

SUPPLEMENTARY INFORMATION

FlavorMiner: unraveling the flavor profile of individual molecules from their structure with Machine Learning techniques.

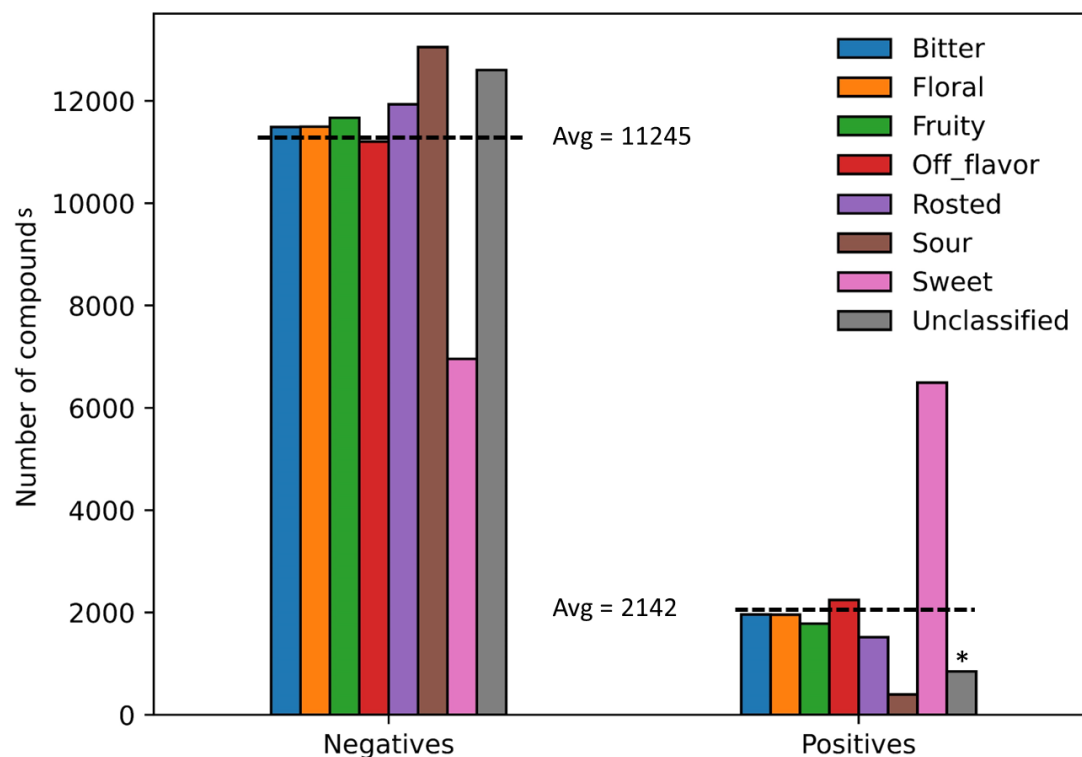
Fabio Herrera-Rocha, Miguel Fernández-Niño, Jorge Duitama, Mónica P. Cala, María José Chica, Ludger

