Computational Screening of Umami Tastants Using Deep Learning

Prantar Dutta, Kishore Gajula, Rakesh Gupta* and Beena Rai

Physical Sciences Research Area, TCS Research, Tata Consultancy Services, Pune, India

* *Correspondence to*: Rakesh Gupta

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

Abstract

Umami, a fundamental human taste modality, refers to the savory flavors in meats and broths, often associated with monosodium glutamate and protein richness. With limited knowledge of umami molecules, the food industry seeks efficient approaches for identifying novel tastants. In this study, we have devised a virtual screening pipeline for identifying potential novel umami tastants from molecular databases. We first curated a comprehensive classification dataset containing 439 umami and 428 non-umami molecules. A transformer-based architecture was trained to differentiate between the two classes, achieving the best performance to date. Additionally, we built a neural network model for predicting the potency of umami compounds, the first effort of its kind. These two models, in conjunction with similarity analysis and toxicity screening, form an end-to-end framework for the rational discovery of novel tastants. We finally applied this framework to the FooDB database as an illustrative use case. This study demonstrates the potential of data-driven methods in predicting the taste of molecules from structural and chemical features.

Keywords: Tastant, Umami, Deep Learning, QSPR, Computational Screening

https://doi.org/10.26434/chemrxiv-2024-spz16 **ORCID:** https://orcid.org/0000-0002-1019-9418 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

1 Introduction

Umami, an essential taste modality, plays a distinctive role in shaping the overall flavor profile of various foods. It is commonly associated with protein-rich food sources that evoke a taste akin to monosodium glutamate (MSG) [1]. In 1908, Kikunae Ikeda, a Japanese chemist, introduced the term 'umami' to describe the savory taste sensation of MSG. In 2002, it officially gained recognition as one of the five fundamental tastes, joining the ranks of sweet, bitter, sour, and salty. The Umami Information Center (https://www.umamiinfo.com/) in Japan maintains an extensive and diverse database encompassing over 800 food items that showcase umami characteristics, spanning meats, fish and shellfishes, mushrooms, spices, seasonings, cheeses, and pulses. In mammals, the sensation of umami primarily relies on the heterodimeric T1R1/T1R3 receptors situated within the taste buds of the tongue, which belong to the Class C of the G protein-coupled receptors (GPCRs) family [2]. To a lesser extent, metabotropic glutamate receptors 1 and 4 (mGluR1 and mGluR4), also classified as GPCRs, contribute to umami perception [3, 4]. Upon exposure to umami tastants, these receptors activate and transmit signals to the brain's gustatory cortex. The foremost stimulants of these receptors include glutamates, aspartates, and nucleotides. For a comprehensive exploration of umami taste receptors and the underlying molecular mechanisms governing umami perception, readers are encouraged to refer to published articles and reviews dedicated to this subject [5, 6, 7].

The development of rapid, dependable, and affordable methods for predicting the taste of compounds and ingredients holds paramount importance in the food industry. Although experimental techniques, such as MALDI-TOF mass spectroscopy and reversed-phase high-performance liquid chromatography, are available for identifying and characterizing umami molecules, these approaches are laborious, costly, and time-intensive [8]. Consequently, there has

https://doi.org/10.26434/chemrxiv-2024-spz16 **ORCID:** https://orcid.org/0000-0002-1019-9418 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

been a rising inclination towards in-silico quantitative structure-property relationship (QSPR) models as a promising alternative to experimental methods. These models establish a correlation between the taste of molecules and their structural characteristics, and their popularity can be attributed to advancements in machine learning (ML) algorithms, particularly deep learning (DL), and the availability of curated datasets. While various studies have focused on developing classification models to differentiate between sweet/non-sweet and bitter/non-bitter compounds [9, 10, 11, 12, 13, 14, 15], aiming to identify novel tastants by screening molecule databases, progress in the realm of umami taste has been comparatively limited. To address this gap, Charoenkwan et al. curated the UMP442 peptide database. They devised two distinct methodologies, iUmamiSCM [16] and UMPredFRL [17], based on amino acid sequences, for classifying peptides as umami or non-umami. Similarly, Pallante et al. leveraged the same database to construct ML models; however, they transformed the peptide amino acid sequences into simplified molecular-input line-entry system (SMILES) representations to ensure broader applicability of the models to general molecules [18]. In another work, using deep learning algorithms, SMILES-based and graph-based representations were used to simultaneously classify sweet, bitter, and umami molecules [19]. However, despite the existing efforts, there remains a significant opportunity to collect additional data, advance model development, and predict the umaminess of compounds, typically assessed in terms of intensity or potency.

In this contribution, we have curated the most extensive umami classification dataset by combining a list of nitrogenous compounds that evoke either umami or sweet taste from a patent document with the UMP442 database. We effectively classified umami and non-umami molecules using a state-of-the-art DL technique, outperforming existing models. Furthermore, in what we believe to be the first endeavor of its kind, we built predictive models to categorize umami molecules based

https://doi.org/10.26434/chemrxiv-2024-spz16 **ORCID:** https://orcid.org/0000-0002-1019-9418 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

on their potency. Finally, we leveraged these models to develop a virtual screening pipeline. We applied it to a large food database, illustrating the pertinence of ML approaches for in-silico taste prediction of compounds. Our work expands the possibilities for rational design and screening of novel tastants, creating opportunities to craft flavor profiles that cater to discerning human taste preferences.

2 Methods

2.1 Dataset

A patent, filed by the American biotechnology company Senomyx Inc. (later acquired by the current patent assignee Firmenich Inc.) in 2004 across various geographies, disclosed the synthesis and in-vitro studies of flavoring agents and enhancers possessing umami and sweet tastes aimed at enhancing the taste and flavor of food, beverages, and oral drug formulations [20]. It detailed a collection of 299 umami and 126 sweet molecules, all under the amide-derivatives category, sharing a common general formula. We extracted the IUPAC names of the compounds from the patent and converted them to SMILES representation using PubChemPy for compounds present in PubChem and the OPSIN server (https://opsin.ch.cam.ac.uk/) [21] for others. Additionally, the patent provided the half maximal effective concentration (EC50) of the umami molecules for activating the human T1R1/T1R3 receptor, determined through in-vitro dose-response analysis. We collected this data to develop a potency prediction model. Readers are advised to refer to the patent document for a comprehensive understanding of the step-by-step synthesis procedures, experimental details of activity assays, and other molecular properties.

https://doi.org/10.26434/chemrxiv-2024-spz16 **ORCID:** https://orcid.org/0000-0002-1019-9418 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

The UMP442 dataset, freely available on GitHub (https://github.com/Shoombuatong/Dataset-Code/tree/master/iUmami), is a compilation of amino acid sequences. It comprises 140 umami peptides, curated from prior literature, and 302 non-umami peptides sourced from a bitter peptide database. We employed the cheminformatics package RDKit to convert these sequences into SMILES representations to achieve consistency and compatibility. Subsequently, we merged and standardized the SMILES representations from the UMP442 dataset and the patent for the classification task. This process yielded a final dataset containing 867 entries, encompassing 439 umami molecules and 428 non-umami molecules. The negative set includes sweet and bitter molecules, whose taste sensations are also mediated by GPCRs. Our dataset is the most expansive and diverse collection of umami compounds to date, ensuring a robust foundation for further investigations in this domain.

