

When Should a Molecular Taste Prediction Be Trusted? Evaluating the Reliability of Fingerprint-Based Taste Classification

Elliot Chang
elliotchang1004@gmail.com

ABSTRACT

Accurate molecular taste prediction is important for the development of drugs and the support of research in food science. Despite increasing use of machine learning to predict the taste of molecules, little attention has been paid to determining when those predictions should be trusted. We built and compared a machine learning model on a dataset of molecular fingerprints from FlavorDB against a baseline model, and then evaluated its ability to identify the taste of a molecule from molecular properties using performance metrics. The fingerprint-based model achieved the strongest performance with a macro-F1 score of 0.83, but the reliability varied between taste classes, with weaker performance in underrepresented tastes and in molecules that were less similar to those in the training data.

These findings indicate that machine learning can predict taste from molecular structure, but that performance metrics alone are insufficient for judging prediction reliability because reliability varies across taste classes, dataset balance, and a molecule's similarity to the training data.

INTRODUCTION

The ability to accurately predict taste is crucial in pharmaceutical development because of the influence taste plays on the willingness for patients to consume different medicines. Being able to predict taste based on molecular structure allows researchers to develop new medicines more efficiently and reduce the need for extensive trial and error formulating. In food science, accurate taste predictability can aid in the creation of food alternatives such as sweeteners or meat. Linking the relationship between molecular characteristics and perceived taste allows for the replication of desirable flavors while decreasing negative health impacts associated with red meat or excess sugar consumption.

Machine learning models are becoming increasingly more capable at predicting taste from molecular properties. Current models such as random forest [1] are applicable for predicting taste when paired with a large database of molecular fingerprints, however; current research [12, 14] primarily states an accuracy of the model and does not emphasize the nuances of when a taste prediction should or should not be considered reliable. One accuracy number is not enough to dictate the trustworthiness of a prediction due

to variability in data bias and leakage. In this study, a trustworthy prediction is not just a high accuracy value but a prediction that remains reliable across taste classes and experimental molecules.

This paper investigates when a fingerprint-based taste prediction can be trusted, and when a reader should be cautious.

PRIOR WORK

FlavorDB2 [3] is a curated database where researchers developed an in-depth repository containing 25,595 flavor molecules which are used as a standard platform for research within the food science and fragrance industries. The database categorizes molecules into taste classes such as “sweet” or “bitter”; however, it does not act as a taste predictor for new or experimental molecules; it simply is a catalog of current molecules. One nuance of the FlavorDB2 catalog is its skewed set of molecules falling mostly under the sweet class while holding little molecules under umami. Because FlavorDB2 is a large portion of the primary dataset, the class imbalances are inherited by the predictive model which may bias the model’s accuracy toward common molecules when compared to more rare taste classes.

Building on datasets such as FlavorDB [2], Bittersweet: Building machine learning models for predicting the bitter and sweet taste of small molecules [12] established the largest dataset of taste verified compounds and developed a machine learning framework called BitterSweet which utilized both 2D and 3D properties to train random forest, ridge logistic regression, and ADABOOST models. Although Bittersweet demonstrated that machine learning could be used to predict molecule taste, its binary method of classifying molecules limited its applicability by making more complex molecules harder to classify if they fall under multiple categories.

Further developments such as A Chemical Language Model for Molecular Taste Prediction [14] presented the Flavor Analysis and Recognition Transformer (FART) and attempted to remedy the issues found in Bittersweet by creating a large language model trained on 15,025 compounds to predict the taste of molecules based on chemical structures derived from SMILES strings. It reached an accuracy of 91 percent and out-performed all previous models in its area; however, the accuracy of the model emphasizes performance metrics rather than examining how reliability varied across different identifiers.

METHODS

This project evaluated the reliability and shortcomings of a machine learning model designed to predict the taste labels (sweet, bitter, sour, umami) of molecules, using a dataset derived from FlavorDB. The goal was to identify which characteristics played the largest role in determining a taste label, and to test whether a random forest classifier could improve on a baseline model that always predicts sweet, since sweet molecules make up the majority of the dataset.