A. Wessjohann, Mehdi D. Davari*, Andrés Fernando González Barrios*

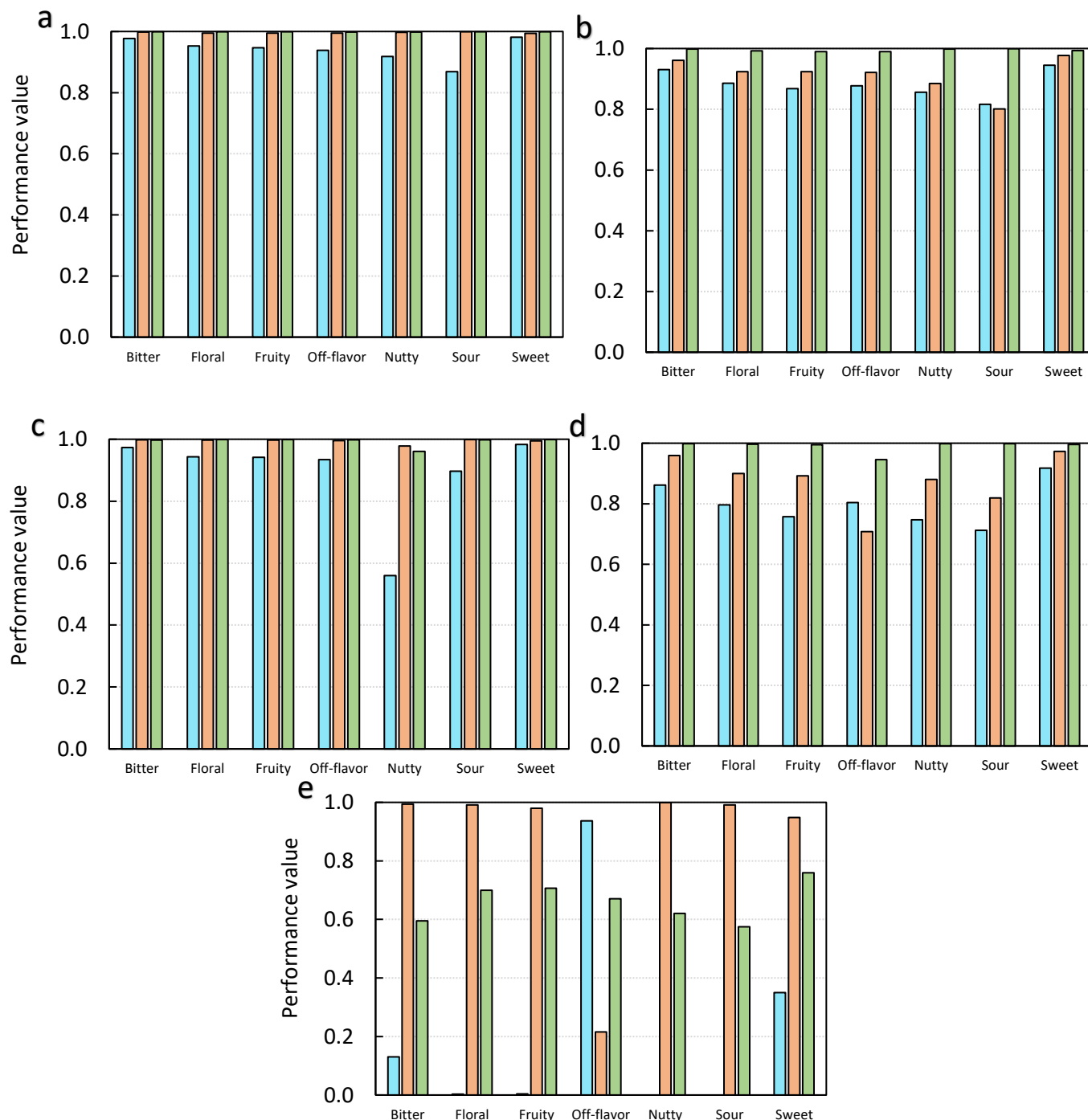
Table of contents

Supplementary Figure 1.	2
Supplementary Figure 2.	3
Supplementary Figure 3.	4
Supplementary Figure 4.	5
Supplementary Figure 5.	6
Supplementary Figure 6.	7
Supplementary Figure 7.	8
Supplementary Figure 8.	8
Supplementary Figure 9.	10
Supplementary Figure 10.	11
Supplementary Figure 11.	12
Supplementary Figure 12.	12

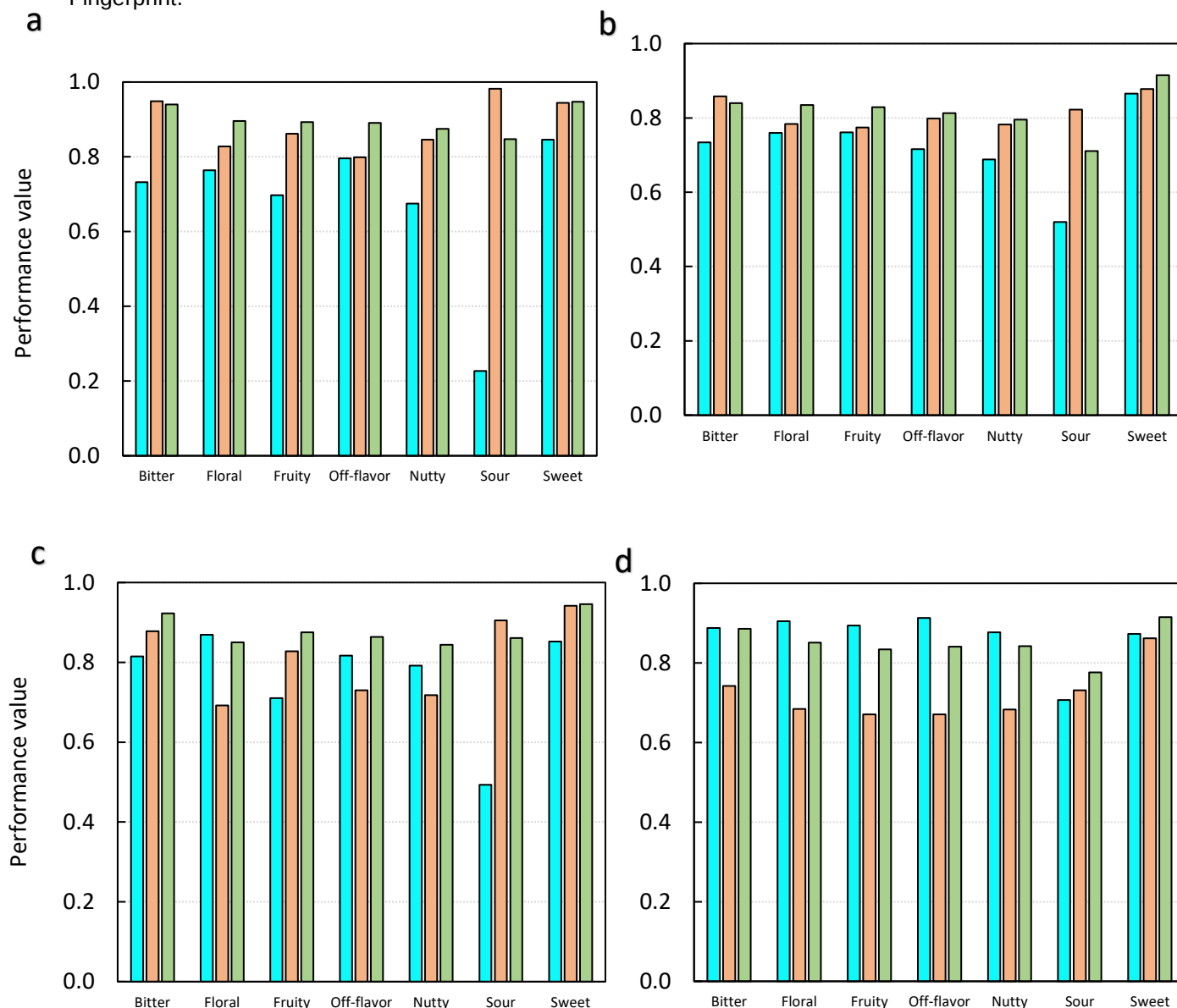
Supplementary Figure 1. Distribution of Positive and Negative Examples in Molecular Flavor Datasets. This figure illustrates the distribution of negative and positive examples in the molecular flavor dataset across all target flavors. The legend also denotes the average number of compounds (Avg). The negative set consists of molecules lacking a specific flavor profile, while the positive set comprises molecules containing the flavor note. Additionally, (*) indicates the presence of an unclassified set containing molecules with flavor profiles not aligning with the target flavor notes.



