

VirtuousMultiTaste: Predicting Multiple Taste Sensations with a Multi-objective ML-driven Method

Lampros Androutsos^{1†}, Lorenzo Pallante^{2†}, Agorakis Bompotas³, Filip Stojceski⁴, Gianvito Grasso⁴, Dario Piga⁴, Giacomo di Benedetto⁵, Christos Alexakos³, Athanasios Kalogeras³, Konstantinos Theofilatos^{1*}, Marco A. Deriu² and Seferina Mavroudi^{1,6}

¹ InSyBio PC, Patras, 265 04, Greece

² Polito^{BIO}MedLab, Department of Mechanical and Aerospace Engineering, Politecnico di Torino, Torino, 10129, Italy

³ Industrial Systems Institute, Athena Research Center, 265 04 Patras, Greece.

⁴ Department of Innovative Technologies, Dalle Molle Institute for Artificial Intelligence, Lugano-Viganello, 6962, Switzerland

⁵ Enginlife Engineering Solutions, Turin, Italy

⁶ Department of Nursing, University of Patras, 265 04 Patras, Greece.

* k.theofilatos@insybio.com

† Joint First Authors

Abstract

Taste perception plays a pivotal role in guiding nutrient intake and aiding in the avoidance of potentially harmful substances through five basic tastes - sweet, bitter, umami, salty, and sour. Taste perception originates from molecular interactions in the oral cavity between taste receptors and chemical tastants. Hence, the recognition of taste receptors and the subsequent perception of taste heavily rely on the physicochemical properties of food ingredients. In recent years, several advances have been made towards the development of machine learning-based algorithms to classify chemical compounds' tastes using their molecular structures. Despite the great efforts, there remains significant room for improvement by developing multi-class models to predict the entire spectrum of basic tastes. Here, we present a multi-class predictor aimed at distinguishing three different tastes, i.e., bitter, sweet, and umami, from other taste sensations. The developed model has been integrated into a publicly accessible web platform. This work lays the groundwork for a comprehensive understanding of the molecular features that drive the perception of tastes, paving the way towards new methodologies in the rational design of foods, such as the pre-determination of specific tastes, the engineering of complementary diets to traditional pharmacological treatments, and many others.

Introduction

Taste and smell play a pivotal role in the chemosensory perception of food since they are fundamental determinants for the food selection and intake process ¹. Biochemical compounds derived from food ingestion trigger the taste perception process through the binding with specific proteins known as taste receptors, located on the tongue's taste buds and dedicated to the recognition of the five basic tastes: sweet, bitter, sour, salty, and umami ^{2,3}. Sweet taste is commonly associated with energy-rich food, to help identify sources of sugars and carbohydrates ⁴. Conversely, bitter taste is normally recognized as an unpleasant flavour and acts as a warning against potentially dangerous compounds ⁵. Sour taste both helps to detect spoiled food and identify the presence of biologically relevant vitamins ⁶. Salty taste is crucial to monitor the uptake of essential electrolytes, which play a central role in maintaining body osmosis ⁷. Finally, umami taste relates to the protein content in food, through the recognition of amino acids and oligopeptides ⁸. Therefore, each taste is associated with critical biological functions and nutritional needs that are important to preserve health status. In this sense, the taste of chemical compounds present in food stimulates an increase in nutrient intake while helping to avoid potentially harmful substances ⁹. Indeed, nutritious foods usually have an appetitive taste, e.g., sweet, umami, or a low concentration of salts and acids. Instead, toxic substances usually present a repulsive flavour, such as bitter tastants, strong sour taste stimuli, and high concentrations of salts ¹⁰. In general, taste sensation relies on the affinity of specific biochemical compounds and their target taste receptors ^{11,12}. Small variations in tastants' chemistry may result in a drastic change in perceived taste. Therefore, shedding light on the physio-chemical features of food ingredients is of primary importance to pinpoint the molecular bases and mechanisms determining the food taste and subsequent food consumption.

In recent years, several studies have developed machine-learning (ML) tools to predict the taste of specific compounds starting from their chemical structure ¹³. In literature, there is a net prevalence of ML tools for predicting sweet and bitter tastes e.g., BitterX ¹⁴, BitterPredict ¹⁵, e-Bitter ¹⁶, iBitter-SCM ¹⁷, BERT4Bitter ¹⁸, iBitter-Fuse ¹⁹, a QSTR-based approach ²⁰, e-Sweet ²¹, Predisweet ²², BoostSweet ²³, BitterSweetForest ²⁴, BitterSweet ²⁵, bitter-sweet classifier by Bo et al. ²⁶, VirtuousSweetBitter ²⁷. Five examples of umami taste predictors are instead present in recent literature, namely, iUmami-SCM ²⁸, UMPred-FRL ²⁹, VirtuousUmami ³⁰, Umami-MRNN ³¹, and Umami-BERT ³². From a technical point of view, several ML algorithms are used for taste prediction, among which Multiple Linear Regression (MLR) and Support Vector Machine (SVM) were the first used models for binary classification ^{13,33}. Thanks to recent scientific progress, these models were outclassed by tree-based models, such as Random Forest (RF) or AdaBoost (AB), and Neural Network (NN) ^{34,35}. Within this framework, the latter algorithms also support multi-class classification and work well in the non-linear classification domain, if the selected database is large enough. Multi-class and multi-labelling techniques have been employed in several applications for the food and agricultural industries, but there are still limited

applications for classifying compounds by taste and predicting the relative taste intensity^{36–39}. Recently, a multi-class classification method based on learning vector quantization NN to classify tea samples of five commercial brands has been proposed⁴⁰. Monforte et al. presented an orthogonal partial least square discriminant analysis (OPLS-DA) and RF-combined multi-class pipeline for the discrimination of white wine ageing based on target oxidation markers⁴¹. Moreover, an SVM multiclass classification demonstrated high efficiency in the classification of 7 different types of raw food⁴². Given the aforementioned context, there is a notable lack of research focused on the simultaneous prediction of multiple tastes, especially in real-world scenarios where foods frequently exhibit a complex blend of tastes.

This lack of tools for predicting multiple tastes represents a significant limitation in the field of food science and technology, specifically in the formulation and optimization of food products. In this study, we address these gaps by developing a machine learning-based tool that predicts not one, but four different tastes, and by focusing on the underlying physicochemical properties that contribute to these tastes. In the present work, we aimed at developing a multi-class taste predictor, named VirtuousMultiTaste, able to distinguish between bitter, sweet, and umami from other taste sensations. We employed a hybrid combination of heuristic optimization and nonlinear machine learning classification methods. Building upon our previous work³⁰, this new four-taste predictor was trained and tested using similar ensemble dimensionality reduction and classification techniques. This approach effectively reduced the number of physicochemical features and identified the important ones for predicting the four different tastes. The simplicity of the model reduces the likelihood of overfitting and makes it more user-friendly through a web interface (<https://virtuous.isi.gr/#/virtuous-multitaste>), expanding the potential audience of users. VirtuousMultiTaste is the first machine learning-based web tool that predicts four different tastes and opens the possibility of analysing different compounds and understanding the chemical-physical factors that contribute to the overall taste perception.

