ranking orders across algorithms. Bold entries indicate features with inconsistent and unstable ranking orders.

Method	CV6 accuracy	CI:95%	Top 6 feature rankings	Top 5 feature rankings
RF	0.503	[0.481, 0.526]	Fungi_Abundance (%), Soil_pH, NH4_Nitrate (μg/g), CO2_Emission (μg/g/day), C_N_Ratio, Urease (μmol/g/h)	**CO2_Emission (μg/g/day),** NH4_Nitrate (μg/g), **Soil_pH, Urease (μmol/g/h), C_N_Ratio**
XGB	0.511	[0.491, 0.53]	CO2_Emission (μg/g/day), β_Glucosidase (μmol/g/h), NH4_Nitrate (μg/g), Urease (μmol/g/h), Fungi_Abundance (%), C_N_Ratio	**Urease (μmol/g/h), C_N_Ratio, NH4_Nitrate (μg/g), β_Glucosidase (μmol/g/h), Fungi_Abundance (%)**
RF-SHAP	0.508	[0.497, 0.519]	Fungi_Abundance (%), NH4_Nitrate (μg/g), C_N_Ratio, CO2_Emission (μg/g/day), β_Glucosidase (μmol/g/h), Organic_C (%)	NH4_Nitrate (μg/g), **CO2_Emission (μg/g/day), C_N_Ratio, Organic_C (%), β_Glucosidase (μmol/g/h)**
XGB-SHAP	0.502	[0.492, 0.512]	Fungi_Abundance (%), Organic_C (%), Soil_pH, C_N_Ratio, β_Glucosidase (μmol/g/h), Urease (μmol/g/h)	**Urease (μmol/g/h), C_N_Ratio, Soil_pH, Organic_C (%), β_Glucosidase (μmol/g/h)**
HVGS	0.505	[0.495, 0.515]	NH4_Nitrate (μg/g), C_N_Ratio, Fungi_Abundance (%), CO2_Emission (μg/g/day), Urease (μmol/g/h), Organic_C (%)	C_N_Ratio, Fungi_Abundance (%), CO2_Emission (μg/g/day), Urease (μmol/g/h), Organic_C (%)
GMM	0.505	[0.495, 0.515]	NH4_Nitrate (μg/g), C_N_Ratio, Fungi_Abundance (%), CO2_Emission (μg/g/day), Urease (μmol/g/h), Organic_C (%)	C_N_Ratio, Fungi_Abundance (%), CO2_Emission (μg/g/day), Urease (μmol/g/h), Organic_C (%)
Spearman	0.511	[0.495, 0.527]	CO2_Emission (μg/g/day), C_N_Ratio, Organic_C (%), β_Glucosidase (μmol/g/h), Fungi_Abundance (%), Land_Use_Type	C_N_Ratio, Organic_C (%), β_Glucosidase (μmol/g/h), Fungi_Abundance (%), Land_Use_Type

In sharp contrast, the unsupervised clustering methods (HVGS and GMM) produced identical feature rankings in both the Top-6 and Top-5 lists — an exact one-to-one correspondence in both feature identity and order (NH4_Nitrate (μg/g), C_N_Ratio, Fungi_Abundance (%), CO2_Emission (μg/g/day), Urease (μmol/g/h), Organic_C (%) for Top-6; identical ordering for Top-5). This perfect concordance between two independently derived unsupervised algorithms strongly suggests that HVGS and GMM are capturing a stable, reproducible underlying structure in the data that is largely invariant to the specific unsupervised technique applied. Furthermore, the Spearman correlation-based ranking exhibited internal consistency between its own Top-6 and Top-5 lists, with the Top-5 features representing a perfectly preserved subset and order of the Top-6 ranking (simply dropping CO2_Emission

(μg/g/day) from the top position while retaining the relative order of the remaining five features). This stability within the Spearman approach, despite achieving accuracy on par with the top-performing supervised model (0.511), reinforces the notion that rank-based and unsupervised structural approaches yield more robust and reproducible feature orderings than supervised classifiers, irrespective of SHAP-based interpretability enhancements.

Taken together, these findings reveal a critical dissociation between predictive accuracy and interpretive stability: while all seven algorithms achieve statistically comparable classification performance, only the unsupervised (HVGS, GMM) and correlation-based (Spearman) methods yield feature rankings that are internally consistent and stable across ranking depths. Supervised models — regardless of whether SHAP is applied — display substantial rank volatility, undermining confidence in any single feature-importance ordering derived from these methods and highlighting the necessity of triangulating feature importance across multiple methodological paradigms rather than relying on any one supervised or SHAP-based ranking in isolation.

Consent to participate

Not applicable.

Ethics approval

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The author has no permission to share data.

Code availability

Python code is publicly available on GitHub.

AI use

Not applicable.

Funding

This research has no fund.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

None.

References

Domingo, J.L., 2026. Artificial intelligence revolution in toxicology: clinical precision, global equity, and the 2030 roadmap. Toxicology Letters 418, 111871. https://doi.org/10.1016/j.toxlet.2026.111871.

