Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Interpretable Ensemble Learning for Multiclass Odor Prediction of Food Flavor Compounds: From Molecular Structure to Sensory Evaluation

Mingxia Zhao
Beijing Life Science Academy
Shiyu Li
Beijing Technology and Business University
Yile Wang
Beijing Technology and Business University
Li Liang
Beijing Technology and Business University
Lili Zhang
Beijing Technology and Business University
Guihong Qi
Beijing Life Science Academy
Yinan Yang
Beijing Life Science Academy
Yuyu Zhang
Beijing Technology and Business University
Wen Duan
15754367187@163.com

Beijing Life Science Academy
Shihao Sun
Beijing Life Science Academy

Article

Keywords:

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11531512/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Interpretable Ensemble Learning for Multiclass Odor Prediction of Food Flavor Compounds: From Molecular Structure to Sensory Evaluation

Mingxia Zhao[a, b, c#], Shiyu Li[b, c#], Yile Wang[d], Li Liang[b, c], Lili Zhang[b, c], Guihong Qi[a], Yinan Yang[a], Yuyu Zhang[b, c], Wen Duan[a, b, c*], Shihao Sun[a*]

[a] Beijing Life Science Academy, Beijing, 102209, China.
[b] Food Laboratory of Zhongyuan, Beijing Technology and Business University, Beijing 100048, China.
[c] Key Laboratory of Flavor Science of China General Chamber of Commerce, Beijing Technology and Business University, Beijing 100048, China.
[d] School of Economics, Beijing Technology and Business University, Beijing 102488, China

*Corresponding author at:

School of Food and health, Beijing Technology and Business University, Beijing 100048, China.
E-mail address: 15754367187@163.com (Wen Duan).
Beijing Life Science Academy, Beijing, 102209, China. E-mail address: sunsh@ztri.com.cn (Shihao Sun).

# These authors contributed equally to this work and are co-first authors.

Email addresses of all authors:
zhaozhang0823@163.com (Mingxia Zhao); lishiyu5130302@163.com (Shiyu Li); wyl13393401611@126.com (Yile Wang); gcfll@126.com (Li Liang); lili921116@163.com (Lili Zhang); qigh@blsa.com.cn (Guihong Qi); yangyn@blsa.com.cn (Yinan Yang); zhangyuyu@btbu.edu.cn (Yuyu Zhang)

Abstract:

This study aimed to develop an interpretable multiclass framework for predicting odor categories of food-relevant flavor compounds from chemical structure, thereby supporting high-throughput flavor screening and ingredient evaluation in food science. A curated dataset of 973 compounds from the Flavor and Extract Manufacturers Association database, covering nine odor categories, was encoded with 1,975 molecular features describing three-dimensional structure, functional groups,

electronic and charge topology, and two-dimensional fingerprints. Variance filtering and mutual information selection reduced the feature set to 300 descriptors within stratified cross-validation. XGBoost, LightGBM, and CatBoost were optimized across multiple random seeds, averaged within each learner family, and combined using a logistic-regression stacking classifier. Out-of-fold cross-validation achieved an accuracy of 84.58%, a macro-averaged F1 score of 0.7718, and a weighted one-vs-rest area under the receiver operating characteristic curve of 0.9649. Ablation analysis showed that removing mutual information selection or multi-seed optimization reduced performance, whereas equal-weight voting slightly outperformed linear stacking. SHapley Additive exPlanations identified ester count as an important feature for fruity odor prediction and nitrogen-containing fragments without attached hydrogen atoms as important features for nutty odor prediction. Molecular docking and density functional theory calculations provided plausible physicochemical interpretations of these attributions. Independent sensory evaluation of 20 food-relevant compounds showed category-dependent agreement between model predictions and human perception, ranging from 50.00% to 78.57%. These results indicate that ensemble learning, feature attribution, molecular simulation, and sensory testing can support interpretable prediction of odor attributes relevant to flavor science and food product development.

Keywords: Odor classification; SHAP; docking; ensemble learning; sensory evaluation

1. Introduction

Odor perception involves the recognition and encoding of volatile chemical signals by the mammalian olfactory system. In food science, this process is central to flavor quality, ingredient selection, off-odor control, and the development of aroma-active compounds for food and beverage products. Understanding structure-odor relationships is also useful for fragrance design, environmental odor monitoring, and the clinical assessment of olfactory dysfunction[1]. Humans possess approximately 400 functional olfactory receptor genes, enabling the detection of a

broad range of structurally diverse odorants[1]. However, molecular structure and odor perception are not related in a straightforward manner: structurally similar compounds can have different odors, whereas structurally dissimilar compounds may share odor attributes[2]. These nonlinear and sometimes ambiguous relationships challenge prediction methods based on empirical rules and expert knowledge[3]. Gas chromatography-olfactometry and trained sensory panels provide direct information on odor properties, but these approaches are labor-intensive, costly, and relatively low-throughput, especially when large candidate libraries must be screened during food-flavor formulation. Their reproducibility can also be affected by differences among evaluators, odor concentration, sample matrix, and experimental conditions. Quantitative structure-activity relationship (QSAR) models offer a computational alternative by linking molecular descriptors, such as logP, molecular weight, polar surface area, and fingerprint patterns, to sensory attributes through regression or classification[4,5]. Conventional linear and shallow machine learning models may nevertheless struggle to capture complex descriptor interactions[5]. More recent approaches incorporate molecular fingerprints, including Morgan fingerprints and MACCS keys, physicochemical descriptors, and three-dimensional conformational information[6]. Machine learning applications have also progressed from support vector machines and random forests[7] to deep learning architectures, including graph neural networks and attention-based models[8,9]. In prospective validation, the graph neural network developed by Lee et al. achieved odor descriptions approaching the accuracy of human evaluators, illustrating the potential of data-driven modeling[2]. Despite this progress, reliance on a single fingerprint or descriptor set can limit the representation of conformation, chirality, charge distribution, and functional-group topology, potentially reducing discrimination among isomers and stereoisomers[4]. Interpretability also remains a challenge for food-flavor applications: predictive accuracy alone does not reveal which chemical relationships a model has learned, and model attributions are rarely examined jointly through intermolecular interactions, electronic structure, and sensory consistency.

To address these limitations, we developed an interpretable ensemble framework for multiclass odor prediction of food-relevant flavor molecules (Figure 1A). The framework integrates four elements: molecular descriptors representing 3D conformation and chirality, functional groups, electronic and charge topology, and 2D topology; variance filtering and mutual information selection fitted within cross-validation to limit information leakage; within-family soft voting and between-family stacking to combine three gradient boosting algorithms; and complementary analyses using SHAP, molecular docking, density functional theory (DFT), and sensory evaluation. This design was intended to balance predictive performance, chemical interpretability, and practical relevance for food-flavor screening. The descriptor set provided chemically diverse information, feature selection reduced dimensionality before model fitting, and the ensemble architecture reduced dependence on any single learner. SHAP, docking, DFT, and sensory validation were then used as complementary layers of evidence, connecting statistical attribution with plausible molecular interactions and human perception. Together, these components connect predictive performance with statistical attribution, physicochemical interpretation, and experimental assessment in a workflow that can assist the prioritization of candidate aroma compounds for food applications.