Results

Dimensionality reduction

As outlined in the Methods section, the statistical analysis to reduce the number of employed molecular descriptors was performed on the training set with the Kruskal–Wallis test⁴³ since not all features followed normal distribution when tested with the Shapiro–Wilk test⁴⁴. Moreover, the correction of p-values for multiple testing to get q-values was applied using the Benjamini–Hochberg FDR adjustment method⁴⁵. By setting the q-value threshold to 0.05, we identified 1306 statistically significant differentiated features. To further highlight the importance of the data dimensionality reduction process, Supplementary Fig. S2 presents a Principal Component Analysis (PCA) using all features and 1306 statistically significant differentiated features. Pairwise comparisons of each taste versus the rest were

further conducted using the Mann-Whitney test. Supplementary Fig. S3 presents a heatmap of the top 5 features differentiating each taste from the rest in our training dataset.

Model performance

The multi-objective evolutionary optimization algorithm employed in our study indicated that the RF method achieved better performance across multiple objectives compared to SVM. Therefore, we developed and evaluated 20 different RF models (see also Supplementary Table S4). Among these models, we selected the best model (RF model #2 in Supplementary Table S4) considering the balance between the achieved performance and the relatively small number of features (15). A summary of the model performance is reported in Table 1. During the training phase, we employed 10-fold cross-validation (CV) on the training set, presenting mean values and standard deviations for the performance metrics. The testing set comprises the 3377 left-out samples not used for training, consisting of 1577 bitter, 1544 sweet and 289 other compounds. The relative ROC curves are represented in Fig. 1A. In Table 2 and Table 3, a summary of the performance of each taste class is presented for the training set and the testing set, respectively. The relative ROC curves from the validation of the testing set are represented in Fig. 1B.

Table 1 | Summary of model performance with the selected RF model in the 10-fold cross-validation (CV) and test sets.

	ACC	F1	F2	Precision	Recall	AUC
CV	76.54%±1.0	76.58%±1.0	76.61%±1.01	76.92%±1.05	76.64%±1.0	0.92±0.02
Test	71.76%	74.32%	73.10%	78.98%	71.76%	0.87

Table 2 | Summary of the 10-fold cross-validation training performance for each class.

	ACC	F1	F2	Precision	Recall	AUC
Bitter	88.56%±0.81	79.09%±1.78	77.31%±3.30	82.42%±0.63	77.21%±4.19	0.92±0.016
Sweet	83.43±0.79	73.26%±0.85	75.59%±2.23	70.03%±1.30	75.65±2.81	0.90±0.01
Other	82.83%±0.43	66.62%±0.13	64.03%±1.27	67%±1.39	63.57%±1.71	0.86±0.10
Umami	95.99%±0.16	88.49%±0.51	88.64%±0.68	87.61%±0.67	87.40%±1.17	0.98±0.005

Table 3 | Summary of model performance using the final trained model for each taste class in the test set.

	ACC	F1	F2	Precision	Recall	AUC
Bitter	81.70%	79.01%	76.24%	84.10%	74.50%	0.89
Sweet	79.53%	75.38%	71.57%	82.73%	69.23%	0.86
Other	84.19%	42.96%	56.01%	30.94%	70.24%	0.86

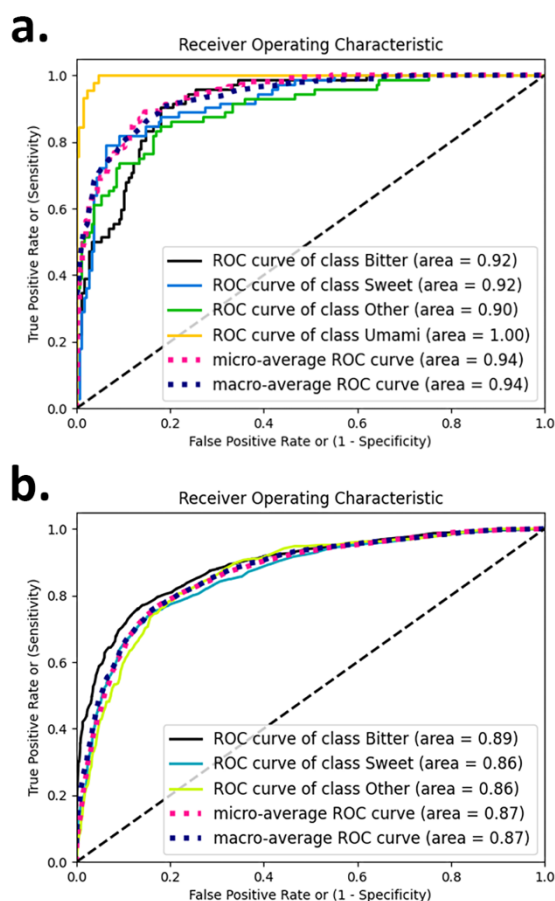


Fig. 1 | Receiver Operating Characteristic (ROC) Curves. ROCs of the four-taste classification for (a) the cross-validation set and (b) the testing set.

Feature Importance

Feature importance is a crucial aspect of ML models as it provides valuable insights into the contribution and relevance of different input features for making accurate predictions. The features selected during the model construction on which the predictions rely are 15 and include the molecular descriptors ATSC0c, ATSC0se, AATS0i, ATSC1p, AATSC2se, AATSC0m, AATSC1Z, AATSC2are, AATSC1pe, SpDiam_A, ATSC1c, ATSC1se, ATSC1Z, ATSC1m, and ATSC4s. The selected features are summarized in Fig. 2 reporting the level of importance according to the SHAP values. The distributions of the 15 features for the bitter, sweet, other and umami samples are also represented in Supplementary Fig. S1.