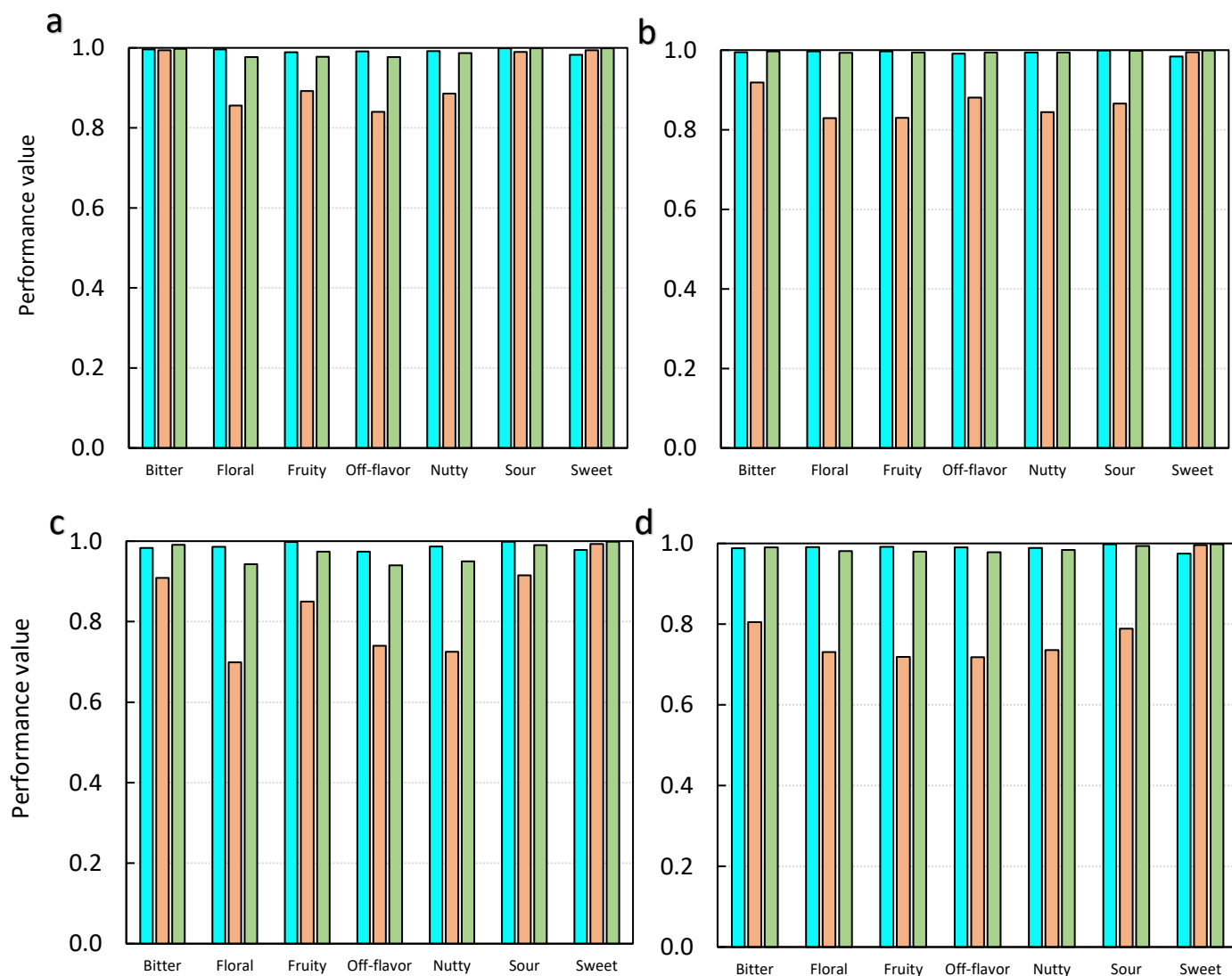
Supplementary Figure 2. Comparative Performance of Machine Learning Algorithms for Flavor Prediction during training. This figure illustrates the performance comparison of various machine learning algorithms for predicting seven target flavors during training using 5-fold cross-validation. Each bar represents Recall (blue bar), Specificity (orange bar), and ROC AUC Score (green bar) for the algorithms across the seven flavors. **(a)** Random Forest with RDKit Molecular Descriptors **(b)** K-Nearest Neighbors with RDKit Molecular Descriptors **(c)** Random Forest with RDKit Extended Connectivity Fingerprint **(d)** K-Nearest Neighbors with RDKit Extended Connectivity Fingerprint **(e)** Graph Neural Network with Molecular Graph.



