iBitter-GRE: A Novel Stacked Bitter Peptide Predictor with ESM-2 and Multi-View Features

Jingwei Lv [1], Aoyun Geng [1], Zhuoyu Pan [2], Leyi Wei [3,4], Quan Zou [5,6], Zilong Zhang [1], and Feifei Cui [1,*]

1 - ***School of Computer Science and Technology,*** *Hainan University, Haikou 570228, China*

2 - ***International Business School,*** *Hainan University, Haikou 570228, China*

3 - ***Centre for Artificial Intelligence Driven Drug Discovery,*** *Faculty of Applied Science, Macao Polytechnic University, Macao SAR, China*

4 - ***School of Informatics,*** *Xiamen University, Xiamen, China*

5 - ***Institute of Fundamental and Frontier Sciences,*** *University of Electronic Science and Technology of China, Chengdu 610054, China*

6 - ***Yangtze Delta Region Institute (Quzhou),*** *University of Electronic Science and Technology of China, Quzhou 324000, China*

Correspondence to Feifei Cui: *feifeicui@hainanu.edu.cn (F. Cui)*

https://doi.org/10.1016/j.jmb.2025.169005
Editor: Rita Casadio

Abstract

Accurate identification of bitter peptides is essential for research. Although models using sequence information have evolved in the context of bitter peptides, there is still room for improvement in their predictive performance. In the present study, we introduced a novel predictive tool, iBitter-GRE, designed to improve the accuracy of bitter peptide identification. Our model uses ESM-2 and traditional descriptors capture the physical and biochemical properties of bitter peptides for feature extraction. To expand the model's learning capabilities, we adopted a stacking approach to integrate multiple learners. Feature contributions were analyzed using SHAP values. Validation by domain experts confirmed that our model effectively identifies the key biochemical characteristics of bitter peptides. Benchmark experiments showed that iBitter-GRE achieves higher accuracy than existing methods. To assist the researchers, we created a web server accessible at http://www.bioai-lab.com/iBitter-GRE. We believe that iBitter-GRE is a valuable tool for the discovery and identification of bitter peptides.

Introduction

Bitterness, one of the five basic human senses, plays a vital role in daily life.[1] By sensing bitterness, humans can detect potential dangers, and through long-term evolution, we have developed the perception that bitter substances are difficult to swallow and are potentially harmful. The ability to sense bitterness helped our ancestors survive, and it has become an integral part of modern life. Most medicines are bitter[2] and the addition of bittering agents to small objects effectively prevents infants from ingesting them. In addition, bitter peptides have been found to offer medicinal benefits, such as regulating blood glucose,[3] acting as antioxidants,[4] and promoting cardiovascular health.[5] Given the crucial role of bitter peptides, it is vital to accurately determine which peptides exhibit bitterness. By doing so, pharmaceutical processes can be improved to reduce the bitterness of medications,[6] thereby encouraging patients to take their medications more willingly. In the food industry, manufacturers can achieve desired taste profiles by adjusting the content of bitter peptides.[7] However, the discovery and identification of bitter peptides is highly complex, time-consuming, and expensive. Therefore, it is essential to develop a model that can predict the bitterness of peptides. Fortunately, owing to detailed research conducted in the field of machine learning, powerful and cost-effective models are continuously being proposed.[8,9] These models have been widely applied across various fields and have significantly advanced the research and development.[10,11] Research on bitter peptides is no exception.