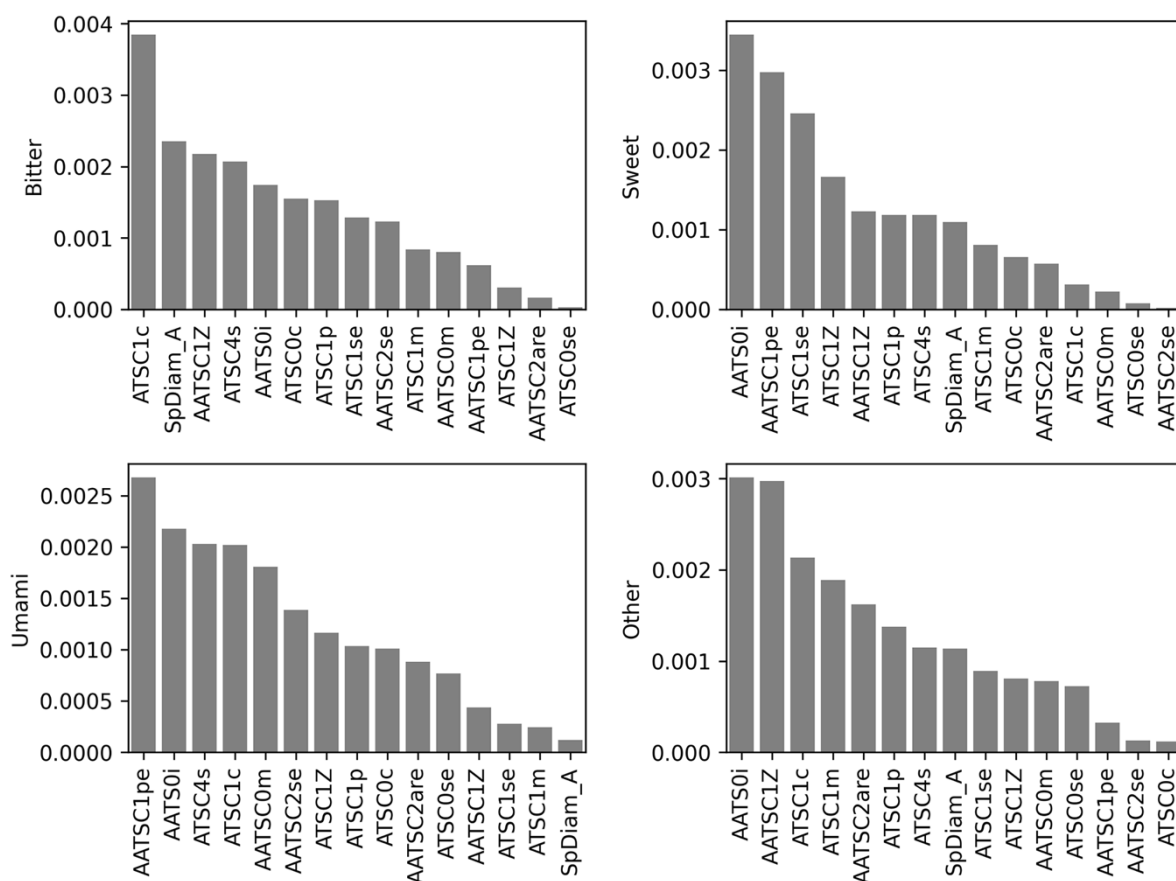


Fig. 2 | Feature importance using SHAP. Feature importance of the selected features using the average of the absolute SHAP values of each taste class and ranking features in order of importance.

In this connection, the correlation between the selected features holds great importance in understanding the underlying relationships and interactions within the taste prediction model. Fig. 3 represents the correlation between the 15 most important features selected by the model during training.

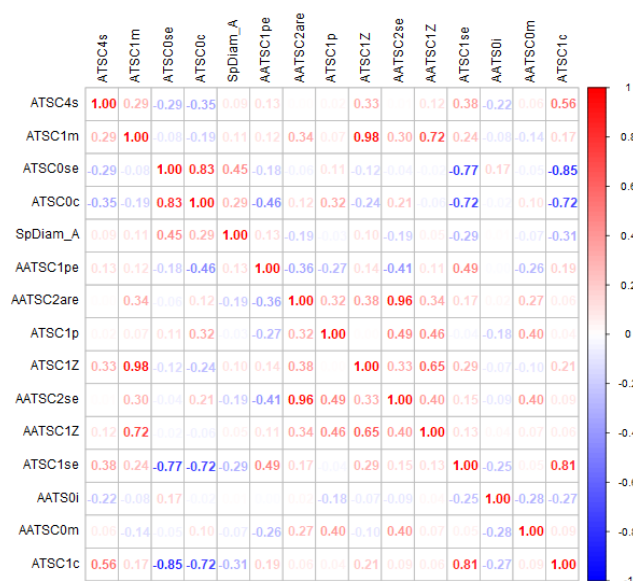


Fig. 3 | Correlation plot of the best 15 selected features. Each square represents the correlation between two features. The colour and value of each square indicate the magnitude of the correlation coefficient, with blue and red values indicating negative and positive correlations, respectively.

External Datasets

The developed model was tested to screen external datasets collecting food-related and natural compounds:

1. FoodDB (<https://foodb.ca/>) is the most extensive and comprehensive resource worldwide concerning food constituents, chemistry and biology, containing over 70000 compounds. After having removed missing SMILES, duplicate compounds, and molecules with structural errors, we ended up with 69309 molecules: 14693 were predicted as Bitter, 5375 as Sweet, 3149 as Umami and 46092 as Other.
2. FlavorDB (<https://cosylab.iiitd.edu.in/flavordb/>) collects information regarding flavour molecules. For the present work, we considered only 2599 molecules related to natural ligands from the dataset: 778 were predicted as Bitter, 1661 as Sweet, 29 as Umami and 131 as Other.
3. PhenolExplorer (<http://phenol-explorer.eu>) is a comprehensive database that compiles information on polyphenols found in foods. We considered only compounds having composition data (SMILES), resulting in 489 compounds: 365 were predicted as Bitter, 23 as Sweet, 9 as Umami and 92 as Other.
4. Natural Product Atlas (<https://www.npatlas.org/>) encompasses naturally occurring compounds derived from microorganisms, as documented in peer-reviewed primary scientific literature. We preserved 32491 compounds after curating the dataset with the ChEMBL structure pipeline: 26653 were predicted as Bitter, 2019 as Sweet, 1880 as Umami and 1939 as Other.

- PhytoHub (<https://phytohub.eu/>) is an openly accessible electronic database containing detailed information about dietary phytochemicals and their metabolites in humans and animals. We preserved 1746 compounds after having removed missing SMILES and compounds with issues in their molecular structures: 1213 were predicted as Bitter, 228 as Sweet, 62 as umami and 243 as Other.

The distributions of predicted tastes for each of the external datasets are represented in Fig. 4 (see also Supplementary Table S5 for detailed numbers).

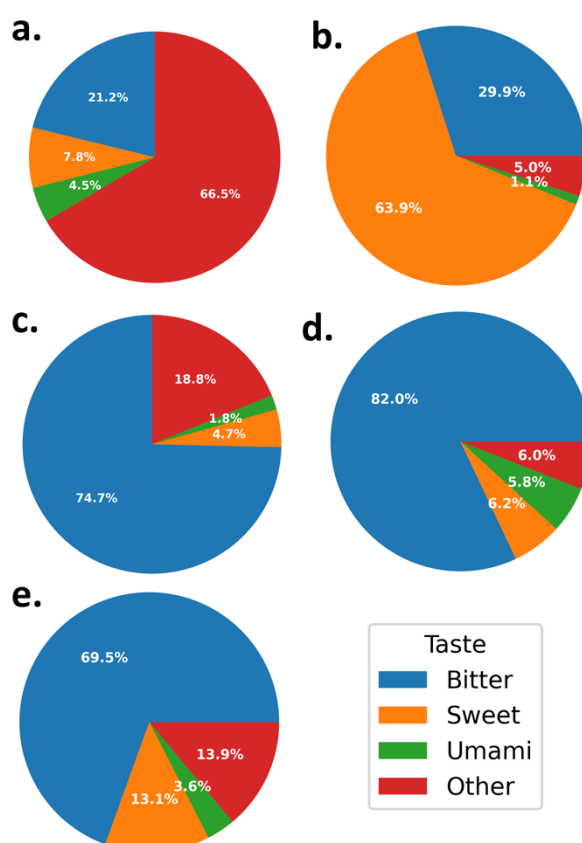


Fig. 4 | Distribution of predicted tastes for external databases. The pie charts represent the distributions of bitter, sweet umami and other tastes in the analysed external datasets, including (a) FooDB, (b) FlavorDB, (c) PhenolExplorer, (d) Natural Product Atlas, and (e) PhytoHub.