The dataset was composed of molecules represented by SMILES strings [13]. Each row represented a single molecule and included the SMILES string, the taste label, and the source column. SMILES strings and taste labels were the primary columns used.

Before modeling, an exploratory data analysis was run to find missing values and duplicate molecules. No missing values were found. There were 402 duplicate molecules, which were flagged and removed to prevent data leakage between the training and test sets, since leakage inflates accuracy beyond what the model would reach in practice [6]. After cleaning the initial data set of 25,595 molecules by filtering out duplicate molecules, the data set contained 14,563 unique molecules.

Table 1. Taste-class distribution of the cleaned dataset before modeling

Taste label	Number of molecules	Percentage
Sweet	9,462	64.97%
Bitter	1,526	10.48%
Sour	1,535	10.54%
Umami	56	0.38%
Undefined / other	1,984	13.62%
Total	14,563	100.00%

The model cannot process SMILES strings in their raw form, because they are text. Each molecule was converted into a Morgan fingerprint [8, 10], a numerical representation the model can read. The RDKit library was used for this conversion.

Using a random train/test split of 0.30 with a fixed random seed of 42, the cleaned dataset of 14,563 unique molecules was divided into 10,194 training molecules and 4,369 test molecules. The training set was used to fit the model, and the test set was held out for evaluation.

Because the train/test split was random, chemically similar molecules may have been placed in both sets. This can make performance appear stronger than it would be on unique molecules. To address this, a Tanimoto similarity analysis was performed on each test molecule and its nearest training molecule. For future work, scaffold-based or cluster-based splitting could be used to reduce chemical similarity in the training and test sets which would provide a more accurate assessment of the model.

To examine the effect of duplicate leakage, a controlled test was done using the same test set with and without duplicate molecules. The model was analyzed on the clean set of data then on the same set with duplicate molecules added back in. This kept changes to a minimum while evaluating the effect of duplicate leakage on the model.

A baseline model that always predicts sweet was used as a point of comparison. The classifier was a random forest [1] trained on the fingerprint features which builds many decision trees based on a series of yes/no questions and then predicts a taste class based on the majority vote between the decision trees.

The model was evaluated using accuracy, precision, recall, and macro-F1 [9, 11].

$$Precision_i = \frac{TP_i}{(TP_i + FP_i)} \quad Recall_i = \frac{TP_i}{(TP_i + FN_i)} \quad (1)$$

$$F1_i = 2 * \frac{(Precision_i * Recall_i)}{(Precision_i + Recall_i)} \quad (2)$$

where TP_i , FP_i , and FN_i are the true positives, false positives, and false negatives for class i . Macro-F1 is the unweighted mean of the per-class F1 scores across the $C = 4$ taste classes,

$$Macro - F1 = \left(\frac{1}{C}\right) * \sum_{i=1}^C F1_i \quad (3)$$

Macro-F1 weights each taste class equally, which matters because the classes are imbalanced [4]. Reliability was checked by examining per-class recall, source patterns, multi-label molecules, a SHAP analysis [7] on named properties, and an applicability-domain check using Tanimoto similarity to the training set. Each test molecule was assigned the Tanimoto similarity of its nearest training molecule, allowing model performance to be evaluated as chemical similarity decreased.

$$T(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{c}{a + b - c} \quad (4)$$

where A and B are two molecular fingerprints, a and b are the numbers of on-bits in each, and c is the number of bits they share.

All analyses were performed in Python 3.11.13 using pandas 2.2.3, NumPy 2.0.2 [5], scikit-learn 1.6.1, RDKit 2025.03.1, SHAP 0.48.0, and Matplotlib 3.10.0 in Google Colab.

RESULTS

The model identified sweet molecules most reliably, with a recall of 0.98, and umami least reliably, with a recall of 0.45.(Figure 1).

The model reached an accuracy of 0.95 against a baseline of 0.75, and a macro-F1 of 0.83 against baseline of 0.21 (Figure 2).