2. Materials and methods

2.1. Data sources and molecular feature extraction

Data were obtained from the FEMA (Flavor and Extract Manufacturers Association) flavor database[10], which contains compounds relevant to flavor and food-aroma research. Mixtures, entries that did not represent single compounds, invalid SMILES strings, ambiguous odor labels, and non-volatile compounds were excluded. The curated machine-learning dataset comprised 973 records assigned to nine odor categories: Caramellic, Floral, Fruit, Green, Herbal, Nuts, Odorless, Spice, and Sulfur (Figure 1B). These categories correspond to common descriptors used in food aroma description and flavor quality assessment. Fruit represented the largest category, followed by Floral and Odorless, whereas Nuts and Caramellic were

minority categories. Each record contained a SMILES structure and its corresponding odor label. Because odor labels in public flavor databases can reflect perceptual overlap, records were treated as single-label classification targets while recognizing that real food odor perception may be multidimensional and concentration dependent.

RDKit descriptors represented four complementary domains: 3D conformation and chirality, including moments of inertia, sphericity, and asymmetry; counts of 85 functional groups, including esters, thiols, phenols, aldehydes, and ketones; electronic and charge topology, including topological polar surface area and partial-charge-related indices; and 2D topology encoded by structural fingerprints. These domains were selected because odor perception can be influenced by both local functional groups and whole-molecule properties such as size, polarity, hydrophobicity, and shape. Descriptor calculation was followed by quality control. Features with excessive missing or infinite values, all-zero values, extreme sparsity, or strong redundancy were removed before model construction. The processed matrix contained 1,975 numerical molecular features for 973 records, together with SMILES and odor-label columns. The high dimensionality of this matrix motivated the subsequent variance and mutual-information screening steps.

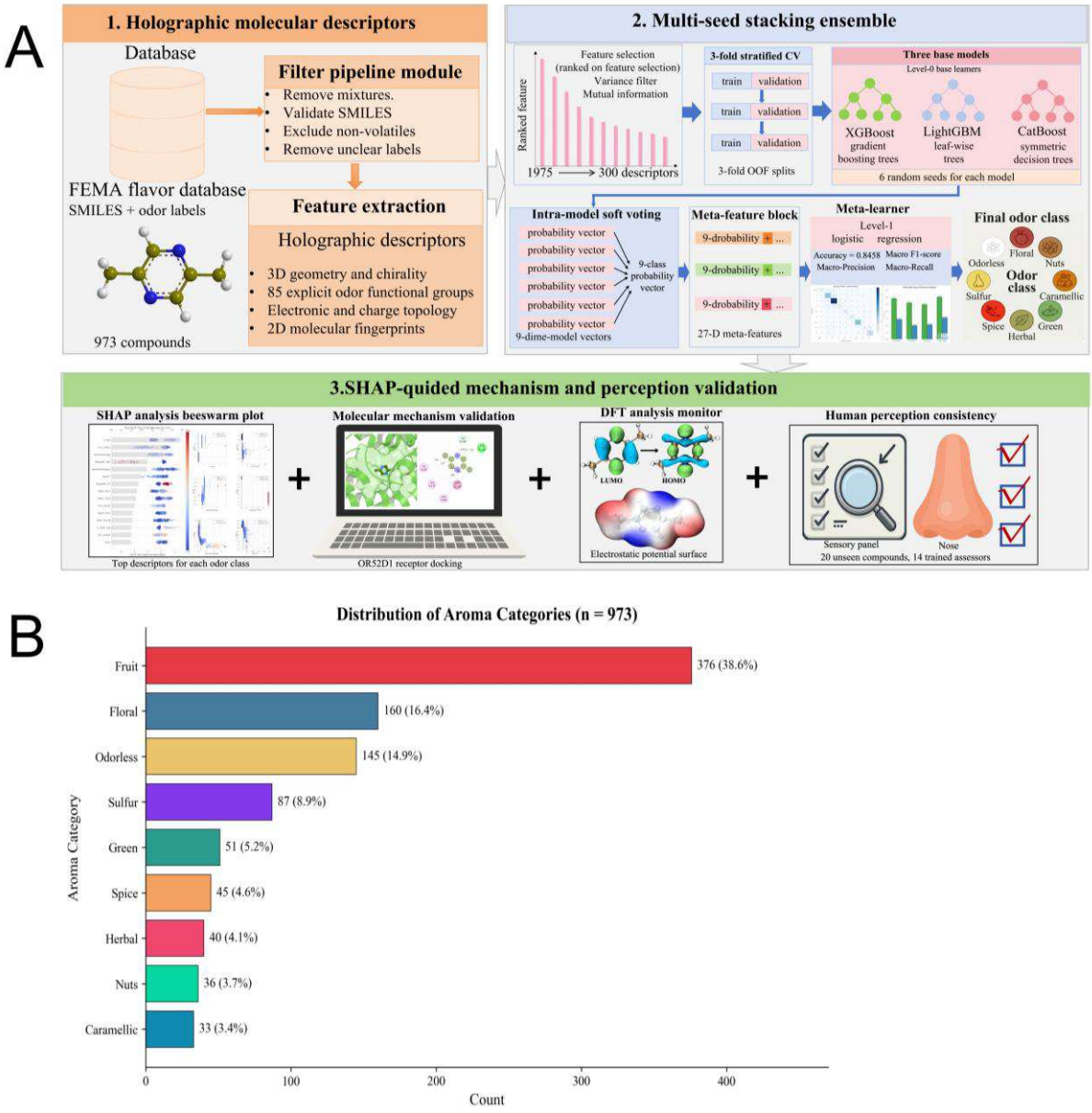


Figure 1. Model Framework(A) and Flavor Category(B)

2.2. Cross-validation and out-of-fold prediction

Model performance was estimated using stratified cross-validation rather than a separate 7:3 held-out test split. Within each Level-0 learner, three-fold stratified cross-validation generated out-of-fold (OOF) probabilities for stacking and approximately preserved class proportions. This OOF design allowed every compound to contribute once to validation-like predictions while still enabling model training on complementary folds. It was particularly useful for the imbalanced

nine-class dataset, where a single hold-out split could yield unstable estimates for minority categories. The Level-1 logistic regression meta-classifier was evaluated using five-fold stratified OOF prediction on the 973-sample dataset. Therefore, the reported model metrics and confusion matrix describe cross-validated OOF performance rather than prediction on an independent held-out test set.

2.3. Feature selection

Feature selection combined variance filtering with mutual information ranking within three-fold stratified cross-validation[11]. Selection statistics were estimated exclusively from each training fold, and the resulting feature indices were applied to the corresponding validation fold. This fold-specific procedure reduced the risk that information from validation samples would influence feature ranking. Variance filtering first removed descriptors with little sample-level variation, because such features contribute limited discriminative information and can increase computational burden. Mutual information was then used to retain features with nonlinear associations with the odor label, allowing potentially informative descriptors to be selected even when their relationship to class membership was not monotonic or linear.

Descriptors with sample variance greater than 0.01 were retained. Mutual information between each retained descriptor and the nine-class odor label was estimated using a nonparametric nearest-neighbor approach. SelectKBest retained the 300 highest-scoring descriptors from the 1,975 initial molecular features to capture nonlinear feature-label associations while reducing dimensionality and computational cost. The selected set was not interpreted as a complete chemical explanation of odor; rather, it served as a compact and data-driven representation for model fitting. This distinction is important because descriptors excluded by mutual information screening may still be chemically meaningful, but less useful for discrimination within this specific dataset.

2.4. Model development

The framework used XGBoost[12], LightGBM[13], and CatBoost as heterogeneous base learners. These algorithms provide complementary modeling strategies: regularized boosting in XGBoost, histogram-based leaf-wise tree growth in LightGBM, and ordered boosting with symmetric trees in CatBoost. All three can model nonlinear decision boundaries and interactions among descriptors, but they differ in tree growth, regularization, and handling of stochasticity. Their combination was expected to improve robustness because different algorithms may emphasize different descriptor patterns even when trained on the same selected feature space.