VirtuousMultiTaste Platform

The multi-taste predictor that was developed has been integrated into a web-based interface (<https://virtuous.isi.gr/#/virtuous-multitaste>). This platform serves as a user-friendly graphical interface for conducting analyses on chemical compounds expressed in different notations such as SMILES, FASTA, InChI, SMARTS, or PubChem compound names. When the user provides a PubChem name, the system searches the PubChem database for the corresponding compound and retrieves its SMILES representation to perform the multi-taste prediction. The platform design follows the separation of

concerns principle between the presentation layer (front-end) and the data access layer (back-end). The front-end is the visible part of the application that users interact with on their devices. It allows users to input compounds directly into a text field or upload a file that consists of several compounds, each one written on a separate line. Furthermore, the users may choose to specify the annotation format of their inputs or let the system recognize it automatically. Once the analysis is completed, the results are displayed in a table format, showing the queried compounds' SMILES, their 2D molecular representation, and the taste prediction result presented both in a textual format and as a spider chart. Moreover, next to each compound there are two buttons for downloading the calculated molecular descriptors or the best 15 descriptors on which the final model relies. The front-end development utilizes the Ionic framework, chosen for its wide range of user interface components suitable for both browsers and mobile devices. The back-end component is a cloud-based web service implemented using the Flask micro-framework, known for its lightweight yet robust nature. It receives input from the front-end, executes the VirtuousMultiTaste analyser, and sends the results back to the front-end. To facilitate this information exchange, a RESTful API is provided, which accepts and transmits data in the form of JavaScript Object Notation (JSON).

Discussion

Machine learning techniques have demonstrated their pivotal role in advancing the development of prediction tools and digital support systems across diverse fields, such as nutrition and agri-food research ⁴⁶⁻⁵². In our previous works related to the prediction of taste, a bitter/sweet predictor, named VirtuousBitterSweet ²⁷, and an umami predictor, named VirtuousUmami ³⁰, were created. Herein, we developed the first machine learning-driven taste predictor capable to identify bitter, sweet, umami from other taste sensations starting from the molecular structures of a query compound.

The data used to train the model was obtained from a variety of previously defined databases, including samples for each of the taste sensations under investigation. Two-dimensional Mordred molecular descriptors were used to extract features from the data. It is noteworthy that previous studies have achieved notable outcomes in taste prediction within the domain of 2D molecular descriptors alone ^{22,25,27,30}. This advancement marks a significant stride, as 2D molecular descriptors are computationally less demanding and remain unaffected by variations in three-dimensional molecular structures compared to 3D molecular descriptors. The extremely high number of molecular features (1613), especially in comparison to the number of compounds in the training dataset, was reduced using the Mann-Whitney statistical analysis, pinpointing 1307 statistically significant descriptors. The PCA analysis showed that statistically significant features can discriminate remarkably better between the four different tastes if compared to the analysis considering all the descriptors (see also Supplementary Fig. S2). The number of features used for the taste prediction has been further refined during the model

development: the employed heuristic multi-objective Pareto-based evolutionary optimization algorithm was able to select the Random Forest (RF) as the most appropriate classifier, to choose the optimal parameters being 95 Decision Trees and the most important and informative features on which the model relies. Ultimately, we ended up with only 15 features: this allows us to not only achieve better performance but also simplify the model and improve its explainability. Interestingly, the most frequent descriptor class among the 15 selected features utilizes the so-called Autocorrelation of a Topological Structure (ATS) which quantifies the spatial arrangement and distribution patterns of atoms or molecular fragments within a molecule. It calculates the correlation between the properties of a specific atom or fragment and the properties of other atoms or fragments within a defined distance in the molecule. This approach provides information about the local environment and structural features of a molecule. In particular, the autocorrelation descriptors were computed using the Moreau-Broto autocorrelation weighted by Sanderson electronegativity (ATSC0se, AATSC2se, ATSC1se), by mass (ATSC1m and AATSC0m), by Gasteiger charge (ATSC0c, ATSC1c), by atomic number (ATSC1Z, AATSC1Z), by ionization potential (AATS0i), by polarizability (ATSC1p), by Allred-Rocow electronegativity (AATSC2are), by Pauling electronegativity (AATSC1pe) and by intrinsic state (ATSC4s). Interestingly, the selected features are mostly associated with charge distribution, electronegativity, and polarizability, which are fundamental characteristics for the effective interaction of tastants with taste receptors capable of recognizing sweet, bitter, and umami tastes. Although there has been a significant reduction in the number of features, it remains challenging to intuitively elucidate the chemical and physical properties of tastants solely based on the 15 most important features. To enhance the model's explainability, future studies should prioritize the use of simpler descriptors or the development of specific methodologies to intuitively correlate the molecular descriptors to the relative structural features or functional groups. This will enable a clearer understanding of the underlying factors contributing to the prediction of multiple taste sensations.

The proposed machine learning model was benchmarked against commonly used machine learning methods and pipelines. RFs, SVMs and XGBoost were applied to the same dataset tuning their parameters with grid search and selecting their features using the minimum Redundancy Maximum Relevance (mRMR) method (see also Supplementary Table S6). VirtuousMultiTaste outperforms the other three classifiers (RF, XGBoost and SVM) across all performance metrics. Moreover, VirtuousMultiTaste has been compared with previous tools dedicated to the prediction of only the bitter and sweet taste, i.e., VirtuousSweetBitter²⁷ and BitterSweet²⁵. These two tools were selected because they were the only ones readily accessible and usable for a proper performance assessment. To have a fair comparison, we selected compounds not included in any of the training sets of the three taste predictors, resulting in a final external test of 869 compounds (409 bitter and 460 sweet). We evaluated the performance for predicting separately the bitter and the sweet taste, thus accessing the ability of each classifier to effectively detect bitter/non-bitter (Supplementary Table S7) or sweet/non-sweet

(Supplementary Table S8) molecules. VirtuousMultiTaste exhibited slightly superior performance in bitter taste prediction compared to the other tools, with all performance metrics (ACC, F1, F2, Precision, and Recall) hovering around 83%. In comparison, VirtuousSweetBitter and BitterSweet achieved performance levels of approximately 80% and 77% across all metrics, respectively (Supplementary Table S7). Regarding the prediction of sweet taste, VirtuousMultiTaste performed at intermediate values between VirtuousSweetBitter and BitterSweet. However, VirtuousMultiTaste maintained a satisfactory level of performance and showed commendable predictive capabilities (Supplementary Table S8). As for the comparison with tools for umami taste prediction, we were limited to comparing VirtuousMultiTaste with our previous tool, VirtuousUmami ³⁰, as all other previously developed predictors, such as iUmami-SCM ²⁸ and UMPred-FRL ²⁹, are based on the peptide sequence of the compound and therefore cannot be applied to more general molecular annotations, such as SMILES. It was not possible to create an independent test set for both VirtuousMultiTaste and VirtuousUmami, due to the limited number of proven umami compounds in the literature. Comparing the performance metrics in the relative cross-validation sets, VirtuousMultiTaste and VirtuousUmami achieved similar accuracy (around 96%) and AUC scores (above 96%) in the cross-validation set, whereas slightly lower values for F1, F2, Precision and Recall was achieved by the present tool (see also Supplementary Table S9). Moreover, testing the two tools with non-umami compounds not used for training, including sweet, bitter, and other compounds, VirtuousMultiTaste and VirtuousUmami both achieved an accuracy of over 99%. Certainly, it is important to note that this comparison only allows us to consider the accuracy in predicting the negative class, i.e., non-bitter, and therefore it represents a partial comparison. Furthermore, we tested three recently proven umami peptides, i.e., FLNQDEEAR (FR-9), FNKEE (FE-5), and EEFLK (EK-5) ⁵³, and all of them have been predicted as umami by VirtuousMultiTaste. In summary, VirtuousMultiTaste has demonstrated remarkable performance compared to previous taste-specific predictors, with the added advantage of being able to predict multiple taste sensations at the same time.

