Classification of Tastants: A Deep Learning Based Approach

Prantar Dutta, Deepak Jain, Rakesh Gupta* and Beena Rai

Physical Sciences Research Area, Tata Research Development and Design Centre, TCS Research, 54-B, Hadapsar Industrial Estate, Pune- 411013, India

* *Correspondence to*: Rakesh Gupta

Email: gupta.rakesh2@tcs.com, Phone: + 91-20-66086422

Abstract

Predicting the taste of molecules is of critical importance in the food and beverages, flavor, and pharmaceutical industries for the design and screening of new tastants. In this work, we have built deep learning models to classify sweet, bitter, and umami molecules— the three basic tastes whose sensation is mediated by G protein-coupled receptors. An extensive dataset containing 1938 bitter, 2079 sweet, and 98 umami tastants was curated from existing literature. We analyzed the chemical characteristics of the molecules, with special focus on the presence of different functional groups. A deep neural network model based on molecular descriptors and a graph neural network model were trained for taste prediction. The class imbalance due to fewer umami molecules was tackled using special sampling techniques, such that the classwise metrics for all the three taste classes are optimized. Both models show comparable performance during evaluation, but the graph-based model can learn task-specific representations from the molecular structure without requiring handcrafted features. We further explain the deep neural network predictions using Shapley additive explanations and connect them to the physics of tastant-receptor binding. This study develops an in-silico

approach to classify molecules based on their taste by leveraging the recent progress in deep learning, which can serve as a powerful tool for tastant design.

Keywords: Tastant, Sweet, Bitter, Umami, Multiclass Classification, Deep Learning, Graph Neural Network, SHAP

1 Introduction

Taste is a sensory modality that governs the interaction of humans with the food they eat. The gustatory system, which is responsible for the sense of taste, differentiates healthy nutrients from harmful toxins, ensuring survival and a high quality of life. The interplay of five basic tastes – sweet, bitter, sour, salty, and umami, constitutes our taste experience [1]. The brain associates each taste with an underlying chemical characteristic. Sweetness and umaminess indicate the presence of carbohydrates and proteins, respectively, while sourness is linked with acids, often due to food spoilage. The unpleasant bitter sensation informs us to avoid the ingestion of rotten food and toxic substances, although safe edibles like cocoa and coffee can be bitter too. Saltiness relates to the minerals like sodium that are essential for regulating body fluids. In combination with olfaction (sense of smell) and somatosensation (sense of touch, temperature, and pain), gustation determines our overall perception of flavors [2].

Taste prediction and the design of tastants play a crucial role in food and beverages, flavor, and pharmaceutical industries. For example, the search for safe artificial low-calorie sweeteners with similar chemosensory profile as sucrose is still an open research problem. The demands for the specialized flavors in the consumer product landscape is evolving rapidly to keep up with

the latest trends, and medicinal chemists are constantly looking for taste modulators to combine with oral drug formulations. Although taste perception shows variability among individuals and demographics due to genetic [3], cultural [4], and medical [5] factors, specific biological mechanisms, which are common to all humans, exist for perceiving each of the basic taste qualities. Thus, rational approaches can be conceived to design and screen tastants based on their chemistry and interactions with the gustatory system.

Thousands of small protuberances called papillae cover the tongue, each of which contains hundreds of taste buds [6]. Each of the taste bud comprises of 50-100 taste cells having specialized sensing receptors. Tastants stimulate the receptor cells, leading to signal transmission to the gustatory cortex in the brain. Sweet, bitter, and umami receptors belong to the family of G protein-coupled receptors (GPCRs) [7]. GPCRs are present on cell surfaces and interact with molecules present outside the cell. They perform critical physiological functions including taste and smell sensing, vision, behavior regulation, immune system regulation, and neurotransmission. The molecule/ligand binding to a GPCR induces a conformational change in it, which in turn activates the associated G protein and downstream signaling pathways [8]. The most common GPCR classification scheme is the A-F system in which the receptors are grouped into six classes based on their similarity of sequence and functions [9], although other schemes exist too. Sweetness and umaminess are recognized by the heterodimeric complexes T1R1/T1R3 and T1R1/T1R2, respectively, which belong to the class C GPCRs, while 25 class A GPCRs called T2Rs mediate the bitter sensation. Sour and salty tastes are detected by transient receptor potential ion channels in the taste cells [10, 11].

For structure-based computational design and screening of sweet, bitter, and umami molecules, elucidation of tastant-GPCR interactions and the subsequent conformational dynamics of the receptor complex play a critical role. The GPCRs consist of seven transmembrane helices, connected by three intracellular and three extracellular loops, with an extracellular N-terminus and intracellular T-terminus. Class C GPCRs are made up of the extracellular Venus flytrap domain (VFD), the seven seven-helix transmembrane domain (TMD), and the connecting cystine-rich domain (CRD) between the two [12, 13]. However, experimental structures of taste receptors are still elusive despite considerable progress in structure determination of GPCRs. To overcome this problem, *in silico* techniques like homology modeling, molecular docking, and molecular dynamics have been applied as surrogates to facilitate virtual screening of tastants [14]. But the absence of accurate receptor structures decreases the reliability of the results, and the computational cost prohibits high-throughput screening of large databases. Moreover, no simple correlation can be established between the tastant-receptor binding affinity and the taste quality or intensity, further complicating the task.

Ligand-based structure-property relationship models offer an exciting alternative to rigorous molecular modeling techniques. Recent advances in computing capabilities and machine learning (ML) algorithms, and availability of curated datasets, make it possible to map molecular features to taste qualities with high accuracy. Initial studies mostly focused on building sweet/non-sweet and bitter/non-bitter classification models using standard molecular descriptors from cheminformatics tools, wherein tastant databases constituted the positive set and random molecules were selected to create the negative set [15, 16, 17, 18, 19]. Later, efforts were made to create the bitter/sweet classifiers entirely using tastant molecules,

eliminating the need for the randomly selected negative set [20, 21]. A few studies also reported predictive models to quantify the taste intensity of molecules using properties like relative sweetness with respect to sucrose [22, 23]. A comprehensive discussion on databases and ML approaches related to tastants can be found in the recent review by Malavolta, Pallante, and co-workers [24]. An integrated data and structure-based modeling framework, combining structure-property relationship, sweet/bitter classification, and molecular docking, was also proposed to screen potential sweeteners [25].

Existing studies on computational taste prediction mostly considers sweet and bitter tastants only. However, the approaches can, in principle, be extended to predict multiple taste qualities simultaneously. The primary hurdle is the lack of large, curated datasets that can be exploited to build predictive models. For example, not many umami molecules are known, although their perception mechanism is like sweet and bitter molecules (via GPCRs). The exceptional progress in molecular ML in recent years, especially deep learning (DL) and graph-based models, offers a range of tools to make better predictions using existing and limited data [26, 27]. Nonetheless, the enigmatic power of DL algorithms limits the interpretability and explainability of models, which leads to skepticism about deployment for industrial use.

In this work, we develop multiclass classifiers for differentiating between bitter, sweet, and umami molecules. We apply deep neural network (DNN), also known as multilayer perceptron, and graph neural network (GNN) models on an extensive dataset curated from multiple sources in literature. GNNs are especially attractive because molecules can be easily represented as graphs, with atoms as nodes and bonds as edges [28]. Additionally, GNNs can work without including expert handcrafted features that require domain knowledge. We further try to make

sense of the results from our DL models using state-of-the-art explainability methods. An analysis of the functional groups in the tastants is presented as well, with an aim to relate the taste quality with chemistry. Our work widens the scope and advances the applicability of in silico taste prediction using data-driven techniques by inheriting latest developments in DL, coupled with insights from chemistry.

2 Methods

2.1 Dataset

A collection of bitter, sweet, and umami molecules was curated using information from the ChemTastesDB database [29], and the datasets made available by Tuwani and co-workers [21]. Both the datasets include tastants from multiple repositories and earlier works on data-based modeling, the details of which can be found in the respective papers [18, 25]. ChemTastesDB has 2944 verified tastants, both organic and inorganic, belonging to nine classes, including the five basic tastes and four additional categories, namely tasteless, multitaste, non-sweet, and miscellaneous. We extracted only the canonical simplified molecular input line entry system (SMILES) representations and the corresponding taste labels of sweet, bitter, and umami compounds from the database. Similarly, the sweet and bitter molecules used by Tuwani et al. to build the BitterSweet models were obtained. The two sets of canonical SMILES were merged, and the duplicates were removed to create our initial raw tastant database. It contained 1966 bitter molecules, 2091 sweet molecules, and 98 umami molecules, making up a total of 4155 tastants.

