

Article

A Machine Learning Method to Identify Umami Peptide Sequences by Using Multiplicative LSTM Embedded Features

Jici Jiang ¹, Jiayu Li ², Junxian Li ¹, Hongdi Pei ^{1,3}, Mingxin Li ¹, Quan Zou ^{4,5,*}  and Zhibin Lv ^{1,*} ¹ College of Biomedical Engineering, Sichuan University, Chengdu 610065, China² College of Life Science, Sichuan University, Chengdu 610065, China³ Wu Yuzhang Honors College, Sichuan University, Chengdu 610065, China⁴ Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610054, China⁵ Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou 324000, China

* Correspondence: zouquan@nclab.net (Q.Z.); lvzhibin@pku.edu.cn (Z.L.)

Abstract: Umami peptides enhance the umami taste of food and have good food processing properties, nutritional value, and numerous potential applications. Wet testing for the identification of umami peptides is a time-consuming and expensive process. Here, we report the iUmami-DRLF that uses a logistic regression (LR) method solely based on the deep learning pre-trained neural network feature extraction method, unified representation (UniRep based on multiplicative LSTM), for feature extraction from the peptide sequences. The findings demonstrate that deep learning representation learning significantly enhanced the capability of models in identifying umami peptides and predictive precision solely based on peptide sequence information. The newly validated taste sequences were also used to test the iUmami-DRLF and other predictors, and the result indicates that the iUmami-DRLF has better robustness and accuracy and remains valid at higher probability thresholds. The iUmami-DRLF method can aid further studies on enhancing the umami flavor of food for satisfying the need for an umami-flavored diet.

Keywords: umami peptide; deep representation learning; SMOTE; ANOVA; light gradient boosting; mutual information; multiplicative LSTM



Citation: Jiang, J.; Li, J.; Li, J.; Pei, H.; Li, M.; Zou, Q.; Lv, Z. A Machine Learning Method to Identify Umami Peptide Sequences by Using Multiplicative LSTM Embedded Features. *Foods* **2023**, *12*, 1498. <https://doi.org/10.3390/foods12071498>

Academic Editor: Christophe Flahaut

Received: 26 February 2023

Revised: 24 March 2023

Accepted: 30 March 2023

Published: 2 April 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Umami taste has been widely accepted as the fifth basic taste, along with the four other basic tastes of sweet, sour, salty, and bitter [1]. Umami substances are important for enhancing the flavor of food and healthy eating [2]. Umami peptides frequently contain aspartic acid, glutamic acid, asparagine, or glutamine residues. Peptides that contain these umami amino acids may or may not have an umami flavor, and may instead have a bitter flavor in the absence of umami amino acids [3]. Umami peptides are a novel class of umami agents with numerous potential uses and a distinctive flavor. Additionally, these peptides act synergistically with other umami compounds to enhance the sweetness of sweet items and the saltiness of salty items, but reduce sour and bitter tastes, thus softening the taste. Umami is a very important factor affecting the quality of food, and increasing the content of umami substances in the food improves the overall palatability.

Wet tests are costly and time-consuming when identifying umami peptides. The post-genomic era's proliferation of peptide sequence databases [4,5] has had a significant impact on the practical application of automated mathematical methods for the discovery of novel umami peptides. The development of umami peptide prediction tools using deep representation learning features has attracted increasing interest in the field of bioinformatics [6–8]. Umami-SCM [9] was developed in 2020 and uses the scoring card method (SCM). It is combined with the propensity score of amino acids and dipeptides for identifying umami

peptides [10]. The independent test accuracy of this method was reported to be 0.865 and the predictor performed better in 10-fold cross-validation tests than in existing methods. Charoenkwan et al. developed UMPred-FRL in 2021 [11], which integrated seven different traditional feature codes for constructing the umami peptide classifier. Jiang et al. proposed iUP-BERT in 2022 [12], which is based on the use of a single deep representational learning feature encoding method (BERT: bidirectional encoder representations from transformer). Compared to Umami-SCM and UMPred-FRL, iUP-BERT has superior performance in both independent testing and cross-validation. Despite notable advancements in the field, particularly in terms of independent testing, machine learning (ML)-based umami peptide detection algorithms that rely exclusively on sequence data still need to significantly improve in terms of performance. The above predictors are still not accurate enough, and there is still room for improvement. For instance, we found that iUP-BERT was not as robust as expected.

Representation learning [13,14] comprises ML techniques that enable the automatic identification of representations from raw data for feature detection or classification. This eliminates the need for manual feature engineering and enables machines to learn the features of protein or peptide sequences, and apply them to perform specific tasks. During depth evaluation in representation learning, ML techniques are used for transforming data from the original representation to a new representation that preserves the information necessary for the object of interest [15–22], while discarding the redundant information [23–31]. Sequence-based deep representation learning has been recognized as an innovative and efficient construction in protein and peptide research for protein feature prediction [32–39], including the unified representation (UniRep) method [40] and BiLSTM [41].

In this study, sequence-based unified representation (UniRep) features based on multiplicative LSTM were solely used for developing an ML-based model, iUmami-DRLF, for the identification of umami peptides. iUmami-DRLF showed exceptional outcomes in the independent tests and 10-fold cross-validation studies. The obtained results had high accuracy, and more importantly, the results of independent testing proved iUmami-DRLF to be far superior to the current techniques and conventional non-deep representation learning techniques. Additionally, iUmami-DRLF has a wider range of applications and excellent umami peptide discrimination potential. The iUmami-DRLF predictor outperformed the conventional forecasting techniques in the 10-fold cross-validation tests ($S_n = 0.959$ and $auROC = 0.957$) and independent tests ($ACC = 0.921$, $MCC = 0.815$, $S_n = 0.821$, $S_p = 0.967$, $auROC = 0.956$, and $BACC = 0.894$). The independent test accuracy of iUmami-DRLF is improved by 2.45% as compared with that of iUP-BERT. The effects of various feature analysis methods and various deep representation learning features for classification results were examined using the unified manifold approximation and projection (UMAP) dimensionality reduction approach. Compared with other SOTA methods, iUmami-DRLF in this study has higher accuracy under various probability thresholds, and shows better robustness and generalization performance. The steps performed for the construction of iUmami-DRLF are depicted in Figure 1.