Various computational approaches based on quantitative structure–activity relationship (QSAR) modeling have emerged to predict peptide bitterness.[12–18] For instance, Yin et al.[13] devised multiple QSAR models to assess the bitterness levels in dipeptides by utilizing support vector regression (SVR) to analyze 48 angiotensin-converting enzyme (ACE) inhibitory dipeptides, 55 ACE inhibitory tripeptides, and 48 bitter dipeptides. They also introduced quantitative multidimensional amino acid descriptors E (E1-E5), where E1 through E5 denote the hydrophobicity, stereochemical attributes, side chain volume/molecular size, amino acid preferences in helices, composition, and net charge, respectively. In 2013, Soltani et al.[15] analyzed 229 experimental bitterness values of 224 peptides and 5 amino acids, focusing specifically on the bitterness threshold (log(1/T)), where T represents the bitterness threshold concentration (unit: M). They employed three machine learning methods—multiple linear regression (MLR), support vector machine (SVM), and artificial neural network (ANN)—for modeling. Subsequently, in 2019, Xu and Chung[16] proposed a QSAR model for predicting bitter peptides by integrating fourteen amino acid descriptors. Their cross-validation dataset comprised 48 dipeptides, 52 tripeptides, and 23 tetrapeptides along with their respective bitterness thresholds. The cross-validation results for dipeptides, tripeptides, and tetrapeptides were 0.941 ± 0.001, 0.742 ± 0.004, and 0.956 ± 0.002, respectively. These machine learning-based methods emerged relatively early, primarily because machine learning was not yet fully developed during its initial stages. However, compared to traditional biological experiments, machine learning models offer advantages such as lower costs, rapid screening capabilities, and enhanced interpretability. Nevertheless, these methods also present certain limitations. Firstly, the models are highly dependent on the distribution of the training data; if the training data is overly concentrated, the model's generalization ability significantly diminishes. Secondly, when confronted with bitter peptides that exhibit structural differences from those in the training set, the model's recognition capability is weakened. Additionally, in QSAR modeling, the selection of descriptors is crucial—an excessive number of descriptors may lead to model overfitting, while too few descriptors may fail to adequately capture the complex relationships between activity and structure. Given that many biological activities and molecular structures exhibit nonlinear relationships, traditional linear regression methods struggle to effectively capture these complexities, thereby necessitating the adoption of more sophisticated nonlinear models. In summary, although early machine learning approaches had certain shortcomings, advancements in technology have enabled the use of more complex nonlinear models, which can better elucidate the intricate relationships between biological activity and molecular structure, thereby enhancing the overall performance and applicability of the models.

In addition, a series of models have been developed for identifying bitter peptides based on sequence information, which means the compositions of amino acid. In 2020, Charoenkwan et al.[19] introduced iBitter-SCM, which uses the scoring card method (SCM) with estimated propensity scores of amino acids and dipeptides, and is the first model that was built based on sequence information. In 2021, Charoenkwan et al.[20] first applied deep learning to bitter peptide prediction and created a model named BERT4Bitter, which uses a BERT-based model that autonomously generates feature descriptors from raw peptide sequences without requiring systematic design and selection of feature encodings. In addition, three NLP-inspired feature encoding methods, namely, TFIDF, Pep2Vec, and FastText, were used to represent the peptide sequences. Just a few months later, in 2021, Charoenkwan et al.[21] used various types of feature encoding schemes and an optimal feature set, which were determined using the GA-SAR method and used as input for the SVM-based classifier. In 2022, Jiang et al.[22] introduced iBitter-DRLF, which primarily uses UniRep + BiLSTM_106 as the best fusion feature set, with additional selection using the LGBM classifier input. In 2023, Zhang et al.[23] introduced bitter-RF, which extracts ten types of feature information and uses 1206 features for model learning with RF after removing all zero items. Furthermore, we considered that in 2023, Yu et al.[24] developed a novel model leveraging an augmented dataset named BTP720, which was derived from the foundational BTP640 dataset. These models utilize sequence information as features, enabling the identification of latent characteristics within bitter peptide sequences through both traditional feature extraction methods and deep learning approaches. Consequently, the extracted features encompass more diverse and comprehensive information. However, considering that the functionality of bitter peptides involves site-specific binding, the use of sequence information alone neglects structural attributes.

We also observed that previously developed models primarily focused on feature extraction. However, during the feature extraction phase, most studies employed traditional methods; even when some models utilized deep learning[20] or leveraged deep learning for feature selection,[22] their outputs remained difficult to interpret. Moreover, the features extracted by prior models were relatively unidimensional, based either on sequence information or structural spatial information. Additionally, these models employed only a single classifier, which may result in the learning of unidimensional features and limit the broad applicability of the models.[25] Furthermore, certain evaluation metrics could be improved. Consequently, we developed a novel, cost-effective, and highly efficient model. Considering that researchers have elucidated the structure of bitter taste receptors[26] and explored the relationship between peptide bitterness and their physical and biochemical properties,[27] the features of our model incorporate both sequence information and structure-related information extracted from sequence data as comprehensively as possible through pre-trained models.