2.2 Featurization and Data Preprocessing

We generated molecular descriptors using the cheminformatics tool RDKit for building DNN models [30]. The Descriptors module of RDKit returns a list of 200 features for each molecule, which represent the structural, physical, and chemical information of the molecule as numerical values. They include a wide variety of molecular properties and fragment counts. However, not all features are relevant for a particular dataset or task and including them can lead to poor model quality. We manually removed five features *(MaxAbsEStateIndex*, *MinAbsEStateIndex*, *ExactMolWt*, *MaxAbsPartialCharge*, and *MinAbsPartialCharge*) that are perfectly correlated to other features *(MaxEStateIndex*, *MinEStateIndex*, *MolWt*, *MaxPartialCharge*, and *MinPartialCharge*) in the dataset. Then, we removed those features for which less than 20 % of the molecules have non-zero values, to prevent overfitting. The remaining 102 descriptors after feature selection were used to build the DNN models. We also dropped a few molecules from the dataset for which RDKit was unable to generate descriptors. The final clean dataset contained 4115 tastants— 1938 bitter, 2079 sweet, and 98 umami.

The dataset was split into training and test sets with a train-test ratio of 85:15 for model building and evaluation. Multiple training and test sets with different random states were created to ensure the reliability of results (see Supporting Information). We applied min-max scaling and one-hot encoding to transform the features and taste labels, respectively. All preprocessing was performed with the Scikit-learn package [31].

For building the GNN model, we obtained the SMILES strings of the molecules from the database for conversion to graph objects, keeping the training and test sets same as previously

discussed. In the GNN framework, molecules are treated as undirected graphs. Each heavy atom (non-hydrogen) in a molecule is considered as a node, and we compute the following node features: one-hot encoding of the element, degree of the atom, whether the atom is aromatic or not, number of attached hydrogen atoms, and the implicit valence. The bonds are homologous to graph edges with the following edge features: one-hot encoding of the bond order or aromaticity, whether the bond is part of a ring, and whether the bond is conjugated. In addition, an adjacency matrix is generated for each molecule which contains information about the neighbors of all the atoms.

2.3 Model Development

We first built a DNN model to classify the 4115 tastants in our database into three taste classes— bitter, sweet, and umami. The architecture consisted of the input layer, two hidden layers, and the output layer. Both the hidden layers were made up of 100 neurons each and activated by a rectified linear unit (ReLU) function. To reduce overfitting, the dropout technique was employed after the first hidden layer with a probability 0.3. Three output neurons, with SoftMax activation, predicted the probability of molecules belonging to each of the three classes. We chose the class with the highest probability as the model output for taste prediction. The input data was fed to the DNN model in batches of 32. The Adam optimizer with a learning rate of 0.0001 and the categorical cross-entropy loss function were used to train the model. 15 % of the training data was kept aside for validation, and the model with lowest validation loss obtained during the training process of 200 epochs was saved as the best model. We arrived at the architecture after experimenting with multiple values of hidden layers and number of units to maximize validation accuracy and ensure minimal overfitting. Finally, the

model was evaluated on the test set by computing the overall accuracy, confusion matrix, and classwise precision, recall, and F1 scores. The Keras API of TensorFlow 2 was used to implement the DNN model [32].

As our dataset includes much fewer umami compounds than sweet or bitter, the problem of class imbalance arises. Imbalanced datasets can lead to inferior performance for the minority class, even though the overall accuracy may be quite high. We attempt to solve this problem using the synthetic minority oversampling technique (SMOTE), a popular method of data augmentation which generates synthetic datapoints based on the original data but not its duplicates [33].

We further built a GNN model based on convolutional neural networks operating directly on molecular graphs, as proposed by Duvenaud and co-workers [34]. The architecture consists of two identical convolution blocks, each of which is made of a graph convolution layer and a graph pooling layer, with batch normalization applied between the two layers. The convolution blocks update the per-atom feature vectors in a non-linear way by incorporating information from its bonds and adjacent atoms. A channel width of 64 and ReLU activation function is used for the graph convolution layer and max-pooling is used for aggregating the neighborhood information of an atom. Finally, a graph gather layer combines the node-level feature vectors into a single graph-level feature vector that represents the entire molecule. This feature vector is then passed through a dense layer of 128 neurons to the output layer that predicts the desired probabilities of the three taste classes. The dropout technique was used after each layer with probability 0.1 to reduce overfitting. Figure 1 presents a schematic of the information flow in the GNN architecture. Additional details regarding the model architecture can be found

in the original paper. Like DNN training, the input data was fed to the model in batches of 32, the categorical cross entropy loss function was employed, and 15 % of the training data was kept aside for validation. The model was trained for 50 epochs and the hyperparameters were optimized using random search to minimize the validation loss. We implemented the GNN model using the DeepChem framework [35, 36].

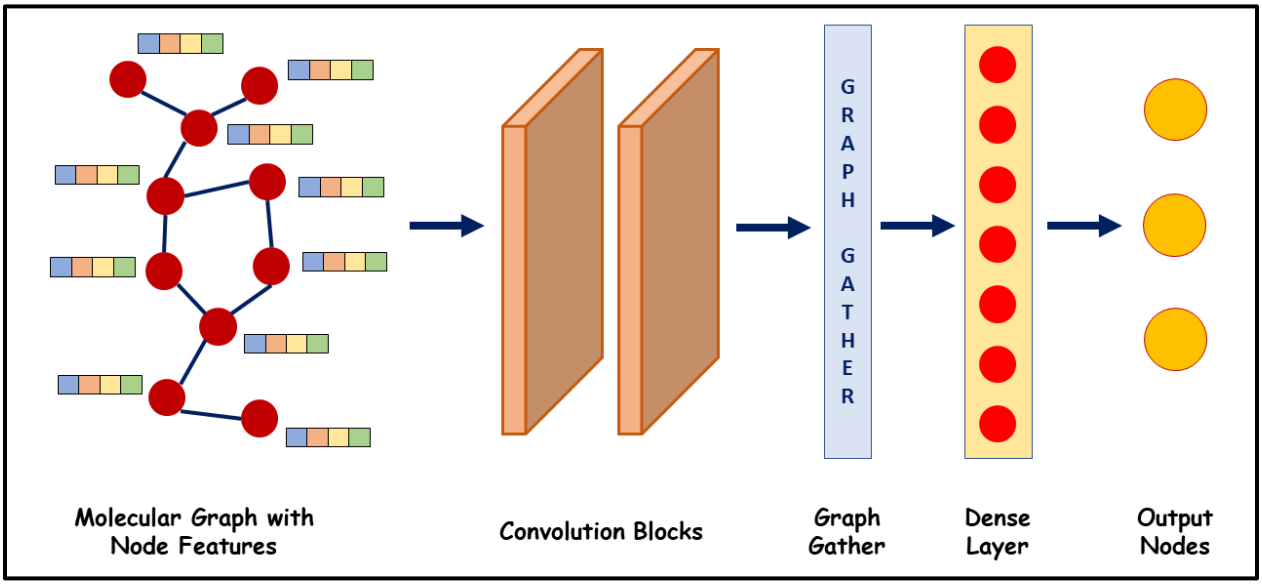

Figure 1: Flow of information in the graph neural network architecture. The input molecular graph with its node features is processed by two convolution blocks. A graph gather layer combines the per-atom representations to generate a molecule-level fingerprint vector, which is processed by a dense layer. The three output nodes predict the probability of a molecule being either bitter, sweet, or umami.

The class imbalance problem is particularly challenging for GNNs to tackle. We experimented with oversampling using SMILES enumeration. 20 variants of each SMILES string were generated for the umami molecules in the training set and augmented with the original training data. It is to be noted that although the initial inputs for the augmented date are different,

GNNs are, by nature, permutation invariant. Hence, the learnt representation after the two convolutional blocks is same for all the SMILES variant of a molecule, which effectively makes the data transformation a simple case of oversampling.