The controlled duplicate test decreased macro-F1 by about 0.02 when adding duplicate molecules back into the test set on the main split and by about an average of 0.016 across multiple random seeds.

Molecular weight ranked as the top named property related to taste (Figure 3).

Macro-F1 fell from 0.96 on the molecules most similar to the training set to 0.71 on the least similar (Figure 4).

Molecular weight, TPSA, NumHAcceptors, and NumHDonors were strongly correlated, with correlation values above 0.8 (Figure 5).

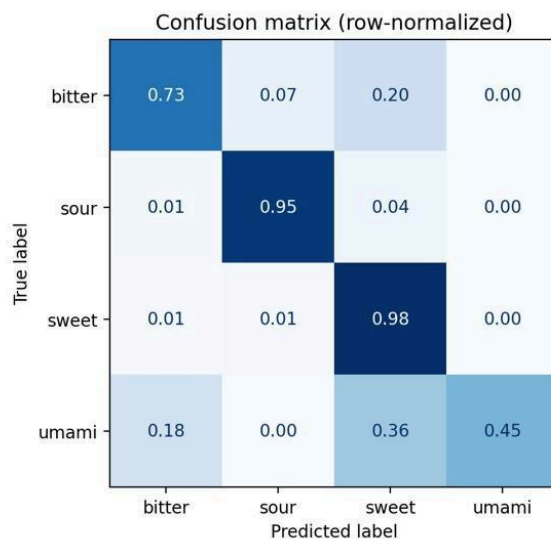


Figure 1: Row-normalized confusion matrix for the random forest taste classifier across the four taste classes.

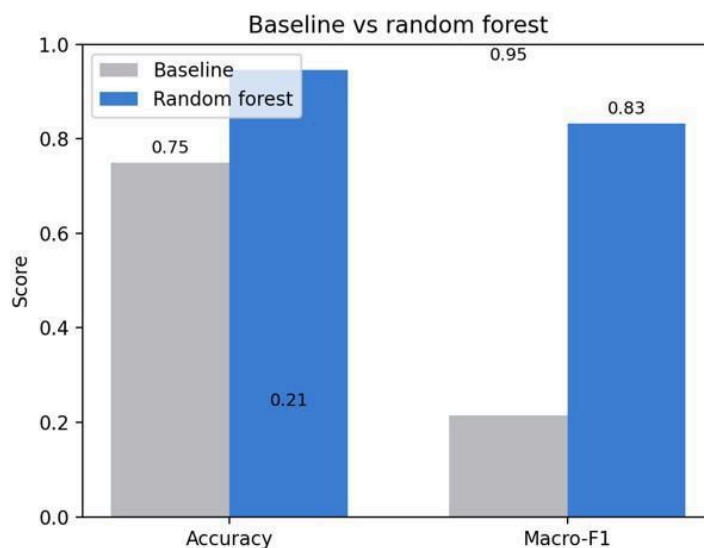


Figure 2: Accuracy and macro-F1 of the random forest compared with the majority-class baseline.