The final models we built relied entirely on the physical and biochemical properties of bitter peptides for feature extraction. We combined features extracted using the ESM-2[8,28] protein language model, which was trained on a masked language modeling objective, with features extracted using traditional methods that focused on protein biochemical properties. For our model, we implemented a stacking methodology that integrates three base models derived from decision trees (DT).[29] Utilizing SHapley Additive exPlanations (SHAP)[30] value analysis, we assessed the contributions of individual features. Validation by domain experts substantiated that our model accurately identifies the critical biochemical characteristics of bitter peptides. To ensure convenience for all users, particularly those who urgently need it, we established a freely available online web server at http://www.bioai-lab.com/iBitter-GRE. The code for building our model is available at https://github.com/EuclidLv/iBitter-GRE.

Materials and Methods

Overall framework of iBitter-GRE

As shown in Figure 1, there are four significant steps to build this model: dataset construction, feature representation, stacking model framework, webserver development. First, the same benchmark dataset was used as the first five models. Second, we used ESM-2, which is a well-known protein sequence pretraining model that efficiently represents amino acid sequences built based on the BERT framework, and it is trained through Masked Language Model Learning pretraining tasks on the UniProt dataset; in this study, the ESM-2, we used six layers and 8 M parameters. We also used seven different traditional feature descriptors to extract features, and then we used recursive feature elimination with cross-validation (RFECV)[31] to reduce the dimension of the features. The features extracted using these two steps were fused before the third step. Third, we selected GB, RF, and ETree as the base classifiers of the stacked model and LR as the meta-classifier. Finally, we established a global web server so that everyone who wants to use our model can access it.

Dataset

A reliable and high-quality benchmark dataset is essential, particularly since the first five models utilize the same benchmark dataset, called BTP640. This dataset is available at https://pmlab.pythonanywhere.com/BERT4Bitter (accessed on 20 May 2024), indicating that it has been validated through practical use and is widely recognized, providing a fair basis for comparison with all previous models. However, while the original dataset BTP640 was extensively validated, the additional data introduced by Yu et al. in the expanded dataset BTP720[24] seems to have added excessive noise, significantly declining the performance of other existing models. Furthermore, BTP720 has yet to gain wide adoption, and its validity and consistency require further verification. Therefore, we believe that using the original dataset BTP640 is more appropriate.

To construct the BTP 640 dataset, Charoenkwan et al.[19] manually curated experimentally validated bitter peptides from multiple research articles.[12–14,16–18,32–40] Peptide chains containing ambiguous residues, such as X, B, U, and Z, were excluded from the collected sequences. Additionally, to ensure experimental rigor, duplicate sequences were removed, resulting in a dataset of 320 unique and experimentally confirmed bitter peptide sequences. Given the limited scientific significance of experimentally verifying that certain peptides are definitively non-bitter, the collection of non-bitter peptide samples was relatively scarce. To enhance experimental accuracy, non-bitter peptides were gathered following established protocols from previous studies.[41–43] Specifically, 320 peptide sequences were randomly selected from the BIOPEP database[44] to construct the negative dataset, with these randomly generated sequences classified as non-bitter peptides. Subsequently, the dataset was randomly split into training and testing sets in an 8:2 ratio. The specific division of the training and testing sets is shown in Table 1, and the distribution of amino acids as well as the length distributions in the training and testing sets are illustrated in Figure 2. In terms of amino acid distribution, as shown in Figure 2(A), there are differences in the distribution of amino acids between the two training and testing sets. Specifically, amino acids such as Alanine (A), Aspartic acid (D), Glutamic acid (E), Histidine (H), Isoleucine (I), Lysine (K), Methionine (M), Asparagine (N), Serine (S), Valine (V), Tryptophan (W), and Tyrosine (Y) exhibit a higher distribution frequency in the training set, while amino acids