2.4.1. Multi-seed optimization and within-family soft voting

For each learner, independent randomized hyperparameter searches used six consecutive seeds (0-5) to control data splitting and stochastic training operations, including sample and feature subsampling. Feature selection and candidate evaluation were performed within cross-validation. The best estimator for each seed was selected using the weighted one-vs-rest area under the receiver operating characteristic curve (weighted OVR-AUC), which accounts for multiclass discrimination while giving larger classes proportionally greater weight. Within each learner family, the nine-class probability vectors from the six selected estimators were averaged to produce XGBoost-Ensemble, LightGBM-Ensemble, and CatBoost-Ensemble. Averaging across seeds was used to stabilize predictions and reduce sensitivity to a single random split or hyperparameter trajectory.

2.4.2. Stacking ensemble

Stacking combined the three within-family ensembles using a multiclass logistic regression meta-classifier[14]. For each training sample, the three-fold OOF predictions from the three ensembles were concatenated into a 27-dimensional vector (three learners x nine class probabilities). These meta-features summarized how each base learner distributed probability mass across the nine odor categories. The meta-classifier learned from these vectors, generated final class probabilities using softmax, and assigned the class with the highest probability. An equal-weight voting

baseline averaged the probability vectors of the three ensembles without meta-learning. Comparing stacking with this voting baseline allowed the analysis to distinguish the value of broad ensemble averaging from the additional contribution of the linear meta-classifier.

2.4.3. Model evaluation

Performance was evaluated from OOF predictions using accuracy, macro-averaged precision, recall, and F1, and weighted OVR-AUC. Macro-averaged metrics assigned equal weight to each class, whereas weighted OVR-AUC used class sample proportions as weights. K-nearest neighbors (KNN), a decision tree, and individual boosting models served as comparators. Ablation experiments examined the effects of removing mutual information selection or multi-seed optimization and replacing stacking with equal-weight voting.

2.5. SHAP analysis

SHAP was used to quantify the contributions of selected molecular descriptors to Level-0 base-learner predictions[15,16]. SHAP values were calculated across the 300 selected descriptors for each odor category. For global interpretation, features were ranked by mean absolute SHAP value, which reflects the average magnitude of contribution regardless of direction. Signed SHAP values were also inspected because they indicate whether a descriptor tends to increase or decrease the model output relative to its baseline. This distinction allowed the analysis to identify both globally important descriptors and the direction of their associations with category-specific predictions.

2.6. Molecular docking and electronic structure analysis

2.6.1. Molecular docking

The compound with the highest predicted class probability in each odor category was selected for exploratory docking against a structure assigned to human OR52D1 (UniProt Q9H346), which was chosen for its reported broad tuning[17]. The receptor structure was obtained from UniProt, water molecules were removed,

hydrogens were added, and Gasteiger charges were assigned. Ligand structures were generated in Chem3D and energy-minimized using MMFF94 to obtain plausible low-energy conformations before docking. Docking was performed in AutoDock Vina 1.2.0 with a rigid receptor and flexible ligands[18]. PyMOL 2.6 was used to define a docking box centered at (0.13, 0.67, 2.60) Å, with dimensions of 70 x 70 x 85 Å. Each ligand underwent 50 independent runs, and the optimal pose was selected on the basis of binding energy and clustering density. Docking was used as a qualitative structural analysis rather than direct proof of receptor activation.

2.6.2. DFT calculations

Representative compounds underwent geometry optimization and energy calculations at the B3LYP/6-31G(d,p) level[19]. Frontier orbital distributions, the energy gap between the highest occupied and lowest unoccupied molecular orbitals (HOMO-LUMO gap; Delta E = E_LUMO - E_HOMO), and molecular electrostatic potential were analyzed to characterize electronic structure and potential hydrogen-bonding or electrostatic interaction sites[20]. HOMO and LUMO energies are reported in eV. These calculations were intended to complement the docking and SHAP results by showing whether model-highlighted functional groups were also located in electronically active regions of the representative molecules.

2.7. Sensory evaluation

Twenty independent compounds that had not been used for model training or validation were assigned randomized three-digit codes and presented in random order. These compounds were selected as food-relevant odorants to provide an external sensory check under controlled evaluation conditions. The panel comprised 14 healthy participants who had received at least 20 hours of basic flavor-recognition training. Evaluation conditions followed ISO 8589[21]. Panelists first familiarized themselves with category references and then classified each sample into the most appropriate odor category by smelling. Evaluations were separated by at least 30 seconds. Between samples, panelists rinsed their mouths with tasteless purified water and smelled odorless test paper to reduce carryover effects. Three sessions separated

by 24 hours yielded 42 assessments per compound[22]. Agreement was calculated by comparing the model-predicted category with the distribution of human panel assignments for each compound and category, excluding the odorless category from the odor-active sensory comparison. The protocol was approved by the Ethics Committee of Beijing Technology and Business University (BTBU202661). All participants provided written informed consent and could withdraw at any time.

3. Results and discussion

3.1. Optimal hyperparameters and training convergence

Representative hyperparameter combinations associated with the highest validation AUC for each seed are summarized in Table 1. After selection of 300 descriptors, all three boosting models favored moderate tree complexity and medium-to-high learning rates. This pattern is consistent with the size and structure of the dataset: overly shallow trees may fail to capture nonlinear descriptor interactions, whereas very deep trees may overfit minority-category patterns. The selected configurations therefore reflected a compromise between flexibility and regularization.

Table 1. Representative Optimal Hyperparameter Settings of Three Types of Base Learners

Model	Best Seed	Key Hyperparameters
XGBoost	Seed_2	n_estimators = 300, max_depth = 7, learning_rate = 0.1, subsample = 0.7
LightGBM	Seed_1	n_estimators = 300, max_depth = 5, learning_rate = 0.1, subsample = 0.9
CatBoost	Seed_0	iterations = 300, depth = 6, learning_rate = 0.1

XGBoost used max_depth = 7 and subsample = 0.7, combining deeper trees with subsampling. LightGBM used max_depth = 5 and subsample = 0.9, restricting leaf-wise tree growth while retaining more samples. CatBoost used depth = 6, 300 iterations, a learning rate of 0.1, and MultiClass loss. These configurations illustrate different approaches to controlling model complexity in the selected descriptor space.

CatBoost training loss decreased rapidly from approximately 2.0 to 0.6 during the early iterations, declined more gradually after approximately 150 iterations, and continued to decrease through iteration 300 (Figure 2A). The curve indicated stable optimization without obvious oscillation or divergence. The slower late-stage

decrease suggests that the model continued refining class boundaries after the largest gains had already been achieved. Because the loss did not show abrupt instability near the final iteration, the selected CatBoost configuration was considered suitable for inclusion in the ensemble.

3.2. Predictive performance across models

The interpretable ensemble framework was compared with KNN, a decision tree, and the individual XGBoost, LightGBM, and CatBoost models using OOF cross-validated predictions (Table 2 and Figure 2B). This comparison was designed to evaluate whether nonlinear boosting and ensemble integration improved discrimination over simpler baseline methods for food-related odor classification. The same processed descriptor matrix and selected feature representation were used so that performance differences mainly reflected modeling strategy rather than differences in input information.

KNN achieved an accuracy of 0.7708 and a macro-averaged F1 score of 0.6765, whereas the decision tree achieved 0.7338 and 0.6581, respectively. Both performed less well than the boosting models, consistent with the findings of Lusito et al.[23]. These results suggest that distance-based classification and a single tree may be insufficient to capture the structure–odor relationships represented by the 300 selected descriptors.

