



Reassessing SHAP-based interpretations in QSAR: Model-centric limits and unsupervised alternatives for fluorocarbon inhalation toxicity

ARTICLE INFO

Keywords:

QSAR
SHAP
Feature importance reliability
Unsupervised descriptor prioritization
Fluorocarbon toxicity

ABSTRACT

Ye et al. (2025) report strong QSAR performance for fluorocarbon inhalation toxicity using ‘SVM-RBF’ and ‘XGBoost’, complemented by SHAP analyses to identify influential molecular descriptors. While predictive accuracy and generalization are commendable, the interpretability claims warrant caution. Supervised models possess two distinct accuracies—target prediction and feature-importance reliability—the latter lacking ground truth validation. Consequently, SHAP, as a model-dependent explainer, can faithfully reproduce and even amplify model biases, is sensitive to model specification, struggles with correlated descriptors, and does not infer causality. High accuracy does not guarantee reliable importances. We recommend augmenting the pipeline with unsupervised, label-agnostic descriptor prioritization (e.g., ‘feature agglomeration’, ‘highly variable feature selection’) followed by non-targeted association screening (e.g., Spearman correlation with p-values) to improve stability and mitigate model-induced interpretative errors.

Ye et al. (2025) explored QSAR-based prediction of acute inhalation toxicity and SHapley Additive exPlanations (SHAP) interpretability analysis of fluorocarbon environmental-friendly insulating gases. Their comprehensive evaluation identified SVM with radial basis function kernel and XGBoost as superior models, demonstrating exceptional predictive capability and robust generalization across diverse fluorocarbon compounds. Through detailed SHAP analysis, Ye et al. successfully identified critical molecular descriptors that significantly influence toxicity profiles, providing valuable mechanistic insights for the rational design of safer insulating gas alternatives with reduced environmental impact.

However, this paper raises significant theoretical and methodological concerns regarding the use of supervised models such as SVM and XGBoost for feature importance assessment. The model-specific nature of these approaches leads to potentially erroneous interpretations and conclusions about molecular descriptor importance. Critically, Ye et al. should acknowledge that supervised models possess two distinct types of accuracy: target prediction accuracy and feature importance reliability. While target prediction accuracy can be validated against ground truth values (labels), feature importances lack corresponding ground truth for accuracy validation, representing a fundamental limitation in their interpretability framework.

Feature importance in supervised models fundamentally refers to contributions to prediction mechanisms rather than true biological or chemical associations between variables. Consequently, feature importances derived from supervised models are inherently biased, leading to potentially misleading interpretations (Adler and Painsky, 2022; Alaimo Di Loro et al., 2023; Dunne et al., 2023; Fisher et al., 2019; Huti et al., 2023; Loecher, 2024; Nalenz et al., 2024; Nazer et al., 2023; Nguyen et al., 2015; Salles et al., 2021; Smith et al., 2024; Steiner and Kim, 2016; Strobl et al., 2007; Ugirumura et al., 2024; Wallace et al., 2023; Zarei et al., 2021). Extensive research has demonstrated that high target

prediction accuracy does not guarantee reliable feature importances, a critical distinction overlooked in this study (Fisher et al., 2019; Lenhof et al., 2024; Lipton, 2018; Mandler and Weigand, 2024; Molnar et al., 2022a,b; Parr et al., 2024; Potharlanka and Bhat, 2024; Watson and Wright, 2021; Wood et al., 2024).

Furthermore, the function of explain = SHAP(model) implies that SHAP solely relies on a given supervised model, inheriting and propagating its inherent limitations and biases. This dependency means SHAP explanations are fundamentally constrained by the quality of the underlying model’s feature representation. SHAP may actually amplify model biases by presenting them as objective feature importance scores, creating a false sense of interpretative certainty. Recent research has identified fundamental mathematical limitations in Shapley-based approaches, including issues with feature interdependence handling and inability to capture causal relationships. Additionally, SHAP values are highly susceptible to model specification effects—the same dataset analyzed with different model architectures can yield contradictory feature importance rankings, undermining the reliability of mechanistic insights drawn from any single model implementation. Even with technically correct SHAP implementation, the resulting explanations remain model-centric interpretations rather than data-centric truths about molecular descriptor significance. While SHAP is a mathematically elegant approach, its exclusive reliance on potentially flawed feature representations from supervised models means it can faithfully produce misleading outcomes despite high prediction accuracy (Wu, 2025; Bilodeau et al., 2024; Huang and Marques-Silva, 2024; Kumar et al., 2021; Hooshyar and Yang, 2024; Lones, 2024; Molnar et al., 2022a,b; Létoffé et al., 2025; Ponce-Bobadilla et al., 2024; Coupland et al., 2025).

There is no algorithm to accurately calculate true associations between variables. This paper advocates for incorporating unsupervised models such as feature agglomeration (FA) and highly variable gene

<https://doi.org/10.1016/j.envres.2025.123125>

Received 17 September 2025; Accepted 12 October 2025

Available online 16 October 2025

0013-9351/© 2025 Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

selection (HVGS), and followed by non-targeted nonlinear nonparametric methods such as Spearman's correlation with p-values instead of solely relying on supervised models with SHAP. While supervised models must suffer from instability in feature importance ranking orders due to model specific nature and label-driven errors, FA, HVGS and Spearman exhibit stronger stability in feature ranking orders due to the absence of label-driven errors.