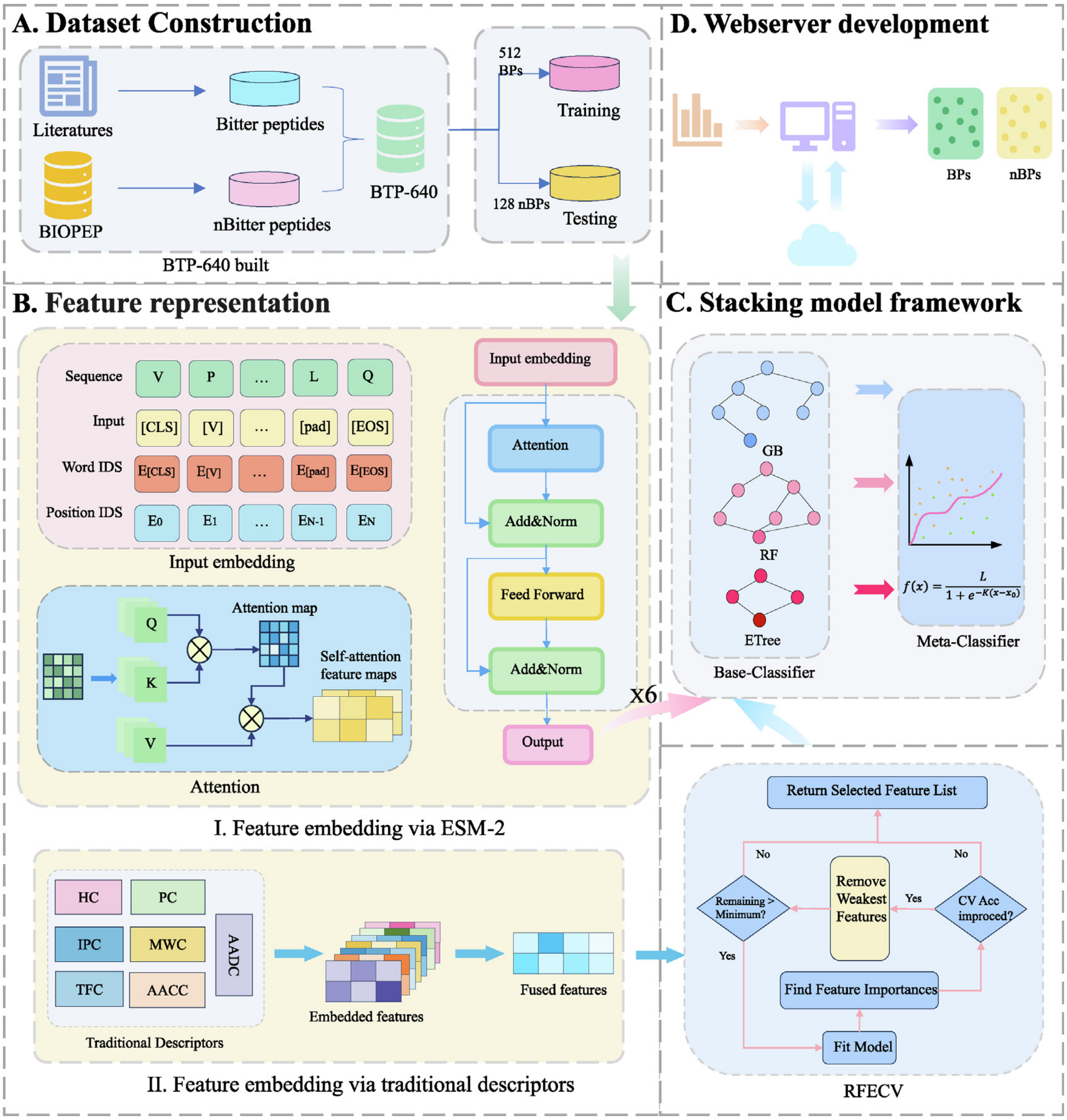

Figure 1. Overall framework of iBitter-GRE. (A) Dataset construction. (B) Feature representation. (C) Stacking model framework (D) Website development.

Table 1 Details of samples in BTP640.

	Positive	Negative	Total
Training set	256	256	512
Testing set	64	64	128

such as Cysteine (C), Phenylalanine (F), Glycine (G), Leucine (L), Proline (P), Glutamine (Q), Arginine (R), and Threonine (T) are more prevalent in the testing set. Nevertheless, both sets share a common feature: the distribution frequency of the amino acid P is significantly higher than that of other amino acids, while C has a notably lower distribution frequency compared to other amino acids. Figure 2 (B) presents the distribution of sample lengths in both the training and testing sets. In both sets, the majority of the samples are concentrated around 10 amino acids in length, which is consistent with the peptide-based dataset used in this study. Furthermore, the length of most samples does not exceed 15 amino acids.

Feature encoding

Two types of feature descriptors are used in this study. Some descriptors have been widely used for several years and are referred to as traditional descriptors. In this study, we calculated the molecular weight, hydrophobicity, polarity, isoelectric point, amino acid composition, transition frequency, and amino acid distribution. The details of the seven features are listed in Table 2.