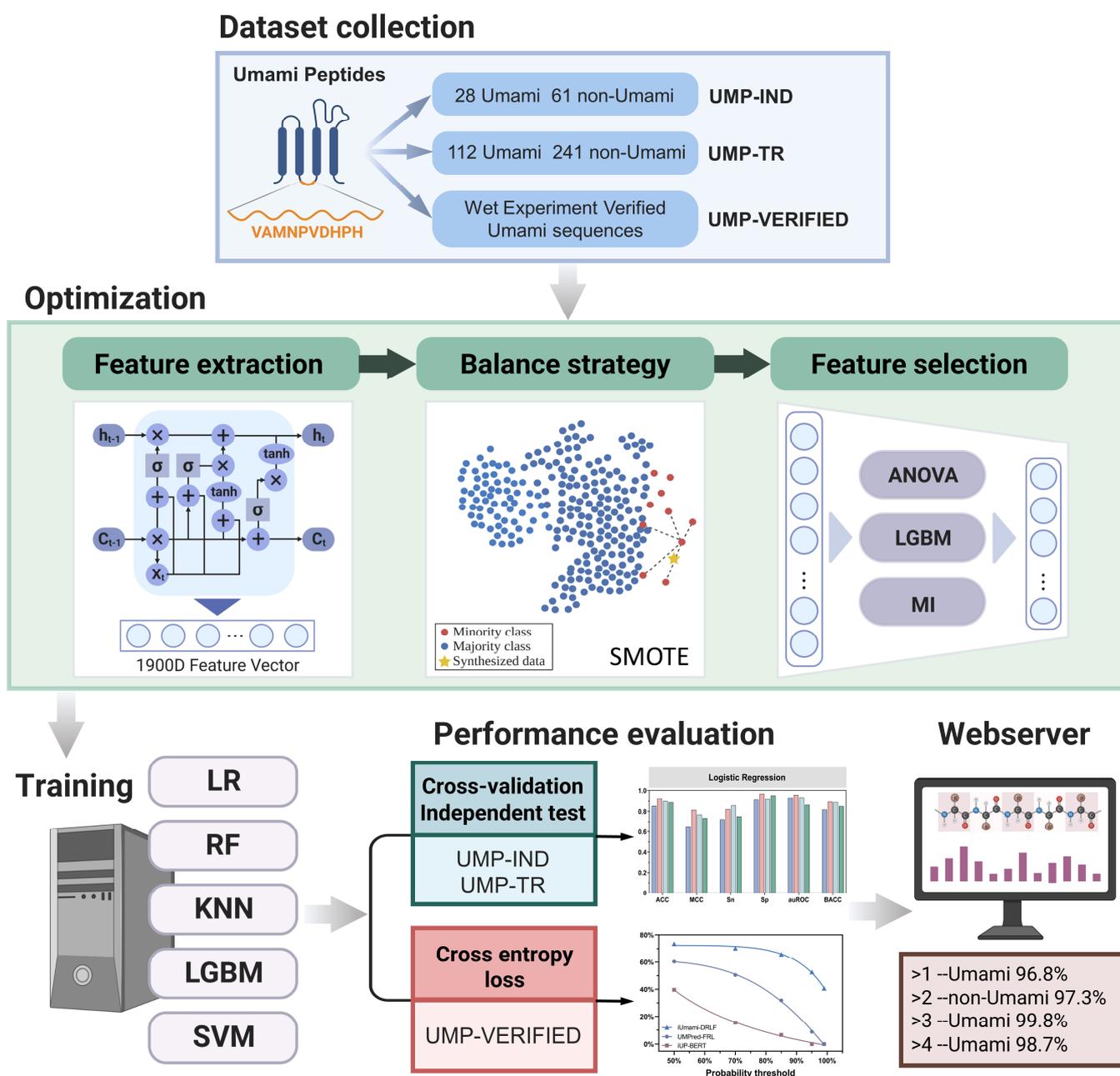


Figure 1. Overview of model development. The pre-trained UniRep sequence embedding model was used to embed the peptide sequences into eigenvectors. The peptide sequences were converted into 1900-dimensional (D) UniRep eigenvectors. The synthetic minority over-sampling technique (SMOTE) was used for balancing the imbalanced data. These features were used as inputs to the k-nearest neighbors (KNN), logistic regression (LR), support vector machine (SVM), random forest (RF), and light gradient boosting machine (LGBM) predictor algorithms. Feature extraction was performed for model optimization using analysis of variance (ANOVA), LGBM, and mutual information (MI). The selected feature sets were subjected to another round of analysis using the three feature extraction algorithms and various hyperparameters. The final optimized model was developed by comparison of model performance in 10-fold cross-validation and independent tests. Based on the 91 wet-test validated umami peptide sequences reported in the latest research (UMP-VERIFIED), we evaluated iUmami-DRLF in comparison to state-of-the-art methods.

2. Materials and Methods

2.1. Benchmark Dataset

In this work, the model was developed using the updated benchmark dataset from iUmami-SCM [9], which also facilitates future comparisons. The BIOPEP-UWM [4] database and experimentally verified umami peptides were included in the positive dataset, while bitter non-umami peptides were included in the negative dataset. The UMP442 benchmark dataset, which contains 304 non-umami peptides and 140 umami peptides, is acquired after data cleaning. To avoid the prediction model becoming overfit, the dataset was arbitrarily split into a training subset UMP-TR, and an independent test peptide subset denoted as UMP-IND. The UMP-TR dataset comprised 112 umami and 241 non-umami peptides, while the UMP-IND dataset comprised 28 umami and 61 non-umami peptides. The URL for both datasets is <http://public.aibiochem.net/peptides/iUmami-DRLF/> (accessed on 1 April 2023). To validate the accuracy and robustness of our model, we also collected 91 wet-experiment verified umami peptide sequences from the latest research (please see Supplementary Table S2). The 91 wet-experiment verified umami peptides dataset was named UMP-VERIFIED.

2.2. Feature Extraction

For UniRep [40], a total of 24 million core amino acid sequences from UniRef50 were used for training the UniRep model. By identifying the subsequent amino acid by reducing cross-entropy losses, the model learns how to accurately express proteins after training. Using the trained model, the input sequence was represented as a single fixed-length vector (hidden state) with a multiplicative long short-term memory (mLSTM) encoder. The ideal ML model was trained using the output vector representation. Supervised learning is achieved in various bioinformatics tasks by using the input sequence as a personality.

First, a matrix containing the sequences of S amino acid residues was integrated using the single thermal code $R^{S \times 10}$. The matrix was then put through into the mLSTM encoder to generate an output hidden state of $R^{1900 \times S}$ as an embedding matrix. The 1900-dimensional (D) UniRep feature vector was finally derived using an average pooling operation.

The equations used by the mLSTM encoder for performing the calculations are provided hereafter (Equations (1)–(7)). Where m_t represents the current intermediate multiplication state, \hat{h}_t is the input before the hidden state, f_t represents the forgotten gate, i_t represents the input gate, o_t stands the output gate, h_t stands the hidden state for output, and C_t is the current unit state.

$$m_t = (X_t W_{xm}) \otimes (W_{hm} h_{t-1}) \quad (1)$$

$$\hat{h}_t = (W_{mh} m_t + W_{xh} X_t) \times \tanh \quad (2)$$

$$f_t = \sigma(X_t W_{xf} + m_t W_{mf}) \quad (3)$$

$$i_t = \sigma(X_t W_{xi} + m_t W_{mi}) \quad (4)$$

$$o_t = \sigma(X_t W_{xo} + m_t W_{mo}) \quad (5)$$

$$C_t = f_t \otimes C_{t-1} + i_t \otimes \hat{h}_t \quad (6)$$

$$h_t = o_t \otimes \tanh(C_t) \quad (7)$$

In this example, \otimes stands for element-by-element multiplication, X_t represents the current input, h_{t-1} remains for the previous hidden state, C_{t-1} represents the previous unit state, σ stands for a sigmoid function, and \tanh represents a tangent function.

2.3. Balancing Strategy

Classifiers were built from unbalanced datasets using the synthetic minority over-sampling technique (SMOTE) [33] methodology. SMOTE is an improved method based on the random oversampling algorithm [42], and primarily combines the analysis of minority class samples, the location of nearby samples, and the creation of artificially created new samples in accordance with the minority class samples. SMOTE first identifies the neighboring samples for all minority class samples using the k-nearest neighbors (KNN) algorithm, and then uses linear random interpolation for realizing sample synthesis. A random interpolation position is selected among the samples, and an equal number of interpolations are considered for each sample point. Such a balancing strategy for achieving data balance not only increases the sample size but also improves sample quality. Classifiers can learn more distinct features after processing with SMOTE, which significantly improves the performance of classifiers.

2.4. Feature Selection Strategy

We used three feature selection techniques, namely, analysis of variance (ANOVA) [43,44], light gradient boosting machine (LGBM) [6,45], and mutual information (MI) [46], for selecting the retrieved features. These techniques were employed in this study for determining the best feature space, and ranking the features based on their relevant ratings. The features with importance values larger than a crucial threshold (average feature importance value) were selected after sorting the features from the “largest” to the “smallest” based on the importance values.

2.4.1. Analysis of Variance (ANOVA)

In this study, the features were sorted in order of importance using the ANOVA score. The mean difference between groups can be efficiently evaluated using ANOVA, which computes the ratio of variance within groups to the variance between groups for each feature [47]. The following formula was used for determining the ANOVA score:

$$S(t) = \frac{S_{\theta}^2(t)}{S_{\omega}^2(t)} \quad (8)$$

where $S(t)$ represents the score of the feature t , $S_{\theta}^2(t)$ stands for the variance between groups, and $S_{\omega}^2(t)$ is the variance within groups. The formulae used for calculating $S_{\theta}^2(t)$ and $S_{\omega}^2(t)$ are as follows:

$$S_{\theta}^2(t) = \frac{1}{K-1} \sum_{i=1}^K m_i \left(\frac{\sum_{j=1}^{m_i} f_t(i, j)}{m_i} - \frac{\sum_{i=1}^K \sum_{j=1}^{m_i} f_t(i, j)}{\sum_{i=1}^K m_i} \right)^2 \quad (9)$$

$$S_{\omega}^2(t) = \frac{1}{N-K} \sum_{i=1}^K \sum_{j=1}^{m_i} \left(f_t(i, j) - \frac{\sum_{j=1}^{m_i} f_t(i, j)}{m_i} \right)^2 \quad (10)$$

where K denotes the quantity of groups, N denotes the entire quantity of instances, and $f_t(i, j)$ denotes the value of the j -th sample in the i -th group of the feature t .

2.4.2. Lighting Gradient Boosting Machine (LGBM)

LGBM [33] is a quick, dispersed, strong gradient boosting framework based on a decision tree technique that is employed in numerous ML applications, including classification and ranking. The gradient boosting decision tree (GBDT), which has the ability to learn the performances of learners, is continuously improving with several computational

iterations. Here, we define $h_c(x)$ as an estimated function in Equation (11) and evaluate the loss function in Equation (12):

$$h_c(x) = \operatorname{argmin}_{h \in H} \sum L(y, F_{c-1}(x) + h(x)) \quad (11)$$

$$r_{ti} = - \frac{\partial L(y, F_{t-1}(x_i))}{\partial F_{t-1}(x_i)} \quad (12)$$

where c means the current iteration, and $F_{c-n}(x)$ means the last n iterations' model achievement. The following formula is used to select the most potential features in the current iteration, and the importance of each feature is obtained by ranking.

$$F_{c+n}(x) = h_{2n}(x) + F_{c-n}(x) \quad (13)$$

2.4.3. Mutual Information (MI)

MI has been widely used for feature selection since its development [48]. The advantage of MI in feature selection lies in its ability to equivalently define multidimensional variables and detect nonlinear relationships between variables. Owing to these advantages, the MI method can fully consider the joint correlation and redundancy of features during feature selection [49].

The entropy estimate for the peptide sequence S is provided in Equation (14):

$$H(S) = - \sum_{i \in \Sigma U} P(\varepsilon_i) \log P(\varepsilon_i) \quad (14)$$

Using this entropy equation, the equation for the MI peptide sequence was deduced as:

$$MI = \sum_{i \in \Sigma U} \sum_{j \in \Sigma U} P(\varepsilon_i, \varepsilon_j) \log \frac{P(\varepsilon_i, \varepsilon_j)}{P(\varepsilon_i)P(\varepsilon_j)} \quad (15)$$

where ΣU is the alphabet of amino acid residues and $P(\varepsilon_i)$ is the marginal probability of residue i .