XGBoost, LightGBM, and CatBoost achieved accuracies of 0.8150, 0.8221, and 0.8191 and macro-averaged F1 scores of 0.7419, 0.7468, and 0.7486, respectively. All three achieved weighted AUC values above 0.96. Their stronger performance suggests that iterative boosting captured nonlinear relationships and interactions among molecular descriptors more effectively than the simpler comparators. LightGBM achieved the highest accuracy among the individual boosting models, whereas CatBoost performed best in macro-averaged F1 and weighted AUC, suggesting some complementarity in their predictions.

Table 2. Comparison of classification performance of different models on the test set

Model	Accuracy	Macro Precision	Macro Recall	Macro F1	Weighted AUC
KNN	0.7708	0.7093	0.6686	0.6765	0.9285

Decision Tree	0.7338	0.6450	0.6744	0.6581	0.8367
XGBoost (single)	0.8150	0.7631	0.7272	0.7419	0.9653
LightGBM (single)	0.8221	0.7817	0.7281	0.7468	0.9616
CatBoost (single)	0.8191	0.7768	0.7362	0.7486	0.9675
Stacking (Full framework)	0.8458	0.8099	0.7496	0.7718	0.9649

3.3. Confusion matrix analysis

The OOF confusion matrix was examined to assess performance across the nine odor categories (Figure 2C). The Fruit category, which contained the largest number of samples, had a recall of 0.9255, with 348 of 376 Fruit samples correctly classified. This result is relevant to food-flavor screening because fruity notes are widely used in beverages, dairy products, confectionery, and formulated flavors. The recall exceeded that for all categories except Odorless, which also showed strong separation from odor-active classes. Off-diagonal entries revealed category-specific misclassification patterns[24]: 18 Floral samples and 14 Green samples were classified as Fruit, and 10 Spice samples were classified as Sulfur. These errors may reflect shared topological or functional-group characteristics that make the corresponding categories difficult to distinguish. For example, floral and fruity odorants often share ester, alcohol, or terpene-like structural motifs, whereas spicy and sulfurous notes can involve pungent or high-impact functional groups. The matrix therefore indicates that most errors occurred between perceptually or chemically neighboring categories rather than uniformly across all classes, a pattern consistent with overlapping aroma descriptors in food sensory evaluation.

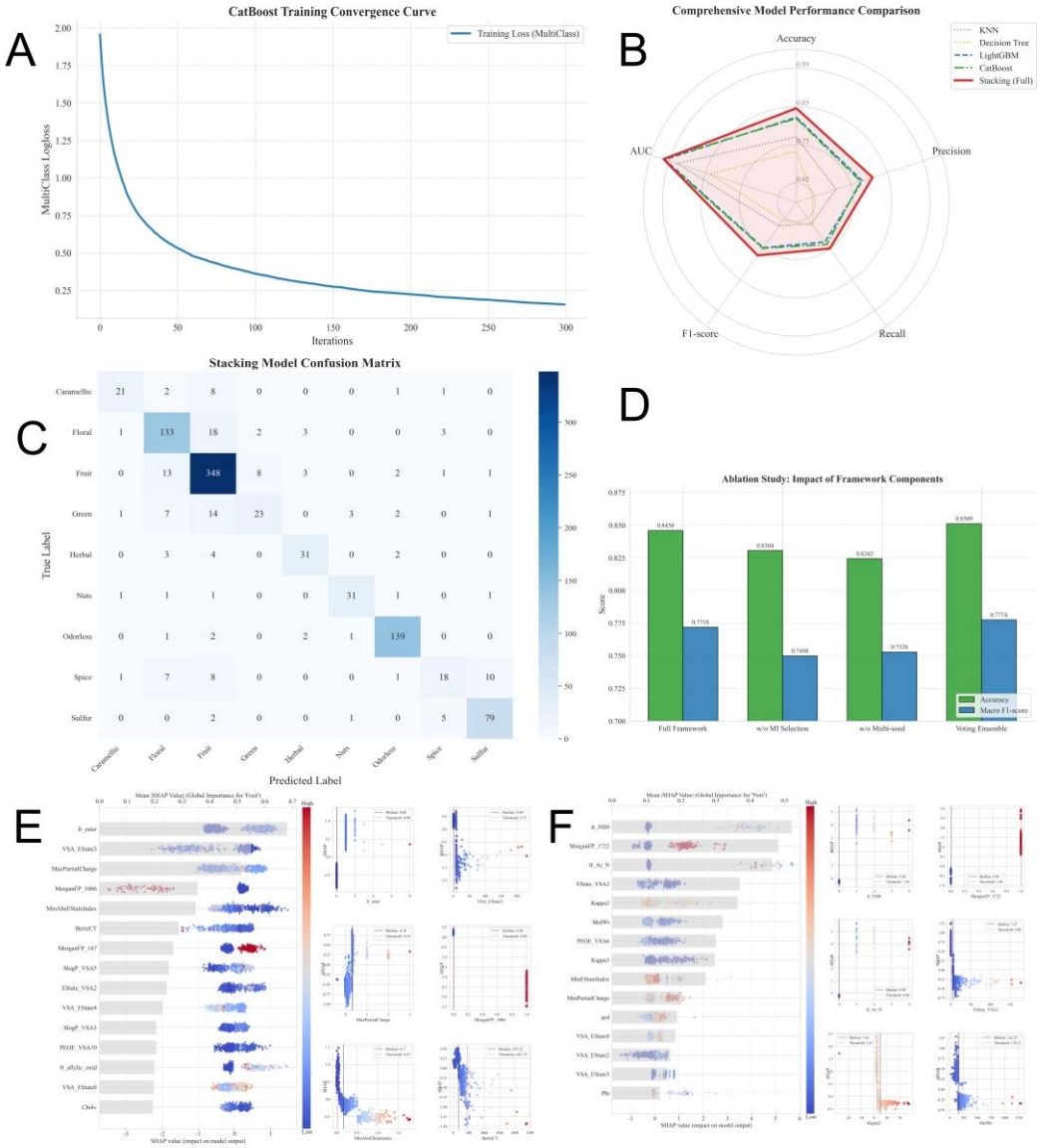


Figure 2. CatBoost Training Convergence Curve (A), Comprehensive Model Performance Comparison (B), Stacking Model Confusion Matrix (C). Ablation Study: Impact of Framework Components (D), Model's ranking of SHAP global feature importance for predicting Fruit (E) and Nuts (F)

3.4. Ablation analysis

Ablation experiments assessed the contributions of mutual information selection and multi-seed optimization. Equal-weight probability averaging also served as a comparator for the stacking meta-learner (Table 3 and Figure 2D).

Removing mutual information selection reduced accuracy to 0.8304 and macro-averaged F1 to 0.7498, corresponding to decreases of 1.54 and 2.20 percentage points, respectively, relative to the complete framework. This finding suggests that

selecting 300 informative descriptors reduced the influence of noisy variables in the original 1,975-dimensional feature space. In high-dimensional molecular data, redundant or weakly informative descriptors can increase variance and make minority-class decision boundaries less stable. Removing multi-seed optimization reduced accuracy to 0.8242 and macro-averaged F1 to 0.7528, decreases of 2.16 and 1.90 percentage points, respectively. These results support the use of multiple seeds to reduce sensitivity to stochastic variation in model optimization. The larger accuracy decrease after removing multi-seed optimization suggests that seed averaging contributed particularly to overall label assignment, whereas mutual information selection contributed strongly to balanced category-level F1.