3 Results and Discussion

3.1 Exploratory Data Analysis

The final set of 4115 tastants, belonging to any one of the three taste qualities— bitter, sweet, or umami, was analyzed for characteristics, patterns, and insights. Figure 2 (a) shows the density distribution plot of molecular weights of the tastants. Molecular weight relates to the size of molecules, and hence affects ligand binding to a receptor. We observe that most tastants have molecular weights within 1000 Daltons, and thus can be considered as small molecules, although a few molecules with higher weights exhibit taste qualities as well. The molecular weights are normally distributed, with umami tastants more likely to be heavier than sweet and bitter compounds. Figure 2 (b) shows the density distribution of the octanol-water partition coefficient logP, a measure of hydrophobicity. We find a good mix of hydrophilic (logP < 0) and hydrophobic (logP > 0) molecules in the sweet and bitter categories, while most umami molecules are hydrophilic. Di Pizio et al. reported in a study with limited datapoints (677 bitter and 312 sweet) that bitter compounds have higher hydrophobicity than sweet ones, while sweet compounds have a wider size range [37]. However, our work, based on a superset of their data, reveal that such conclusions may not be drawn with high certainty. Figures 2 (c), (d), and (e) highlight the hydrogen bond-forming tendencies of the three classes of tastants.

Hydrogen bond stabilizes protein-ligand complexes and thus crucially affects the binding affinity. In agreement with our expectations, most tastant molecules have hydrogen bond donors and acceptors, which enable them to interact with the residues in the binding pockets of the receptors.

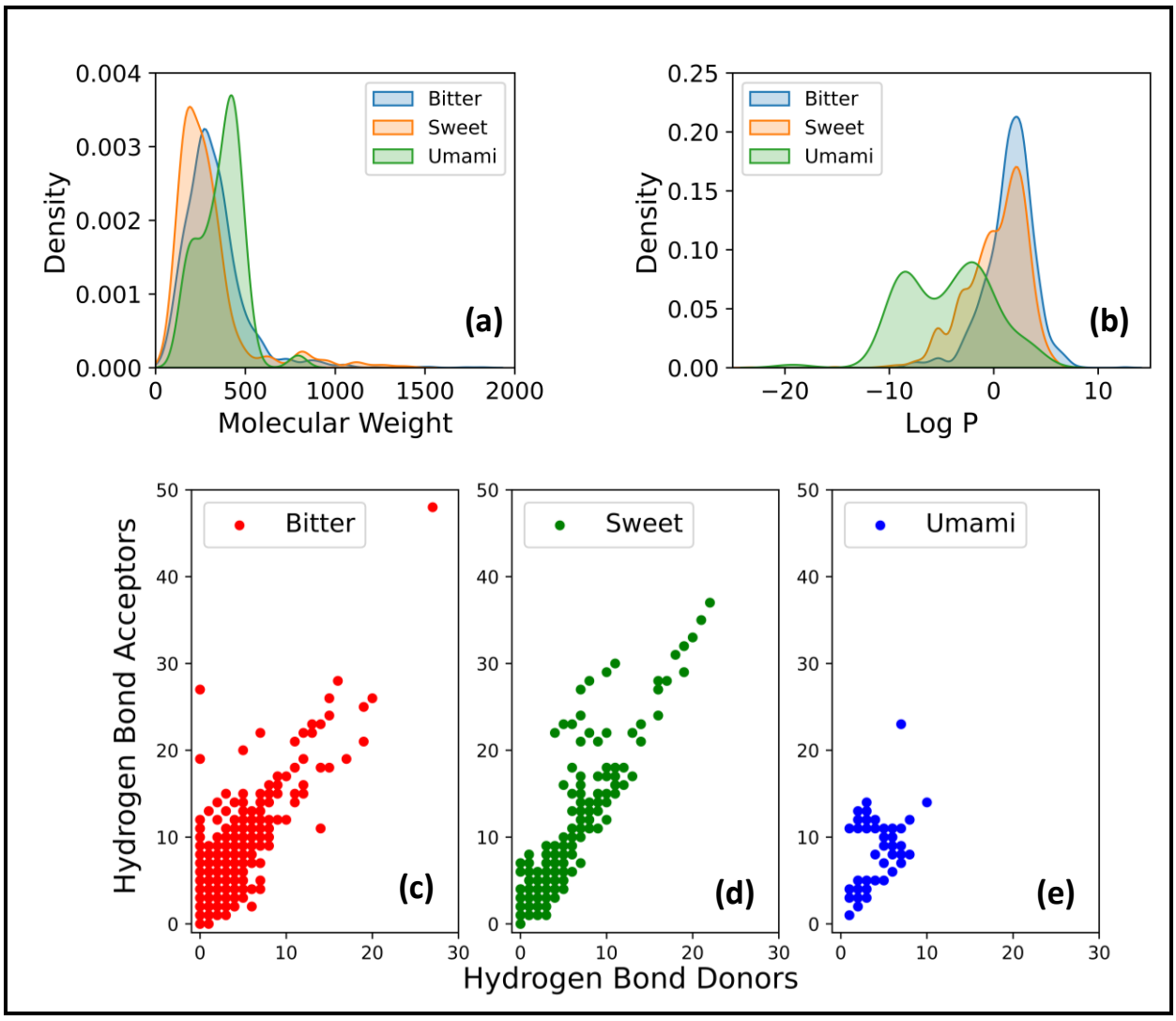

Figure 2: Dataset characteristics based on key molecular properties. Density distribution of (a) molecular weight and (b) octanol-water partition coeffcient logP of bitter, sweet, and umami compounds in the dataset. Hydrogen bond donors and acceptors in the (c) bitter, (d) sweet, and (e) umami molecules. A single point in (c), (d), and (e) can be an overlap of multiple tastants.

To further explore the dataset visually, we performed principal component analysis (PCA) to reduce the high dimensional data. PCA is an unsupervised learning technique that transforms a large set of correlated variables into a smaller set of uncorrelated variables, while maintaining the variation of the original dataset. Figure 3 (a) shows the relationship between the first and second principal components. As evident from the PCA plot, the tastants have ample structural diversity. But the overlap between the three taste qualities is significant and no trivial way to separate them is apparent, which complicates the classification task. The first principal component explains 30 % of the total variance, and the second one explains around 11 %. We also generated t-distributed stochastic neighbor embedding (t-SNE) plots for our dataset, as shown in Figure 3 (b). t-SNE is also an unsupervised dimensionality reduction technique, more powerful than PCA for visualizing complex data in two-dimensional space [38]. It minimizes the divergence between the distribution that measures pairwise similarities of input objects and the distribution that measures pairwise similarities of the corresponding embeddings. We assessed the performance of the algorithm for different values of the two key hyperparameters, perplexity and learning rate, by visually comparing the generated plots after optimizing for 2000 iterations. A perplexity of 80 and a learning rate of 800 was found to be suitable. t-SNE concurs with PCA regarding the structural diversity of the tastants in the dataset and the difficulty of the classification task. Although small clusters of similar tasting compounds can be observed, the overlap between the classes is high. No clear distinction between the taste qualities is apparent from the t-SNE visualization. It is to be noted that cluster size and distance between clusters in t-SNE plots bear no significance.