2.2 Featurization and Data Preprocessing

We extracted features from the SMILES representation of the tastants using the Descriptors module of RDKit, resulting in 208 2-D features per molecule, encompassing a wide range of structural, physical, and chemical properties. However, not all descriptors are relevant to the dataset or task at hand; thus, eliminating irrelevant features was necessary to enhance the quality and generalizability of our model. We initially removed all constant and quasi-constant features for the umami classification problem using the entire dataset. Subsequently, we filtered out features with non-zero values in less than 20% of the molecules, reducing the feature set length to 124. Next, we assessed Pearson's correlation coefficient between the 124 descriptors to evaluate their linear association. If the coefficient exceeded 0.7, one of the features was randomly discarded. Ultimately, our refined dataset comprised 28 informative features for the 867 molecules.

https://doi.org/10.26434/chemrxiv-2024-spz16 **ORCID:** https://orcid.org/0000-0002-1019-9418 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

The UMP442 dataset incorporates pre-defined training (UMP-TR) and test (UMP-IND) sets, with a train-test split of 80:20, to facilitate consistent model comparison. UMP-TR comprises 112 umami and 241 non-umami molecules, whereas UMP-IND is composed of 28 umami and 61 non-umami molecules. Previous works employing the UMP442 dataset [16, 17, 18] used this fixed split to train and validate their models. We divided the data obtained from the patent into training and test sets using the same split ratio. The resultant training set consists of 239 umami and 101 non-umami molecules, while 60 umami and 25 non-umami molecules comprise the test set. Subsequently, we combined UMP-TR and UMP-IND with their corresponding counterparts from the patent data to construct overall training and test sets for model building and validation. The combined training set comprises 693 records, while the test set comprises 174. Finally, we applied z-score normalization to scale all 28 features before model construction.

The task of predicting potency focused exclusively on the patent dataset, which provides EC50 values for activating the human T1R1/T1R3 receptor. This was conceptualized as a binary classification problem, categorizing molecules with EC50 values below 2 μM as highly potent. The rationale for this methodology is detailed in section 3.1. Initially, we generated 208 features using RDKit and applied a data-cleaning procedure similar to the one employed for the umami/non-umami classification problem discussed earlier. Then, we calculated the Pearson's correlation coefficient between each feature and the EC50 values. In instances where the coefficient was less than 0.05, the respective feature column was excluded from the dataset. The final cleaning step involved removing one of a pair of correlated features with a Pearson's coefficient greater than 0.7. The cleaned dataset consisted of 49 features for the 299 umami tastants.

https://doi.org/10.26434/chemrxiv-2024-spz16 **ORCID:** https://orcid.org/0000-0002-1019-9418 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

The selected EC50 threshold of 2 μM classified 246 molecules into the high potency category, while the remaining 53 were designated as having low umami potency. Employing an 80:20 train-test ratio, the resulting training set comprised 198 high-potency and 41 low-potency tastants, while the test set included 48 high-potency and 12 low-potency molecules. To normalize the features, we applied min-max scaling. However, this led to an imbalance in the dataset, with fewer molecules exhibiting low than high potency. To address this class imbalance and prevent potential model bias, we implemented the Synthetic Minority Oversampling Technique (SMOTE), a widely used data augmentation method [22]. SMOTE randomly selects instances from the minority classes, identifies their five nearest neighbors within the same class, and randomly selects one. Subsequently, it generates synthetic samples by creating a convex combination of the two points in the feature space. The resampled training set, post-SMOTE application, was the basis for constructing the potency classification model.

2.3 Model Development

We employed TabPFN, a transformer-based architecture designed specifically for supervised classification on small tabular datasets, for the umami classification task [23]. Notably, TabPFN excels without hyperparameter tuning and has demonstrated superior performance to traditional classification methods across numerous small datasets containing fewer than 1000 data points. TabPFN uses a pre-trained transformer to approximate Bayesian inference for tabular datasets in a single pass, eliminating the necessity for parameter updates. Given a labeled training set, the model predicts the test set through in-context learning. An additional advantage lies in its ability to achieve high-accuracy classification within seconds, rendering it suitable for high-throughput screening pipelines. For a deeper understanding of the theoretical foundations and training protocol of TabPFN, readers are encouraged to refer to the original paper.

https://doi.org/10.26434/chemrxiv-2024-spz16 **ORCID:** https://orcid.org/0000-0002-1019-9418 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

The pre-trained TabPFN architecture was provided with the training data for the umami classification problem to approximate probabilistic inference, and subsequently, the test set was used for predictions. The model performance was evaluated by calculating a series of metrics – accuracy, balanced accuracy, sensitivity, specificity, F1-score, and the area under the receiver operating characteristic curve (ROC-AUC) for the two datasets. We also computed the above metrics for UMP-IND, a subset of the original test set, to compare the classification performance with existing models in the literature.

A deep neural network (DNN), also called multilayer perceptron, was constructed for the potency classification task. This network comprised an input layer, followed by two hidden layers, each consisting of 100 neurons and an output neuron. The neurons of the hidden layers were activated by a rectified linear unit (ReLU) function. In contrast, the sigmoid function was used in the output neuron to predict the probability of a molecule possessing high umami potency. A probability threshold 0.5 served as the differentiating criteria between the two classes. To reduce overfitting, we employed the dropout regularization technique after both the hidden layers with a probability of 0.3. The training data was provided to the model in batches of 32. The Adam optimizer with a learning rate 0.0001 and the binary cross-entropy loss function were utilized for model training. 15 % of the training set was reserved for validation, and the model with the lowest validation loss during the 1000 epochs of the training process was identified and saved as the optimal model. We finalized the chosen architecture after conducting numerous experiments involving varying numbers of hidden layers, a range of unit values for each hidden layer, and diverse dropout strategies. The aim was to optimize validation accuracy while minimizing the risk of overfitting. Finally, the accuracy, precision, recall, and F1-score performance metrics were calculated for the training and test sets to evaluate model performance.

https://doi.org/10.26434/chemrxiv-2024-spz16 **ORCID:** https://orcid.org/0000-0002-1019-9418 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

2.4 Screening Pipeline

The umami classification and potency prediction ML models were leveraged to develop a virtual screening pipeline for identifying potential umami tastants from molecule databases. Figure 1 shows the overall structure and flow of the pipeline. While deploying ML models in real-world scenarios, specifying an applicability domain (AD) is imperative for the targeted exclusion of molecules with structures markedly different from those present in the training set. The AD of the umami classification model was established using the Tanimoto similarity coefficient. Suppose the median Tanimoto coefficient between a new molecule and its three most structurally similar counterparts in the training set is below 0.6. In that case, the molecule falls outside the model's AD. Consequently, the initial step in the pipeline involves the selection of molecules based on their structural similarity. It is important to highlight that the similarity threshold is not rigid and can be adjusted according to the desired level of model reliability and the specific requirements of the industrial use case.