Du, M., Xue, P., Minasny, B., Tang, Y., Tran, T.K.A., Jang, H.J., McBratney, A., 2026. Climate and anthropogenic activities regulate genome-inferred bacterial growth potential. Soil Biology and Biochemistry 222, 110295. https://doi.org/10.1016/j.soilbio.2026.110295.

GitHub. cvshap3.py. https://github.com/y-takefuji/soil_microbe.

Hill, A.B., 1965. The environment and disease: association or causation? Proceedings of the Royal Society of Medicine 58 (5), 295–300. https://doi.org/10.1177/003591576505800503.

Loos, A., Khazneh, E., Kübler, J., 2026. Enhancing safety evaluations: a comprehensive framework for evidence-based safety assessment using the Bradford hill criteria. Therapeutic Innovation & Regulatory Science 60 (2), 349–359. https://doi.org/10.1007/s43441-025-00910-y.

Loughman, J., Bullimore, M.A., Flitcroft, D.I., 2026. Myopia and eye disease: from correlation to causation. Progress in Retinal and Eye Research 114, 101497. https://doi.org/10.1016/j.preteyeres.2026.101497.

Pengo, M.F., Javaheri, S., Maiolino, G., Redline, S., Lombardi, C., 2026. Insomnia and cardiovascular disease: untangling a complex relationship. Journal of Sleep Research 35 (4), e70299. https://doi.org/10.1111/jsr.70299.

Qian, M., Li, L., Cheng, Y., Hou, Y., Fu, S., Xu, Z., 2026. Predicting and explaining poor prognosis in diabetic kidney disease using SHAP-based interpretable machine learning. iScience 29 (6), 116280. https://doi.org/10.1016/j.isci.2026.116280.

Schiller, J., Stiller, S., Ryo, M., 2025. Artificial intelligence in environmental and Earth system sciences: explainability and trustworthiness. Artificial Intelligence Review 58, 316. https://doi.org/10.1007/s10462-025-11165-2.

Stock, A., Gregr, E.J., Chan, K.M.A., 2023. Data leakage jeopardizes ecological applications of machine learning. Nature Ecology & Evolution 7 (11), 1743–1745. https://doi.org/10.1038/s41559-023-02162-1.

Takefuji, Y., 2025a. Model-specific feature importances: distinguishing true associations from target-feature relationships. Journal of Affective Disorders 369, 390–391. https://doi.org/10.1016/j.jad.2024.10.019.

Takefuji, Y., 2025b. Beyond XGBoost and SHAP: unveiling true feature importance. Journal of Hazardous Materials 488, 137382. https://doi.org/10.1016/j.jhazmat.2025.137382.

Takefuji, Y., 2026a. The reliability gap: why high predictive accuracy doesn't guarantee stable feature importance. Marine Pollution Bulletin 226, 119398. https://doi.org/10.1016/j.marpolbul.2026.119398.

Takefuji, Y., 2026b. Evaluating linear parametric poisson regression vs nonparametric unsupervised learning in ulcerative colitis data. Gastroenterology. https://doi.org/10.1053/j.gastro.2026.04.011. S0016-5085(26)06761-06762. Advance online publication.

Takefuji, Y., 2026c. When high accuracy misleads: stability limits of supervised feature importance in QSAR biodegradation. Chemosphere 396, 144846. https://doi.org/10.1016/j.chemosphere.2026.144846.

Takefuji, Y., 2026d. Beyond prediction: assessing stability in feature selection methods for materials science applications. Computational Materials Science 267, 114609. https://doi.org/10.1016/j.commatsci.2026.114609.

Takefuji, Y., 2026e. Ensuring reliable feature importance in food chemistry AI. Food Chemistry 509, 148515. https://doi.org/10.1016/j.foodchem.2026.148515.

Takefuji, Y., 2026f. The absence of true association validation in supervised feature importance: evidence from battery capacity prediction for second-life applications. Journal of Energy Storage 180, 124146. https://doi.org/10.1016/j.est.2026.124146.

Thapa, S.K., Li, J., Mach, J., Hilmer, S.N., Kirkpatrick, C.M.J., 2026. Applying the Bradford Hill criteria to assess the independent causal roles of aging and medication in frailty progression: a systematic review. Drugs & Aging 43 (1), 9–28. https://doi.org/10.1007/s40266-025-01273-7.

Zhao, J., Chen, Y., Du, C., Zhang, W., Liu, W., Chen, J., Ma, S., Yao, Y., 2026. Understanding senna risks: evidence from the FDA adverse event reporting system. Journal of Ethnopharmacology 361, 121170. https://doi.org/10.1016/j.jep.2026.121170.

Yoshiyasu Takefuji

SciencePark Corporation, 3-24-9 Iriya-Nishi, Zama-shi, Kanagawa, 252-0029, Japan

E-mail address: takefuji@keio.jp.