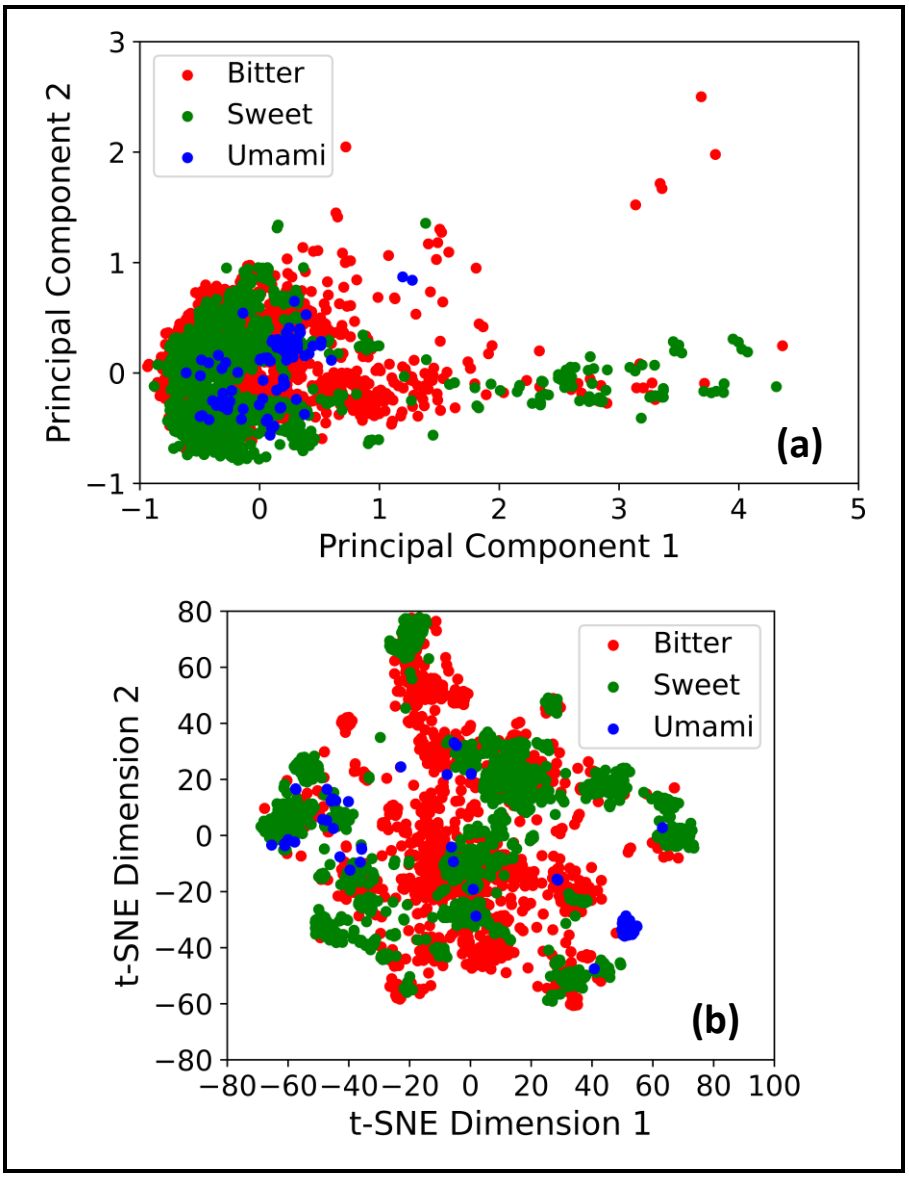

Figure 3: Unsupervised dimensionality reduction using (a) PCA and (b) t-SNE techniques. Molecules are colored according to their taste.

All results from our exploratory data analysis confirm the huge diversity of molecules, mostly within the small molecule chemical space with a few exceptions. Distributions of important molecular properties, as well as low-dimensional representations of the dataset, point to significant overlap between the three taste qualities. The structural similarities between many

sweet and bitter compounds have been long known in the scientific community, with multiple cases of taste alteration on slight modifications in the structure [39]. Our work demonstrates this complication using data analytics and adds an added layer of complexity by including umami tastants within its scope.

3.2 Functional Group Analysis

To look deeper into the chemistry of tastants, we analyzed the functional groups present in the bitter, sweet, and umami compounds in our database. In-built modules in RDKit can compute the frequency of a predefined list of 85 substructures in a molecule, and we evaluated only those fragments for our analysis. Figure 4 shows the most common functional groups in the three classes of tastants along with their frequencies of occurrence. Carbonyl oxygen, which can belong to aldehydes, ketones, carboxylic acids, esters, amides, and other functional groups, is the most common fragment in both bitter (59.9 %) and sweet (66.2 %) molecules. In Figure 4, we consider only non-overlapping groups, and hence carbonyl is not shown— aldehyde, ketones, and other unique groups are treated as separate entities. It is visible from our analysis that many functional groups occur often in both bitter and sweet molecules, which agrees with our earlier discussion on the structural similarities between the two tastes. Apart from benzene, ether, and tertiary amine, which are among the five most frequently occurring groups in both sweet and bitter compounds as shown in Figure 4, we also find aliphatic hydroxyl (36.2 % bitter and 31.9 % sweet), carboxylic acid (12.1 % bitter and 33.3 % sweet), methoxy (17.1 % bitter and 19 % sweet), primary amine (9.6 % bitter and 22.4 % sweet), secondary amine (26.2 % bitter and 33.3 % sweet), and tertiary amine (34.7 % bitter and 14.4 % sweet) in many

molecules of the two classes. About 45.6 % bitter and 16.2 % sweet compounds have bicyclic rings in their structure.

Umami molecules are rich in nitrogen and phosphorous containing functional groups, as evident from Figure 4. Primary, secondary, and tertiary amines are present in 31.6 %, 76.5 %, and 64.3 % of umami compounds, respectively, the latter two being among the five most common groups. Imidazole (57.1 %), amide (27.6 %), and aniline (22.4 %) exist widely as well. Imidazole and phosphate ester strikingly distinguish umami as these two groups are scarcely found in sweet (0.2 % contain imidazole, 0.05 % contain phosphate ester) and bitter (2.3 % contain imidazole, 0.05 % contain phosphate ester) compounds. Like sweet and bitter, ether and aliphatic hydroxyl can be found frequently in umami molecules too, as shown in Figure 4. Sulfide group occur in 16.3 % of umami compounds, compared to 1.6 % and 4.3 % for bitter and sweet, respectively. Salts of reactive metals like calcium, potassium, sodium, and magnesium make up slightly more than half of the umami molecules in our database, with disodium salts being the most common.

Only five fragments, among the 85 calculated by RDKit, are not present in any of the molecules in our entire database. The bitter class shows the greatest diversity in terms of groups present in at least one compound. Even rare functional groups (present in less than 5 % of tastants for all three classes) occur more often in bitter molecules than in sweet or umami. This diversity can be attributed to the 25 T2Rs that can bind to a wide variety of ligands and elicit a taste sensation. In contrast, we observe the least number of functional groups in umami molecules, which can possibly be the cause or consequence of the smaller set of known umami-tasting compounds. Overall, the functional group analysis sheds light on the chemical make-up of

tastants and corroborates the inferences from exploratory data analysis regarding the complexity of the classification problem.