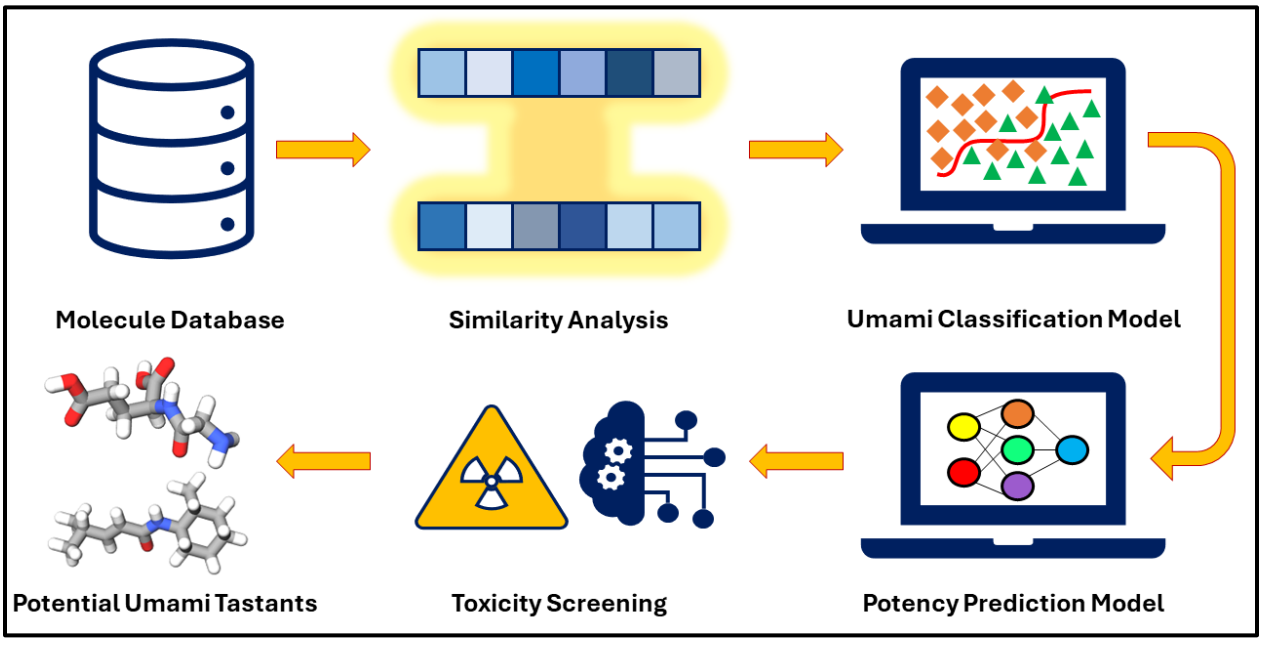

Figure 1: Virtual screening pipeline for identifying novel umami tastants from molecule databases

https://doi.org/10.26434/chemrxiv-2024-spz16 **ORCID:** https://orcid.org/0000-0002-1019-9418 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Following the initial shortlisting, the umami classification model is employed to identify molecules with umami taste. Subsequently, the umami potency prediction model is utilized to select highly potent tastants among the molecules identified as umami in the preceding step. Finally, the selected potent compounds undergo toxicity screening, refining the screening process to identify safe edibles. Many toxicity prediction tools and models are already available in the literature, and no new models were developed for this study. We used the ProTox-II webserver (https://tox-new.charite.de/protox_II/) to estimate the median lethal dose (LD50) of the screened compounds, measured in milligrams per kilogram of body weight [24]. The server classifies molecules, given their SMILES strings, into one of the following six classes based on LD50 values: Class I (fatal if swallowed; LD50 ≤ 5), Class II (fatal if swallowed; 5 < LD50 ≤ 50), Class III (toxic if swallowed; 50 < LD50 ≤ 300), Class IV (harmful if swallowed; 300 < LD50 ≤ 2000), Class V (may be harmful if swallowed; 2000 < LD50 ≤ 5000), and Class VI (non-toxic; LD50 > 5000). We consider molecules predicted to belong to Class VI as potential umami tastants. The entire pipeline was applied to discover novel taste molecules from the FooDB database (https://foodb.ca/) to illustrate a representative use case.

3 Results and Discussion

3.1 Exploratory Data Analysis

The umami classification dataset, composed of 439 umami and 428 non-umami molecules, was analyzed to unveil underlying chemical features and patterns. Since taste sensation arises from ligand-receptor binding interactions, molecules' structural and chemical properties significantly influence their ability to elicit taste responses. Figure 2 (a) shows the density distribution of

https://doi.org/10.26434/chemrxiv-2024-spz16 ORCID: https://orcid.org/0000-0002-1019-9418 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

molecular weights of the tastants in our dataset. Nearly all umami molecules exhibit molecular weights below 1000 Daltons, with most falling below 500 Daltons. As molecular weight is determined by the size of molecules—a crucial factor influencing receptor-ligand binding—the majority of tastants in the dataset can be categorized as small molecules, given their respective weights. The non-umami molecules, characterized by sweetness and bitterness, predominantly belong to the small molecule category. Figure 2 (b) shows the octanol-water partition coefficient density distribution (logP). Within this dataset, 70% of the umami tastants exhibit hydrophobic characteristics ($logP > 0$), with the remaining being hydrophilic ($logP < 0$). Thus, we cannot conclude with certainty that the hydrophobicity of molecules plays a role in triggering taste response in humans. The logP distribution for non-umami molecules, akin to their molecular weight distribution, exhibits substantial overlap with that of umami tastants, underscoring the complexity of the classification task. Figure 2 (c) shows the density distribution of hydrogen bond donors and acceptors in the 439 umami molecules in our dataset. The stabilization of protein-ligand complexes is crucially influenced by hydrogen bonding, which significantly impacts binding affinity. Most umami compounds exhibit a notable propensity for hydrogen bonding, with the presence of both hydrogen bond donors and acceptors. Hence, they interact strongly with the amino acid residues in the binding pocket.

We employed principal component analysis (PCA) to visualize and gain insights into the structure of our dataset. PCA serves as a technique for dimensionality reduction, creating new uncorrelated variables from initially correlated features while maintaining the variance in the original data. In Figure 2 (d), the two-dimensional feature space, derived from the 28 features obtained post-cleaning, illustrates the first and second principal components, explaining approximately 26% and 12% of the variance in the data, respectively. The plot underscores the considerable structural and

https://doi.org/10.26434/chemrxiv-2024-spz16 **ORCID:** https://orcid.org/0000-0002-1019-9418 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

chemical diversity within the database. Notably, the evident overlap between umami and non-umami molecules highlights the complicated nature of the classification task, emphasizing the necessity for ML models capable of capturing complex, non-linear relationships and dependencies.