Regarding the range of applicability for the proposed model, we conducted an evaluation to assess whether the model's performance exhibited variations based on the similarity between the tested compounds and the chemicals used during the training phase. Accordingly to our previous works ^{27,30}, we evaluated the average similarity score between test and training compounds: (i) the Morgan Fingerprints (1024 bits, radius 2) were calculated using RDKit for all the compounds in the dataset; (ii) the Tanimoto similarity index was computed between each molecule in the test set and the previously-defined fingerprints; (iii) then the average similarity score was calculated by averaging the similarity scores of the 5 most similar couple of compounds. Compounds in the test set have been divided into 10 quartiles according to the average similarity score and the model performance has been evaluated on each quartile separately (see also Supplementary Fig. S4). Since the performance remained remarkably stable for each similarity quartile, we concluded that the model was able to preserve its performance

regardless of the similarity of the query compounds with the ones in the training set, thus ensuring a general applicability domain. This fact can also be attributed to the composition of the training database used for the model, which encompasses three of the five fundamental taste sensations and includes a fourth category, i.e. "other taste", including a wide spectrum of compounds with distinct tastes. As a result, the curated dataset might embrace a substantial portion of the chemical space that underlies the taste sensations of tastants. It is important to underscore that a primary limitation of the current study is the relatively limited number of umami compounds in the training dataset, particularly in comparison to the number of compounds representing other taste sensations. As previously noted in existing literature^{28–30}, augmenting the number of experimentally identified umami samples would not only enhance the predictive performance of the algorithm but also bolster its robustness.

VirtuousMultiTaste predictor also demonstrates its versatility by accommodating various types of molecular structure notations, such as SMILES, FASTA, InChI, SMARTS, or PubChem name. This capability enables the screening of diverse compound types, thereby creating opportunities to explore extensive molecular databases for the identification of tastes. In this context, we applied the VirtuousUmami predictor to five distinct external databases associated with food or natural compounds, namely FooDB, FlavorDB, PhenolExplorer, Natural Product Atlas, and PhytoHub, to pinpoint the distributions of tastes. The representation of umami taste is relatively lower compared to sweet or bitter tastes across the screened databases. This outcome aligns reasonably with expectations, as umami taste is closely associated with the protein content of food, which, in turn, is underrepresented within the selected external databases and in general if compared to sweet or bitter. In contrast, the bitter taste predominates in the PhenolExplorer, Natural Product Atlas, and PhytoHub databases. This finding aligns with the outcomes of previous machine learning-based classifiers, indicating that a remarkable proportion of natural compounds exhibit a bitter taste^{24,25,54}. Concerning FlavorDB, a substantial majority of the molecules were projected to possess sweet (63.9%) or bitter (29.9%) characteristics. It is noteworthy that the screening of FlavorDB yielded a distribution similar to that reported in previous literature, albeit with a considerably larger number of processed compounds (approximately 400 more)²⁵.

As a further example of applicability, this platform can be used also for food screening, enabling the analysis of the chemical composition of desired foods. Therefore, we examined coffee and chocolate using their relative composition as reported in FooDB. The taste analysis was conducted directly from our platform in the Foods tab (<https://virtuous.isi.gr/#/foods>), where we provided an interactive exploration of the FooDB and enabled the direct execution of the multi-taste prediction for each of the compounds present in the investigated foods. Regarding the screening of coffee, 130 compounds were predicted as bitter, 44 as sweet, 4 as umami and 14 as other tastes, thus pointing out that coffee is characterized by mainly bitter compounds. Chocolate also presented a predominance of bitter

compounds, with a taste profile consisting of 96 bitter compounds, 33 sweet compounds, 4 umami compounds and 13 other taste compounds. Both food screenings seem in accordance with the primary taste perceptions for these foods, which are indeed bitter. This type of analysis could be further refined in future works using models able to consider not only the chemical composition of foods but also their concentrations and their potency to move towards in silico sensory analysis.

The creation of a user-friendly web interface (accessible at <https://virtuous.isi.gr/#/virtuous-multitaste>) was driven by the objective of ensuring the accessibility of the multi-taste prediction model to users who may lack experience or familiarity with technical coding. Additionally, the web interface is complemented by a corresponding GitHub repository (<https://github.com/lorenzopallante/VirtuousMultiTaste>) that provides access to the technical Python codes for those interested in further exploration and customization.

In conclusion, a novel machine learning-based taste predictor, named VirtuousMultiTaste, was developed for identifying bitter, sweet, and umami from other taste sensations by combining heuristic optimization and nonlinear machine learning methods. VirtuousMultiTaste represents a paramount tool for rapidly screening compound databases to identify a diverse array of potential candidate compounds with anticipated taste properties. In a broader sense, it is worth mentioning that the future perspectives beyond the present work also include the possibility of effectively predicting the remaining two taste sensations, i.e., sour and salty, to obtain a singular and comprehensive taste predictor capable of predicting all five fundamental tastes at once. This research lays the foundations for future works aimed at developing specific models capable of predicting the sensory profile of foods and engineering novel products with desired tastes or properties, potentially impacting various fields, such as nutrition, precision medicine, the food market, and beyond.

Methods

Database and data curation

Data was mined from publicly available databases to construct a dataset with compounds of nine different classes (sweet, bitter, non-sweet, umami, tasteless, sour, salty, multitaste and other tastes). In detail, we utilized the sweet/bitter database curated for our previous VirtuousSweetBitter classifier²⁷, umami and non-umami samples from the UMP442 database²⁸ used for our VirtuousUmami classifier³⁰, as well as, compounds with miscellaneous tastes from the ChemTastesDB⁵⁵. A summary of the collected compounds from these taste databases is reported in the Supplementary Information (Table S1). After removing compounds labelled as Multitaste, since they were not enough to justify a multi-labelling approach, we used only three classes (sweet, bitter and umami) and all other classes were

considered as Other because of relatively low sample size. Therefore, this work is focused on the prediction of four classes i.e., Sweet, Bitter, Umami and Other.

Regarding the Bitter and Sweet compounds, a total of 5290 compounds (2741 sweet and 2549 bitter) with their SMILES description were used as the starting database. All the SMILES were checked using the RDKit library⁵⁶, searching for the relative correct SMILES in the PubChem database (if necessary) and removing duplicates, as well as compounds with incorrect SMILES.

Umami compounds were gathered from the UMP442 dataset and the ChemTastesDB. The UMP442 dataset collects 140 umami molecules from previous literature^{57,58} and the BIOPEP-UWM database⁵⁹, whereas the ChemTastesDB contains 87 Umami peptides. These peptides were collected using their amino acid sequences and converted into their SMILES representation using the RDKit package, as done in previous work³⁰.

Regarding the samples assigned to the “Other” class, which encompasses compounds that are neither bitter, sweet, nor umami, a total of 208 compounds from previous literature were collected^{60–62}. Additionally, 203 tasteless compounds and a further 370 taste compounds which were neither bitter, sweet, nor umami were gathered from the ChemTasteDB. Therefore, we ended up with 649 compounds labelled as Other.

In summary, our initial database was composed of 6166 compounds - 2741 sweet, 2549 bitter, 227 umami and 649 other - all with their SMILES representation. A schematic summary of the initial database is reported in Supplementary Table S2.

Following pre-processing protocols used in previous literature^{25,27,30}, all SMILES were processed with the ChEMBL Structure Pipeline⁶³. This step is aimed to identify and address any potential issues in the molecular structures, as well as standardize the SMILES representation across the entire dataset. In detail, the protocol runs a molecule checker on the compound structure, standardizes chemical structures and generates the parent molecule representation based on a set of predefined rules. Incorrect SMILES and duplicates were then removed obtaining a final dataset of 4717 compounds (1904 sweet, 1937 bitter, 227 umami and 649 other).