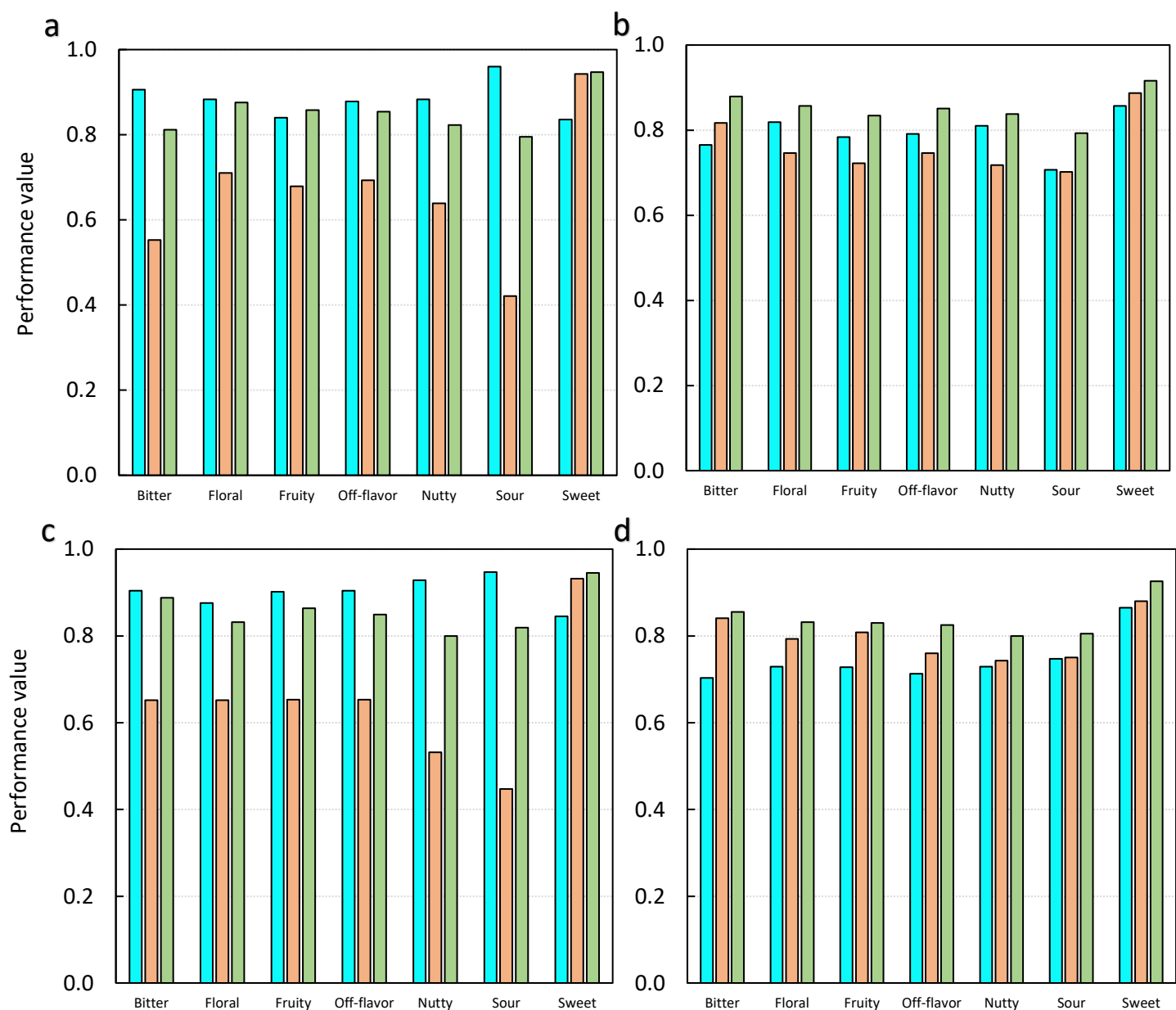
Supplementary Figure 3. Comparative Performance of Machine Learning Algorithms for Flavor Prediction using oversampled data. This figure illustrates the performance comparison of various machine learning algorithms for predicting seven target flavors during test using data oversampled with Synthetic Minority Over-sampling Technique (SMOTE). Each bar represents Recall (blue bar), Specificity (orange bar), and ROC AUC Score (green bar) for the algorithms across the seven flavors. **(a)** Random Forest with RDKit Molecular Descriptors **(b)** K-Nearest Neighbors with RDKit Molecular Descriptors **(c)** Random Forest with RDKit Extended Connectivity Fingerprint **(d)** K-Nearest Neighbors with RDKit Extended Connectivity Fingerprint.



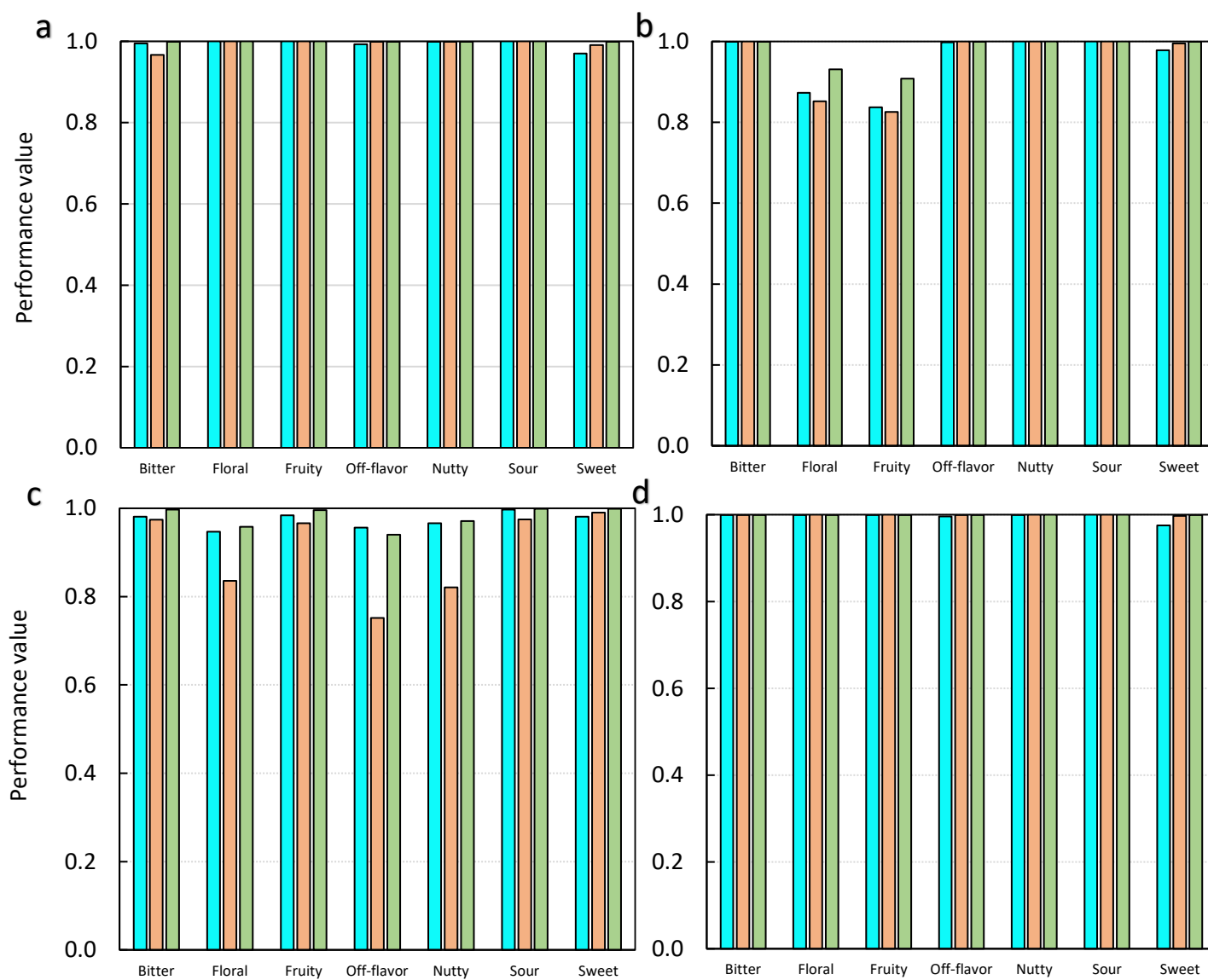
Supplementary Figure 4. Comparative Performance of Machine Learning Algorithms for Flavor Prediction using oversampled data during training. This figure illustrates the performance comparison of various machine learning algorithms for predicting seven target flavors during test using data oversampled with Synthetic Minority Over-sampling Technique (SMOTE) during training using 5-fold cross-validation. Each bar represents Recall (blue bar), Specificity (orange bar), and ROC AUC Score (green bar) for the algorithms across the seven flavors. **(a)** Random Forest with RDKit Molecular Descriptors **(b)** K-Nearest Neighbors with RDKit Molecular Descriptors **(c)** Random Forest with RDKit Extended Connectivity Fingerprint **(d)** K-Nearest Neighbors with RDKit Extended Connectivity Fingerprint.