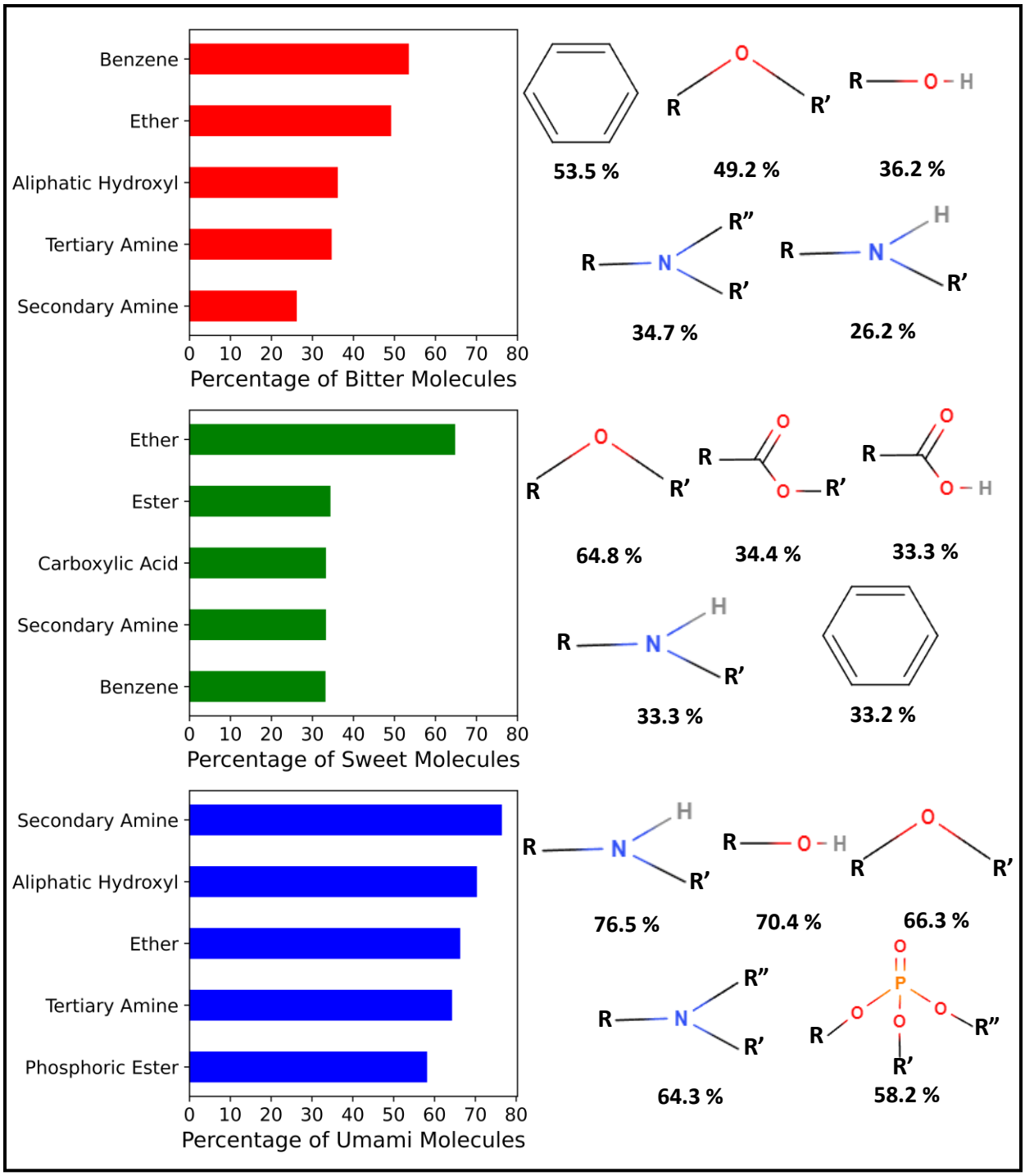

Figure 4: Most common functional groups in bitter, sweet, and umami compounds. The bar plots denote the percentage of compounds in each of the taste classes having the functional groups. The chemical structure of the functional groups and their exact percentage are provided beside each of the bar plots. Red, green, and blue in the bar plots correspond to bitter, sweet, and umami tastants, respectively.

3.3 DNN Model Performance

The predicted taste classes from the DNN model were compared to the actual taste labels in our dataset to evaluate its performance. Figure 5 shows the confusion matrices of the DNN model with and without SMOTE for both the training and test sets. The hold-out test set contains 285 bitter, 318 sweet, and 15 umami molecules, which preserves the classwise distribution of tastants in the entire dataset. The overall prediction accuracies of the DNN model without SMOTE are 0.90 and 0.87 for the training and test sets, respectively. On applying SMOTE, the corresponding accuracies rise to 0.91 and 0.89. An interesting observation is that the model rarely mislabels sweet or bitter compounds as umami. Although these results are satisfactory, accuracy is not always the appropriate measure of model performance in classification problems, especially with imbalanced datasets. We calculate more insightful metrics— precision, recall, and F1 score, for the three taste classes, as shown in Table 1. Precision is defined as the ratio of true positives to total predicted positives, while recall is the ratio of true positives to total actual positives. F1 score, which is the harmonic mean of precision and recall, strikes a balance between the two, and takes uneven class distribution into account. For multiclass classification tasks, especially with class imbalance, classwise F1 score, and confusion matrix gives a complete description of the prediction performance. We see that the F1 scores of all three classes improve on applying SMOTE. All precision and recall values

increase as well. In particular, the umami F1 score changes significantly, which was the main motivation for using SMOTE to resample the dataset.

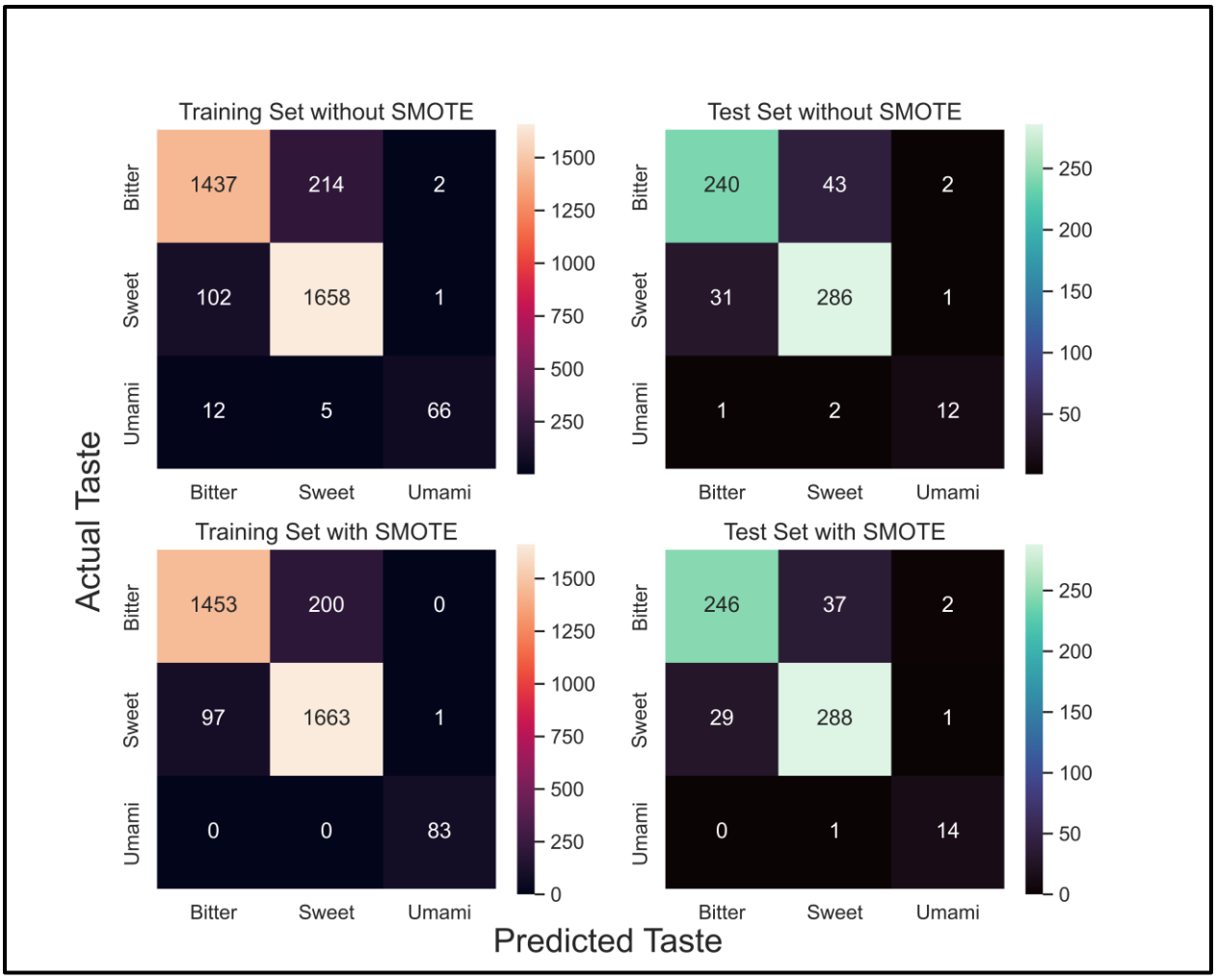

Figure 5: Confusion matrices based on the predictions of the DNN model with and without SMOTE for training and test sets. The numbers in each cell denote the absolute number of datapoints that satisfy the condition of the cell.

The confusion matrices show us that many sweet molecules are mislabeled as bitter and vice versa. The model is not able to entirely resolve the complexity of classification due to the

structural similarity between sweet and bitter compounds. Although direct comparisons with existing literature is not possible due to the novelty of the multiclass problem and disparity between datasets, the accuracies and F1 scores are similar to or better than those obtained in the simpler sweet/non-sweet and bitter/non-bitter classification models, many of which further suffer from additional limitations like small dataset size, random negative set, unverified taste information, and lack of chemical diversity [16, 17, 18, 21]. Hence, our results indicate that neural networks can tackle complex classification problems in the biochemical domain by learning representations, provided that properly curated datasets are available.

