Matter

CellPress

Letter

Feature assessment of MOF biocompatibility should consider statistical approaches beyond machine learning models

Yoshiyasu Takefuji[1,*]
[1]Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo 135-8181, Japan
*Correspondence: takefuji@keio.jp
https://doi.org/10.1016/j.matt.2025.102171

Menon and Fairen-Jimenez investigated guiding the rational design of biocompatible metal-organic frameworks for drug delivery.[1] In their computational pipeline workflow for assessment of metal organic framework (MOF) biocompatibility, the second workflow (ii) contains three fatal assumptions. The first is that feature selection is carried out to find the optimum subset of features that can maximize model performance. The second is that these subsets are then used to train different models—namely random forests, support vector machines, and gradient boosted machines—on a coarse grid of hyperparameters evaluated across multiple metrics. The best-performing model is further optimized on a fine grid of hyperparameters. The third is that the final model is first validated on unseen test data and then used to screen the Cambridge Structural Database (CSD) for potentially biocompatible MOF candidates for drug delivery.

This paper highlights significant concerns regarding three critical assumptions: the distinction between target prediction accuracy and feature importance accuracy is crucial. Achieving high target prediction accuracy does not necessarily ensure the reliability of feature importances. Menon and Fairen-Jimenez must address fundamental theoretical principles of machine learning, particularly the importance of ground truth values. While supervised machine learning can leverage ground truth values for validating target prediction accuracy, feature importances derived from machine learning models do not have corresponding ground truth values. The lack of these values in feature importance calculations results in disparate methodologies employed by different models, which can inherently lead to biased assessments of feature importance.[2–7] Over 100 peer-reviewed articles documented non-negligible biases in feature importances from models.[2–7]

Despite machine learning models achieving high prediction accuracy, the reliability of feature importance remains questionable because of the absence of ground truth values. While these values effectively validate target prediction accuracy, they fail to provide a robust framework for assessing feature importance. In the absence of ground truth values, accurate determination of target-feature associations requires consideration of three critical statistical elements: data distribution characteristics, statistical relationships between variables, and validation through statistical significance (p values).

This paper therefore recommends augmenting their pipeline with information-theoretic tools—specifically total correlation (TC)[8] and transfer entropy (TE)[9]—to capture complex, higher-order dependencies that standard feature-importance methods miss. TC quantifies the joint synergy and redundancy among multiple descriptors, while TE measures directional, potentially nonmonotonic, information flow between features and the target. Together, these model-agnostic metrics provide a statistically rigorous framework for uncovering intricate target-feature relationships and thus yield more reliable guidance in the *in silico* screening of biocompatible MOFs for drug delivery.

DECLARATION OF INTERESTS

The author declares no conflicting interests.

REFERENCES

1. Menon, D., and Fairen-Jimenez, D. (2025). Guiding the rational design of biocompatible metal-organic frameworks for drug delivery. Matter *8*, 101958. https://doi.org/10.1016/j.matt.2025.101958.

2. Yang, L., Li, H., Wu, Q., Meng, F., Qiu, H., and Xu, L. (2023). Bias-Correction Feature Learner for Semi-Supervised Instance Segmentation. IEEE Trans. Multimedia *25*, 5852–5863. https://doi.org/10.1109/TMM.2022.3199922.

3. Raza, S., Shaban-Nejad, A., Dolatabadi, E., and Mamiya, H. (2024). Exploring Bias and Prediction Metrics to Characterise the Fairness of Machine Learning for Equity-Centered Public Health Decision-Making: A Narrative Review. IEEE Access *12*, 180815–180829. https://doi.org/10.1109/ACCESS.2024.3509353.

4. Fisher, A., Rudin, C., and 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.

5. Ugirumurera, J., Bensen, E.A., Severino, J., and Sanyal, J. (2024). Addressing bias in bagging and boosting regression models. Sci. Rep. *14*, 18452. https://doi.org/10.1038/s41598-024-68907-5.

6. Strobl, C., Boulesteix, A.L., Zeileis, A., and 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.

7. Salles, T., Rocha, L., and Gonçalves, M. (2021). A bias-variance analysis of state-of-the-art random forest text classifiers. Adv. Data Anal. Classif. *15*, 379–405. https://doi.org/10.1007/s11634-020-00409-4.

8. Tserkis, S., Assad, S.M., Lam, P.K., and Narang, P. (2025). Quantifying total correlations in quantum systems through the Pearson correlation coefficient. Phys. Lett. *543*, 130432. https://doi.org/10.1016/j.physleta.2025.130432.

9. Ekhlasi, A., Nasrabadi, A.M., and Mohammadi, M. (2023). Improving transfer entropy and partial transfer entropy for relative detection of effective connectivity strength between time series. Commun. Nonlinear Sci. Numer. Simul. *126*, 107449. https://doi.org/10.1016/j.cnsns.2023.107449.