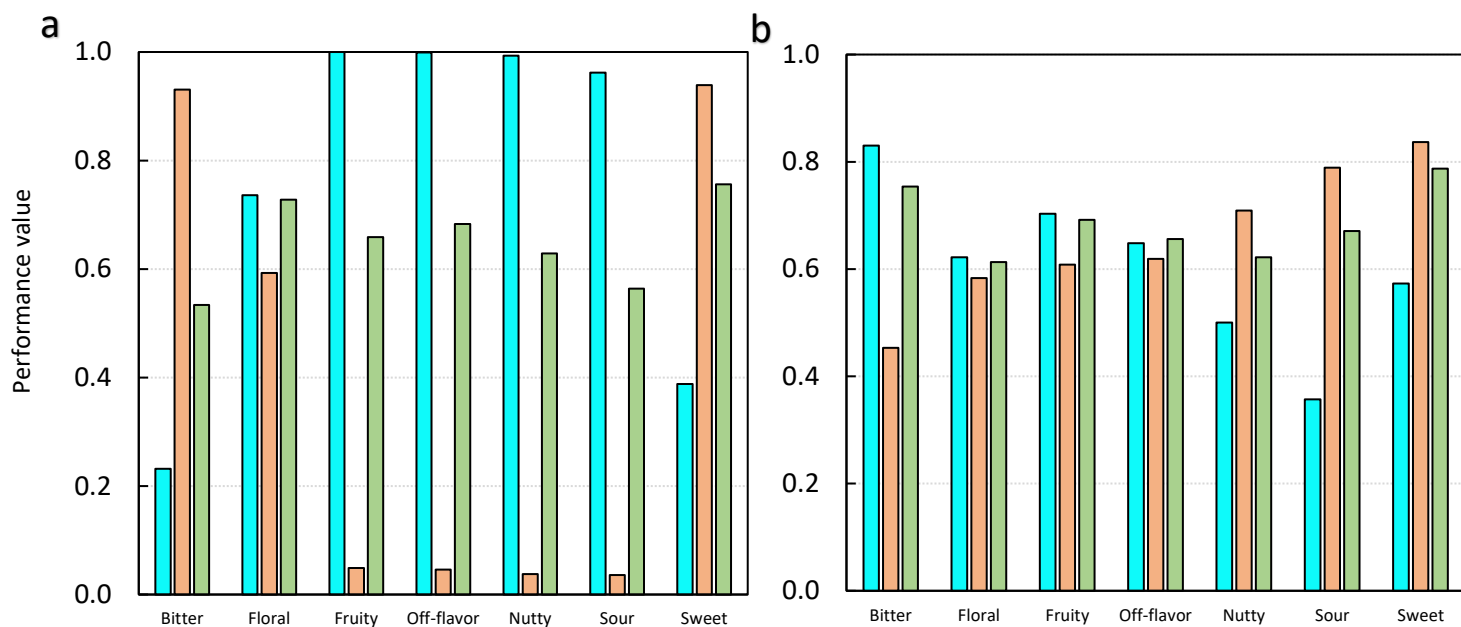
Supplementary Figure 5. Comparative Performance of Machine Learning Algorithms for Flavor Prediction using undersampled data. This figure illustrates the performance comparison of various machine learning algorithms for predicting seven target flavors during test using data oversampled with Cluster Centroid Technique. Each bar represents Recall (blue bar), Specificity (orange bar), and ROC AUC Score (green bar) for the algorithms across the seven flavors. **(a)** Random Forest with RDKit Molecular Descriptors **(b)** K-Nearest Neighbors with RDKit Molecular Descriptors **(c)** Random Forest with RDKit Extended Connectivity Fingerprint **(d)** K-Nearest Neighbors with RDKit Extended Connectivity Fingerprint.