Table 1: Classwise precision, recall, and F1 scores of the training and test sets for the DNN model with and without SMOTE

Sampling	Dataset	Metric	Taste Class		
			Bitter	Sweet	Umami
Without SMOTE	Training	Precision	0.92	0.88	0.96
		Recall	0.87	0.94	0.79
		F1 Score	0.90	0.91	0.87
	Test	Precision	0.88	0.86	0.80
		Recall	0.84	0.90	0.80
		F1 Score	0.86	0.88	0.80
With SMOTE	Training	Precision	0.94	0.89	0.99
		Recall	0.88	0.94	1.00
		F1 Score	0.91	0.92	0.99
	Test	Precision	0.89	0.88	0.82
		Recall	0.86	0.91	0.93
		F1 Score	0.88	0.89	0.88

3.4 Explaining DNN Predictions using SHAP

Despite exceptional predicting capabilities, neural networks are infamous for their lack of interpretability and explainability. The two terms are often used interchangeably, but a subtle yet crucial difference exists. Interpretability is the extent to which a cause-and-effect relationship can be determined to consistently predict how the output changes given a change in input. High interpretability often comes at the cost of performance, as it is difficult to establish cause-effect relationships beyond simple ML models like linear regression and decision trees. For the DNN model, we are concerned with explainability, which aims to understand the behavior of ML algorithms in human terms. Much effort in ML research has been directed towards explainability in recent years, and we chose the Shapley additive explanations (SHAP) technique because of its unified framework for interpreting predictions [40]. SHAP uses a cooperative game-theoretic approach to compute the contribution of each feature towards the prediction and provides Shapley values as output. Specifically, we leverage the DeepExplainer method, which is suitable for neural networks. Figure 6 shows the average of absolute SHAP values over the entire test set for the ten key features based on the prediction of the DNN model with and without SMOTE. The features are ranked according to their overall relative importance, as explained by SHAP. The average SHAP values are computed separately for each taste class and plotted together by stacking. Please see supporting information for more detailed visualization for the three taste classes. Individual SHAP values corresponding to each test example and each feature is also shown in supporting information.

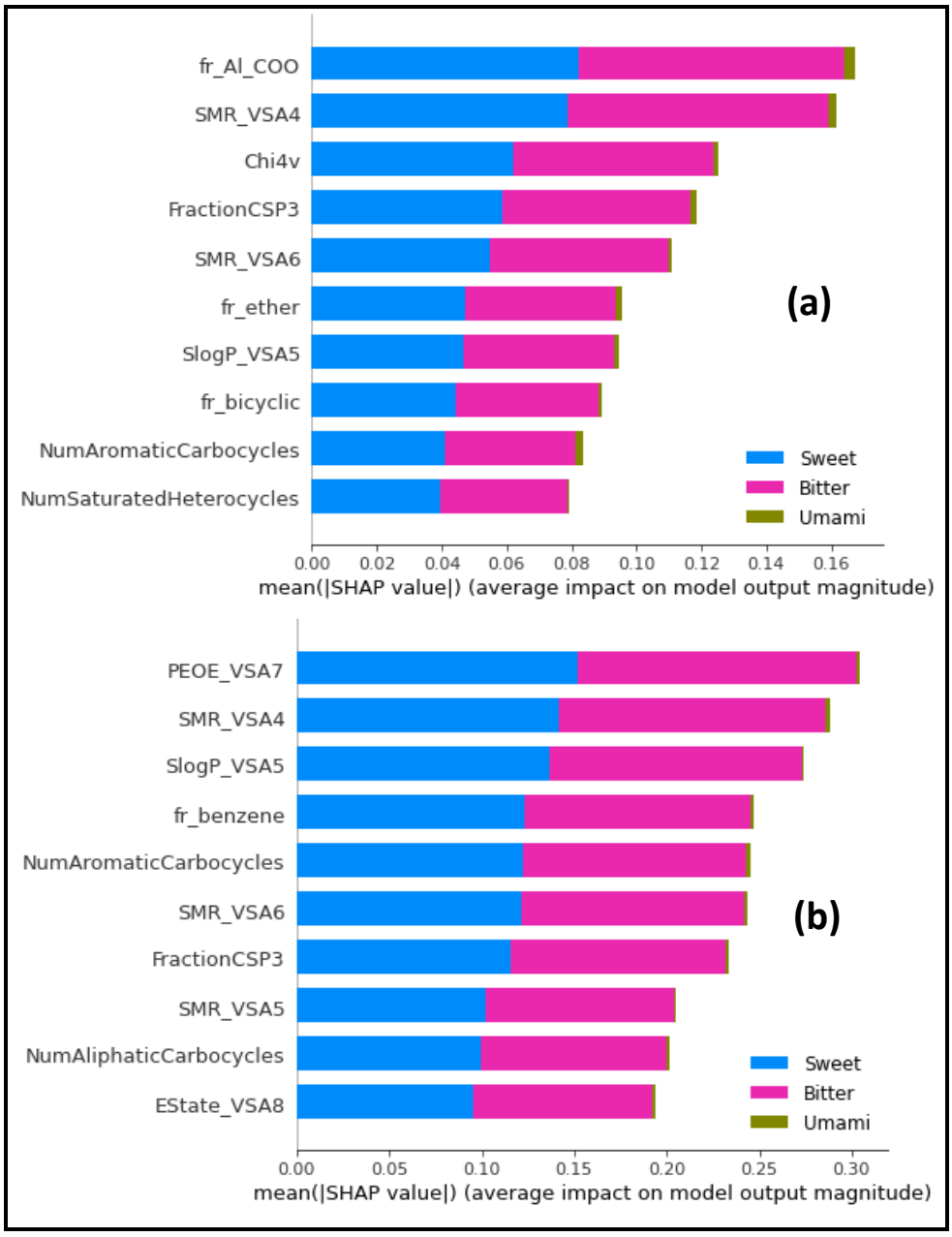

Figure 6: Bar plots of the mean absolute SHAP values for the test set of the ten important features, ranked in order of their relative importance, based on the predictions of the DNN model (a) without SMOTE and (b) without SMOTE. Stacking is used to visualize the SHAP values of all three taste classes in a single chart.

Comparing the 20 main features (only ten are shown in Figure 6) of the DNN model with and without SMOTE, 11 are common to both, although their relative importance changes. Resampling with SMOTE helps the model to learn the relevant information to adequately represent all three classes. We observe that electrostatic (PEOE_VSAs), polarizability (SMR_VSAs), hydrophobicity/hydrophilicity (SlogP_VSAs), and electro-topological (EState_VSAs) properties, along with Lipinski parameters (FractionCSP3, NumAliphaticCarbocyles, NumAromaticCarbocyles, NumSaturatedHeterocycles) and counts of fragments like benzene and aliphatic carboxylic acid, are the crucial features that determine the taste of molecules. The SHAP explanation agrees with our current understanding of the physics of receptor-ligand interactions, where these properties (electrostatic, polarizability, hydrophobicity, etc.) determine the affinity of molecules towards a binding pocket. The DNN model discovers this physics without being explicitly programmed and makes predictions accordingly. Our work demonstrates the capability of neural networks in learning complex structure-property relationships in molecules, while still being explainable to some extent with assistance from techniques like SHAP. Incorporating explainability within the realm of DL, especially in biochemical applications, can play a paramount role in bolstering its acceptance in industrial settings and bridging the gap with physics-based theoretical understanding of various phenomena.