From the original dataset, a random subset was used for training including 360 sweet, 360 bitter, 360 umami and 360 other chemicals. The umami class was oversampled with additional 133 umami samples using the Adaptive Boosting (AdaBoost) algorithm, which will be described in detail in the Model construction and performance evaluation section. The remaining 1544 sweet, 1577 bitter and 289 other peptides were used for external testing to examine the generalization properties of the trained models.

A table summarising the division between the training and the test set is reported in the Supplementary Data (Table S3).

Molecular descriptors and dimensionality reduction

Following our previous works^{27,30}, the molecular features calculation was achieved using the 1613 2D Mordred descriptors⁶⁴ for each compound. The comprehensive compilation of employed Mordred descriptors can be accessed at <https://mordred-descriptor.github.io/documentation/master/descriptors.html>. The 2D Mordred descriptors furnish valuable insights into compounds, encompassing essential molecular information such as molecular weight, the count of distinct atom types, bond types, hybridization degree, spectral diameter, detour index, count of hydrogen donors and acceptors, molecular distance edge between distinct atom types, the polarizability of atoms and bonds, as well as the topological polar surface. Furthermore, these descriptors encompass additional features derived from symbolic representations, including the Zagreb index, adjacency matrix descriptors, Moreau–Mroto descriptors, Moran coefficients, Geary coefficients, and descriptors delineating the Burden matrix and Barysz matrix⁶⁴. Furthermore, the dataset underwent pre-processing steps to prepare it for input into the ML model. Features with a high percentage of missing values (>30%) were discarded, while the remaining missing values were imputed using the kNN-impute method with k=20⁶⁵. Subsequently, the data were arithmetically normalized in the range of [0-1]. Data pre-processing, statistical analysis, and the generation of additional plots, such as ROC curves and bean plots, were performed using the InSyBio Biomarkers Tool⁶⁶.

Model construction and performance evaluation

Based on our previous work³⁰, we used a hybrid approach of heuristic optimization and nonlinear machine learning to develop classification models. Specifically, we used an ensemble dimensionality reduction technique that employed a heuristic multi-objective Pareto-based evolutionary optimization algorithm to (a) identify the optimal feature subset to use as input to the classifiers, (b) select the most appropriate classifier among SVM and RF, and (c) select the optimal parameters for the classifier (e.g., C and gamma for SVM, the number of trees for RF). By utilizing the multi-objective Pareto-based approach, we sought to optimize prediction performance, minimize the selected features, and simplify the classification model. The weights used for the optimization objectives were Selected Features Number Minimization 1, Accuracy (ACC) 10, F1 score 10, F2 score 1, Precision (PRC) 1, Recall (REC) 10, ROC-AUC (AUC) 1, Number of SVs or Trees Minimization 1, and Manhattan Distance 1. These weights allowed us to effectively address the imbalanced nature of our classification problem. To get a better handle on the unbalanced nature of the multi-class taste prediction problem and improve the accuracy of the prediction models, we additionally used the Adaptive Boosting (AdaBoost) algorithm as an additional pre-processing step before the training⁶⁷. This algorithm performs boosting, assigning weights over the internal training set in the cross-validation iterations. This algorithm assigns higher

weights to the minority class and then generates synthetic copies of the minority class samples until all the classes are balanced. This is especially helpful for the unbalanced class, i.e., the umami one. The outcome is the generation of multiple models that exhibit similar performance concerning the user-defined objectives, which correspond to the Pareto set of optimal solutions.

The evolutionary algorithm was applied to a population of 100 individuals, and the termination criteria was set to a maximum of 200 generations. Ten different runs were conducted to deal with the stochastic nature of the algorithm, and the results presented are the average performance of these runs. Convergence of the algorithm (average performance differing less than 5% from that of the best-performing individual) was found after 50 generations for each independent run. This finding affirmed that the selected maximum number of generations was appropriate for this problem. Additional parameters of the evolutionary algorithm were set to their default values as suggested by the InSyBio Biomarkers tool user manual (arithmetic crossover probability: 0, mutation probability: 0.01, two-point crossover probability: 0.9). Stratified 10-fold cross-validation was used to train and test the prediction models. Further details on the implementation of the trained models and a summary of the performance metrics used are available in the Supplementary Data.

The pipeline of the adopted methods is presented in Fig. 5.

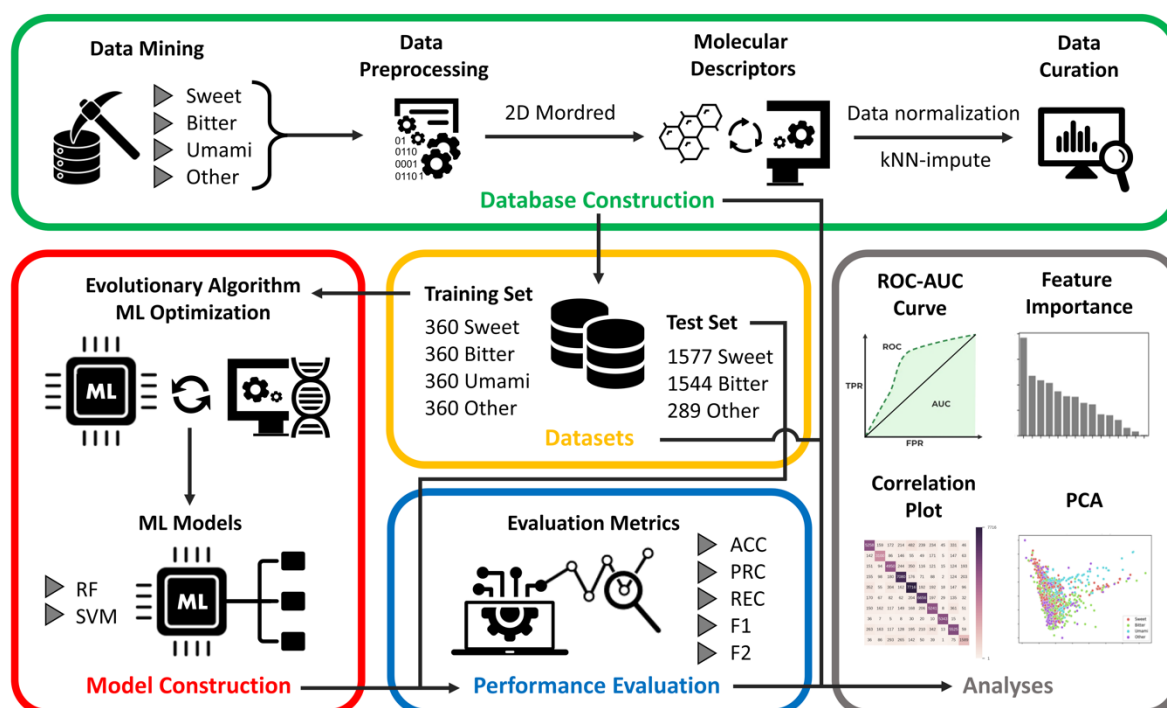


Fig. 5 | Flowchart of VirtuousMultiTaste. The flowchart represents the major stages in the workflow of the proposed taste prediction tool.