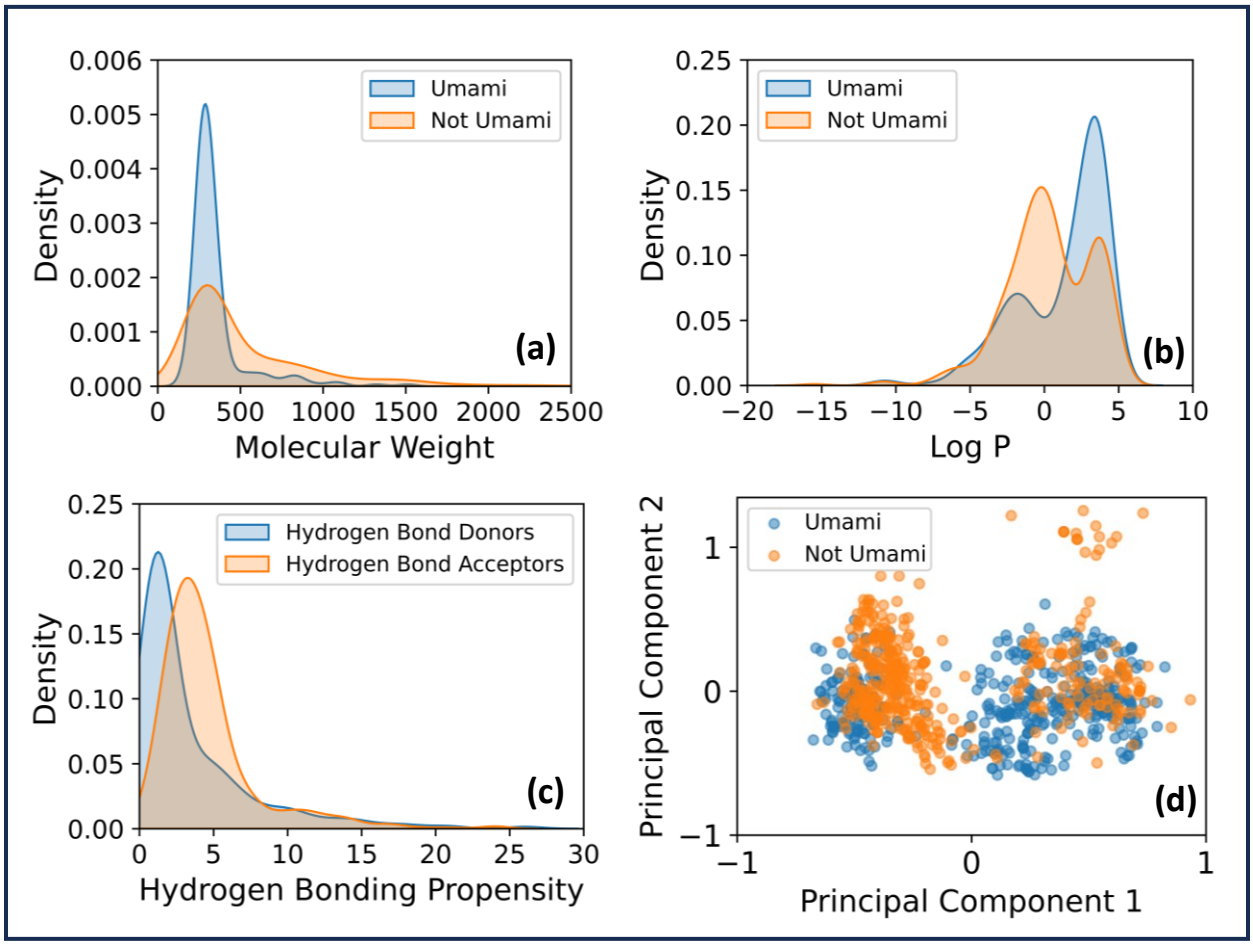

Figure 2: Dataset analysis based on molecular properties. Density distribution of (a) molecular weight and (b) octanol-water partition coefficient of umami and non-umami molecules in the dataset. (c) Count of hydrogen bond donors and acceptors in the umami molecules. (d) Unsupervised dimensionality reduction using principal component analysis. Transparency rendering is applied to show overlap.

To delve into the correlation between the presence of various functional groups and taste, we conducted a detailed analysis of the functional groups in the umami molecules. RDKit was used

https://doi.org/10.26434/chemrxiv-2024-spz16 **ORCID:** https://orcid.org/0000-0002-1019-9418 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

to quantify the occurrence of 85 specific substructures within the molecules. The findings revealed a consistent presence of the amide group in all umami molecules. This outcome aligns with expectations, given that our dataset comprises peptides from UMP442 and amide derivatives from the patent. Due to the same underlying cause, over 99% of the umami compounds feature secondary amine groups. Approximately 67% of the molecules exhibit aromaticity, as evidenced by the presence of benzene rings in their structure. 46% of umami tastants feature an ether group, with 36% specifically incorporating methoxy. Carboxylic acid is present in 32% of the compounds. Other nitrogenous functional groups, such as primary amine, tertiary amine, and pyridine, are present in 31%, 24%, and 12% of the molecules, respectively.

Subsequently, we examined the 299 umami compounds within the patent in the context of the potency prediction problem. Figure 3 (a) depicts the distribution of EC50 values for activating the human T1R1/T1R3 receptor, revealing a highly skewed distribution. Approximately 54% of the tastants exhibit EC50 values below 1 μM, and a significant majority, around 82%, showcase values below 2 μM. The characteristics of the data, marked by its low volume, sparsity, and skewness, pose challenges for developing robust and generalizable regression models to predict precise EC50 values. Additionally, the absence of a standard and widely accepted scale for quantifying the intensity of umaminess diminishes the practical significance of knowing exact EC50 values for umami tastants. Consequently, we reframe the potency prediction problem as a classification task.

The patent document specifies that, for a compound to exhibit umaminess, it should possess an EC50 value below 10 μM, and more preferably below 5 μM, 3 μM, 2 μM, 1 μM, or 0.5 μM. Thus, there is no precisely defined threshold for categorizing the potency of umami tastants. In this study, we adopt a threshold of 2 μM to delineate two classes—high and low—based on the data distribution. However, alternative thresholds within the 0.5 μM to 5 μM range may also be

https://doi.org/10.26434/chemrxiv-2024-spz16 **ORCID:** https://orcid.org/0000-0002-1019-9418 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

considered. Figure 3 (b) shows the PCA plot derived from transforming the 49 features obtained post-data cleaning. The first and second principal components explain about 19 % and 16 % of the variance in the data, respectively. The significant overlap between the two classes underlines the non-trivial nature of their classification, emphasizing the need for complex models.