Consent to participate

Not applicable.

Ethics approval

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

Not applicable.

Code availability

Not applicable.

AI use

Not applicable.

Authors' contributions

Yoshiyasu Takefuji completed this research and wrote this article.

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.envres.2025.123125>.

Data availability

No data was used for the research described in the article.

References

- Adler, A.I., Painsky, A., 2022. Feature importance in gradient boosting trees with cross-validation feature selection. *Entropy* 24 (5), 687. <https://doi.org/10.3390/e24050687>.
- Alaimo Di Loro, P., Scacciatelli, D., Tagliaferri, G., 2023. 2-step gradient boosting approach to selectivity bias correction in tax audit: an application to the VAT gap in Italy. *Stat. Methods Appl.* 32, 237–270. <https://doi.org/10.1007/s10260-022-00643-4>.
- Bilodeau, B., Jaques, N., Koh, P.W., Kim, B., 2024. Impossibility theorems for feature attribution. *Proc. Natl. Acad. Sci.* 121 (2), e2304406120. <https://doi.org/10.1073/pnas.2304406120>.
- Coupland, H., Scheidwasser, N., Katsiferis, A., Davies, M., Flaxman, S., Hulvej Rod, N., Mishra, S., Bhatt, S., Unwin, H.J.T., 2025. Exploring the potential and limitations of deep learning and explainable AI for longitudinal life course analysis. *BMC Public Health* 25 (1), 1520. <https://doi.org/10.1186/s12889-025-22705-4>.
- Dunne, R., Reguant, R., Ramarao-Milne, P., Szul, P., Sng, L.M.F., Lundberg, M., Twine, N. A., Bauer, D.C., 2023. Thresholding Gini variable importance with a single-trained random forest: an empirical Bayes approach. *Comput. Struct. Biotechnol. J.* 21, 4354–4360. <https://doi.org/10.1016/j.csbj.2023.08.033>.
- Fisher, A., Rudin, C., Dominici, F., 2019. All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.* 20, 177.
- Hooshyar, D., Yang, Y., 2024. Problems with SHAP and LIME in interpretable AI for education: a comparative study of post-hoc explanations and neural-symbolic rule extraction. *IEEE Access* 12, 137472–137490. <https://doi.org/10.1109/ACCESS.2024.3463948>.
- Huang, X., Marques-Silva, J., 2024. On the failings of Shapley values for explainability. *Int. J. Approx. Reason.* 171, 109112. <https://doi.org/10.1016/j.ijar.2023.109112>.
- Huti, M., Lee, T., Sawyer, E., King, A.P., 2023. An investigation into race bias in random forest models based on breast DCE-MRI derived radiomics features. In: *Clinical Image Based Procedure Fairness AI Med Imaging Ethical Philos Issues Med Imaging*, pp. 225–234. https://doi.org/10.1007/978-3-031-45249-9_22.
- Kumar, I., Scheidegger, C., Venkatasubramanian, S., Friedler, S., 2021. Shapley residuals: quantifying the limits of the shapley value for explanations. *Adv. Neural Inf. Process. Syst.* 34, 26598–26608.
- Lenhof, K., Eckhart, L., Rolli, L.M., Lenhof, H.P., 2024. Trust me if you can: a survey on reliability and interpretability of machine learning approaches for drug sensitivity prediction in cancer. *Briefings Bioinf.* 25 (5), bbae379. <https://doi.org/10.1093/bib/bbae379>.
- Létoffé, O., Huang, X., Marques-Silva, J., 2025. Towards trustable SHAP scores. *Proc. AAAI Conf. Artif. Intell.* 39 (17), 18198–18208. <https://doi.org/10.1609/aaai.v39i17.34002>.
- Lipton, Z.C., 2018. The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. *ACM Queue* 16 (3), 31–57. <https://doi.org/10.1145/3236386.3241340>.
- Loecher, M., 2024. Debiasing SHAP scores in random forests. *AStA Adv. Statist. Anal.* 108, 427–440. <https://doi.org/10.1007/s10182-023-00479-7>.
- Lones, M.A., 2024. Avoiding common machine learning pitfalls. *Patterns* 5 (10), 101046. <https://doi.org/10.1016/j.patter.2024.101046>.
- Mandler, H., Weigand, B., 2024. A review and benchmark of feature importance methods for neural networks. *ACM Comput. Surv.* 56 (12), 318. <https://doi.org/10.1145/3679012>.
- Molnar, C., König, G., Herbringer, J., Freiesleben, T., Dandl, S., Scholbeck, C.A., et al., 2022a. General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models. Springer International Publishing. https://doi.org/10.1007/978-3-031-04083-2_4.
- Molnar, C., et al., 2022b. General pitfalls of model-agnostic interpretation methods for machine learning models. In: Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K. R., Samek, W. (Eds.), *Xxai - Beyond Explainable AI*, vol 13200. Springer, p. 4. https://doi.org/10.1007/978-3-031-04083-2_4.
- Nalenz, M., Rodemann, J., Augustin, T., 2024. Learning de-biased regression trees and forests from complex samples. *Mach. Learn.* 113, 3379–3398. <https://doi.org/10.1007/s10994-023-06439-1>.
- Nazer, L.H., Zatarah, R., Waldrup, S., et al., 2023. Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS Digital Health* 2 (6), e0000278. <https://doi.org/10.1371/journal.pdig.0000278>.
- Nguyen, T.T., Huang, J.Z., Nguyen, T.T., 2015. Unbiased feature selection in learning random forests for high-dimensional data. *Sci. World J.* 2015. <https://doi.org/10.1155/2015/471371>. Article 471371.
- Parr, T., Hamrick, J., Wilson, J.D., 2024. Nonparametric feature impact and importance. *Inf. Sci.* 653, 119563. <https://doi.org/10.1016/j.ins.2023.119563>.
- Ponce-Bobadilla, A.V., Schmitt, V., Maier, C.S., Mensing, S., Stodtmann, S., 2024. Practical guide to SHAP analysis: explaining supervised machine learning model predictions in drug development. *Clin. Translat. Sci.* 17 (11), e70056. <https://doi.org/10.1111/cts.70056>.
- Potharlanka, J.L., Bhat, M.N., 2024. Feature importance feedback with Deep Q process in ensemble-based metaheuristic feature selection algorithms. *Sci. Rep.* 14 (1), 2923. <https://doi.org/10.1038/s41598-024-53141-w>.
- Salles, T., Rocha, L., Gonçalves, M., 2021. A bias-variance analysis of state-of-the-art random forest text classifiers. *Adv. Data Anal. Classificat.* 15, 379–405. <https://doi.org/10.1007/s11634-020-00409-4>.
- Smith, H.L., Biggs, P.J., French, N.P., et al., 2024. Lost in the forest: encoding categorical variables and the absent levels problem. *Data Min. Knowl. Discov.* 38, 1889–1908. <https://doi.org/10.1007/s10618-024-01019-w>.
- Steiner, P.M., Kim, Y., 2016. The mechanics of omitted variable bias: Bias amplification and cancellation of offsetting biases. *J. Causal Inference* 4 (2), 20160009. <https://doi.org/10.1515/jci-2016-0009>.
- Strobl, C., Boulesteix, A.L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinf.* 8, 25. <https://doi.org/10.1186/1471-2105-8-25>.
- Ugrumurera, J., Bensen, E.A., Severino, J., Sanyal, J., 2024. Addressing bias in bagging and boosting regression models. *Sci. Rep.* 14 (1), 18452. <https://doi.org/10.1038/s41598-024-68907-5>.
- Wallace, M.L., Mentch, L., Wheeler, B.J., Lyons, M., Reichmann, W.M., 2023. Use and misuse of random forest variable importance metrics in medicine: demonstrations through incident stroke prediction. *BMC Med. Res. Methodol.* 23 (1), 144. <https://doi.org/10.1186/s12874-023-01965-x>.