Equal-weight voting achieved slightly higher accuracy and macro-averaged F1 (0.8509 and 0.7775) than stacking (0.8458 and 0.7718). One possible explanation is that the base learners generated similar OOF probabilities, leaving limited additional information for the linear meta-classifier to exploit. The logistic regression meta-classifier may also have been unable to capture nonlinear relationships among the 27 probability features[25]. However, these explanations remain hypotheses. The absolute difference in macro-averaged F1 was 0.0056; future work could evaluate nonlinear meta-classifiers, such as multilayer perceptrons or XGBoost, using independent validation and food-matrix-based sensory datasets.

Overall, the ablation results support the contributions of feature selection and multi-seed optimization. The current linear stacking model did not outperform equal-weight voting on all metrics, so the findings support the broader ensemble approach rather than a clear advantage of stacking itself[26-28]. This result is useful because it shows that model combination improved performance, but that the simplest probability-averaging strategy remained highly competitive. For datasets of this size, a linear meta-classifier may not always extract more information than calibrated base-learner probabilities already contain.

Table 3 Comparison of Performance of Each Model in Ablation Experiments

Setting	Accuracy	Macro Recall	Macro F1	Weighted AUC
Full framework (stacking)	0.8458	0.7496	0.7718	0.9649
Without MI Selection (LGBM proxy)	0.8304	0.7376	0.7498	0.9679
Without multi-seed optimization	0.8242	0.7323	0.7528	0.9573
Replace Stacking with Voting	0.8509	0.7564	0.7774	0.9739

3.5. SHAP-based interpretation

SHAP analysis examined the molecular descriptors contributing most strongly to the base learners' predictions for each odor category[29,30]. The results identified functional-group features with chemically interpretable associations (Figure 2E and Figure 2F). For the Fruit category, ester count (fr_ester) had the highest mean absolute SHAP value, and the presence of an ester group generally contributed positively to the prediction. This agrees with established flavor-chemistry knowledge[31]: fruity odorants such as ethyl acetate, ethyl butyrate, and isopentyl esters contain ester groups that influence molecular volatility and polarity and can participate in receptor interactions through their oxygen atoms. For the Nuts category, fr_NH0, which represents nitrogen-containing fragments without attached hydrogen atoms, was the highest-ranked feature by mean absolute SHAP value. Nitrogen-containing heterocycles, including pyrazines, thiazoles, and pyridines, are associated with nutty, roasted, and burnt odors and commonly occur among products of Maillard reactions or thermal processing. Ring nitrogen atoms can contribute to polarity and, where chemically appropriate, act as hydrogen-bond acceptors. The prominence of fr_NH0 is therefore consistent with known structure-odor associations. SHAP values nevertheless describe statistical contributions to model predictions and do not establish a causal mechanism. Representative fruity and nutty odorants were subsequently examined by molecular docking and DFT to explore possible physicochemical explanations for these attributions.

3.6. Molecular docking and DFT results

Compounds with the highest prediction confidence in each odor category were selected for molecular docking and DFT calculations (Table 4). The fruity and nutty

odorants are discussed here because these notes are common and practically important in food-aroma design, including fruit-flavored products, roasted foods, and nut-containing flavor systems. The analysis examined whether the functional groups highlighted by SHAP also corresponded to plausible receptor-interaction sites and distinctive electronic features. This analysis was not intended to prove a universal receptor mechanism for each odor class. Instead, it asked whether the statistical features emphasized by the model could be linked to chemically interpretable interaction patterns in representative food-relevant odorants.

3.6.1. Predicted receptor interactions

Isopropyl 2-methylbutyrate (CAS 66576-71-4) had the highest prediction confidence for the Fruit category, at 0.9538. Docking yielded a pose within the receptor pocket with a predicted binding energy of -3.80 kcal/mol (Figure 3). In this pose, the ester carbonyl oxygen formed a predicted hydrogen bond with GLY A:19 at a distance of 2.98 Å. This interaction provides a plausible structural interpretation of the positive contribution of fr_ester: the ester group may serve as a receptor-recognition site as well as a classification feature. The hydrophobic portion of the molecule may also help stabilize the ligand within the binding pocket, while the carbonyl oxygen provides a localized polar site for directional interaction. The docking result supports this interpretation but does not independently establish receptor activation or odor specificity.

For the Nuts category, 2,5-dimethylpyrazine (CAS 123-32-0) had the highest prediction confidence, at 0.9443. Its predicted docking pose showed a hydrogen bond between a pyrazine ring nitrogen and TYR A:178, together with hydrogen-bond-related interactions involving TYR A:258 (Figure 3). These contacts may help orient the ligand within the receptor pocket. The ring nitrogen atoms correspond to the nitrogen sites represented by fr_NH0, providing a possible physicochemical explanation for this feature's positive contribution to nutty odor predictions. The aromatic pyrazine scaffold also provides a compact, electron-rich structure that is common among roasted and nutty odorants. As with the fruity odorant,

these computational associations require experimental confirmation of receptor binding and activation.

3.6.2. Frontier orbitals and electrostatic potential

Frontier molecular orbitals and electrostatic potential were analyzed to further interpret the functional groups highlighted by SHAP and docking[32]. As shown in Figure 4, the representative fruity compound exhibited HOMO and LUMO distributions mainly around the ester group and adjacent structural regions (Figure 4A). After conversion from Hartree to electron volts, the LUMO and HOMO energies were -0.082 eV and -7.374 eV, respectively, corresponding to a HOMO-LUMO gap of 7.292 eV. These distributions suggest that the ester-containing region is electronically relevant to potential intermolecular interactions. The electrostatic potential map of the fruity compound further showed local polarity around the ester group (Figure 4C), with the blue region indicating an ester-associated positive electrostatic region. Together, these results are consistent with the contribution of fr_ester to classification and suggest the possible involvement of ester oxygen atoms in hydrogen-bonding and dipolar interactions. Thus, the electronic analysis supports the same molecular region highlighted by both SHAP and docking.

For the representative nutty compound, the HOMO and LUMO were mainly distributed across the nitrogen-containing aromatic ring, including the nitrogen atoms and adjacent regions (Figure 4B). The converted LUMO and HOMO energies were -1.469 eV and -6.857 eV, respectively, giving a HOMO-LUMO gap of 5.388 eV. Compared with the fruity compound, the nutty compound showed a lower LUMO energy, a higher HOMO energy, and a smaller HOMO-LUMO gap, suggesting comparatively greater electronic responsiveness within the nitrogen-containing aromatic system. The electrostatic potential map also showed pronounced local positive regions around the heteroatom-containing ring (Figure 4D), supporting the potential role of ring nitrogen atoms in receptor recognition. The correspondence among these electronically active sites, the predicted docking contacts, and the SHAP-highlighted fr_NH0 descriptor provides a coherent physicochemical

interpretation of the model attribution. However, orbital energies and docking poses alone cannot establish the biological mechanism underlying odor perception.