Model Explainability and Applicability

Model explainability was performed using SHAP (SHapley Additive exPlanations) ⁶⁸. SHAP values offer insights into the contribution of individual features to the model's final prediction, facilitating the interpretation of its decision-making process. To compute the SHAP values, we employ the TreeExplainer method to generate an explainer object for the Random Forest (RF) model, since RFs were selected over SVMs from the multi-objective optimization framework deployed. The explainer object enables the computation of the SHAP values for each feature in the dataset. Considering the nature of the classification problem, we obtain the SHAP values for each class, providing insight into the contribution of each feature to the probability of that class. The SHAP values analysis of each taste reveals the features with the highest importance for each taste and their contribution to the final prediction.

The developed model was also tested on external databases related to foods or natural products, including FooDB, (<https://foodb.ca/>), FlavorDB (<https://cosylab.iiitd.edu.in/flavordb/>), PhenolExplorer (<http://phenol-explorer.eu>), Natural Product Atlas (<https://www.npatlas.org/>), and PhytoHub (<https://phytohub.eu/>). Coherently with the pre-processing of the dataset used for training and testing, each external database was first checked for missing SMILES or data, standardised with the ChEMBL Structure Pipeline, and featurized using Mordred descriptors.

Data and code availability

A public GitHub repository is available at <https://github.com/lorenzopallante/VirtuousMultiTaste> containing all source code and data used for the present work.

References

1. Kershaw, J. C. & Mattes, R. D. Nutrition and taste and smell dysfunction. *World J. Otorhinolaryngol. - Head Neck Surg.* **4**, 3–10 (2018).
2. Pallante, L. *et al.* On the human taste perception: Molecular-level understanding empowered by computational methods. *Trends Food Sci. Technol.* **116**, 445–459 (2021).
3. Töle, J. C., Behrens, M. & Meyerhof, W. Taste receptor function. in *Handbook of Clinical Neurology* vol. 164 173–185 (2019).

4. Lim, J. & Pullicin, A. J. Oral carbohydrate sensing: Beyond sweet taste. *Physiol. Behav.* **202**, 14–25 (2019).
5. Behrens, M. & Meyerhof, W. Bitter taste receptors and human bitter taste perception. *Cell. Mol. Life Sci.* **63**, 1501–1509 (2006).
6. Neta, E. R. D. C., Johanningsmeier, S. D. & McFeeters, R. F. The chemistry and physiology of sour taste - A review. *J. Food Sci.* **72**, R33–R38 (2007).
7. Roper, S. D. The taste of table salt. *Pflugers Arch.* **467**, 457–463 (2015).
8. Wang, W., Zhou, X. & Liu, Y. Characterization and evaluation of umami taste: A review. *TrAC - Trends Anal. Chem.* **127**, 115876 (2020).
9. Breslin, P. A. S. An Evolutionary Perspective on Food and Human Taste. *Curr. Biol.* **23**, R409–R418 (2013).
10. Tan, S. Y. & Tucker, R. M. Sweet taste as a predictor of dietary intake: A systematic review. *Nutrients* **11**, 94 (2019).
11. Kinnamon, S. C. Taste receptor signalling - from tongues to lungs. *Acta Physiol.* **204**, 158–168 (2012).
12. Servant, G., Kenakin, T., Zhang, L., Williams, M. & Servant, N. The function and allosteric control of the human sweet taste receptor. in *Advances in Pharmacology* vol. 88 59–82 (2020).
13. Malavolta, M. *et al.* A survey on computational taste predictors. *Eur. Food Res. Technol.* **248**, 2215–2235 (2022).
14. Huang, W. *et al.* BitterX: A tool for understanding bitter taste in humans. *Sci. Rep.* **6**, 1–8 (2016).

15. Dagan-Wiener, A. *et al.* Bitter or not? BitterPredict, a tool for predicting taste from chemical structure. *Sci. Rep.* **7**, 12074 (2017).
16. Zheng, S. *et al.* e-Bitter: Bitterant prediction by the consensus voting from the machine-learning methods. *Front. Chem.* **6**, (2018).
17. Charoenkwan, P. *et al.* iBitter-SCM: Identification and characterization of bitter peptides using a scoring card method with propensity scores of dipeptides. *Genomics* **112**, 2813–2822 (2020).
18. Charoenkwan, P., Nantasenamat, C., Hasan, M. M., Manavalan, B. & Shoombuatong, W. BERT4Bitter: a bidirectional encoder representations from transformers (BERT)-based model for improving the prediction of bitter peptides. *Bioinformatics* (2021) doi:10.1093/bioinformatics/btab133.
19. Charoenkwan, P. *et al.* iBitter-Fuse: A Novel Sequence-Based Bitter Peptide Predictor by Fusing Multi-View Features. *Int. J. Mol. Sci.* **22**, 8958 (2021).
20. Rojas, C. *et al.* A QSTR-based expert system to predict sweetness of molecules. *Front. Chem.* **5**, 1–12 (2017).
21. Zheng, S., Chang, W., Xu, W., Xu, Y. & Lin, F. e-Sweet: A machine-learning based platform for the prediction of sweetener and its relative sweetness. *Front. Chem.* **7**, (2019).
22. Bouysset, C., Belloir, C., Antonczak, S., Briand, L. & Fiorucci, S. Novel scaffold of natural compound eliciting sweet taste revealed by machine learning. *Food Chem.* **324**, 126864 (2020).
23. Lee, J., Song, B. S., Chung, Y. K., Jang, J. H. & Huh, J. BoostSweet: Learning molecular perceptual representations of sweeteners. *Food Chem.* **383**, 132435 (2022).
24. Banerjee, P. & Preissner, R. BitterSweetForest: A Random Forest Based Binary Classifier to Predict Bitterness and Sweetness of Chemical Compounds. *Front. Chem.* **6**, 93 (2018).

25. Tuwani, R., Wadhwa, S. & Bagler, G. BitterSweet: Building machine learning models for predicting the bitter and sweet taste of small molecules. *Sci. Rep.* **9**, 7155 (2019).
26. Bo, W. *et al.* Prediction of bitterant and sweetener using structure-taste relationship models based on an artificial neural network. *Food Res. Int.* **153**, 110974 (2022).
27. Maroni, G. *et al.* Informed classification of sweeteners/bitterants compounds via explainable machine learning. *Curr. Res. Food Sci.* **5**, 2270–2280 (2022).
28. Charoenkwan, P., Yana, J., Nantasenamat, C., Hasan, M. M. & Shoombuatong, W. IUmami-SCM: A Novel Sequence-Based Predictor for Prediction and Analysis of Umami Peptides Using a Scoring Card Method with Propensity Scores of Dipeptides. *J. Chem. Inf. Model.* **60**, 6666–6678 (2020).
29. Charoenkwan, P. *et al.* UMPred-FRL: A New Approach for Accurate Prediction of Umami Peptides Using Feature Representation Learning. *Int. J. Mol. Sci.* **22**, 13124 (2021).
30. Pallante, L. *et al.* Toward a general and interpretable umami taste predictor using a multi-objective machine learning approach. *Sci. Rep.* **12**, 21735 (2022).
31. Qi, L. *et al.* Umami-MRNN: Deep learning-based prediction of umami peptide using RNN and MLP. *Food Chem.* **405**, 134935 (2023).
32. Zhang, J. *et al.* Umami-BERT: An interpretable BERT-based model for umami peptides prediction. *Food Res. Int.* **172**, 113142 (2023).
33. Eberly, L. E. Multiple linear regression. in *Methods in molecular biology (Clifton, N.J.)* vol. 404 165–187 (2007).
34. Shanmugasundar, G. *et al.* A comparative study of linear, random forest and adaboost regressions for modeling non-traditional machining. *Processes* **9**, 2015 (2021).