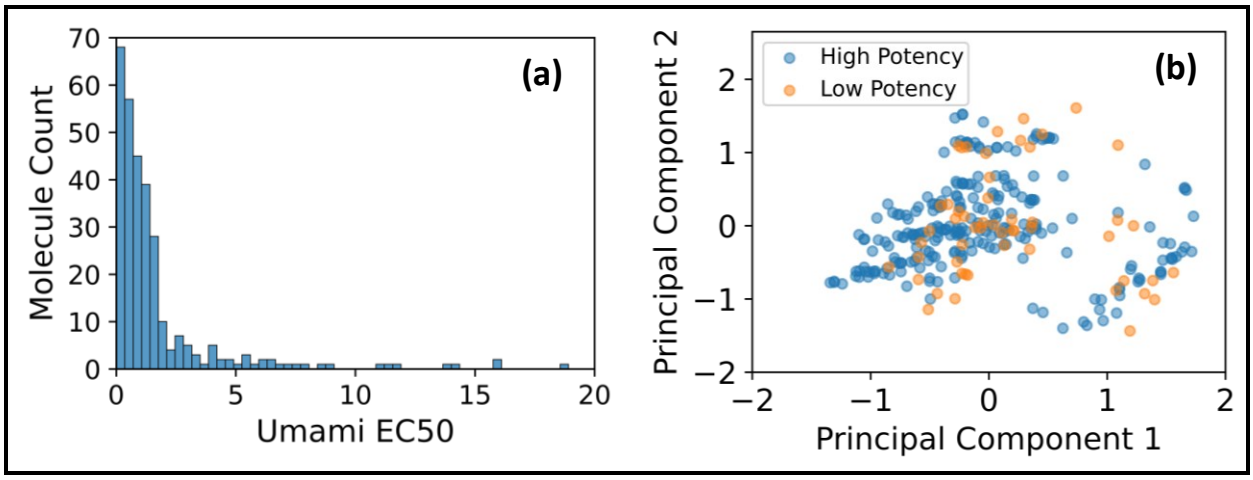

Figure 3: (a) The distribution of half maximal effective concentration to activate the human T1R1/T1R3 receptor of umami molecules in the patent. (b) Unsupervised dimensionality reduction using principal component analysis of the patent data. Transparency rendering is applied to show overlap.

The insights derived from exploratory data analysis underscore the richness of the dataset in terms of structural features and chemical properties, revealing a diverse array of molecules. The examination of key molecular properties and dimensionality reduction highlights a substantial overlap between umami compounds and the negative set, encompassing both sweet and bitter tastants. These findings affirm the intricate nature of the umami classification task and validate the robustness of our approach. Additionally, the potency prediction task poses significant challenges given the characteristics of the EC50 distribution, and the transformation into a binary classification problem enhances its tractability and applicability in industrial contexts.

https://doi.org/10.26434/chemrxiv-2024-spz16 **ORCID:** https://orcid.org/0000-0002-1019-9418 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Collectively, these results emphasize the imperative for sophisticated statistical models capable of learning pertinent molecular representations and making accurate predictions.

3.2 Umami Classification Model

The umami classification model was built using 693 training records and 174 hold-out test records. Given that the number of umami molecules in the combined dataset is approximately equal to that of non-umami molecules, and this balance is maintained in both the training and test splits, no further data augmentation was deemed necessary. The TabPFN classifier underwent in-context learning with the training set and was subsequently assessed on the test set using six widely utilized metrics. The classifier's performance for both the training and test sets is presented in Table 1. We observe that the values for all the metrics—accuracy, balanced accuracy, sensitivity, specificity, F1-score, and ROC-AUC—are above 0.99 for the training set and 0.93 for the test set. While the table may imply identical values for all the metrics in both training and test sets, they differ in their third or higher decimal places. Sensitivity, also known as recall, and specificity reflect the true positive and true negative rates, respectively. On the other hand, the F1 score offers a balance between precision and recall. The ROC-AUC metric evaluates the capacity of the model to discriminate between positive and negative instances across various threshold values. The comprehensive array of metrics collectively provides a holistic view of the performance, and the results showcase the high accuracy and robustness of the umami classifier. Furthermore, the training and inference processes occur instantaneously, rendering it well-suited for flexible reconfiguration with additional data and deployment in high-throughput screening applications.

https://doi.org/10.26434/chemrxiv-2024-spz16 **ORCID:** https://orcid.org/0000-0002-1019-9418 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Table 1: Performance metrics of the umami classification model for the training and test sets

Metric	Training Set	Test Set
Accuracy	0.99	0.93
Balanced Accuracy	0.99	0.93
Sensitivity	0.99	0.93
Specificity	0.99	0.93
F1-score	0.99	0.93
ROC-AUC	0.99	0.93

Our model demonstrates superior performance compared to previously reported umami classification models in the literature: iUmami-SCM [16], UMPred-FRL [17], and VirtuousUmami [18]. However, it is important to note that these models were trained on a subset of the training data utilized in this study. To facilitate an accurate comparison, we assessed the performance of the TabPFN classifier on the UMP-IND set, which constitutes the hold-out data in the three studies, as mentioned earlier, and is a subset of our test set. We computed the four metrics shared across all the studies—accuracy, sensitivity, specificity, and ROC-AUC. Table 2 presents a comparative analysis of all models, focusing on predictions for the UMP-IND dataset. It is evident from the results that our model surpasses the three other models in both accuracy and sensitivity. Furthermore, the specificity is the same as the two models and exceeds that of the third. However, our model ranks third in terms of ROC-AUC, albeit with a marginal difference of only 0.04 compared to the best-performing model. However, iUmami-SCM and UMPred-FRL, the two models exhibiting higher ROC-AUC than ours for the UMP-IND set, rely on features derived from peptide sequence information, making them inapplicable to other classes of molecules. In contrast, VirtuousUmami, which utilizes SMILES-derived descriptors, demonstrates lower values for all metrics in comparison to our model. Therefore, our work introduces a high-performance, generalizable model for identifying umami-tasting compounds that surpasses previous contributions in literature.

https://doi.org/10.26434/chemrxiv-2024-spz16 **ORCID:** https://orcid.org/0000-0002-1019-9418 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Table 2: Performance comparison of the present work with existing literature for the UMP-IND set

Model	Metric			
	Accuracy	**Sensitivity**	**Specificity**	**ROC-AUC**
iUmami-SCM	0.865	0.714	**0.934**	0.898
UMPred-FRL	0.888	0.786	**0.934**	**0.919**
VirtuousUmami	0.876	0.786	0.918	0.85
This work	**0.899**	**0.821**	**0.934**	0.878

A noteworthy observation from Table 1 and Table 2 is that the umami classification model exhibits superior performance on the entire test set compared to its subset, UMP-IND. This discrepancy suggests that molecules obtained from the patent are markedly simpler to classify than those in the UMP442 database. Consequently, the test set shows metric values higher than the UMP-IND set. The TabPFN classifier achieves outstanding results for the entire test set and given its exceptional speed and largest training dataset to date, the model proves suitable for deployment in virtual screening frameworks in the food flavors industry.

3.3 Umami Potency Prediction Model

The potency prediction problem was reframed from a regression task to a binary classification task to enhance practical applicability and handle the skewed data distribution. The DNN model was trained on 239 records and evaluated on 60 hold-out records. To mitigate the class imbalance, SMOTE was applied for resampling the data. Table 3 shows the performance metrics of the optimal DNN architecture, determined through multiple experiments, for both training and test sets. We explored various configurations for the DNN and identified that the reported model exhibited the lowest validation loss and minimal overfitting. The model demonstrates remarkable performance across various metrics for training and test sets, including accuracy, precision, recall, and F1-score. However, some degree of overfitting is apparent in its predictions. Given the

https://doi.org/10.26434/chemrxiv-2024-spz16 **ORCID:** https://orcid.org/0000-0002-1019-9418 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

inherently subjective nature of categorizing taste potency, where the two defined classes lack clear categorical differences due to adopting a sharp and arbitrary boundary, a certain overfitting is inevitable. Nevertheless, the model still exhibits high precision and recall values, close to 0.9 for the high potency class in the test set, making it well-suited for identifying potent tastants. Our work represents the first effort to quantify the umaminess of molecules and develop statistical tools for its prediction. The potency prediction problem could transition into a regression task as more experimental data becomes available.