2.5. Machine Learning Methods

Five widely used high-performance ML methods were used in this study, namely, KNN, linear regression (LR), support vector machine (SVM), random forest (RF), and LGBM [50].

KNN is one of the most straightforward machine learning algorithms that is better suited for automatic class categorization in studies with high sample sizes. Data are said to belong to a class if the minority of the K most comparable data in the feature space, or the feature space's closest neighbors, also do. This approach only selects the class of the data to be based mainly on the classification of the data or the data nearest to it.

LR is categorized as a supervised learning method in ML. The concept of LR is that if data obey a certain distribution, then the parameters are estimated by maximum likelihood estimation. This method is actually a classification model and is often used for binary and multi-class classification problems. It is widely used owing to its simplicity, parallelizability, and strong interpretability.

SVM is applied for solving binary classification problems in bioinformatics.

RF is a bagging-based technique that uses random feature selection during node splitting in addition to sampling at random.

LGBM is a gradient boosting framework that employs methods for learning from trees.

2.6. Evaluation Metrics and Methods

Five widely used measures were used for evaluating the performance of the models, and were calculated using Equations (16)–(20):

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{FP} + \text{FN} + \text{TP} + \text{TN}} \quad (16)$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \quad (17)$$

$$\text{Sn} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (18)$$

$$\text{Sp} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (19)$$

$$\text{BACC} = \frac{\text{Sn} + \text{Sp}}{2} \quad (20)$$

where TP denotes the amount of umami peptides successfully identified as umami, and TN denotes the quantity of non-peptides successfully identified as non-umami. FP denotes the amount of non-umami peptides falsely identified as umami, while FN denotes the quantity of umami peptides incorrectly identified as non-umami. The developed models were also contrasted with one another and with previously stated models based on the receiver operating characteristic curve (ROC). The area under the ROC curve (auROC) was also used for evaluating the predictive performance, where the values of auROC ranging between 0.5 and 1 stand for random and perfect models, respectively. The BACC approach is used for describing data imbalances, and the values of ACC and BACC are equal in a balanced sample.

K-fold cross-validation and independent testing methods are commonly used to evaluate ML models [51]. The raw data are separated into k-folds in K-fold cross-validation. The remaining $K - 1$ subsets are utilized as training sets, while one subset is used for model validation. In the validation set, K models are evaluated separately, and the final values of the evaluation measures are averaged to obtain the cross-validated values. In this investigation, we employed the 10-fold ($K = 10$) cross-validation approach. The samples used in stand-alone testing were fresh for the trained model, and the test dataset used was completely different from the training set.

2.7. Cross-Entropy Loss

When performing a binary classification task, there are only positive and negative examples, and their probabilities add up to 1. Therefore, we simply need to predict a probability rather than a vector.

The loss function is defined simply as follows:

$$\text{Loss} = -(y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y})) \quad (21)$$

where y is the sample label, which takes the value of 1 if the sample is a positive case and 0 otherwise, and \hat{y} is the probability that the model predicts that the sample is a positive case. In general, the lower the value of the cross-entropy loss function, the higher the classification effect [52–55].

3. Results and Discussion

3.1. Effect of SMOTE

We first extracted a 1900-dimensional feature vector using UniRep. The model was developed and initially trained using five different ML techniques, namely, KNN, LR, SVM, LGBM, and RF, for investigating the effect of SMOTE on the automatic identification of

umami peptides. The outcomes of independent testing and 10-fold cross-validation of the five ML models optimized with SMOTE and five ML models optimized without SMOTE were obtained based on the aforementioned hypotheses, and are depicted in Figure 2 and Supplementary Table S1. The values in the tables and figures indicate model performance measures following the optimization of model parameters.

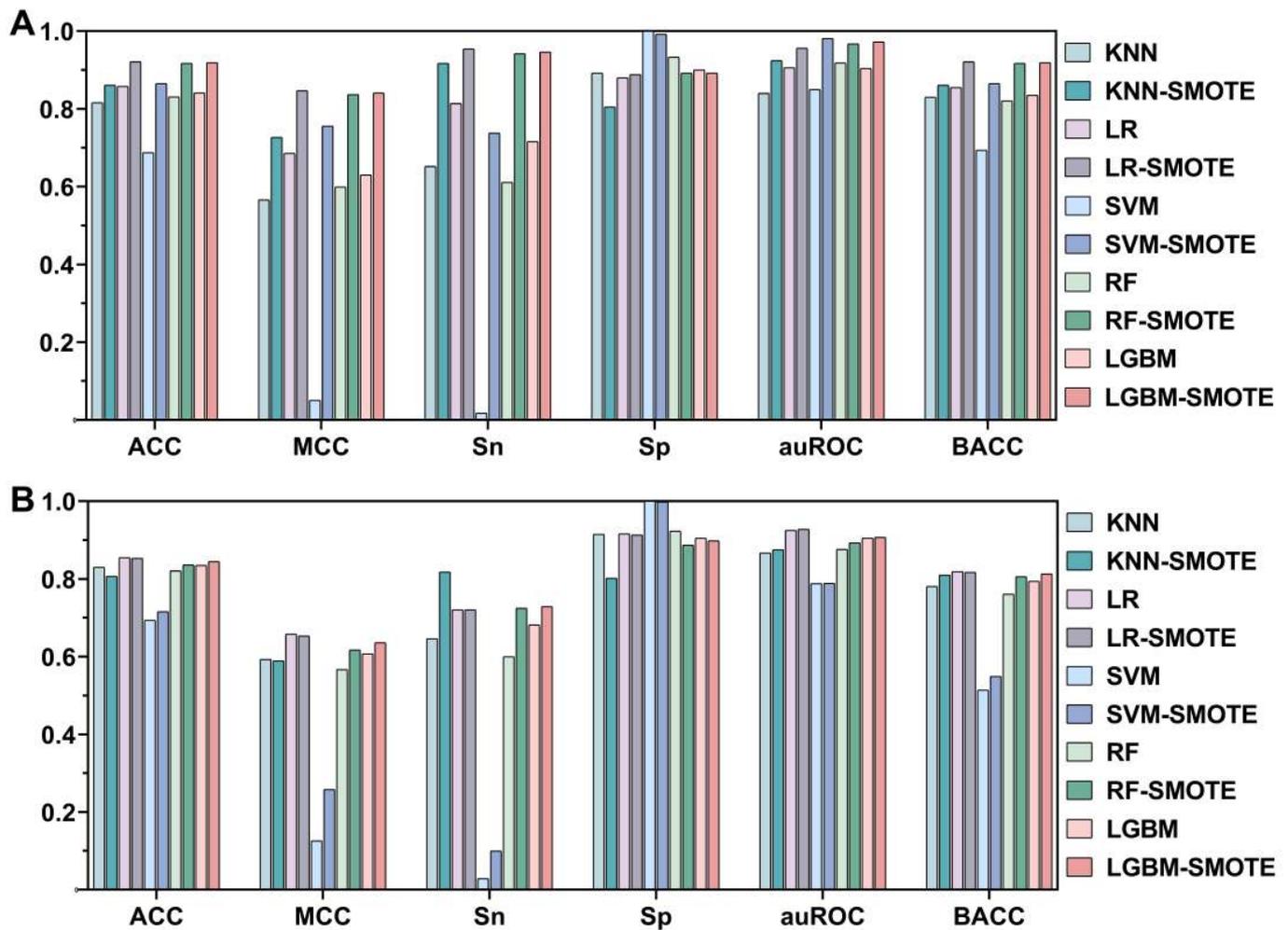


Figure 2. Results of 10-fold cross-validation (A) and independent testing (B) of the five ML models balanced with SMOTE and the five ML models balanced without SMOTE. As illustrated in Figure 2 and Supplementary Table S1, the features of models following optimization with SMOTE were clearly superior to the features of models developed without SMOTE optimization. Using the LR-based prediction model as an example, the LR-SMOTE model outperformed or equaled the LR model without SMOTE optimization in 66.7% of the metrics in 10-fold cross-validation and independent tests. Of the SVM-based models, the SVM-SMOTE model outperformed the SVM model developed without SMOTE optimization in 83.3% of the indicators.

In some models, the Sp values were high while the Sn and other indicators were very poor owing to the bias of the unbalanced dataset bias towards the negative class, which negatively affected the recognition ability of the positive class. These findings also emphasize the value and significance of optimizing imbalanced datasets using SMOTE. However, it can be inferred from the UMAP display in Figure 3 that the improvement in the datasets using SMOTE improved the predictive ability of the models in identifying umami peptides.