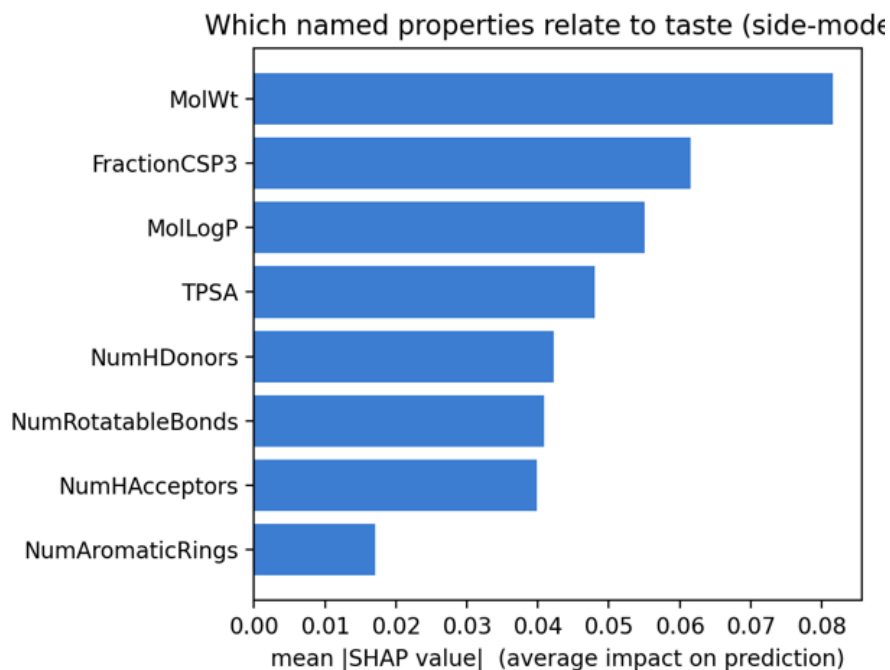


Figure 3: Mean absolute SHAP values for the named molecular properties in the side-model, ranking their average impact on taste prediction.

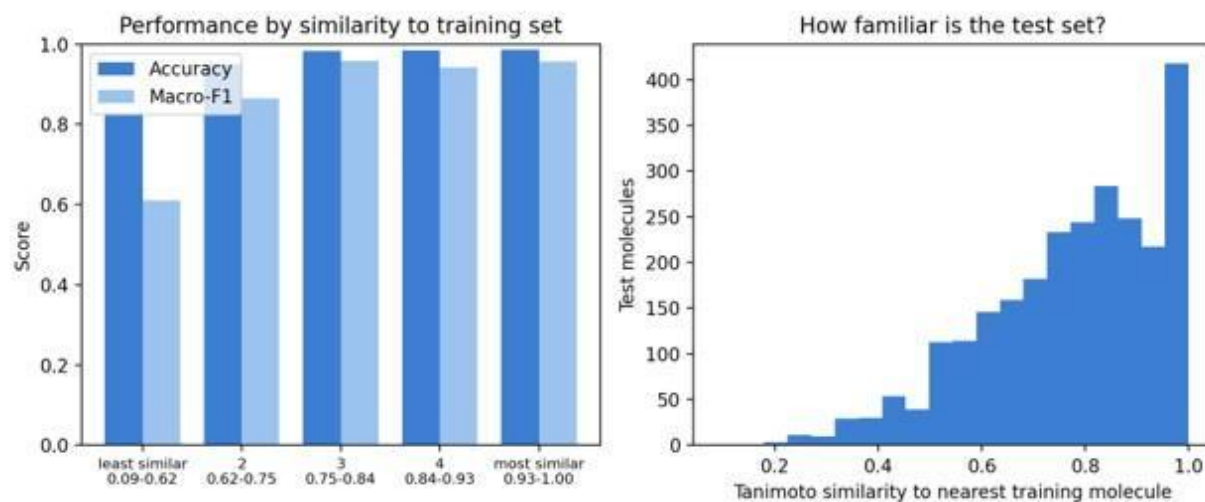


Figure 4: Applicability-domain check. Left: accuracy and macro-F1 grouped by Tanimoto similarity to the nearest training molecule. Right: distribution of test-molecule similarity to the training set.