Table 3: Performance metrics of the umami potency prediction model for the training and test sets

Metric	Training Set	Test Set
Accuracy	0.99	0.82
Precision	1.00	0.88
Recall	0.99	0.90
F1-score	0.99	0.89

The selection of ML models often involves a crucial tradeoff between predictive performance and interpretability or explainability. Neural networks, in particular, are known for their black-box approach, making it challenging to establish cause-and-effect relationships. This lack of understanding of the inner workings of models in human terms limits their trustworthiness in business and industrial settings. Significant research has been devoted to enhancing the explainability of ML models in recent years. In this study, we addressed the challenge by applying the Shapley additive explanations (SHAP) technique to explain the output of the deep neural network (DNN) model [25]. SHAP leverages concepts from cooperative game theory to allocate a contribution to each feature in an ML model, providing insights into how individual features contribute to the final prediction. In particular, we employed the Deep-Explainer method, tailored for neural networks. Figure 4 illustrates the average SHAP values for the entire test set based on

https://doi.org/10.26434/chemrxiv-2024-spz16 **ORCID:** https://orcid.org/0000-0002-1019-9418 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

the predictions of the DNN model. The features are ranked in the figure according to their relative importance, as determined by SHAP.

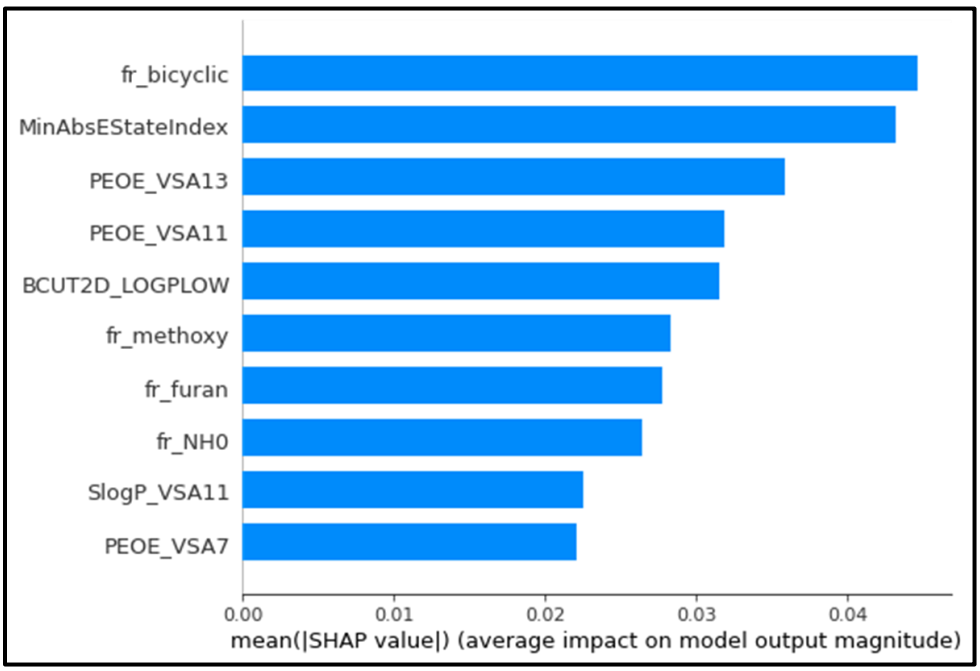

Figure 4: Bar plot illustrating the mean absolute SHAP values for the top ten features, ranked by their relative importance, derived from the predictions of the DNN model on the test set

We observe that electrostatic properties (PEOE_VSAx), hydrophobicity (SlogP_VSAx, BCUT2D_LOGPLOW), and electronic structure indices (EStateIndex), along with counts of fragments such as bicyclic rings, methoxy, and amine, strongly influence the predictions of the DNN model. These findings share some similarities with our understanding of the physics of receptor-ligand binding, which underlies the sensation of taste. However, caution should be exercised in drawing strong associations between the model's workings and binding physics. In the initial feature cleaning process, several features that could significantly influence the affinity of a tastant to the umami taste receptor might have been dropped due to strong correlations with other features. For instance, factors like molecule size, polarizability, and electro-topological

https://doi.org/10.26434/chemrxiv-2024-spz16 **ORCID:** https://orcid.org/0000-0002-1019-9418 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

properties, crucial in receptor-ligand binding, were absent in our cleaned dataset. Therefore, the SHAP values are intended to explain how the DNN model generates its predictions and should not be used to make definitive conclusions about binding physics. Incorporating explainability in complex ML models enhances their trustworthiness for industrial applications, and our work demonstrates the powerful combination of predictive accuracy and interpretability.

3.4 Computational Screening

The primary objective in constructing ML models for molecules is their deployment in practical applications to enhance screening efficiency and minimize costs. Based on the models reported in this study, we developed a computational pipeline for identifying potential novel umami tastants from molecule databases. The pipeline, incorporating similarity analysis, the umami classification model, the umami potency prediction model, and the ProTox-II toxicity tool, was utilized for virtual screening FooDB. The FooDB database is the most extensive and comprehensive repository of food constituents, encompassing about 70,000 food-related compounds. It has been selected as a representative application to demonstrate the pipeline's capabilities; nevertheless, any database can be screened using a similar approach. In addition to SMILES, alternative formats like SMARTS, FASTA, and INChI are also compatible with the pipeline, owing to the capability of RDKit to generate features from these representations.

Prior to applying the pipeline to the database, we eliminated a small number of invalid SMILES strings, resulting in 69,577 valid molecules. The first step in screening involved performing a similarity analysis to select molecules within the AD of the classification model, and only 367 molecules fell within our specified AD. This reaffirms the rarity of umami-tasting compounds, emphasizing the importance of exploring new tastants. The set of 28 features in the classification model was computed for the 367 compounds within AD. The TabPFN classifier identified 171 of

https://doi.org/10.26434/chemrxiv-2024-spz16 **ORCID:** https://orcid.org/0000-0002-1019-9418 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

those as umami. Subsequently, the 49 features for potency prediction were calculated for the 171 compounds obtained after classification. Ultimately, our predictions indicated that 153 molecules exhibited high umami potency, defined by having EC50 values lower than 2 μM for activating the human T1R1/T1R3 receptor.