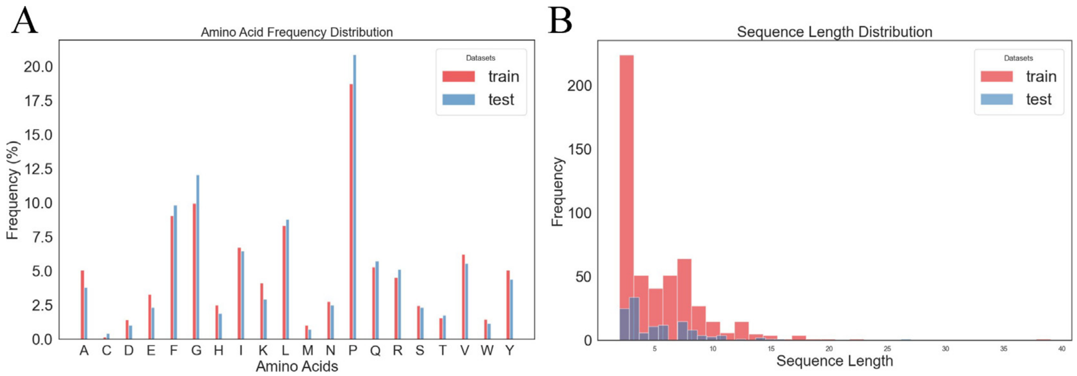

Figure 2. Distribution of training and testing dataset. (A) Amino acid frequency distribution in the training and testing sets. (B) Sequence length distribution in the training and testing sets.

Table 2 Summary of seven different traditional descriptors with their corresponding descriptions and dimensions.

Order	Descriptors	Description	Dimension	References
1	MWC	Calculates total molecular weight of an amino acid sequence	1	45
2	HC	Calculates average hydrophobicity index of an amino acid sequence	1	46
3	PC	Calculates proportion of polar and nonpolar amino acids in a sequence	2	47
4	IPC	Calculates isoelectric point (pI) of an amino acid sequence, the pH at which the molecule has no net charge	1	48
5	AACC	Calculates relative abundance of each amino acid in a sequence	20	49
6	TFC	Calculates frequency of adjacent amino acid pairs in a sequence	400	50
7	AADC	Calculates distribution of each amino acid in different segments (first, 25%; middle, 50%; and last, 25%) of a sequence	60	51

Second, we used ESM-2, which is a well-known protein sequence pre-training model that was built based on the BERT framework and is trained through Masked Language Model Learning pre-training tasks on the UniProt dataset. ESM-2 is an efficient model for representing amino acid sequences; the user can even improve performance by fine-tuning it on the required dataset and efficiently represent amino acid sequences by fine-tuning the ESM-2 model. Therefore, as a pre-trained model, ESM-2 provides several levels of the model, each with different parameters and layers. In this study, we used the first level of the model, which contained eight million parameters and six layers.

Stacking ensemble learning framework of iBitter-GRE

In this study, considering the singularity of the previous generations of model construction, we selected an ensemble learning approach. We ultimately selected stacking[52] as our ensemble method because of its ability to improve the prediction accuracy by integrating diverse models, enhancing the recognition capability of different data, reducing the risk of overfitting, and offering strong adaptability and robustness. The method involves two steps.

The first step involves separate input of the merged data into Random Forest (RF), Extra Trees (ETree), and Gradient Boosting (GB). The three base classifiers then output their results, which are subsequently used as new features for learning by the meta-classifier. The second step involves new features generated in the first step, which are input into the meta-classifier. The meta-classifier output the final prediction results of the model. In this study, we used LR as the meta-classifier. The details of the four models are provided in Table 3.

Performance evaluation strategies

To evaluate the performance of our model fairly and compare it with other former models, we used five common statistical metrics[57,58]: accuracy (ACC), Matthews correlation coefficient (MCC), area under the curve (AUC), sensitivity (Sn), and specificity (Sp). The equations for calculating these metrics are as follows:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{1}$$

$$\mathrm{MCC} = \frac{(\mathrm{TP} \times \mathrm{TN}) - (\mathrm{FP} \times \mathrm{FN})}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \tag{2}$$

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{4}$$

Table 3 Summary of the four models used with their descriptions.