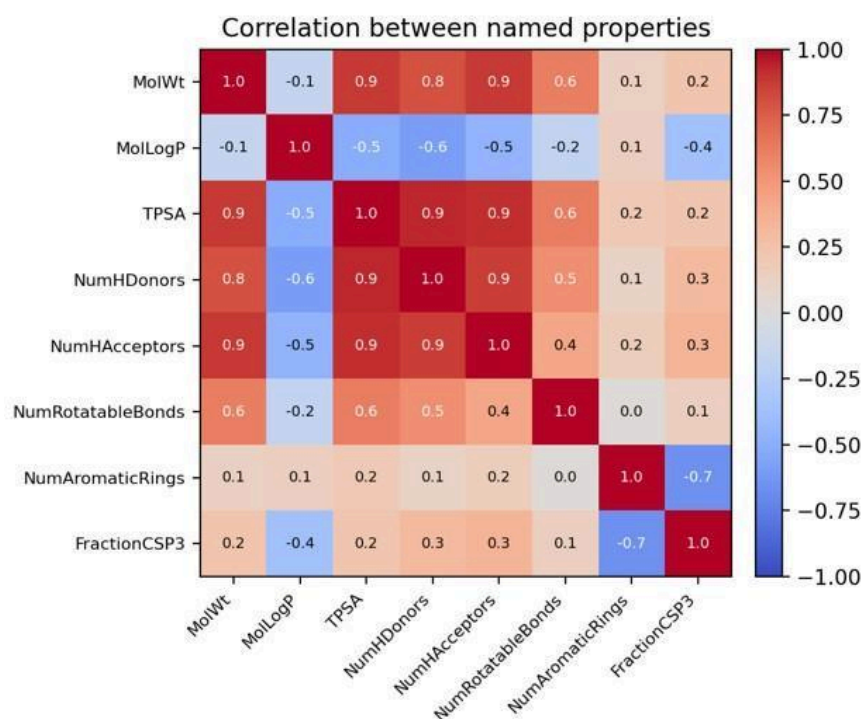


Figure 5: Correlation matrix between the named molecular properties.

DISCUSSION

The model predicted taste from molecular structure with a much higher macro-F1 than the baseline, 0.83 against 0.21 (Figure 2). That gap is clear evidence that molecular structure carries real information about taste, and that the prediction can be trusted in some cases. The SHAP analysis suggests molecular weight plays a role in taste (Figure 3), though it ran on the side-model rather than the fingerprint model, so it points to a likely influence rather than a confirmed one.

The headline number hides an uneven story. Sweet and sour molecules were predicted reliably, while umami was predicted with much less reliability and was mostly absorbed into sweet (Figure 1). Trust should follow the taste class, not the overall number. The model beats the baseline, but with real caveats for the rare classes.

Performance dropped on molecules unlike the training set (Figure 4), so a more novel molecule calls for more caution. Macro-F1 fell across the similarity groups while accuracy stayed roughly flat, which is itself a reason to weight macro-F1 over accuracy, since accuracy does not account for the class imbalance.

Reliable taste prediction matters beyond the metrics. In pharmaceuticals, the taste of a medicine affects whether patients will take it, so a reliable prediction helps researchers formulate medicines that are not unpleasant. In food science, knowing which molecular features drive taste is useful for designing new compounds, such as sugar substitutes.

LIMITATIONS

The dataset is mostly sweet molecules, which may bias the model toward predicting sweet. Even when accuracy is high, the model may not be reliable, so macro-F1 should carry more weight. With more time, the larger dataset could be broken down by source.

The data came from FlavorDB, which may introduce source bias from similarities among molecules. This could weaken the model on molecules that do not share those similarities, which the applicability-domain result hints at (Figure 4).

The SHAP analysis ran on a side-model of named properties, not on the fingerprint model that reached the best performance. So it may not reflect the fingerprint model's actual decisions, which adds caution to any feature-importance claim.

Several named properties were strongly correlated (Figure 5), so the model may lean on fewer independent signals than it appears to. Rare tastes such as umami have few examples, so the model has little to learn from, which is why umami predictions suffer and get absorbed into sweet (Figure 1).

After running the controlled duplicate test, it was observed that adding leakage back into a clean test set only decreased macro-F1 by about 0.02 on the main split and by an average of about 0.016 across multiple random seeds. The cause of the initial changing performance metric may have been due largely

to other factors such as the class imbalance changing when removing leakage. By fixing the test set, the prediction score becomes more trustworthy because the model is evaluated on the same molecules only with and without duplicates.