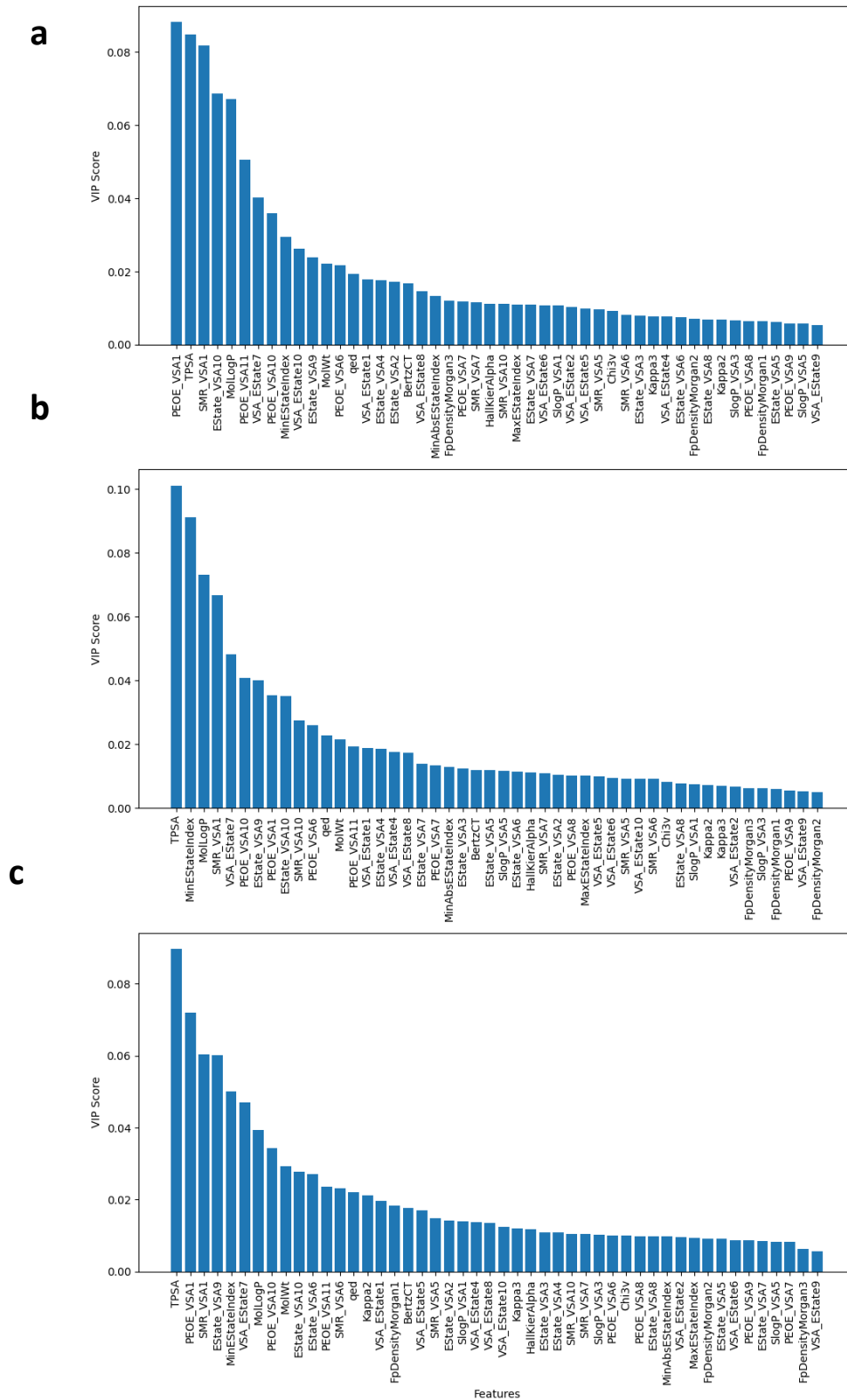
Supplementary Figure 6. Comparative Performance of Machine Learning Algorithms for Flavor Prediction using undersampled data during training. This figure illustrates the performance comparison of various machine learning algorithms for predicting seven target flavors during test using data undersampled with Cluster Centroid Technique during training using 5-fold cross-validation. Each bar Recall (blue bar), Specificity (orange bar), and ROC AUC Score (green bar) for the algorithms across the seven flavors. **(a)** Random Forest with RDKit Molecular Descriptors **(b)** K-Nearest Neighbors with RDKit Molecular Descriptors **(c)** Random Forest with RDKit Extended Connectivity Fingerprint **(d)** K-Nearest Neighbors with RDKit Extended Connectivity Fingerprint.