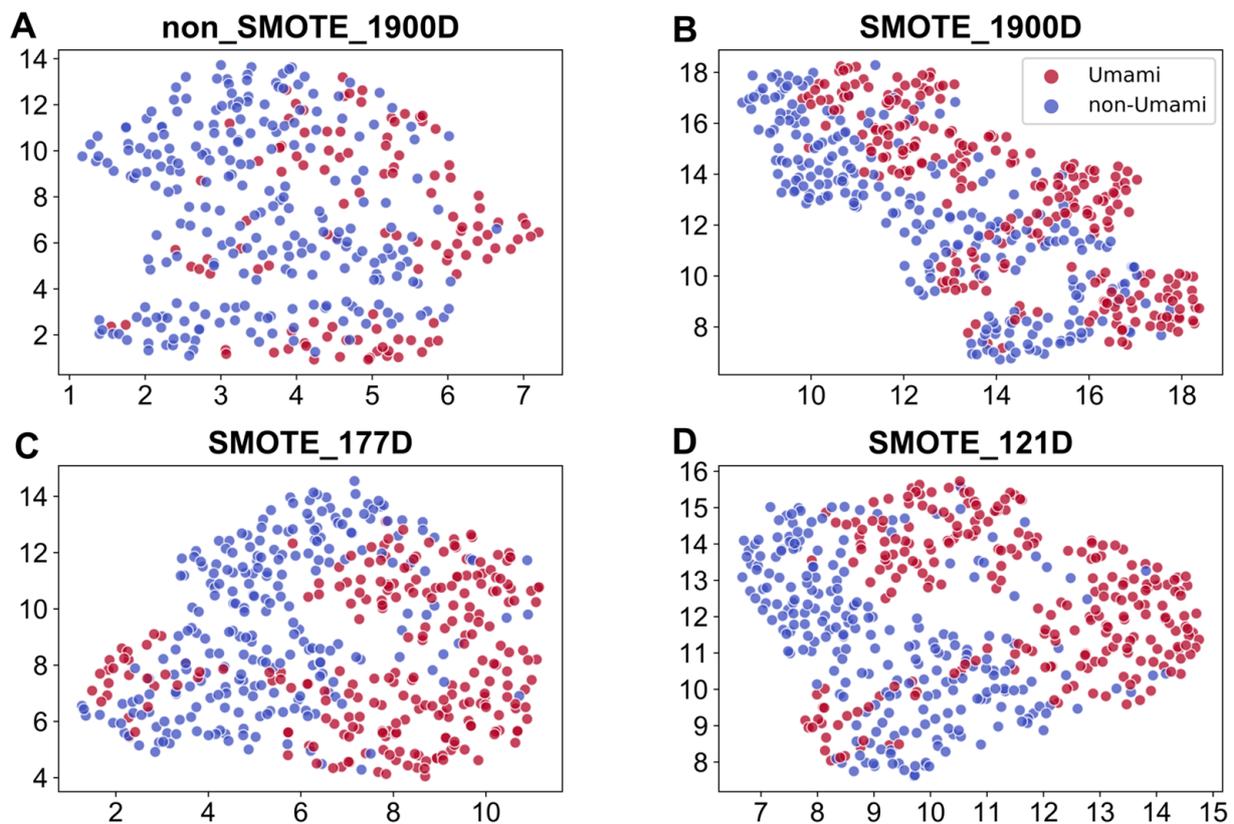


Figure 3. UMAP was used for visualizing the dimension-reduced features. (A) UniRep features without SMOTE balancing, (B) UniRep features following SMOTE balancing, (C) data of the top 177 features selected from the SMOTE-balanced UniRep feature set, and (D) data obtained using the top 121 features selected from the SMOTE-balanced UniRep feature set.

3.2. Effects of Different ML Models

The results of Section 3.1 revealed that the SMOTE algorithm optimized the unbalanced data to a certain extent. The consequents of 10-fold cross-validation and independent tests of the models created using SMOTE-balanced features with the five ML algorithms are depicted in Table 1.

Table 1. Results of 10-fold cross-validation and independent testing based on the five ML algorithms developed using SMOTE-balanced features.

Model	10-Fold Cross-Validation						Independent Test					
	ACC	MCC	Sn	Sp	auROC	BACC	ACC	MCC	Sn	Sp	auROC	BACC
LR ^c	<u>0.921</u> ^a	0.847	0.954	0.888	0.956	<u>0.921</u>	0.853	0.653	0.721	0.913	<u>0.928</u>	0.817
KNN ^c	<u>0.861</u> ^b	0.727	0.917	0.805	0.924	<u>0.861</u>	0.807	0.589	0.818	0.802	0.875	0.810
SVM ^c	<u>0.865</u>	0.756	0.738	0.992	0.981	<u>0.865</u>	0.716	0.258	0.100	0.998	0.789	0.549
RF ^c	<u>0.917</u>	0.837	0.942	0.892	0.967	<u>0.917</u>	0.836	0.617	0.725	0.887	0.893	0.806
LGBM ^c	<u>0.919</u>	0.841	0.946	0.892	0.972	<u>0.919</u>	0.845	0.636	0.729	0.898	0.907	0.813

^a The best performance values are indicated in bold and underlined. ^b Blue indicates equal values of ACC and BACC. ^c LR: logistic regression; KNN: k-nearest neighbors; SVM: support vector machine; LGBM: light gradient boosting machine; RF: random forest.

As depicted in Table 1, the recognition of umami peptides by the LR model outperformed that of the other ML models in 66.7% of the metrics. The consequents of 10-fold cross-validation revealed that the LR model, iUmami-DRLF, exceeded all other ML models in four metrics. The ACC and BACC of iUmami-DRLF were 0.22–6.97% superior to that of the other models, while the MCC and Sn increased by 0.71–16.51% and 0.85–29.27%,

respectively. The results of the independent tests revealed that the LR model, iUmami-DRLF, outscored the other ML models in four metrics. The ACC, MCC, auROC, and BACC efficiency levels of iUmami-DRLF were superior to those of the other models by 0.95–19.13%, 2.67–153.10%, 2.32–17.62%, and 0.49–48.82%, respectively. Although the SVM model achieved the best indicators for the identification of umami peptides in certain aspects, the results of the independent tests revealed that the SVM model will show more unbalanced data (MCC = 0.258, Sn = 0.100, and BACC = 0.549). We, therefore, selected the LR model for developing the umami peptide predictor. Additionally, the results of the 10-fold cross-validation of the five models revealed that the values of ACC and BACC were equal, indicating that the dataset was balanced following optimization with SMOTE. The equal values of ACC and BACC have been indicated in blue in Table 1.

3.3. Effects of Different Feature Selection Methods

As described in Section 3.1, the balanced SMOTE-optimized data encoding method significantly outperformed the unprocessed data encoding approach in the tests. The feature vector that was recovered using UniRep had 1900 dimensions as opposed to the 353 dimensions of the sequence vector that was used in the training set. The use of high-dimensional feature vectors frequently leads to over-fitting or redundancy of feature information. In order to solve this issue, we used three feature selection methods, namely, ANOVA, LGBM, and MI, for selecting the high-dimensional feature vectors. An incremental feature strategy and a hyperparameter grid search approach were employed in this study, and the GridSearchCV module in the scikit-learn library was used for searching the hyperparameters for each model. Table 2 summarizes the outcomes of 10-fold cross-validation and independent testing of the five ML models developed based on the UniRep features selected using the three feature selection methods. The results of independent testing of the aforementioned models with selected features and the models without selected features are compared in Figure 4.

Table 2. Results of 10-fold cross-validation and independent testing of the five ML models developed using UniRep features selected with the three feature selection methods (LGBM, ANOVA, and MI).

Model	Feature Selection Method	Dim	10-Fold Cross-Validation						Independent Test					
			ACC	MCC	Sn	Sp	auROC	BACC	ACC	MCC	Sn	Sp	auROC	BACC
LR ^c	LGBM ^d	177	0.925 ^b	0.853	0.959	0.892	0.957	0.925	<u>0.921</u> ^a	<u>0.815</u>	0.821	<u>0.967</u>	<u>0.956</u>	0.894
	ANOVA ^d	102	0.882	0.764	0.896	0.867	0.938	0.882	0.899	0.768	0.857	0.918	0.930	0.888
	MI ^d	136	0.888	0.777	0.913	0.863	0.942	0.888	0.888	0.733	0.750	0.951	0.864	0.850
KNN ^c	LGBM ^d	33	0.892	0.788	0.938	0.846	0.955	0.892	0.899	0.782	0.929	0.885	0.911	0.907
	ANOVA ^d	15	0.873	0.748	0.896	0.851	0.934	0.873	0.865	0.703	0.857	0.869	0.907	0.863
	MI ^d	58	0.888	0.783	0.954	0.822	0.927	0.888	0.888	0.773	<u>0.964</u>	0.852	0.931	<u>0.908</u>
SVM ^c	LGBM ^d	121	<u>0.944</u>	<u>0.889</u>	<u>0.971</u>	<u>0.917</u>	0.980	<u>0.944</u>	0.888	0.739	0.821	0.918	0.913	0.870
	ANOVA ^d	48	0.925	0.854	0.967	0.884	0.977	0.925	0.865	0.678	0.679	0.951	0.906	0.815
	MI ^d	16	0.919	0.841	0.959	0.880	0.968	0.919	0.888	0.735	0.786	0.934	0.921	0.860
RF ^c	LGBM ^d	88	0.915	0.830	0.934	0.896	0.975	0.915	0.876	0.716	0.821	0.902	0.920	0.862
	ANOVA ^d	118	0.898	0.797	0.913	0.884	0.961	0.898	0.865	0.694	0.821	0.885	0.911	0.853
	MI ^d	8	0.902	0.806	0.921	0.884	0.952	0.902	0.888	0.753	0.893	0.885	0.923	0.889
LGBM ^c	LGBM ^d	35	0.938	0.877	<u>0.971</u>	0.905	<u>0.988</u>	0.938	0.888	0.739	0.821	0.918	0.912	0.870
	ANOVA ^d	19	0.902	0.807	0.942	0.863	0.945	0.902	0.876	0.706	0.714	0.951	0.929	0.833
	MI ^d	18	0.888	0.777	0.917	0.859	0.953	0.888	0.865	0.682	0.750	0.918	0.916	0.834

^a The best performance values are indicated in bold and underlined. ^b Blue indicates equal values of ACC and BACC. ^c LR: logistic regression; KNN: *k*-nearest neighbors; SVM: support vector machine; LGBM: light gradient boosting machine; RF: random forest. ^d LGBM: light gradient boosting machine; ANOVA: analysis of variance; MI: mutual information.