Table 4 Docking Results of Representative Compound

Class	CAS	Confidence Score	Binding Energy (kcal/mol)
Caramellic	3658-77-3	0.9536	-3.67
Floral	2344-70-9	0.9432	-4.87
Fruit	66576-71-4	0.9538	-3.80
Green	30689-75-9	0.9314	-4.27
Herbal	89-80-5	0.9448	-4.82
Nuts	123-32-0	0.9443	-4.19
Spice	22014-48-8	0.9194	-3.44
Sulfur	107-03-9	0.9332	-2.65

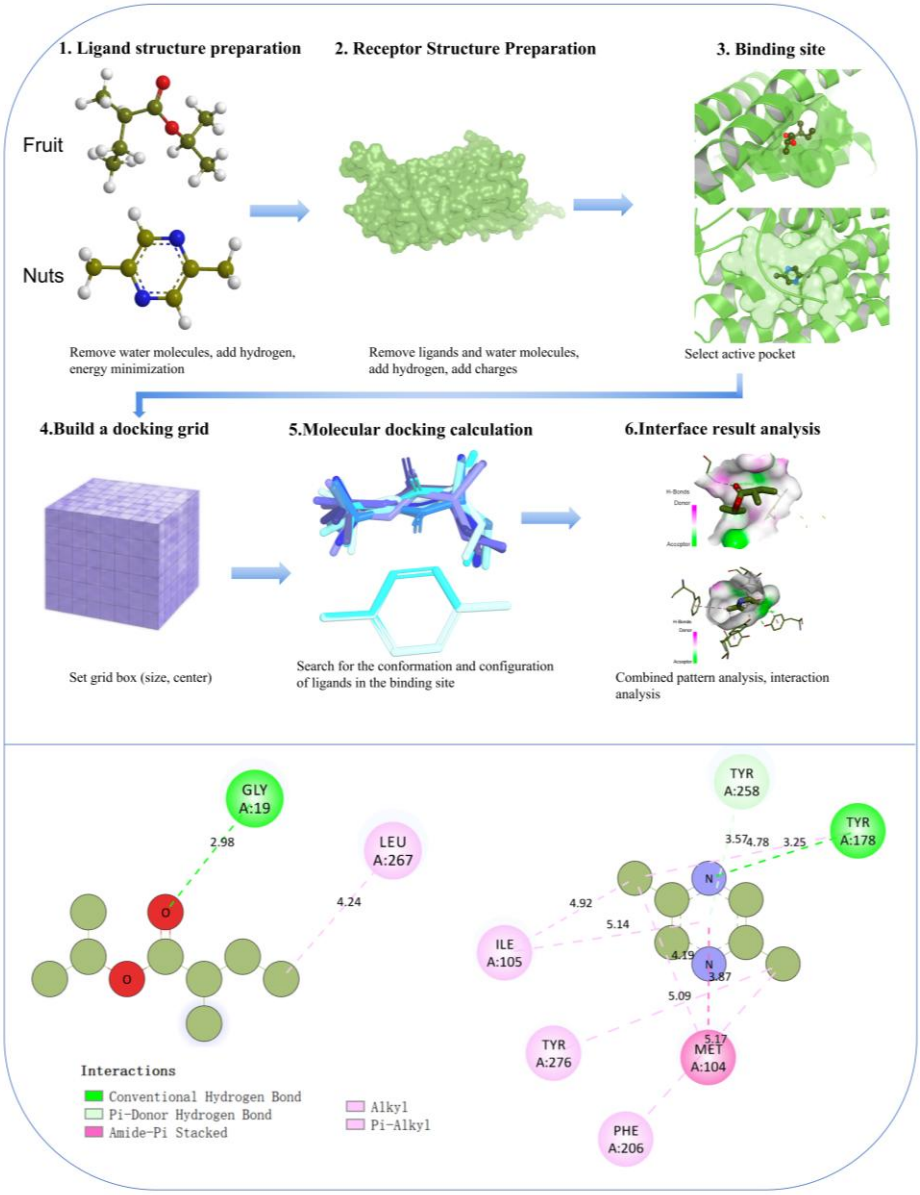


Figure 3. Molecular Docking Analysis of Representative Compounds

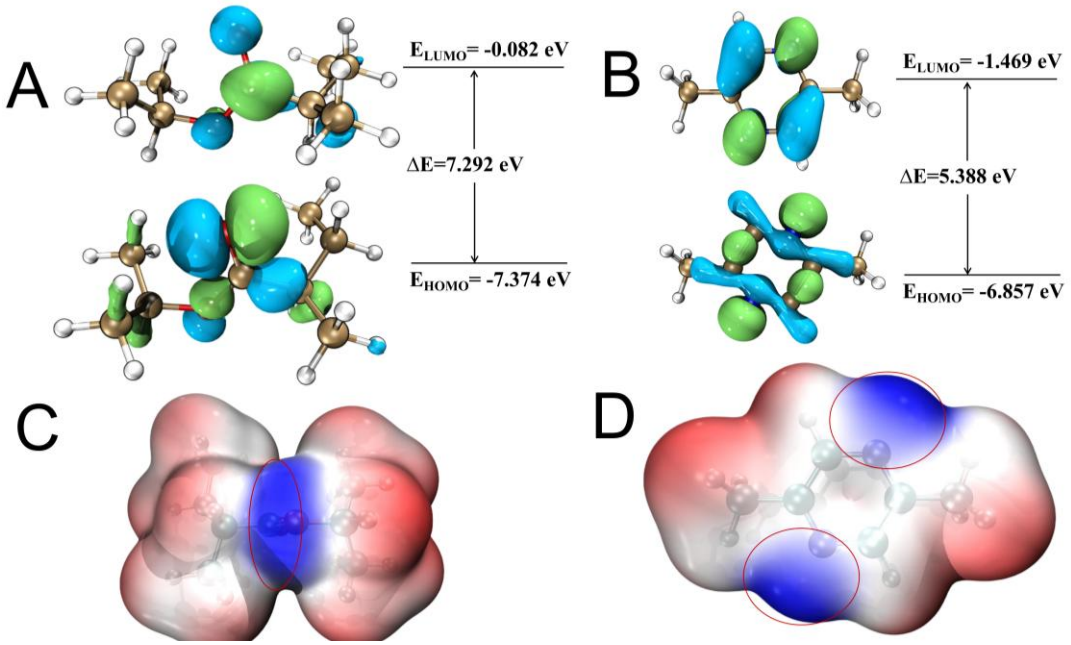
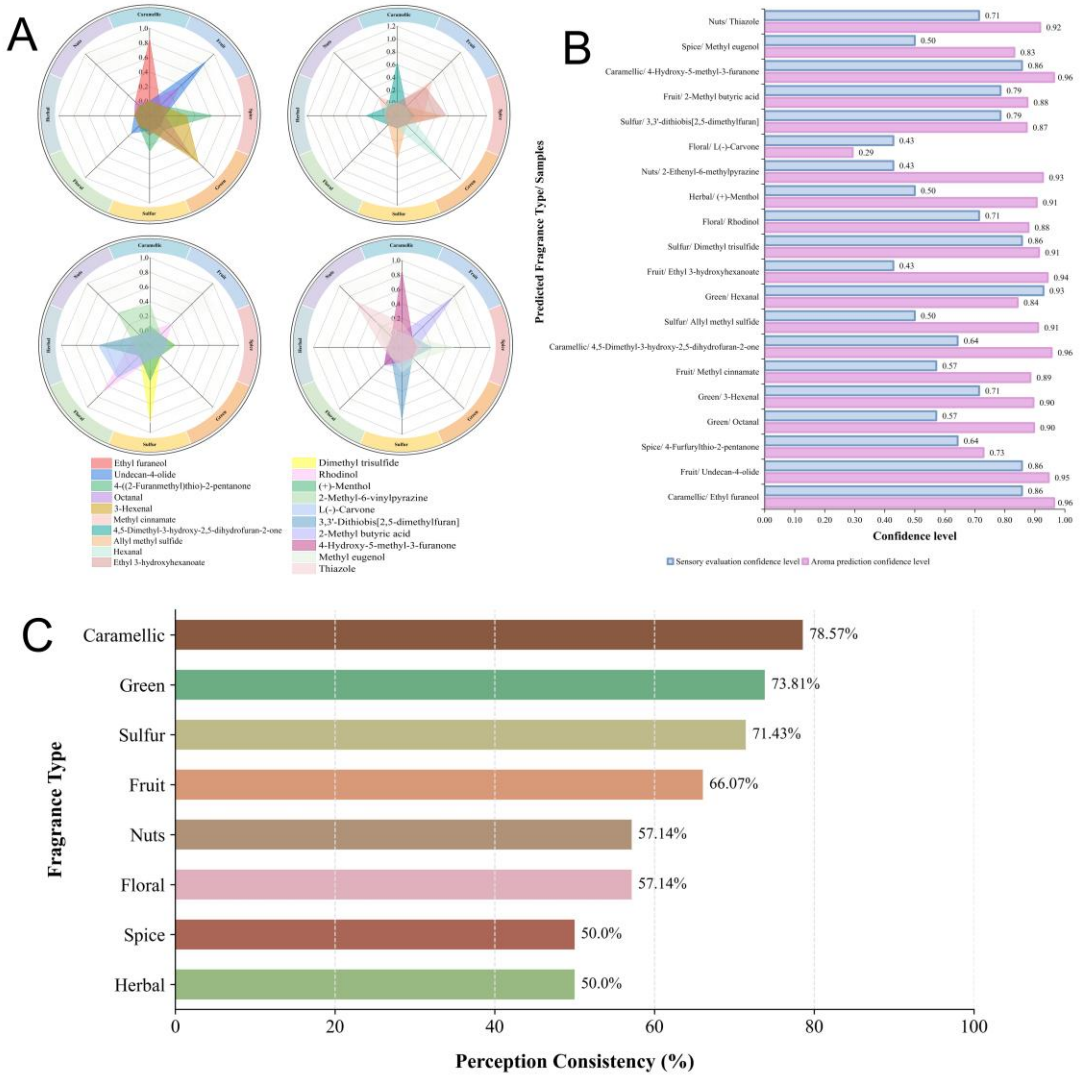