35. Tiriyaki, S. & Aydin, A. An artificial neural network model for predicting compression strength of heat treated woods and comparison with a multiple linear regression model. *Constr. Build. Mater.* **62**, 102–108 (2014).
36. Bolaños, M., Ferrà, A. & Radeva, P. Food Ingredients Recognition Through Multi-label Learning. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* vol. 10590 LNCS 394–402 (2017).
37. Cappellin, L. *et al.* Multiclass methods in the analysis of metabolomic datasets: The example of raspberry cultivar volatile compounds detected by GC-MS and PTR-MS. *Food Res. Int.* **54**, 1313–1320 (2013).
38. Reda, A., Fakharany, E. & Hazman, M. Early prediction of wheat diseases using SVM multiclass. in *Advances in Intelligent Systems and Computing* vol. 639 257–269 (2018).
39. Saha, P., Ghorai, S., Tudu, B., Bandyopadhyay, R. & Bhattacharyya, N. Multi-class support vector machine for quality estimation of black tea using electronic nose. in *Proceedings of the International Conference on Sensing Technology, ICST* 571–576 (IEEE, 2012). doi:10.1109/ICSensT.2012.6461744.
40. Damarla, S. & Kundu, M. Classification of Tea Samples using Learning Vector Quantization Neural Network. in *Proceedings of 2020 IEEE Applied Signal Processing Conference, ASPCON 2020* 99–103 (IEEE, 2020). doi:10.1109/ASPCON49795.2020.9276662.
41. Monforte, A. R., Martins, S. I. F. S. & Silva Ferreira, A. C. Discrimination of white wine ageing based on untarget peak picking approach with multi-class target coupled with machine learning algorithms. *Food Chem.* **352**, 129288 (2021).
42. Tsakanikas, P., Karnavas, A., Panagou, E. Z. & Nychas, G. J. A machine learning workflow for raw food spectroscopic classification in a future industry. *Sci. Rep.* **10**, 11212 (2020).

43. Kruskal, W. H. & Wallis, W. A. Use of Ranks in One-Criterion Variance Analysis. *J. Am. Stat. Assoc.* **47**, 583–621 (1952).
44. Shapiro, S. & Wilk, M. An analysis of variance test for normality (complete samp l es) t. in (2007).
45. Ferreira, J. A. & Zwinderman, A. H. On the Benjamini–Hochberg method. *Ann. Stat.* **34**, 1827–1849 (2006).
46. Hasan, M. M., Manavalan, B., Shoombuatong, W., Khatun, M. S. & Kurata, H. i6mA-Fuse: improved and robust prediction of DNA 6 mA sites in the Rosaceae genome by fusing multiple feature representation. *Plant Mol. Biol.* **103**, 225–234 (2020).
47. Hasan, M. M. *et al.* Meta-i6mA: an interspecies predictor for identifying DNA N 6-methyladenine sites of plant genomes by exploiting informative features in an integrative machine-learning framework. *Brief. Bioinform.* **22**, 1–16 (2021).
48. Mota-Merlo, M. & Martos, V. Use of machine learning to establish limits in the classification of hyperaccumulator plants growing on serpentine, gypsum and dolomite soils. *Mediterr. Bot.* **42**, e67609 (2021).
49. Michelucci, U., Sperti, M., Piga, D., Venturini, F. & Deriu, M. A. A Model-Agnostic Algorithm for Bayes Error Determination in Binary Classification. *Algorithms* **14**, 301 (2021).
50. Venturini, F. *et al.* Exploration of Spanish Olive Oil Quality with a Miniaturized Low-Cost Fluorescence Sensor and Machine Learning Techniques. *Foods* **10**, 1010 (2021).
51. Ahmad, A., Ordoñez, J., Cartujo, P. & Martos, V. Remotely Piloted Aircraft (RPA) in Agriculture: A Pursuit of Sustainability. *Agronomy* **11**, 7 (2020).
52. Martos, V., Ahmad, A., Cartujo, P. & Ordoñez, J. Ensuring Agricultural Sustainability through Remote Sensing in the Era of Agriculture 5.0. *Appl. Sci.* **11**, 5911 (2021).

53. Fu, B. *et al.* Three novel umami peptides derived from the alcohol extract of the Pacific oyster (*Crassostrea gigas*): identification, characterizations and interactions with T1R1/T1R3 taste receptors. *Food Sci. Hum. Wellness* **13**, 146–153 (2024).
54. Fritz, F., Preissner, R. & Banerjee, P. VirtualTaste: A web server for the prediction of organoleptic properties of chemical compounds. *Nucleic Acids Res.* **49**, W679–W684 (2021).
55. Rojas, C. *et al.* ChemTastesDB: A curated database of molecular tastants. *Food Chem. Mol. Sci.* **4**, 100090 (2022).
56. Landrum, G. RDKit Documentation. *Read. Writ.* **1**, 4 (2013).
57. Yu, Z. *et al.* Taste, umami-enhance effect and amino acid sequence of peptides separated from silkworm pupa hydrolysate. *Food Res. Int.* **108**, 144–150 (2018).
58. Zhang, J., Zhao, M., Su, G. & Lin, L. Identification and taste characteristics of novel umami and umami-enhancing peptides separated from peanut protein isolate hydrolysate by consecutive chromatography and UPLC–ESI–QTOF–MS/MS. *Food Chem.* **278**, 674–682 (2019).
59. Minkiewicz, P., Iwaniak, A. & Darewicz, M. BIOPEP-UWM database of bioactive peptides: Current opportunities. *Int. J. Mol. Sci.* **20**, 5978 (2019).
60. Burdock, G. A. *Fenaroli's Handbook of Flavor Ingredients*. *Fenaroli's Handbook of Flavor Ingredients* (CRC Press, 2004). doi:10.1201/9781420037876.
61. Rojas, C. *et al.* Quantitative structure–activity relationships to predict sweet and non-sweet tastes. *Theor. Chem. Acc.* **135**, 1–13 (2016).
62. ToxNet.
63. Bento, A. P. *et al.* An open source chemical structure curation pipeline using RDKit. *J. Cheminformatics* **12**, 51 (2020).

64. Moriwaki, H., Tian, Y. S., Kawashita, N. & Takagi, T. Mordred: A molecular descriptor calculator. *J. Cheminformatics* **10**, 4 (2018).
65. Zhang, S. Nearest neighbor selection for iteratively kNN imputation. *J. Syst. Softw.* **85**, 2541–2552 (2012).
66. InSyBio - Biomarkers.
67. CAO, Y., MIAO, Q.-G., LIU, J.-C. & GAO, L. Advance and Prospects of AdaBoost Algorithm. *Acta Autom. Sin.* **39**, 745–758 (2013).
68. Lundberg, S. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. (2017) doi:10.48550/ARXIV.1705.07874.

Acknowledgements

The present work has been developed as part of the VIRTUOUS project (<https://www.virtuoussh2020.com/>), funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie-RISE (Grant Agreement No 872181).

Author Contributions

Conceptualization: L.A., L.P., M.A.D., K.T., S.M.; Model development: L.A., L.P.; Data Curation: L.P., A.K., L.A., F.S.; Web Interface Development: A.B., C.A.; Supervision: G.G., A.K., C.A., G.d.B, D.P., M.A.D., K.T., S.M.; Coordination: S.M., K.T. and M.A.D.; Project Administration: K.T. and M.A.D.; Funding Acquisition: K.T. and M.A.D. All authors wrote the paper and critically commented on the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.