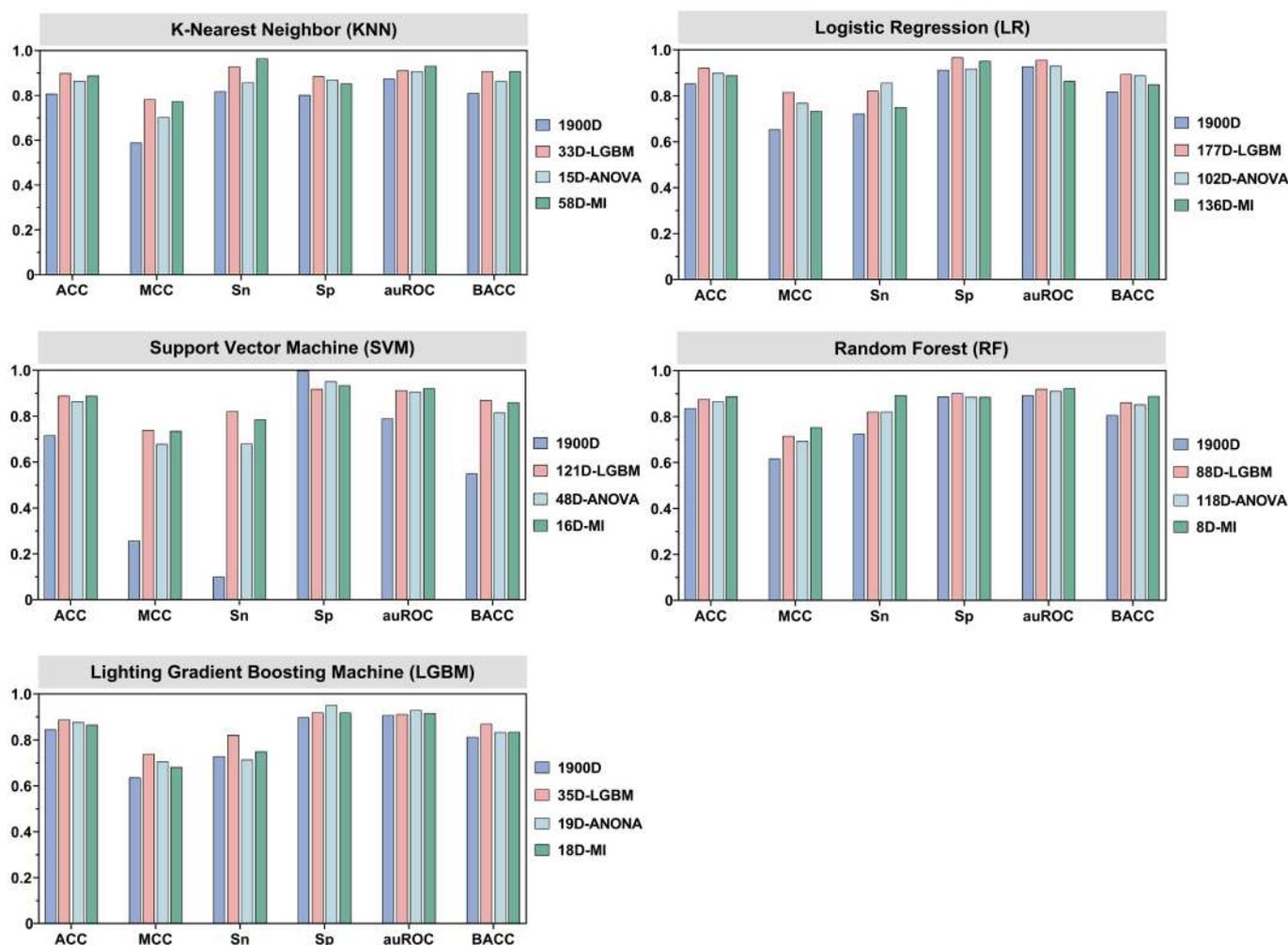


Figure 4. Comparison of the results of independent testing of the models with selected features and the models without selected features.

The outcomes of the independent testing are shown in Figure 4, which amply demonstrates that the chosen fusion feature sets outperformed the unselected fusion features. In the independent tests, the Sp of the 1900D models without feature selection was lower than all the models with feature selection, with the exception of the SVM-based model (5.00–8.75% higher). These results clearly demonstrated that the selection of feature descriptors effectively resolves information redundancy, and helps optimize the prediction performance of the umami peptide prediction model. Figure 4 and Table 2 clearly depict that of the three feature selection methods, and the overall performance of LGBM was superior to that of the other feature selection methods used for the identification of umami peptides. Considering the LR model as an example, the LGBM feature selection method outperformed the other methods (ANOVA and MI) in all six metrics in the 10-fold cross-validation studies. The performance of ACC, MCC, Sn, Sp, auROC, and BACC efficiency improved by 4.17–4.88%, 9.78–11.65%, 5.04–7.03%, 2.88–3.36%, 1.59–2.03%, and 4.17–4.88%, respectively, when the LGBM method was used. The LGBM feature selection method outperformed ANOVA and MI in the independent tests in five metrics. The performance of ACC, MCC, Sp, auROC, and BACC synergy improved by 2.45–3.72%, 6.12–11.19%, 1.68–5.34%, 2.80–10.65%, and 0.68–5.18%, respectively, when the LGBM method was used.

Based on the aforementioned results (Sections 3.1–3.3), we believe that the LR model developed based on the first 177 dimensions of UniRep using SMOTE-optimized data was superior in predicting umami peptides, and corroborates with the results of visual analysis discussed hereafter in Section 3.4. Based on the aforementioned analyses, the first 177D

features of UniRep were selected for constructing the iUmami-DRLF predictor based on the LGBM model for subsequent studies.

3.4. Comparison with Existing Methods

In order to evaluate the efficacy and application of our technique in comparison to other predictors, we assessed and compared the predictive performance of iUmami-DRLF with that of other methods, including iUmami-SCM and UMPred-FRL. Table 3 compares the results of 10-fold cross-validation and independent testing of iUmami-DRLF with those of other existing methods.

Table 3. Results of 10-fold cross-validation and independent testing of iUmami-DRLF and other existing methods.

Classifier	10-Fold Cross-Validation						Independent Test					
	ACC	MCC	Sn	Sp	auROC	BACC	ACC	MCC	Sn	Sp	auROC	BACC
iUmami-DRLF(LR)	0.925 ^b	0.853	0.959	0.892	0.957	0.925	0.921 ^a	0.815	0.821	0.967	0.956	0.894
iUmami-DRLF(SVM)	0.944	0.889	0.971	0.917	0.980	0.944	0.888	0.739	0.821	0.918	0.913	0.870
iUP-BERT	0.940	0.881	0.963	0.917	0.971	0.940	0.899	0.774	0.893	0.902	0.933	0.897
UMPred-FRL	0.921	0.814	0.847	0.955	0.938	0.901	0.888	0.735	0.786	0.934	0.919	0.860
iUmami-SCM	0.935	0.864	0.947	0.930	0.945	0.939	0.865	0.679	0.714	0.934	0.898	0.824

^a The best performance values are indicated in bold and underlined. ^b Blue indicates equal values of ACC and BACC.

Table 3 clearly demonstrates that iUmami-DRLF(SVM) outperformed the other classifiers in all metrics except Sp in the 10-fold cross-validation test. The ACC, MCC, Sn, auROC, and BACC of iUmami-DRLF(SVM) were superior to those of the other methods by 2.02–2.50%, 4.31–9.25%, 1.30–14.63%, 2.37–4.43%, and 2.02–4.77%, respectively. More significantly, the results of independent testing revealed that iUmami-DRLF(LR) performed better than the current predictors in every aspect. The ACC, MCC, Sp, auROC, and BACC were superior to those of the other methods by 3.76–6.51%, 10.86–20.00%, 4.51–15.05%, 4.04–6.47%, and 3.99–8.53%, respectively.

Comparison of the two iUmami-DRLF predictors revealed that the results of the 10-fold validation of iUmami-DRLF(LR) were slightly worse than the results for the SVM model (ACC and BACC, MCC, Sn, Sp, and auROC were 2.02%, 4.31%, 1.30%, 2.79%, and 2.37% lower, respectively). However, the results of independent testing were superior for iUmami-DRLF(LR) (ACC, MCC, Sp, auROC, and BACC were 3.66%, 9.25%, 5.08%, 4.53%, and 2.75% higher, respectively). This revealed that the generalization ability of LR was stronger. The results of comparative analyses demonstrated the superiority of iUmami-DRLF in umami peptide prediction. The umami prediction ability of iUmami-DRLF was more reliable than the existing methods.

3.5. Feature Visualization

Feature visualization can intuitively convey feature information through images to clearly represent the dataset. UMAP is a popular uniform approximation projection algorithm for dimensionality reduction, and was used in this study for visual analyses of the features in the umami peptide dataset. The differences in feature representation are clearly highlighted in UMAP visualization. The results of dimensionality reduction for feature visualization with UMAP are depicted in Figure 3.