Figure 4. Frontier Molecular Orbitals (A: Fruit; B: Nuts) and Electrostatic Potential Analysis (C: Fruit; D: Nuts) of Representative Compounds

3.7. Agreement between model predictions and sensory evaluation

Sensory evaluation of 20 independent compounds assessed agreement between model predictions and human perception across eight odor-active dimensions, excluding the odorless category. Category-specific agreement ranged from 50.00% to 78.57% (Figures 5). Caramel, green, and sulfurous odors showed relatively high agreement, possibly because their distinctive sensory attributes reduced disagreement among panelists and are commonly recognized in food-aroma evaluation. Fruity odors showed intermediate agreement, potentially reflecting overlap with green and caramel-like notes. Nutty, floral, spicy, and herbal odors showed lower agreement, with the lowest values for spicy and herbal odors. Spicy notes can overlap with sulfurous attributes, whereas herbal notes can overlap with green attributes. This pattern suggests that agreement depends partly on the perceptual distinctiveness of each category and is lower when odor profiles overlap or category boundaries are less clear [3]. For food applications, these results indicate that the model may be most useful as a preliminary screening tool to prioritize candidate flavor compounds before confirmatory sensory testing. The sensory results provide an independent assessment of correspondence with human perception, but the moderate agreement and limited

sample of 20 compounds warrant broader validation. The results should therefore be interpreted as a preliminary external check rather than a definitive sensory benchmark. Future models should consider concentration-dependent perception, food-matrix effects, overlapping odor labels, and interindividual variation.



Figures 5. Analysis of the Consistency Between Model Predictions and Actual Sensory Perception of 20 Independent Compounds. A Prediction confidence analysis, B Comparison of prediction confidence and sensory confidence analysis, C Percentage of consistency between prediction model and perception

4. Conclusion

The present framework demonstrates that multiclass odor prediction for food-relevant flavor compounds can benefit from combining chemically diverse molecular descriptors, ensemble learning, and post hoc interpretability analysis. Using 1,975 initial molecular features and cross-validated feature selection, the model

retained 300 informative descriptors and achieved an out-of-fold accuracy of 84.58% with a macro-averaged F1 score of 0.7718. The ablation results indicate that mutual information selection and multi-seed optimization both contributed to model stability and predictive performance, although equal-weight voting slightly outperformed the linear stacking meta-classifier. This suggests that the broader ensemble strategy was valuable, while the specific meta-learning layer may require further refinement. The interpretation analyses linked model behavior to chemically meaningful features associated with food aroma. SHAP highlighted ester count for fruity odor prediction and nitrogen-containing fragments without attached hydrogen atoms for nutty odor prediction. Molecular docking and DFT calculations provided plausible physicochemical explanations for these attributions by locating relevant interaction sites and electronic features in representative odorants. However, these computational analyses should be interpreted as supportive evidence rather than proof of receptor activation or causal odor mechanisms. Independent sensory evaluation further showed that model-human agreement varied across odor categories, with agreement rates of 50.00% – 78.57%, emphasizing both the promise of the framework and the difficulty of predicting perceptually overlapping odor classes. Overall, this study provides an interpretable workflow for molecular odor prediction that connects statistical learning, molecular simulation, and sensory validation in a manner relevant to food-flavor screening and aroma-active compound prioritization. Future work should expand the sensory sample size, incorporate concentration-dependent and food-matrix effects, account for multilabel odor perception, evaluate nonlinear meta-classifiers, and integrate receptor-informed modeling to improve both predictive accuracy and practical interpretation for food applications.

Data and Software Availability

The data underlying this study and the code used to generate and analyze the results are openly available at Zenodo: https://zenodo.org/records/22932520. The archive (version 1.0) contains the full data pipeline from raw SMILES to the final feature matrix, all Python scripts, configuration files, and the complete set of results

(tables as CSV, 12 figures, and the Level-0 cache). The package reproduces the exact experimental results reported in this paper (Level-1 OOF accuracy 0.8458, Macro-F1 0.7718). The software is released under the MIT License; the data is released under CC BY 4.0. All files are in machine-readable format.

Author Contributions

Mingxia Zhao: Writing -original draft, Writing -editing, Conceptualization, Data Processing, Formal analysis. Shiyu Li and Yile Wang: Data Processing and Model Construction. Li Liang, Lili Zhang, Guihong Qi and Yinan Yang revised this manuscript. Shihao Sun and Yuyu Zhang: Funding acquisition and revised this manuscript. Wen Duan: Writing – review & editing, Conceptualization.

Funding

This study has been financially supported by National Natural Science Foundation of China [No. 22578008], Beijing Life Science Academy (BLSA) Foundation [No. 2023600CA0080, No. 2024601QPID08].