Order	Models	Description	Reference
1	RF	RF is an ensemble learning method that builds multiple DT for making predictions. Each tree is trained on different subsets of the original data, and at each split, a random subset of features is considered. The final prediction is obtained by averaging the predictions of all the trees or by taking the majority vote.	53
2	GB	GB is an ensemble learning method that incrementally constructs DT. Each subsequent tree is trained on the residuals, or errors, of the predictions of prior trees. By focusing on these residuals, each tree attempts to correct the mistakes made by the preceding trees. The cumulative effect of the predictions of these trees is combined to form the final model, which iteratively minimizes the prediction error.	54
3	ETree	ETree is a variation of the RF algorithm. Similar to RF, it builds multiple DT but introduces additional randomness using random splits. At each node, a random subset of features and random thresholds for those features are selected to create the splits. The final prediction is derived from the average (in regression tasks) or the majority vote (in classification tasks) of the predictions from all trees.	55
4	LR	LR is a statistical model primarily used for binary classification tasks. It operates by fitting a logistic function to the data, which estimates the probability of an input belonging to a specific class. The logistic function generates output values ranging 0–1, indicating probabilities. A decision boundary is usually set at a probability threshold (conventionally 0.5), and predictions are made based on whether the data points fall above or below this threshold. LR also be adapted for multiclass classification using methods such as one-vs-rest or softmax regression.	56

where TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively. In addition, the areas under the receiver operating characteristic (ROC) curves were calculated.

Results and Discussion

Evaluation of feature descriptors for identifying bitter peptides

In this section, we evaluate our ensemble model through a series of experiments designed to assess its performance under different conditions. First, we examine various parameter configurations of ESM-2 by employing five evaluation metrics: ACC, AUC, Sn, Sp, MCC. This comprehensive evaluation allows us to determine the optimal settings for bitter peptide identification. Next, we investigate the impact of feature selection by analyzing the features retained before and after (RFECV. Acknowledging that features extracted using traditional descriptors may contain redundancy and noise detrimental to the model's performance, we applied RFECV with five-fold cross-validation (cv = 5) and a LR model post-feature extraction to mitigate these issues. Importantly, RFECV was applied exclusively to the traditional descriptors and not to the features derived from large models. This approach stems from the understanding that traditional descriptors may include redundant or irrelevant information that RFECV can effectively filter out, thereby enhancing model performance. In contrast, large models inherently incorporate mechanisms for feature selection during training, focusing on the most pertinent features. Additionally, the high dimensionality and intrinsic optimization present in features from large models make further manual selection less practical and potentially superfluous. We also assess the model's performance using different sets of features: those extracted by ESM-2, those derived from traditional descriptors, the traditional descriptors after RFECV, and a combination of all extracted features. This comparative analysis enables us to understand the contributions of each feature set to the overall performance and to identify the most effective combination for accurate bitter peptide identification.

Figure 3(A) presents a comparative analysis of ESM-2 models with varying parameter sizes for the identification of bitter peptides. The solid lines represent the mean results obtained from 10-fold cross-validation, while the dashed lines denote the median values. In terms of ACC, the results under the 8 M parameters are the most concentrated, followed by 150 M, 650 M, and 35 M. Overall, the results under the 8 M parameters exhibit the highest median and average, while the results under the 650 M parameters are slightly lower. The results under the 150 M parameters show that the median is almost identical to the average, suggesting that the cross-validation results mostly fall within a higher range. This indicates that the 8 M parameters are more likely to yield results with higher accuracy, while the results under the 150 M parameters are more stable. Regarding Sn, the results under the 8 M parameters are more concentrated, followed by 35 M, 150 M, and 8 M. Although the results under the 8 M parameters vary, they still show a higher average and median, indicating that the features extracted under these parameters allow the model to capture positive class samples more effectively. For Sp, the results under the 650 M parameters are the most