Figure 3 demonstrates that compared with the UniRep feature vector without SMOTE optimization (Figure 3A), the SMOTE-optimized 1900D UniRep feature vector (Figure 3B) was better at distinguishing umami peptides from non-umami peptides. Compared with the SMOTE-optimized UniRep features (Figure 3B), the top 177D features (Figure 3C) and top 121D features of UniRep (Figure 3D) were further optimized after feature selection.

3.6. Web Server Development

For other researchers to anticipate umami peptides, we created the user-friendly iUmami-DRLF web server, which is freely accessible online at <https://www.aibiochem.net/servers/iUmami-DRLF/> (accessed on 1 April 2023). The web server is easy to use. The user only needs to enter the peptide sequence in the text box, click the run button, and wait for a few minutes, then the user can identify and judge whether the input peptide sequence is an umami peptide, and the result will be displayed on the web page. The output results include the input sequence, whether it is an umami peptide, and the confidence level. See the web server interface on the website or Supplementary Figures S1–S3. Additionally, please contact the corresponding authors if users need to predict a significant number of sequences.

3.7. Methods' Robustness

To further verify the effectiveness and robustness of the model, we collected 91 wet-experiment verified umami peptide sequences reported in the latest literature [56–70]. These empirical umami peptide sequences constituted the dataset UMP-VERIFIED, which was then used to test state-of-the-art methods, including UMPred-FRL [11] and iUP-BERT [12] for comparison to iUmami-DRLF. Here, the accuracy of models under different prediction probability threshold conditions was adopted to make comparisons. The probability threshold T referred to the fact that, for a peptide sequence, if the probability threshold predicted by the machine learning model was greater than T , the model would determine that the sequence was an umami peptide; otherwise, it was a non-umami peptide.

Figure 5A shows the relationship between the accuracy of the three models and the probability threshold. It can be seen from the figure that our model iUmami-DRLF has the best accuracy under any probability threshold. It was particularly noteworthy that the accuracy rate of iUP-BERT is 0 at 95% threshold probability, indicating that the model has failed. While the value of iUmami-DRLF is 52.7%, which is nearly six times the UMPred-FRL accuracy (8.8%), when the probability threshold was set to 99%, the prediction accuracy of iUP-BERT and UMPred-FRL is 0. It meant that both methods were invalid. In sharp contrast, iUmami-DRLF can still maintain the prediction accuracy of 40.7%, and the model still worked. These results proved that iUmami-DRLF is with better robustness and better model generalization performance than other methods.

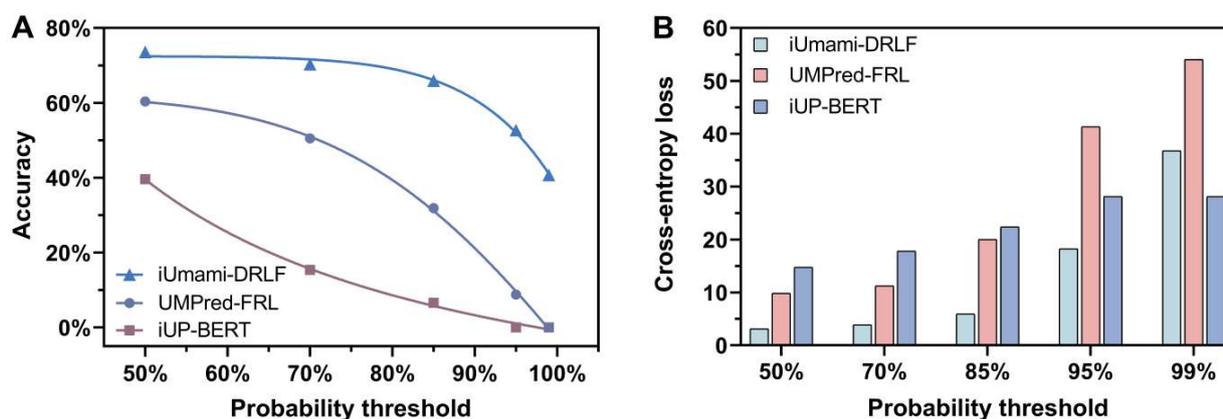


Figure 5. Under varying probability thresholds, the prediction results of iUmami-DRLF (this work), UMPred-FRL, and iUP-BERT are shown using the UMP-VERIFIED dataset. (A) is the relationship between prediction accuracy and probability threshold. (B) is the cross-entropy loss of the predicted outcome about the probability threshold. The smaller the cross-entropy loss, the better the robustness and accuracy of the model. Note that at the probability thresholds of 95% and 99%, the prediction accuracy of iUP-BERT and UMPred-FRL is 0, and their corresponding cross-entropy losses can be calculated, but they are not meaningful.

The robustness and effectiveness of iUmami-DRLF come from the fact that it was an optimized model with minimum cross-entropy loss. For binary classification machine learning models, the closer the prediction output is to the real sample label, the smaller the cross-entropy loss is, resulting in better accuracy [71].

It could be proven by the data shown in Figure 5B. Figure 5B displays models' cross-entropy loss under different probability thresholds. Obviously, iUmami-DRLF has the minimum cross entropy loss of the three models under the probability threshold of 50%, 70%, and 85%. At the probability threshold of 95%, the cross-entropy loss of iUmami-DRLF is significantly smaller than that of UMPred-FRL. For 95% and 99% probability thresholds, UMPred-FRL and iUP-BERT models have failed, and the calculated cross-entropy is meaningless. For example, the cross-entropy loss of iUP-BERT remains unchanged in probability thresholds of 95% and 99% of cases.

4. Conclusions and Future Work

In this research, we proposed a predictor, iUmami-DRLF, for the successful prediction of umami peptides solely based on sequence information. The imbalanced dataset was processed with SMOTE, and the latent umami peptide information was obtained using the UniRep deep representation learning feature embedding approach. Our predictor was strengthened by the use of three feature selection techniques, namely, LGBM, ANOVA, and MI, and the combination of five ML algorithms (KNN, LR, SVM, RF, and LGBM) for model development. Following testing and optimization, the top 177D features of UniRep were selected as the optimal feature set, and then integrated with the LR model for developing the final predictor. The results of 10-fold cross-validation and independent testing revealed that iUmami-DRLF markedly outperformed the existing methods in the independent tests. The latest umami peptide sequences verified by wet experiment were used to validate the method, and the results show that iUmami-DRLF could more reliably, robustly and accurately predict (independent tests: ACC = 0.921, MCC = 0.815, Sn = 0.821, Sp = 0.967, auROC = 0.956) umami peptides than the reported state-of-the-art methods. It hopes that the user-friendly webserver could be useful for researchers in the area. The following areas can still be improved, despite the fact that iUmami-DRLF has significantly increased the accuracy of umami peptide prediction: First, as our feature extraction model requires lots of computation, webserver without GPU configuration will take a long time to complete this task. Users can contact the corresponding authors if they need to predict a large number of sequences. Furthermore, using the most recent empirical data when training the model might produce better outcomes. Ultimately, using the method of model distillation can simplify the feature extraction model and lessen its computational complexity.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/foods12071498/s1>, Figures S1–S3: iUmami-DRLF web server interface; Table S1: Results of 10-fold cross-validation and independent testing about SMOTE; Table S2: Prediction- probability results of the three models and 91 wet-experiment validated umami peptide sequences [5,6,11,56,59,60,62–70].

Author Contributions: Conceptualization, Z.L.; Data curation, J.J. and Z.L.; Formal analysis, J.J. and Z.L.; Funding acquisition, Q.Z. and Z.L.; Methodology, Z.L.; Software, Z.L.; Supervision, Z.L.; Validation, J.J. and Z.L.; Visualization, J.J.; Writing—original draft, J.J.; Writing—review and editing, J.J., J.L. (Jiayu Li), J.L. (Junxian Li), H.P., M.L., Q.Z. and Z.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (No. 62001090, No. 62250028, No. 62131004) the Sichuan Provincial Science Fund for Distinguished Young Scholars (2021JDJQ0025), the Municipal Government of Quzhou (No. 2022D040), and Fundamental Research Funds for the Central Universities of Sichuan University (No. YJ2021104).