References

1. de March, C. A., Ma, N., Billesbølle, C. B., Tewari, J., Llinas Del Torrent, C., van der Velden, W. J. C., Ojiro, I., Takayama, I., Faust, B., Li, L., Vaidehi, N., Manglik, A. & Matsunami, H. Engineered odorant receptors illuminate the basis of odour discrimination. Nature, 635(8038), 499–508 https://doi.org/10.1038/s41586-024-08126-0 (2024).
2. Lee, B. K., Mayhew, E. J., Sanchez-Lengeling, B., Wei, J. N., Qian, W. W., Little, K. A., Andres, M., Nguyen, B. B., Moloy, T., Yasonik, J., Parker, J. K., Gerkin, R. C., Mainland, J. D. & Wiltschko, A. B. A principal odor map unifies diverse tasks in olfactory perception. Science, 381(6661), 999–1006 https://doi.org/10.1126/science.ade4401 (2023).
3. Zhong, R., Ji, Z., Wang, S. & Chen, H. Bridging odorants and olfactory perception through machine learning: A review. Trends in Food Science & Technology, 153, 104700 https://doi.org/10.1016/j.tifs.2024.104700 (2024).
4. Suh, J., Hong, Y. & Park, C. A comparative study of machine learning models on molecular fingerprints for odor decoding. Communications Chemistry, 8, 278 https://doi.org/10.1038/s42004-025-01651-7 (2025).
5. Wen, T., Cai, X. & Li, J. Graph neural networks vs. traditional QSAR: A comprehensive comparison for multi-label molecular odor prediction. Molecules, 30(23), 4605 https://doi.org/10.3390/molecules30234605 (2025).
6. Debnath, T., Badreddine, S., Kumari, P. & Spranger, M. Comparing molecular representations, e-nose signals, and other featurization, for learning to smell aroma molecules. PloS One, 18(8), e0289881 https://doi.org/10.1371/journal.pone.0289881 (2023).
7. Saini, K. & Ramanathan, V. Predicting odor from molecular structure: A multi-label

classification approach. Scientific Reports, 12(1), 13863 https://doi.org/10.1038/s41598-022-18086-y (2022).
8. Jiang, Y., Xie, X., Yang, Y., Liu, Y., Gong, K. & Li, T. Dual-branch graph neural network for predicting molecular odors and discovering the relationship between functional groups and odors. Journal of Computational Chemistry, 46, e70069 https://doi.org/10.1002/jcc.70069 (2025).
9. Iwata, H. Interpretable multitask deep learning models for odor perception based on molecular structure. Current Research in Food Science, 11, 101219 https://doi.org/10.1016/j.crfs.2025.101219 (2025).
10. Medina-Franco, J. L., Martínez-Mayorga, K., Peppard, T. L. & Rio, A. D. Chemoinformatic analysis of GRAS (generally recognized as safe) flavor chemicals and natural products. PloS One, 7(11), e50798 https://doi.org/10.1371/journal.pone.0050798 (2012).
11. Vergara, J. R. & Estévez, P. A. A review of feature selection methods based on mutual information. Neural Computing and Applications, 24(1), 175–186 https://doi.org/10.1007/s00521-013-1368-0 (2014).
12. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining https://doi.org/10.1145/2939672.2939785 (2016).
13. Liu, J., Yan, X., Li, W., Xue, S.-H., Wang, Z. & Su, R. Genomic selection for cashmere traits in Inner Mongolian cashmere goats using random forest, gradient boosting decision tree, extreme gradient boosting and light gradient boosting machine methods. Animals: An Open Access Journal from MDPI, 15(20), 2940 https://doi.org/10.3390/ani15202940 (2025).
14. Wolpert, D. H. Stacked generalization. Neural Networks, 5(2), 241–259 https://doi.org/10.1016/S0893-6080(05)80023-1 (1992).
15. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30 https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html (2017).
16. Fichtelmann, P. & Westermayr, J. Machine learning for smell: Ordinal odor strength prediction of molecular perfumery components. RSC Advances, 16(28), 25696–25704 https://doi.org/10.1039/d6ra01805j (2026).
17. Choi, C., Bae, J., Kim, S., Lee, S., Kang, H., Kim, J., Bang, I., Kim, K., Huh, W.-K., Seok, C., Park, H., Im, W. & Choi, H.-J. Understanding the molecular mechanisms of odorant binding and activation of the human OR52 family. Nature Communications, 14, 8105 https://doi.org/10.1038/s41467-023-43983-9 (2023).
18. Eberhardt, J., Santos-Martins, D., Tillack, A. F. & Forli, S. AutoDock Vina 1. 2.0: New docking methods, expanded force field, and Python bindings. Journal of Chemical Information and Modeling, 61(8), 3891–3898 https://doi.org/10.1021/acs.jcim.1c00203 (2021).
19. Tirado-Rives, J. & Jorgensen, W. L. Performance of B3LYP density functional methods for a large set of organic molecules. Journal of Chemical Theory and Computation, 4(2), 297–306 https://doi.org/10.1021/ct700248k (2008).
20. Chen, L., Medrano Sandonas, L., Traber, P., Dianat, A., Tverdokhleb, N., Hurevich, M.,

Yitzchaik, S., Gutierrez, R., Croy, A. & Cuniberti, G. MORE-Q, a dataset for molecular olfactorial receptor engineering by quantum mechanics. Scientific Data, 12, 324 https://doi.org/10.1038/s41597-025-04616-6 (2025).
21. Turek, P. Recruiting, training and managing a sensory panel in odor nuisance testing. PloS One, 16(10), e0258057 https://doi.org/10.1371/journal.pone.0258057 (2021).
22. Zhang, L.-L., Zhao, L., Wang, H.-Y., Zhi, R.-C., Shi, B.-L. & Xie, N. Determination of recognition threshold and just noticeable difference in the sensory perception of pungency of Zanthoxylum bangeanum. International Journal of Food Properties, 19(5), 1044–1052 https://doi.org/10.1080/10942912.2015.1054037 (2016).
23. Lusito, S., Pugnana, A. & Guidotti, R. Solving imbalanced learning with outlier detection and features reduction. Machine Learning, 113(8), 5273–5330 https://doi.org/10.1007/s10994-023-06448-0 (2024).
24. Ameta, D., Kumar, S., Mishra, R., Behera, L., Chakraborty, A. & Sandhan, T. Odor classification: Exploring feature performance and imbalanced data learning techniques. PloS One, 20(5), e0322514 https://doi.org/10.1371/journal.pone.0322514 (2025).
25. Džeroski, S. & Ženko, B. Is combining classifiers with stacking better than selecting the best one? Machine Learning, 54(3), 255–273. (2004).
26. Li, T., Zhou, J.-L., Xu, J.-C. & Yang, Q. A novel stacked ensemble algorithm based on feature subspace extraction and label correction for multi-label classification. International Journal of Machine Learning and Cybernetics, 17(2), 67 https://doi.org/10.1007/s13042-025-02828-1 (2026).
27. Lodh, E., Chowdhury, T., Shaw, A., Bedajna, D., Bera, S., Majumder, S. & De, M. A stacked ensemble meta-classifier framework for PAM50 breast cancer subtype prediction from gene expression signatures. Network Modeling Analysis in Health Informatics and Bioinformatics, 15(1), 78 https://doi.org/10.1007/s13721-026-00755-x (2026).
28. Sampson, A., James, A. & Tripathi, V. Optimizing breast cancer prediction through stacking ensemble machine learning models: A comparative analysis. Journal of Electrical Systems and Information Technology, 13(1), 1–24 https://doi.org/10.1186/s43067-025-00313-x (2026).
29. Rodríguez-Pérez, R. & Bajorath, J. Interpretation of compound activity predictions from complex machine learning models using local approximations and Shapley values. Journal of Medicinal Chemistry, 63(16), 8761–8777 https://doi.org/10.1021/ACS.JMEDCHEM.9B01101 (2020).
30. Wang, Z., Yang, S., Wang, X., Qiu, J., Cao, J. & Hao, X. Prediction and visual analysis of flue-cured tobacco aroma types based on machine learning and feature derivation. RSC Advances, 16(30), 27308–27318 https://doi.org/10.1039/d6ra00543h (2026).
31. Roth, J. P. & Bajorath, J. Machine learning models with distinct Shapley value explanations decouple feature attribution and interpretation for chemical compound predictions. Cell Reports Physical Science, 5(8), 102110 https://doi.org/10.1016/j.xcrp.2024.102110 (2024).
32. Bhatia, M. An overview of conceptual-DFT based insights into global chemical reactivity of volatile sulfur compounds (VSCs). Computational Toxicology, 29, 100295 https://doi.org/10.1016/j.comtox.2023.100295 (2024).