Subsequently, we employed the ProTox-II server to forecast the LD50 values for the potential 153 umami tastants. We categorized them into one of the six classes (Class I to Class VI, arranged in descending order of toxicity). The model projected that 20 compounds (13%) fell into Class IV (harmful if swallowed), 109 compounds (71%) were assigned to Class V (may be harmful if swallowed), and 24 compounds (16%) were placed in Class VI (non-toxic). None of the molecules belonged to the fatal and toxic categories, denoted by Class I, II, and III. The predictions affirmed the dependability of the toxicity prediction model, given that FooDB encompasses food-related compounds expected to be non-orally toxic. It is important to note that ProTox-II was utilized here solely to showcase the capabilities of our screening pipeline, and alternative tools could be employed. Since toxicity prediction is not the primary focus of our study, and numerous in-silico models are already documented in the literature, we have refrained from developing a new model. The accuracy of the predictions presented here is contingent upon the model within ProTox-II, and it should be acknowledged that many compounds categorized as Class V may indeed be safe for consumption.

Figure 5 depicts the 24 molecules identified through the computational screening pipeline applied to the FooDB database. Notably, all compounds featured nitrogenous groups, with many being dipeptides. A PubChem search revealed that one of the molecules is naturally found in *Vigna radiata* (mung bean/green gram) and *Saccharomyces cerevisiae* (brewer's yeast). This discovery is promising, given that pulses and yeast extracts are recognized for eliciting umami flavor.

https://doi.org/10.26434/chemrxiv-2024-spz16 **ORCID:** https://orcid.org/0000-0002-1019-9418 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Additionally, another molecule serves as a natural product present in *Arabidopsis thaliana* (thale cress), *Trypanosoma brucei* (parasitic kinetoplastid), and *Bos taurus* (cattle). It is important to emphasize that these findings illustrate the pipeline's capabilities, and the screened molecules should only be considered novel tastants with experimental validation. Similarly, the pipeline can be applied to any other molecular database for further exploration.

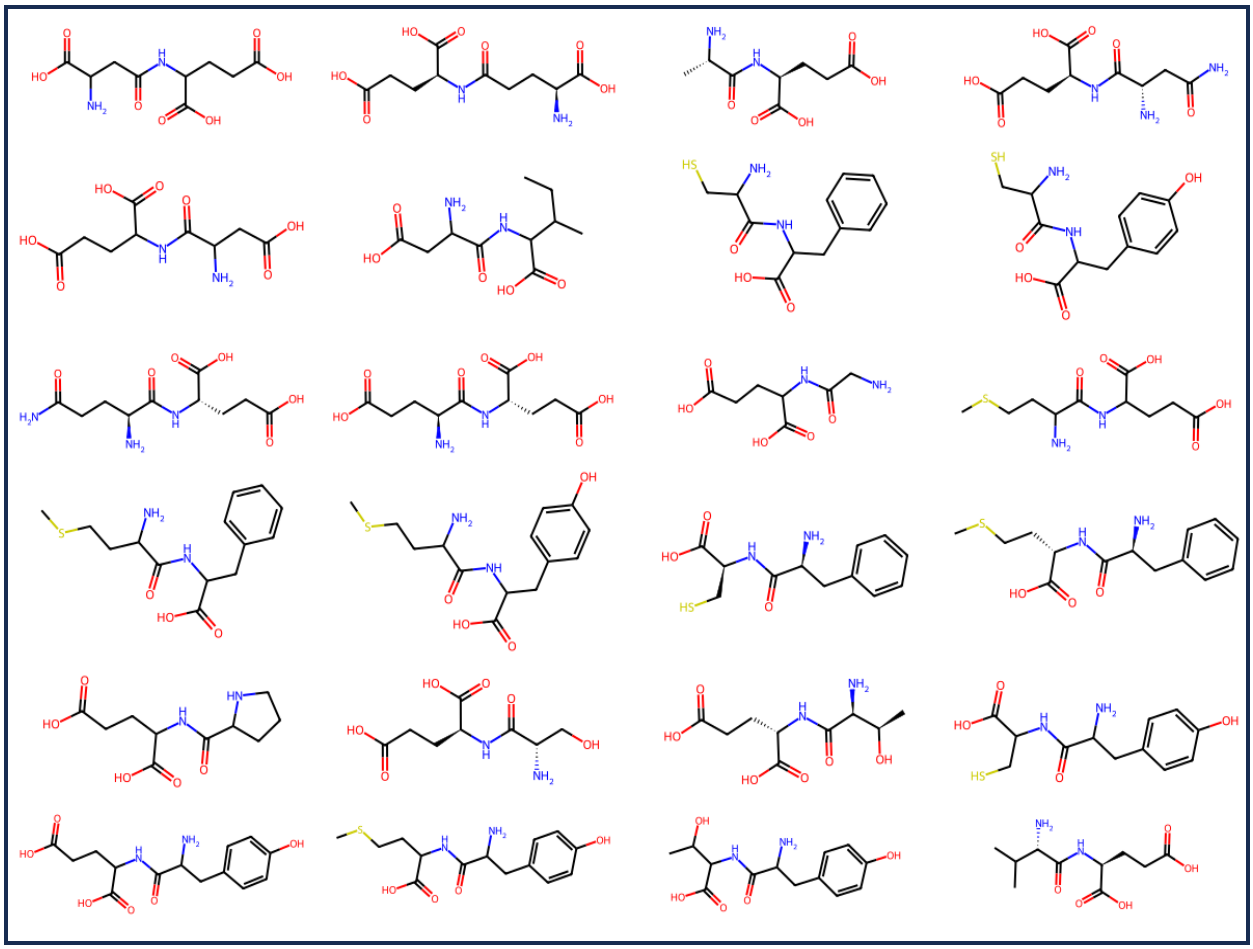

Figure 5: Potent and non-toxic umami tastants identified through the application of the computational pipeline on the FooDB database

https://doi.org/10.26434/chemrxiv-2024-spz16 **ORCID:** https://orcid.org/0000-0002-1019-9418 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

4 Conclusions

This paper introduces a data-driven approach to discovering novel umami tastants. We curated the most comprehensive umami classification dataset and employed a transformer-based architecture for distinguishing between umami and non-umami molecules. We extensively discuss the dataset's characteristics, delving into the distributions of key chemical properties and the prevalence of functional groups. Furthermore, we trained a neural network model to predict the potency of umami compounds, defined by the EC50 value for activating human umami taste receptors. The SHAP method was applied to provide insights into the network's predictions. Leveraging these two deep learning models, we devised an end-to-end computational screening pipeline incorporating similarity analysis and toxicity estimation, facilitating the identification of novel umami tastants from databases. We tested our pipeline by applying it to the FooDB database. Compared to previous data-based works on umami, our contribution stands out for its broader scope, superior performance, and better suitability for industrial applications.

Despite the progress achieved in this study, the dataset for umami tastants remains notably limited when compared to the vast datasets in other fields, such as drug discovery, catalysis, and adsorbent design. Consequently, the full potential of DL models can only be harnessed for real-world applications once we amass a more substantial amount of data. This presents a formidable challenge due to the rarity of umami-tasting molecules in nature, impeding the exploration of new compounds. Additionally, there needs to be a standardized, widely accepted scale for quantifying the umaminess of molecules complicates the search for novel tastants. In this study, we utilized potency as a measure, but the sparse data hinders highly accurate predictions. Moreover, the generalization of the potency prediction model across different data ranges remains to be

https://doi.org/10.26434/chemrxiv-2024-spz16 **ORCID:** https://orcid.org/0000-0002-1019-9418 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

determined. Therefore, meaningful progress in estimating tastants' umami intensity or potency necessitates further experimental work and consensus across the scientific community.