3.5 GNN Model Performance

Similar to the DNN model, we compared the predictions of the GNN model with the actual taste labels and computed the performance metrics. Figure 7 shows the confusion matrices for the GNN model with and without oversampling. The overall prediction accuracies without

oversampling are 0.94 and 0.88 for the training and test sets, respectively, which are comparable to the DNN model accuracies. Like DNN, the GNN model also mislabels many sweet molecules as bitter and vice versa. We experimented with different dropout probabilities and found 0.1 for all layers to be an optimum value that reduced the extent of overfitting without compromising on the test set performance. SMILES enumeration for handling class imbalance is especially helpful to create valid synthetic data that is different from the original data, when string-based featurization is used for building ML models. However, for graph-based featurization, different SMILES for the same molecule lead to node feature and adjacency matrices with different row ordering, which ultimately generates the same embedding as permutation invariancy is a precondition of all GNNs. But oversampling of the minority class is also a valid resampling technique for dealing with class imbalance. The prediction accuracies with oversampling are 0.94 and 0.90 for the training and test set, respectively. Table 2 shows the classwise precision, recall, and F1 score of the GNN model for the training and test sets. We observe that most metrics are comparable to those of the DNN model. Oversampling significantly improves the predictions of the minority umami class, while moderate refinement can also be observed for the sweet and bitter classes during evaluation of the test set. Hence, the performances of both of our DL models are similar on all metrics.

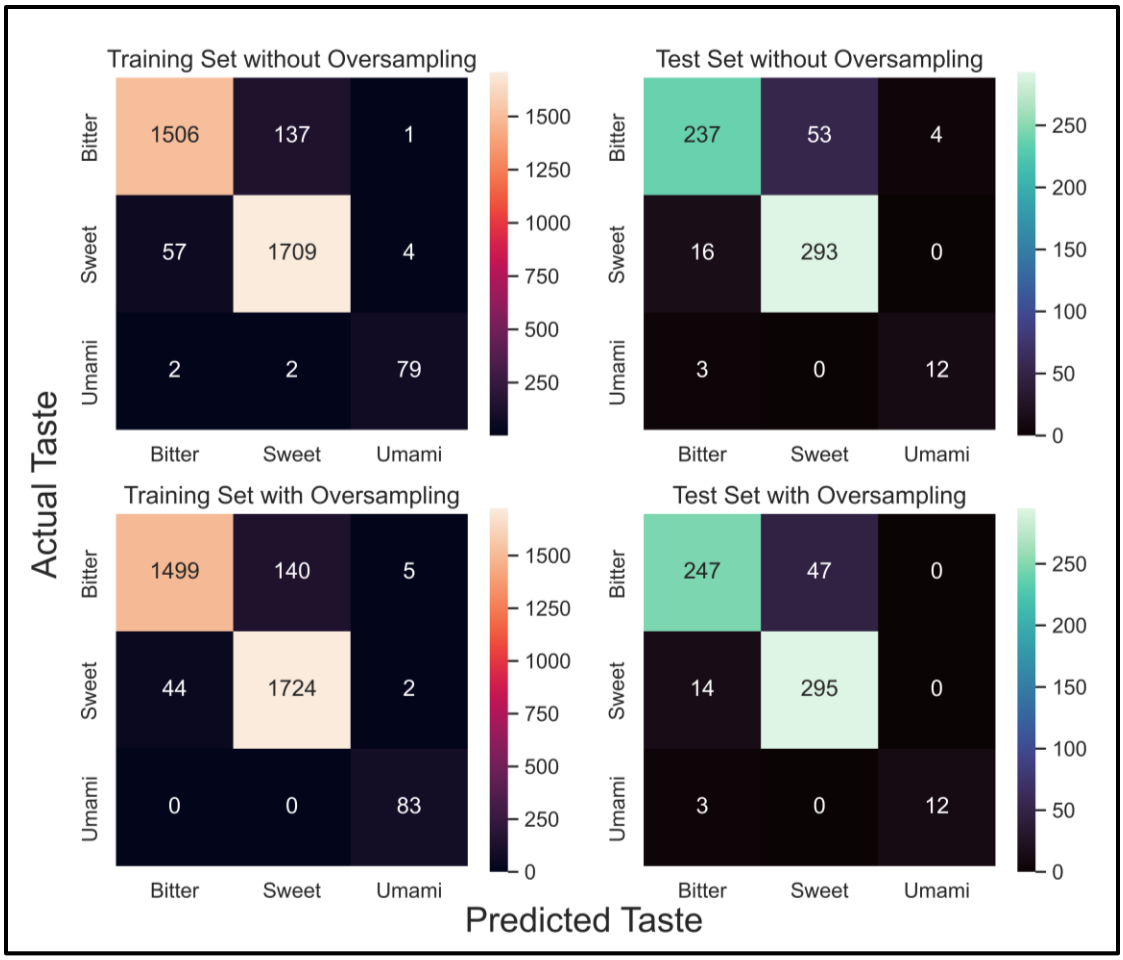

Figure 7: Confusion matrices based on the predictions of the GNN model with and without oversampling using SMILES enumeration for training and test sets. The numbers in each cell denote the absolute number of datapoints that satisfy the condition of the cell.

Although graph-based DL techniques have proved to be powerful tools for various molecular property prediction tasks, they often require large datasets to surpass the performance of traditional ML models. Otherwise, the likelihood of overfitting and poor test set performance is significant. Our tastant dataset only consists of a few thousand datapoints, which is typically not enough for GNNs. But we show that the graph convolutional approach proposed by Duvenaud et al. performs satisfactorily on the dataset and classifies bitter, sweet, and umami

molecules with high accuracy, precision, and recall. As discussed earlier, GNNs have the additional advantage of learning directly from the molecular structure and the associated chemical information like atom type, valance, aromaticity, etc., without the need for software-based expert featurization. The number of descriptors, their values and calculation methodology, as well as the range of information they provide can vary widely among the various commercial and free cheminformatics tools. Hence, if better or comparable performance is achieved, it is desirable to use GNNs in automated workflows over DNNs or traditional ML models. If larger datasets are obtained in the future from experiments, simulations, and data curations, GNNs are likely to outperform other alternatives.

Table 2: Classwise precision, recall, and F1 scores of the training and test sets for the GNN model with and without oversampling

Sampling	Dataset	Metric	Taste Class		
			Bitter	Sweet	Umami
Without Oversampling	Training	Precision	0.96	0.92	0.94
		Recall	0.92	0.97	0.95
		F1 Score	0.94	0.94	0.95
	Test	Precision	0.93	0.85	0.75
		Recall	0.81	0.95	0.80
		F1 Score	0.86	0.89	0.77
With Oversampling	Training	Precision	0.97	0.92	0.92
		Recall	0.91	0.97	1.00
		F1 Score	0.94	0.95	0.96
	Test	Precision	0.94	0.86	1.00
		Recall	0.84	0.95	0.80
		F1 Score	0.89	0.91	0.89

4 Conclusions

In this paper, we present a data-driven approach for analysis and classification of tastants. Among the five basic tastes, bitter, sweet, and umami are sensed via GPCRs and hence are considered for a multiclass classification problem. We curated an extensive dataset of verified tastants from literature, which included a diverse class of molecules. The characteristics of the dataset including key chemical properties and functional groups are discussed in detail. Significant structural similarities between sweet and bitter molecules are observed from our data analysis, which revalidates the existing ideas on taste. To classify the molecules, we built and trained a descriptor-based DNN model and a graph-based GNN model. Both showed comparable performance in terms of multiple metrics. The GNN model has a notable advantage of being able learn from the molecular structures without requiring handcrafted features. As the number of umami molecules in the dataset is much lower compared to bitter and sweet, we applied special techniques to handle the class imbalance. Additionally, the SHAP method was utilized to explain the predictions of the DNN model. We observed that the neural network attributed more importance to features that correspond to physically relevant properties for molecule binding to a receptor.

Future directions in computational taste prediction include expanding the ideas we have established in this paper to all the five basic tastes, multitaste, and tastelessness. Achieving this would require curating datasets with significant number of molecules belonging to all the different classes, as the performance of ML techniques strongly depends on the data. Estimating the relative taste intensities of molecules with respect to a reference is productive as well, but, except for sweetness, the progress is limited due to lack of experimental results.

For industrial applications, simultaneously predicting one or more taste labels for a molecule, along with the taste intensity, will be of great economic advantage. Taste prediction can be further combined with cheminformatics approaches for oral bioavailability and toxicity analysis to screen databases for potential tastants with desired properties. Finally, an integrated computational framework can include a molecular docking or molecular dynamics module at the end of the pipeline to validate the screened molecules. This work contributes a foundation to this framework and demonstrates that DL can be a powerful tool for food and flavor applications.

Acknowledgements

The authors would like to thank Mr. K Ananth Krishnan, CTO, Tata Consultancy Services and Dr. Gautam Shroff, Head of Research, Tata Consultancy Services, for their constant encouragement and support during this project.