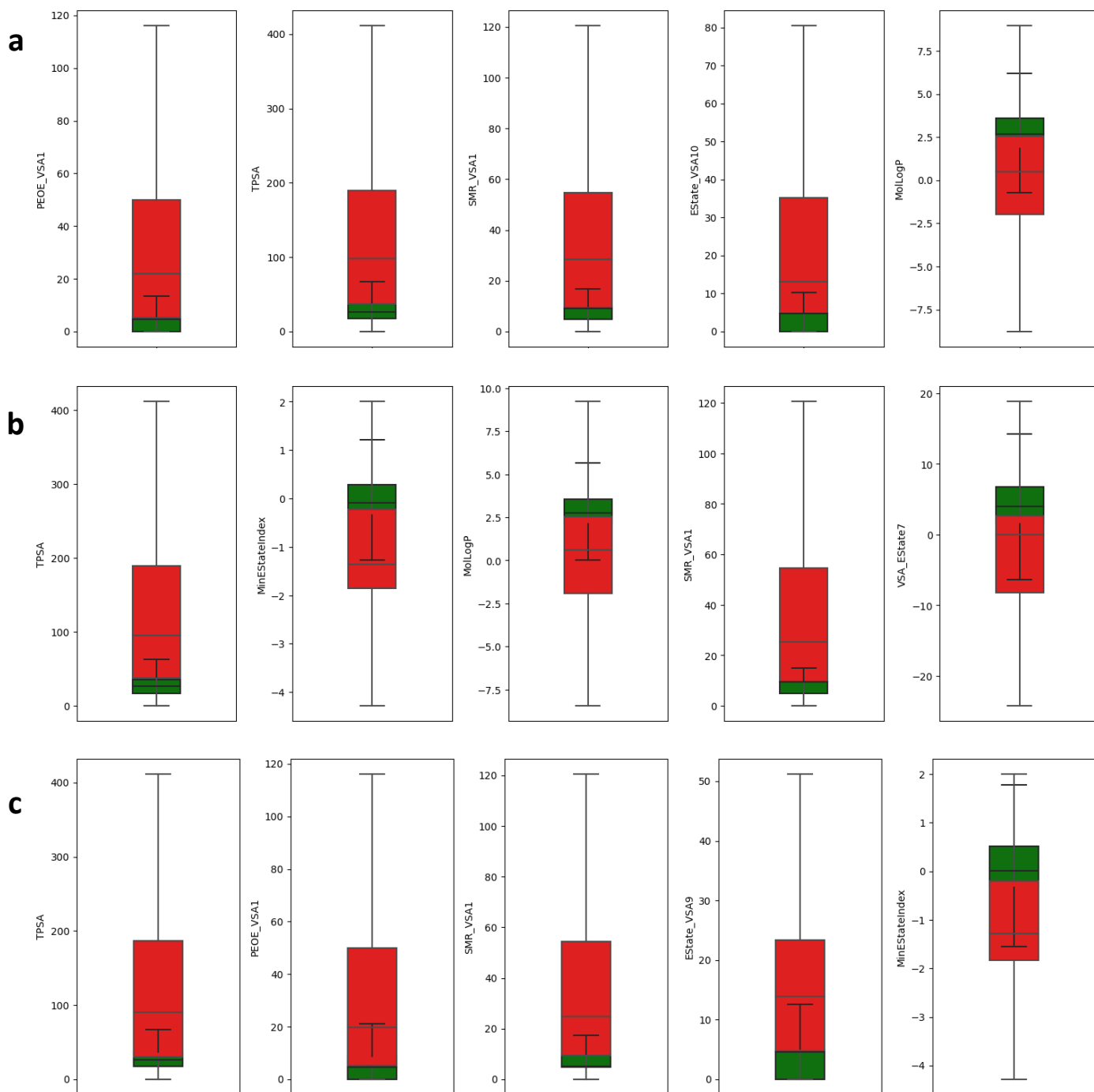
Supplementary Figure 7. Comparative Performance of Machine Learning Algorithms for Flavor Prediction using a Graph Neural Network. This figure illustrates the performance comparison of various machine learning algorithms for predicting seven target flavors during test using data oversampled with resampled Molecular Graph data. Each bar represents Recall (blue bar), Specificity (orange bar), and ROC AUC Score (green bar) for the algorithms across the seven flavors. **(a)** performance during test **(b)** performance during training.



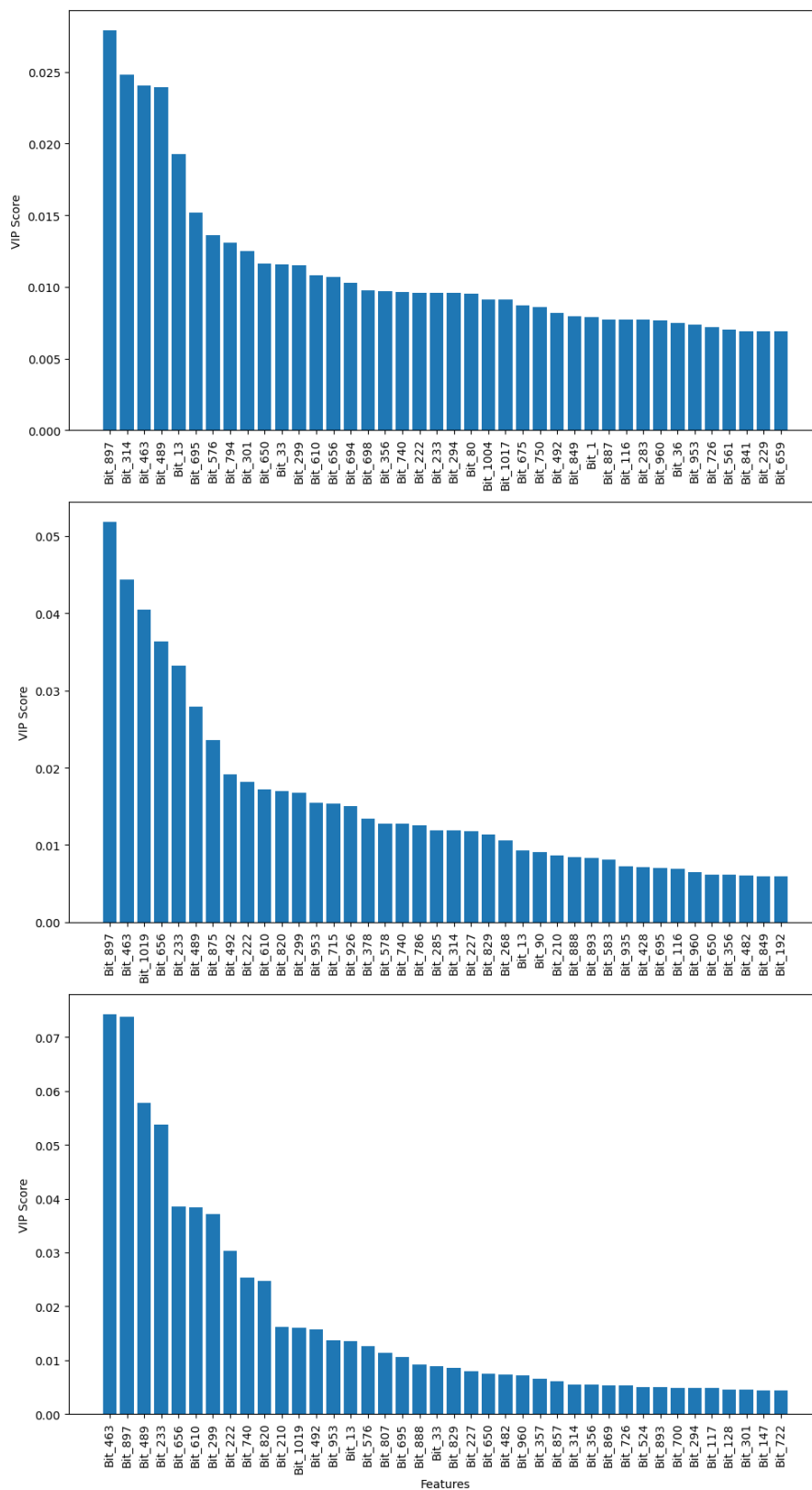
Supplementary Figure 8. Variable Importance Plot (VIP) for Random Forest Predictors with RDKit Molecular Descriptors. This figure showcases VIPs representing the final Random Forest predictors trained with RDKit molecular descriptors. **(a)** Off-flavor **(b)** Floral **(c)** Nutty. For an explanation of each descriptor go to RDKit documentation (<http://www.rdkit.org/docs/GettingStartedInPython.html#list-of-available-descriptors>).