The virtual screening pipeline proposed in this study can be enhanced by incorporating molecular docking, molecular dynamics, and other binding energy calculation methods. These physics-based simulations offer a preliminary validation of the screened molecules before their synthesis or acquisition for laboratory and human panel testing. Our upcoming endeavors will focus on advancing toward *de novo* tastant design by combining statistical techniques with molecular dynamics simulations. This work establishes the groundwork for such an integrated computational framework, showcasing the effectiveness of DL as a formidable tool in the digital transformation of the food flavors industry.

Author Contributions

PD, KG, and RG conceived the project idea. PD and KG curated and analyzed the data. PD developed the deep learning models. All authors contributed to the interpretation and discussion of the results and preparation of this manuscript.

Acknowledgments

The authors thank Dr. Harrick Vin, Chief Technology Officer at Tata Consultancy Services, and Dr. Gautam Shroff, Head of Research at Tata Consultancy Services, for their constant encouragement and support during this project.

https://doi.org/10.26434/chemrxiv-2024-spz16 **ORCID:** https://orcid.org/0000-0002-1019-9418 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Conflict of Interest

The authors have no conflict of interest to declare.

Data Availability Statement

The primary unprocessed molecule data used for this work was curated from the UMP442 database (https://github.com/Shoombuatong/Dataset-Code/tree/master/iUmami) and a patent (https://patents.google.com/patent/CA2900181A1/en), both of which are publicly available online. Due to organizational policies, the authors cannot share the processed data and codes. However, the manuscript includes comprehensive details on data processing, model construction, and analysis, essential for reproducing the results.

References

[1] N. Chaudhari, H. Yang, C. Lamp, E. Delay, C. Cartford, T. Than and S. Roper, "The Taste of Monosodium Glutamate: Membrane Receptors in Taste Buds," *Journal of Neuroscience,* vol. 16, pp. 3817-3826, 1996.

[2] G. Nelson, J. Chandrashekar, M. A. Hoon, L. Feng, G. Zhao, N. J. P. Ryba and C. S. Zuker, "An amino-acid taste receptor," *Nature,* vol. 416, pp. 199-202, 2002.

[3] N. Chaudhari, Landin, A. Marie and S. D. Roper, "A metabotropic glutamate receptor variant functions as a taste receptor," *Nature Neuroscience,* vol. 3, pp. 113-119, 2000.

[4] A. S. Gabriel, T. Maekawa, H. Uneyama and K. Torii, "Metabotropic glutamate receptor type 1 in taste tissue," *The American Journal of Clinical Nutrition,* vol. 90, pp. 7435-7465, 2009.

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

[6] F. Zhang, B. Klebansky, R. M. Fine, H. Xu, A. Pronin, H. Liu, C. Tachdjian and X. Li, "Molecular mechanism for the umami taste synergism," *Proceedings of the National Academy of Sciences,* vol. 105, pp. 20930-20934, 2008.

https://doi.org/10.26434/chemrxiv-2024-spz16 **ORCID:** https://orcid.org/0000-0002-1019-9418 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

[7] J. Diepeveen, Moerdijk-Poortvliet, C. W. Tanja and F. R. Van Der Leij, "Molecular insights into human taste perception and umami," *Journal of Food Science,* vol. 87, pp. 1449-1465, 2022.

[8] W. Wang, X. Zhou and Y. Liu, "Characterization and evaluation of umami taste: A review," *Trends in Analytical Chemistry,* vol. 127, p. 115876, 2020.

[9] 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.

[10] 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.

[11] 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, p. 23450, 2016.

[12] 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.

[13] 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, p. 7155, 2019.

[14] G. Maroni, L. Pallante, G. Di Benedetto, M. A. Deriu, D. Piga and G. Grasso, "Informed classification of sweeteners/bitterants compounds via explainable machine learning," *Current Research in Food Science,* vol. 5, pp. 2270-2280, 2022.

[15] W. Bo, D. Qin, X. Zheng, Y. Wang, B. Ding, Y. Li and G. Liang, "Prediction of bitterant and sweetener using structure-taste relationship models based on an artificial neural network," *Food Research International,* vol. 153, p. 110974, 2022.

[16] P. Charoenkwan, J. Yana, C. Nantasenamat, M. M. Hasan and W. Shoombuatong, "iUmami-SCM: A Novel Sequence-Based Predictor for Prediction and Analysis of Umami Peptides Using a Scoring Card Method with Propensity Score of Dipeptides," *Journal of Chemical Information and Modeling,* vol. 60, pp. 6666-6678, 2020.

[17] P. Charoenkwan, C. Nantasenamat, M. M. Hasan, M. A. Moni, B. Manavalan and W. Shoombuatong, "UMPred-FRL: A New Approach for Accurate Prediction of Umami Peptides Using Feature Representation Learning," *International Journal of Molecular Sciences,* vol. 22, p. 13124, 2021.

https://doi.org/10.26434/chemrxiv-2024-spz16 **ORCID:** https://orcid.org/0000-0002-1019-9418 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

[18] L. Pallante, A. Korfiati, L. Androutsos, F. Stojceski, A. Bompotas, I. Giannikos, C. Raftopoulos, M. Malavolta, G. Grasso, S. Mavroudi, A. Kalogeras, V. Martos and D. Amoroso, "Toward a general and interpretable umami taste predictor using a multi-objective machine learning approach," *Scientific Reports,* vol. 12, p. 21735, 2022.

[19] P. Dutta, D. Jain, R. Gupta and B. Rai, "Classification of tastants: A deep learning based approach," *Molecular Informatics,* vol. 42, no. 12, 2023.

[20] C. Tachdjian, A. P. Patron, S. L. Adamski-Werner, F. Bakir, Q. Chen, V. Darmohusodo, S. T. Hobson, X. Li, M. Qi, D. H. Rogers, M. Rinnova, G. Servant, X.-Q. Tang, M. Zoller and D. Wallace, "Novel flavors, flavor modifiers, tastants, taste enhancers, umami or sweet tastants, and/or enhancers and use thereof". Cananda Patent CA2900181A1, 12 May 2005.

[21] D. M. Lowe, P. T. Corbett, P. Murray-Rust and R. C. Glen, "Chemical Name to Structure: OPSIN, an Open Source Solution," *Journal of Chemical Information and Modeling,* vol. 51, pp. 739-753, 2011.

[22] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Aritificial Intelligence Research,* vol. 16, pp. 321-357, 2002.

[23] N. Hollmann, S. Müller, K. Eggensperger and F. Hutter, "TabPFN: A Tranformer that solves small Tabular Classification Problems in a second," in *International Conference on Learning Representations*, 2023.

[24] P. Banerjee, A. O. Eckert, A. K. Schrey and R. Preissner, "ProTox-II: a webserver for the prediction of toxicity of chemicals," *Nucleic Acids Research,* vol. 46, pp. 257-263, 2018.

[25] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predicitions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

https://doi.org/10.26434/chemrxiv-2024-spz16 **ORCID:** https://orcid.org/0000-0002-1019-9418 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0