Data Availability Statement: The data used to support the findings of this study can be made available by the corresponding author upon request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Torii, K.; Uneyama, H.; Nakamura, E. Physiological roles of dietary glutamate signaling via gut–brain axis due to efficient digestion and absorption. *J. Gastroenterol.* **2013**, *48*, 442–451. [[CrossRef](#)] [[PubMed](#)]
2. Zhang, Y.; Venkatasamy, C.; Pan, Z.; Liu, W.; Zhao, L. Novel Umami Ingredients: Umami Peptides and Their Taste. *J. Food Sci.* **2017**, *82*, 16–23. [[CrossRef](#)]
3. Dang, Y.; Hao, L.; Cao, J.; Sun, Y.; Zeng, X.; Wu, Z.; Pan, D. Molecular docking and simulation of the synergistic effect between umami peptides, monosodium glutamate and taste receptor T1R1/T1R3. *Food Chem.* **2019**, *271*, 697–706. [[CrossRef](#)] [[PubMed](#)]
4. Minkiewicz, P.; Iwaniak, A.; Darewicz, M. BIOPEP-UWM Database of Bioactive Peptides: Current Opportunities. *Int. J. Mol. Sci.* **2019**, *20*, 5978. [[CrossRef](#)] [[PubMed](#)]
5. Cao, C.; Wang, J.; Kwok, D.; Cui, F.; Zhang, Z.; Zhao, D.; Li, M.J.; Zou, Q. webTWAS: A resource for disease candidate susceptibility genes identified by transcriptome-wide association study. *Nucleic Acids Res.* **2022**, *50*, D1123–D1130. [[CrossRef](#)] [[PubMed](#)]
6. Jiang, J.; Lin, X.; Jiang, Y.; Jiang, L.; Lv, Z. Identify Bitter Peptides by Using Deep Representation Learning Features. *Int. J. Mol. Sci.* **2022**, *23*, 7877. [[CrossRef](#)]
7. Yan, N.; Lv, Z.; Hong, W.; Xu, X. Editorial: Feature Representation and Learning Methods With Applications in Protein Secondary Structure. *Front. Bioeng. Biotechnol.* **2021**, *9*, 748722. [[CrossRef](#)]
8. Zhao, Q.; Ma, J.; Wang, Y.; Xie, F.; Lv, Z.; Xu, Y.; Shi, H.; Han, K. Mul-SNO: A Novel Prediction Tool for S-Nitrosylation Sites Based on Deep Learning Methods. *IEEE J. Biomed. Health Informatics* **2021**, *26*, 2379–2387. [[CrossRef](#)]
9. 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.* **2020**, *60*, 6666–6678. [[CrossRef](#)]
10. Charoenkwan, P.; Yana, J.; Schaduagrath, N.; Nantasenamat, C.; Hasan, M.M.; Shoombuatong, W. iBitter-SCM: Identification and characterization of bitter peptides using a scoring card method with propensity scores of dipeptides. *Genomics* **2020**, *112*, 2813–2822.
11. Charoenkwan, P.; Nantasenamat, C.; Hasan, M.; Moni, M.A.; Manavalan, B.; Shoombuatong, W. UMPred-FRL: A New Approach for Accurate Prediction of Umami Peptides Using Feature Representation Learning. *Int. J. Mol. Sci.* **2021**, *22*, 13124. [[CrossRef](#)] [[PubMed](#)]
12. Jiang, L.; Jiang, J.; Wang, X.; Zhang, Y.; Zheng, B.; Liu, S.; Zhang, Y.; Liu, C.; Wan, Y.; Xiang, D.; et al. IUP-BERT: Identification of Umami Peptides Based on BERT Features. *Foods* **2022**, *11*, 3742. [[CrossRef](#)] [[PubMed](#)]
13. Bengio, Y.; Courville, A.; Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [[PubMed](#)]
14. Feifei, C. DeepMC-iNABP: Deep learning for multiclass identification and classification of nucleic acid-binding proteins. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 2020–2028.
15. Caro, M.C.; Huang, H.-Y.; Cerezo, M.; Sharma, K.; Sornborger, A.; Cincio, L.; Coles, P.J. Generalization in quantum machine learning from few training data. *Nat. Commun.* **2022**, *13*, 4919. [[CrossRef](#)]
16. Cunningham, J.M.; Koytiger, G.; Sorger, P.K.; AlQuraishi, M. Biophysical prediction of protein–peptide interactions and signaling networks using machine learning. *Nat. Methods* **2020**, *17*, 175–183. [[CrossRef](#)]
17. Lei, Y.; Li, S.; Liu, Z.; Wan, F.; Tian, T.; Li, S.; Zhao, D.; Zeng, J. A deep-learning framework for multi-level peptide–protein interaction prediction. *Nat. Commun.* **2021**, *12*, 5465. [[CrossRef](#)]
18. Abbasi, A.; Miah, E.; Mirroshandel, S.A. Effect of deep transfer and multi-task learning on sperm abnormality detection. *Comput. Biol. Med.* **2021**, *128*, 104121. [[CrossRef](#)]
19. Arora, V.; Ng, E.Y.-K.; Leekha, R.S.; Darshan, M.; Singh, A. Transfer learning-based approach for detecting COVID-19 ailment in lung CT scan. *Comput. Biol. Med.* **2021**, *135*, 104575. [[CrossRef](#)]
20. Cao, C.; He, J.; Mak, L.; Perera, D.; Kwok, D.; Wang, J.; Li, M.; Mourier, T.; Gavriluc, S.; Greenberg, M.; et al. Reconstruction of Microbial Haplotypes by Integration of Statistical and Physical Linkage in Scaffolding. *Mol. Biol. Evol.* **2021**, *38*, 2660–2672. [[CrossRef](#)]
21. Ao, C.; Jiao, S.; Wang, Y.; Yu, L.; Zou, Q. Biological Sequence Classification: A Review on Data and General Methods. *Research* **2022**, *2022*, 0011. [[CrossRef](#)]
22. Harini, K.; Kihara, D.; Gromiha, M.M. PDA-Pred: Predicting the binding affinity of protein–DNA complexes using machine learning techniques and structural features. *Methods* **2023**, *213*, 10–17. [[CrossRef](#)] [[PubMed](#)]
23. Wang, X.; Alnabati, E.; Aderinwale, T.W.; Maddhuri Venkata Subramaniya, S.R.; Terashi, G.; Kihara, D. Detecting protein and DNA/RNA structures in cryo-EM maps of intermediate resolution using deep learning. *Nat. Commun.* **2021**, *12*, 2302.
24. Meier, F.; Köhler, N.D.; Brunner, A.-D.; Wanka, J.-M.H.; Voytik, E.; Strauss, M.T.; Theis, F.J.; Mann, M. Deep learning the collisional cross sections of the peptide universe from a million experimental values. *Nat. Commun.* **2021**, *12*, 1185. [[CrossRef](#)]
25. Wilhelm, M.; Zolg, D.P.; Graber, M.; Gessulat, S.; Schmidt, T.; Schnatbaum, K.; Schwencke-Westphal, C.; Seifert, P.; de Andrade Krätzig, N.; Zerweck, J.; et al. Deep learning boosts sensitivity of mass spectrometry-based immunopeptidomics. *Nat. Commun.* **2021**, *12*, 3346.