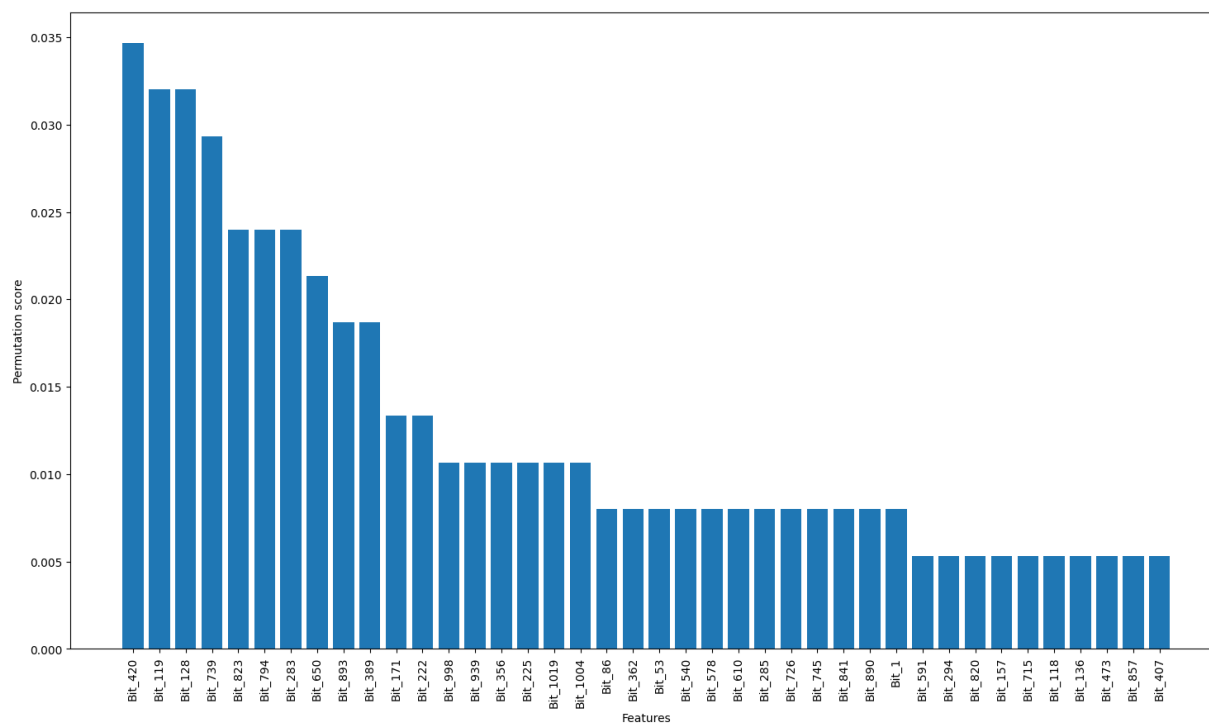
Supplementary Figure 9. Quartile Box Plot of Variable Importance for Models with RDKit Molecular Descriptors. This figure presents Quartile Box Plots illustrating the importance of the first 5 features in the Variable Importance Plot for models trained with RDKit molecular descriptors. Each subplot focuses on a specific flavor category. **(a)** Off-flavor, **(b)** Floral, **(c)** Nutty. Positive examples are denoted in green, while negative examples are represented in red. (For abbreviations like PEOE_VSA1 etc. see Suppl. Fig. 8)



Supplementary Figure 10. Variable Importance Plot (VIP) for Random Forest Predictors with RDKit Molecular Descriptors. This figure showcases VIPs representing the final Random Forest predictors trained with Extended Connectivity Fingerprints (ECFP). **(a)** Off-flavor **(b)** Floral **(c)** Nutty



Supplementary Figure 11. Permutation Score Plot specifically for the Sour predictor.



Supplementary Figure 12. Fragments Representing High Permutation Scores in K-Nearest Neighbors Predictor with ECFP. This figure displays fragments corresponding to the 5 bits with the highest Variable Importance (VIP) scores in K-Nearest Neighbors predictors trained with Extended Connectivity Fingerprint (ECFP) for Sour prediction. (Sour prediction has to be taken with care – see discussion)