- Watson, D.S., Wright, M.N., 2021. Testing conditional independence in supervised learning algorithms. *Mach. Learn.* 110 (8), 2107–2129. <https://doi.org/10.1007/s10994-021-06030-6>.
- Wood, D., Papamarkou, T., Benatan, M., et al., 2024. Model-agnostic variable importance for predictive uncertainty: an entropy-based approach. *Data Min. Knowl. Discov.* 38, 4184–4216. <https://doi.org/10.1007/s10618-024-01070-7>.
- Wu, L., 2025. A review of the transition from Shapley values and SHAP values to RGE. *Statistics* 1–23. <https://doi.org/10.1080/02331888.2025.2487853>.
- Ye, F., Xiao, F., Zhan, A., Chu, Y., Tian, S., Zhang, X., 2025. QSAR-based prediction of acute inhalation toxicity and SHAP interpretability analysis of fluorocarbon environmental-friendly insulating gases. *Environ. Res.* 285 (Pt 1). <https://doi.org/10.1016/j.envres.2025.122340>. Article 122340.
- Zarei, M., Najarchi, M., Mastouri, R., 2021. Bias correction of global ensemble precipitation forecasts by random forest method. *Earth Sci. Inform.* 14, 677–689. <https://doi.org/10.1007/s12145-021-00577-7>.

According to ScholarGPS, Yoshiyasu Takefuji holds notable global rankings in several fields. He ranks 54th out of 395,884 scholars in neural networks (AI), 23rd out of 47,799 in parallel computing, and 14th out of 7222 in parallel algorithms. Furthermore, he ranks the highest in AI tools and human-induced error analysis, underscoring his significant contributions to these domains.

Yoshiyasu Takefuji 
Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku,
Tokyo, 135-8181, Japan
E-mail address: takefuji@keio.jp.