Author Contributions

D.J. and R.G. conceived the project idea. P.D. and R.G. carried out the functional group analysis. P.D. and D.J. developed the deep learning models. All authors contributed to the interpretation and discussion of the results and preparation of this manuscript.

References

[1] R. L. Doty and S. M. Bromley, "Taste," in *Encyclopedia of the Neurological Sciences*, Amsterdam, Elsevier, 2014, pp. 394-396.

[2] C. Spence, "Multisensory flavor perception," *Cell,* vol. 161, pp. 24-35, 2015.

[3] J. Diószegi, E. Llanaj and R. Ádány, "Genetic background of taste perception, taste preferences, and its nutritional implications: a systematic review," *Frontiers in Genetics,* vol. 10, p. 1272, 2019.

[4] S. Jeong and J. Lee, "Effects of cultural background on consumer perception and acceptability of foods and drinks: a review of latest cross-cultural studies," *Current Opinion in Food Science,* vol. 42, pp. 248-256, 2021.

[5] T. Hummel, B. N. Landis and K.-B. Hüttenbrink, "Smell and taste disorders," *GMS Current Topics in Otorhinolaryngology- Head and Neck Surgery,* vol. 10, 2011.

[6] B. P. Trivedi, "Gustatory system: The finer points of taste," *Nature,* vol. 486, pp. S2-S3, 2012.

[7] R. Ahmad and J. E. Dalziel, "G Protein-Coupled Receptors in Taste Physiology and Pharmacology," *Frontiers in Pharmacology,* vol. 11, 2020.

[8] N. Tuteja, "Signaling through G protein coupled receptors," *Plant Signaling and Behavior,* vol. 4, p. 942.947, 2009.

[9] F. Horn, E. Bettler, L. Oliveira, F. Campagne, F. E. Cohen and G. Vriend, "GPCRDB information system for G protein-coupled receptors," *Nucleic Acids Research,* vol. 31, pp. 294-297, 2003.

[10] R. B. Chang, H. Waters and E. R. Liman, "A proton current drives action potentials in genetically identified sour taste cells," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 107, pp. 22320-22325, 2010.

[11] J. Chandrashekar, C. Kuhn, Y. Oka, D. A. Yarmolinsky, E. Hummler, N. J. P. Ryba and C. S. Zuker, "The cells and peripheral representation of sodium taste in mice," *Nature,* vol. 464, pp. 297-301, 2010.

[12] H. Wu, C. Wang, K. J. Gregory, G. W. Han, H. P. Cho, Y. Xia, C. M. Niswender, V. Katritch, J. Meiler, V. Cherezov, P. J. Conn and R. C. Stevens, "Structure of a Class C GPCR Metabotropic Glutamate Receptor 1 Bound to an Allosteric Modulator," *Science,* vol. 344, pp. 58-64, 2014.

[13] L. Chun, W.-h. Zhang and J.-f. Liu, "Structure and ligand recognition of class C GPCRs," *Acta Pharmacologica Sinica,* vol. 33, pp. 312-323, 2012.

[14] G. Spaggiari, A. Di Pizio and P. Cozzini, "Sweet, umami and bitter taste receptors: State of the art of in silico molecular modeling approaches," *Trends in Food Science & Technology,* vol. 96, pp. 21-29, 2020.

[15] C. Rojas, R. Todeschini, D. Ballabio, A. Mauri, V. Consonni, P. Tripaldi and F. Grisoni, "A QSTR-Based Expert System to Predict Sweetness of Molecules," *Frontiers in Chemistry,* vol. 5, 2017.

[16] S. Zheng, W. Chang, W. Xu, Y. Xu and F. Lin, "e-Sweet: A Machine-Learning Based Platform for the Prediction of Sweetener and Its Relative Sweetness," *Frontiers in Chemistry,* vol. 7, 2019.

[17] W. Huang, Q. Shen, X. Su, M. Ji, X. Liu, Y. Chen, S. Lu, H. Zhuang and J. Zhang, "BitterX: a tool for understanding bitter taste in humans," *Scientific Reports,* vol. 6, 2016.

[18] A. Dagan-Wiener, I. Nissim, N. Ben Abu, G. Borgonovo, A. Bassoli and M. Y. Niv, "Bitter or not? BitterPredict, a tool for predicting taste from chemical structure," *Scientific Reports,* vol. 7, 2017.

[19] S. Zheng, M. Jiang, C. Zhao, R. Zhu, Z. Hu, Y. Xu and F. Lin, "e-Bitter: Bitterant Prediction by the Consensus Voting From the Machine-Learning Methods," *Frontiers in Chemistry,* vol. 6, 2018.

[20] P. Banerjee and R. Preissner, "BitterSweetForest: A Random Forest Based Binary Classifier to Predict Bitterness and Sweetness of Chemical Compounds," *Frontiers in Chemistry,* vol. 6, 2018.

[21] R. Tuwani, S. Wadhwa and G. Bagler, "BitterSweet: Building machine learning models for predicting the bitter and sweet taste of small molecules," *Scientific Reports,* vol. 9, 2019.

[22] J.-B. Chéron, I. Casciuc, J. Golebiowski, S. Antonczak and S. Fiorucci, "Sweetness prediction of natural compounds," *Food Chemistry,* vol. 221, pp. 1421-1425, 2017.

[23] A. Goel, K. Gajula, R. Gupta and B. Rai, "In-silico prediction of sweetness using structure-activity relationship models," *Food Chemistry,* vol. 253, pp. 127-131, 2018.

[24] M. Malavolta, L. Pallante, B. Mavkov, F. Stojceski, G. Grasso, A. Korfiati, S. Mavroudi, A. Kalogeras, C. Alexakos, V. Martos, D. Amoroso, G. Di Benedetto and D. Piga, "A survey on computational taste predictors," *European Food Research and Technology,* 2022.

[25] A. Goel, K. Gajula, R. Gupta and B. Rai, "In-silico screening of database for finding potential sweet molecules: A combined data and structure based modeling approach," *Food Chemistry,* vol. 343, 2021.

[26] W. P. Walters and R. Barzilay, "Applications of Deep Learning in Molecule Generation and Molecular Property Prediction," *Accounts of Chemical Research,* vol. 54, pp. 263-270, 2021.

[27] O. Wieder, S. Kohlbacher, M. Kuenemann, A. Garon, P. Ducrot, T. Seidel and T. Langer, "A compact review of molecular property prediction with graph neural networks," *Drug Discovery Today: Technologies,* vol. 37, pp. 1-12, 2020.

[28] D. Jiang, Z. Wu, C.-Y. Hsieh, G. Chen, B. Liao, Z. Wang, C. Shen, D. Cao, J. Wu and H. Tingjun, "Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models," *Journal of Cheminformatics,* vol. 13, 2021.

[29] C. Rojas, D. Ballabio, K. P. Sarmiento, E. P. Jaramillo, M. Mendoza and F. García, "ChemTastesDB: A curated database of molecular tastants," *Food Chemistry: Molecular Sciences,* vol. 4, 2022.

[30] *RDKit, Open Source Cheminformatics Software.*

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau and M. Brucher,

"Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research,* vol. 12, pp. 2825-2830, 2011.

[32] F. Chollet, "Keras," 2015.

[33] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research,* vol. 16, pp. 321-357, 2002.

[34] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik and R. P. Adams, "Convolutional Networks on Graphs for Learning Molecular Fingerprints," in *Advances in Neural Information Processing Systems*, 2015.

[35] *DeepChem: Deep Learning Models for Drug Discovery and Quantum Chemistry.*

[36] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing and V. Pande, "MoleculeNet: a benchmark for molecular machine learning," *Chemical Science,* vol. 9, pp. 513-530, 2018.

[37] A. Di Pizio, Y. B. Shoshan-Galeczki, J. E. Hayes and M. Y. Niv, "Bitter and sweet tasting molecules: It's complicated," *Neuroscience Letters,* vol. 700, pp. 56-63, 2019.

[38] L. Van Der Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research,* vol. 9, pp. 2579-2605, 2008.

[39] S. Damodaran and K. L. Parkin, Fennema's Food Chemistry, CRC Press, 2017.

[40] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, 2017.