26. He, J.; Lin, P.; Chen, J.; Cao, H.; Huang, S.-Y. Model building of protein complexes from intermediate-resolution cryo-EM maps with deep learning-guided automatic assembly. *Nat. Commun.* **2022**, *13*, 4066. [[CrossRef](#)] [[PubMed](#)]
27. Kobayashi, H.; Cheveralls, K.C.; Leonetti, M.D.; Royer, L.A. Self-supervised deep learning encodes high-resolution features of protein subcellular localization. *Nat. Methods* **2022**, *19*, 995–1003. [[CrossRef](#)] [[PubMed](#)]
28. Yildirim, K.; Bozdog, P.G.; Talo, M.; Yildirim, O.; Karabatak, M.; Acharya, U. Deep learning model for automated kidney stone detection using coronal CT images. *Comput. Biol. Med.* **2021**, *135*, 104569. [[CrossRef](#)] [[PubMed](#)]
29. Xiao, Y.; Wu, J.; Lin, Z. Cancer diagnosis using generative adversarial networks based on deep learning from imbalanced data. *Comput. Biol. Med.* **2021**, *135*, 104540. [[CrossRef](#)]
30. Jain, A.; Terashi, G.; Kagaya, Y.; Subramaniya, S.R.M.V.; Christoffer, C.; Kihara, D. Analyzing effect of quadruple multiple sequence alignments on deep learning based protein inter-residue distance prediction. *Sci. Rep.* **2021**, *11*, 7574. [[CrossRef](#)]
31. Wang, X.; Wang, S.; Fu, H.; Ruan, X.; Tang, X. DeepFusion-RBP: Using Deep Learning to Fuse Multiple Features to Identify RNA-binding Protein Sequences. *Curr. Bioinform.* **2021**, *16*, 1089–1100. [[CrossRef](#)]
32. Charoenkwan, P.; Nantasenamat, C.; Hasan, M.; Manavalan, B.; Shoombuatong, W. BERT4Bitter: A bidirectional encoder representations from transformers (BERT)-based model for improving the prediction of bitter peptides. *Bioinformatics* **2021**, *37*, 2556–2562. [[CrossRef](#)]
33. Bao, W.; Cui, Q.; Chen, B.; Yang, B. Phage_UniR_LGBM: Phage Virion Proteins Classification with UniRep Features and LightGBM Model. *Comput. Math. Methods Med.* **2022**, *2022*, 9470683. [[CrossRef](#)]
34. Wang, Y.; Xu, L.; Zou, Q.; Lin, C. prPred-DRLF: Plant R protein predictor using deep representation learning features. *Proteomics* **2022**, *22*, e2100161. [[CrossRef](#)]
35. Villegas-Morcillo, A.; Gomez, A.M.; Sanchez, V. An analysis of protein language model embeddings for fold prediction. *Briefings Bioinform.* **2022**, *23*, bbac142. [[CrossRef](#)]
36. Wei, Y.; Zou, Q.; Tang, F.; Yu, L. WMSA: A novel method for multiple sequence alignment of DNA sequences. *Bioinformatics* **2022**, *38*, 5019–5025. [[CrossRef](#)] [[PubMed](#)]
37. Zhang, Z.; Cui, F.; Lin, C.; Zhao, L.; Wang, C.; Zou, Q. Critical downstream analysis steps for single-cell RNA sequencing data. *Briefings Bioinform.* **2021**, *22*, bbab105. [[CrossRef](#)]
38. Zhang, Z.; Cui, F.; Su, W.; Dou, L.; Xu, A.; Cao, C.; Zou, Q. webSCST: An interactive web application for single-cell RNA-sequencing data and spatial transcriptomic data integration. *Bioinformatics* **2022**, *38*, 3488–3489. [[CrossRef](#)] [[PubMed](#)]
39. Zhang, Z.; Cui, F.; Cao, C.; Wang, Q.; Zou, Q. Single-cell RNA analysis reveals the potential risk of organ-specific cell types vulnerable to SARS-CoV-2 infections. *Comput. Biol. Med.* **2022**, *140*, 105092. [[CrossRef](#)] [[PubMed](#)]
40. Alley, E.C.; Khimulya, G.; Biswas, S.; AlQuraishi, M.; Church, G.M. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* **2019**, *16*, 1315–1322. [[CrossRef](#)] [[PubMed](#)]
41. Graves, A.; Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw.* **2005**, *18*, 602–610. [[CrossRef](#)] [[PubMed](#)]
42. Fernandez, A.; Garcia, S.; Herrera, F.; Chawla, N.V. SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *J. Artif. Intell. Res.* **2018**, *61*, 863–905. [[CrossRef](#)]
43. Kumar, M.; Rath, N.K.; Swain, A.; Rath, S.K. Feature Selection and Classification of Microarray Data using MapReduce based ANOVA and K-Nearest Neighbor. *Procedia Comput. Sci.* **2015**, *54*, 301–310. [[CrossRef](#)]
44. Shaw, R.G.; Mitchell-Olds, T. Anova for Unbalanced Data: An Overview. *Ecology* **1993**, *74*, 1638–1645. [[CrossRef](#)]
45. Lv, Z.; Cui, F.; Zou, Q.; Zhang, L.; Xu, L. Anticancer peptides prediction with deep representation learning features. *Briefings Bioinform.* **2021**, *22*, bbab008. [[CrossRef](#)] [[PubMed](#)]
46. He, W.; Jiang, Y.; Jin, J.; Li, Z.; Zhao, J.; Manavalan, B.; Su, R.; Gao, X.; Wei, L. Accelerating bioactive peptide discovery via mutual information-based meta-learning. *Briefings Bioinform.* **2022**, *23*, bbab499. [[CrossRef](#)]
47. Zhao, D.; Teng, Z.; Li, Y.; Chen, D. iAIPs: Identifying Anti-Inflammatory Peptides Using Random Forest. *Front. Genet.* **2021**, *12*, 773202. [[CrossRef](#)]
48. Battiti, R. Using mutual information for selecting features in supervised neural net learning. *IEEE Trans. Neural Networks* **1994**, *5*, 537–550. [[CrossRef](#)]
49. Tripathi, V.; Tripathi, P. Detecting antimicrobial peptides by exploring the mutual information of their sequences. *J. Biomol. Struct. Dyn.* **2020**, *38*, 5037–5043. [[CrossRef](#)]
50. Ao, C.; Zou, Q.; Yu, L. NmRF: Identification of multispecies RNA 2'-O-methylation modification sites from RNA sequences. *Briefings Bioinform.* **2022**, *23*, bbab480. [[CrossRef](#)]
51. Chen, L.; Yu, L.; Gao, L. Potent antibiotic design via guided search from antibacterial activity evaluations. *Bioinformatics* **2023**, *39*, btad059. [[CrossRef](#)] [[PubMed](#)]
52. Spindelböck, T.; Ranftl, S.; von der Linden, W. Cross-Entropy Learning for Aortic Pathology Classification of Artificial Multi-Sensor Impedance Cardiography Signals. *Entropy* **2021**, *23*, 1661. [[CrossRef](#)] [[PubMed](#)]
53. Miao, F.; Yao, L.; Zhao, X. Adaptive Margin Aware Complement-Cross Entropy Loss for Improving Class Imbalance in Multi-View Sleep Staging Based on EEG Signals. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2022**, *30*, 2927–2938. [[CrossRef](#)] [[PubMed](#)]
54. Egusquiza, I.; Picon, A.; Irusta, U.; Bereciartua-Perez, A.; Eggers, T.; Klukas, C.; Aramendi, E.; Navarra-Mestre, R. Analysis of Few-Shot Techniques for Fungal Plant Disease Classification and Evaluation of Clustering Capabilities Over Real Datasets. *Front. Plant Sci.* **2022**, *13*, 813237. [[CrossRef](#)]

55. Yu, L.; Xia, M.; An, Q. A network embedding framework based on integrating multiplex network for drug combination prediction. *Briefings Bioinform.* **2022**, *23*, bbab364. [[CrossRef](#)]
56. Zhang, T.; Hua, Y.; Zhou, C.; Xiong, Y.; Pan, D.; Liu, Z.; Dang, Y. Umami peptides screened based on peptidomics and virtual screening from *Ruditapes philippinarum* and *Macra veneriformis* clams. *Food Chem.* **2022**, *394*, 133504. [[CrossRef](#)]
57. Liang, L.; Zhou, C.; Zhang, J.; Huang, Y.; Zhao, J.; Sun, B.; Zhang, Y. Characteristics of umami peptides identified from porcine bone soup and molecular docking to the taste receptor T1R1/T1R3. *Food Chem.* **2022**, *387*, 132870. [[CrossRef](#)]
58. Bu, Y.; Liu, Y.; Luan, H.; Zhu, W.; Li, X.; Li, J. Characterization and structure–activity relationship of novel umami peptides isolated from Thai fish sauce. *Food Funct.* **2021**, *12*, 5027–5037. [[CrossRef](#)]
59. Chen, W.; Li, W.; Wu, D.; Zhang, Z.; Chen, H.; Zhang, J.; Wang, C.; Wu, T.; Yang, Y. Characterization of novel umami-active peptides from *Stropharia rugoso-annulata* mushroom and in silico study on action mechanism. *J. Food Compos. Anal.* **2022**, *110*, 104530. [[CrossRef](#)]
60. Zhu, X.; Sun-Waterhouse, D.; Chen, J.; Cui, C.; Wang, W. Comparative study on the novel umami-active peptides of the whole soybeans and the defatted soybeans fermented soy sauce. *J. Sci. Food Agric.* **2021**, *101*, 158–166. [[CrossRef](#)]
61. Wang, W.; Huang, Y.; Zhao, W.; Dong, H.; Yang, J.; Bai, W. Identification and comparison of umami-peptides in commercially available dry-cured Spanish mackerels (*Scomberomorus niphonius*). *Food Chem.* **2022**, *380*, 132175. [[CrossRef](#)]
62. Song, S.; Zhuang, J.; Ma, C.; Feng, T.; Yao, L.; Ho, C.-T.; Sun, M. Identification of novel umami peptides from *Boletus edulis* and its mechanism via sensory analysis and molecular simulation approaches. *Food Chem.* **2023**, *398*, 133835. [[CrossRef](#)]
63. Yu, Z.; Kang, L.; Zhao, W.; Wu, S.; Ding, L.; Zheng, F.; Liu, J.; Li, J. Identification of novel umami peptides from myosin via homology modeling and molecular docking. *Food Chem.* **2021**, *344*, 128728. [[CrossRef](#)]
64. Wang, Y.; Luan, J.; Tang, X.; Zhu, W.; Xu, Y.; Bu, Y.; Li, J.; Cui, F.; Li, X. Identification of umami peptides based on virtual screening and molecular docking from Atlantic cod (*Gadus morhua*). *Food Funct.* **2023**, *14*, 1510–1519. [[CrossRef](#)]
65. Zhu, W.; Luan, H.; Bu, Y.; Li, X.; Li, J.; Zhang, Y. Identification, taste characterization and molecular docking study of novel umami peptides from the Chinese anchovy sauce. *J. Sci. Food Agric.* **2021**, *101*, 3140–3155. [[CrossRef](#)] [[PubMed](#)]
66. Gao, B.; Hu, X.; Xue, H.; Li, R.; Liu, H.; Han, T.; Ruan, D.; Tu, Y.; Zhao, Y. Isolation and screening of umami peptides from preserved egg yolk by nano-HPLC-MS/MS and molecular docking. *Food Chem.* **2022**, *377*, 131996. [[CrossRef](#)] [[PubMed](#)]
67. Shiyan, R.; Liping, S.; Xiaodong, S.; Jinlun, H.; Yongliang, Z. Novel umami peptides from tilapia lower jaw and molecular docking to the taste receptor T1R1/T1R3. *Food Chem.* **2021**, *362*, 130249. [[CrossRef](#)] [[PubMed](#)]
68. Zhu, W.; He, W.; Wang, F.; Bu, Y.; Li, X.; Li, J. Prediction, molecular docking and identification of novel umami hexapeptides derived from Atlantic cod (*Gadus morhua*). *Int. J. Food Sci. Technol.* **2021**, *56*, 402–412. [[CrossRef](#)]
69. Liu, Q.; Gao, X.; Pan, D.; Liu, Z.; Xiao, C.; Du, L.; Cai, Z.; Lu, W.; Dang, Y.; Zou, Y. Rapid screening based on machine learning and molecular docking of umami peptides from porcine bone. *J. Sci. Food Agric.* **2022**. [[CrossRef](#)]
70. Liu, Z.; Zhu, Y.; Wang, W.; Zhou, X.; Chen, G.; Liu, Y. Seven novel umami peptides from Takifugu rubripes and their taste characteristics. *Food Chem.* **2020**, *330*, 127204. [[CrossRef](#)]
71. Rajaraman, S.; Zamzmi, G.; Antani, S.K. Novel loss functions for ensemble-based medical image classification. *PLoS ONE* **2021**, *16*, e0261307. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.