CONCLUSION

This study demonstrates that molecular structure can be used to predict the taste of a molecule with a greater accuracy than a baseline model, but a high-performance metric alone should not be treated as proof of reliability. The true reliability of a taste prediction relies on taste class, the similarity between training and test sets as well as the balance of the dataset used to train the model. Models perform well within common taste labels like sweet but are shown to be less reliable on rare taste labels such as umami or molecules that fall outside of the applicability domain. These findings highlight the importance of digging beyond an accuracy measure and why properly evaluating the limitations of a machine learning prediction is meaningful to society. By highlighting the ideal conditions where taste predictions are the most trustworthy, this paper provides the framework for using machine learning in medicine development and food science.

REFERENCES

1. Breiman, Leo. "Random Forests." *Machine Learning* 45, no. 1 (2001): 5–32. <https://doi.org/10.1023/A:1010933404324>.
2. Garg, Neelansh, Apuroop Sethupathy, Rudraksh Tuwani, et al. "FlavorDB: A Database of Flavor Molecules." *Nucleic Acids Research* 46, no. D1 (2018): D1210–16. <https://doi.org/10.1093/nar/gkx957>.
3. Goel, Mansi, Nishant Grover, Devansh Batra, et al. "FlavorDB2: An Updated Database of Flavor Molecules." *Journal of Food Science* 89, no. 11 (2024): 7076–82. <https://doi.org/10.1111/1750-3841.17298>.
4. He, Haibo, and E. A. Garcia. "Learning from Imbalanced Data." *IEEE Transactions on Knowledge and Data Engineering* 21, no. 9 (2009): 1263–84. <https://doi.org/10.1109/TKDE.2008.239>.
5. Harris, Charles R., K. Jarrod Millman, Stefan J. van der Walt, et al. "Array Programming with NumPy." *Nature* 585, no. 7825 (2020): 357–62. <https://doi.org/10.1038/s41586-020-2649-2>.
6. Kaufman, Shachar, Saharon Rosset, Claudia Perlich, and Ori Stitelman. "Leakage in Data Mining: Formulation, Detection, and Avoidance." *ACM Transactions on Knowledge Discovery from Data* 6, no. 4 (2012): 1–21. <https://doi.org/10.1145/2382577.2382579>.
7. Lundberg, Scott, and Su-In Lee. "A Unified Approach to Interpreting Model Predictions." *arXiv*, 2017. <https://doi.org/10.48550/arXiv.1705.07874>.
8. Morgan, H. L. "The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service." *Journal of Chemical Documentation* 5, no. 2 (1965): 107–13. <https://doi.org/10.1021/c160017a018>.

9. Powers, David M. W. "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation." *arXiv*, 2020. <https://doi.org/10.48550/arXiv.2010.16061>.
10. Rogers, David, and Mathew Hahn. "Extended-Connectivity Fingerprints." *Journal of Chemical Information and Modeling* 50, no. 5 (2010): 742–54. <https://doi.org/10.1021/ci100050t>.
11. Sokolova, Marina, and Guy Lapalme. "A Systematic Analysis of Performance Measures for Classification Tasks." *Information Processing & Management* 45, no. 4 (2009): 427–37. <https://doi.org/10.1016/j.ipm.2009.03.002>.
12. Tuwani, Rudraksh, Somin Wadhwa, and Ganesh Bagler. "BitterSweet: Building Machine Learning Models for Predicting the Bitter and Sweet Taste of Small Molecules." *Scientific Reports* 9, no. 1 (2019): 7155. <https://doi.org/10.1038/s41598-019-43664-y>.
13. Weininger, David. "SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules." *Journal of Chemical Information and Computer Sciences* 28, no. 1 (1988): 31–36. <https://doi.org/10.1021/ci00057a005>.
14. Zimmermann, Yoel, Leif Sieben, Henrik Seng, Philipp Pestlin, and Franz Görlich. "A Chemical Language Model for Molecular Taste Prediction." *npj Science of Food* 9, no. 1 (2025): 122. <https://doi.org/10.1038/s41538-025-00474-z>.