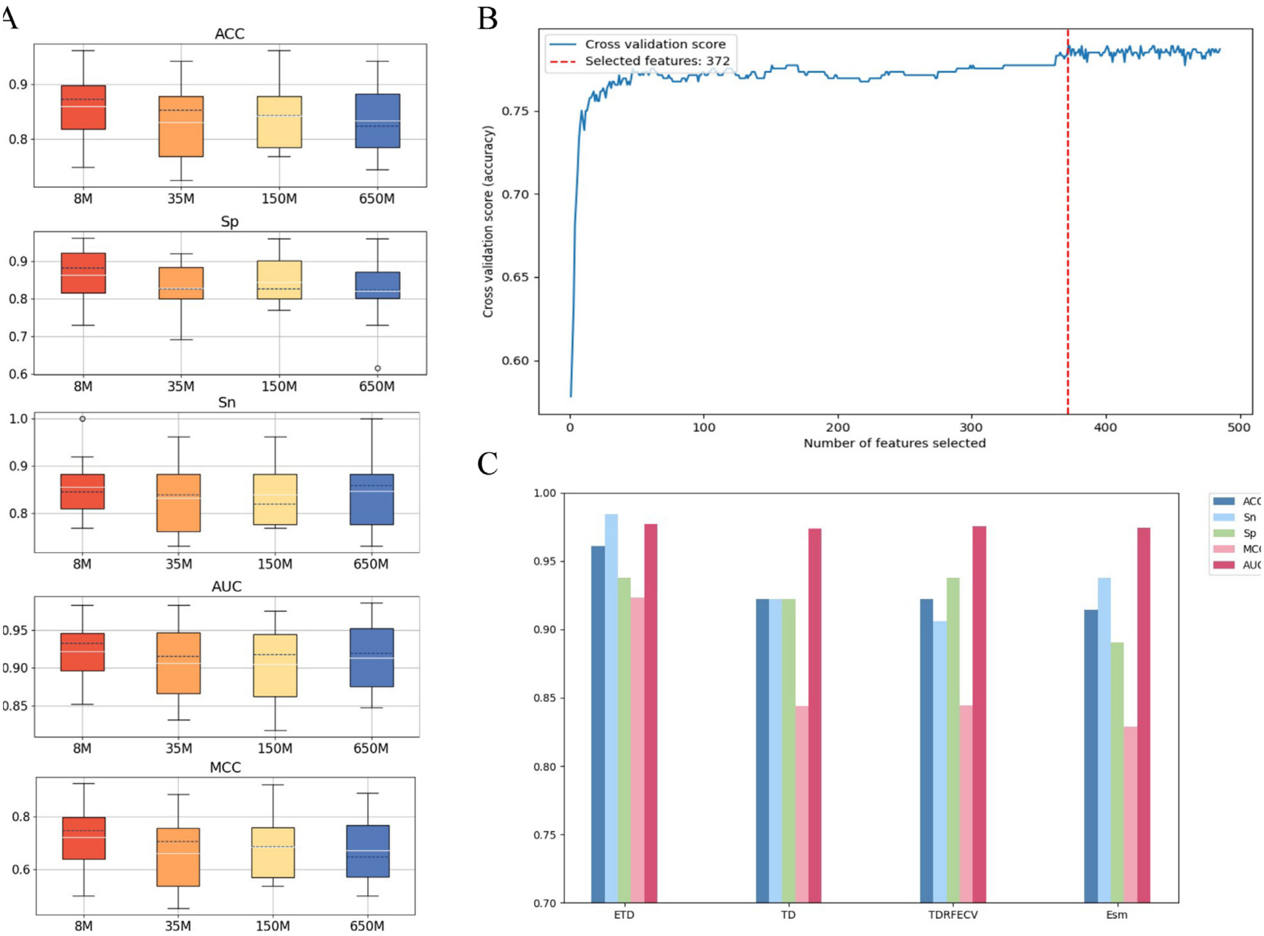

Figure 3. Assessment of feature descriptors for bitter peptide identification. (A) Selection of ESM-2 version. 8 M, 35 M, 150 M and 650 M represent for different version of ESM-2 with different parameter numbers. (B) RFECV plot of selection of traditional descriptors extracted features. (C) Evaluate of different features. The y-axis represents the scores. ETD represents the features extracted from ESM-2 and traditional descriptors, TD represents the features extracted from traditional descriptors, TDRFECV represents the features extracted from traditional descriptors and used RFECV, Esm represents the features extracted from ESM-2.

concentrated, followed by 35 M, 150 M, and 8 M. The average and median under the 8 M parameters remain high, showing that the features extracted under these parameters enable the model to capture positive class samples well, with the median also being higher under the 8 M version. Meanwhile, the results under the 650 M parameters have one outlier below the lower quartile (Q1). In the AUC section, the results under the 650 M parameters show that the median and average are nearly identical, suggesting that the model performs well and is stable under this condition. However, the results under the 8 M parameters show a better average value and higher median. We believe that with appropriate tuning, the results under the 8 M parameters could also achieve minimal differences between the median and average. For MCC, the results under the 8 M parameters still show a higher average and median. Despite some issues with the 8 M parameters, such as a wider distribution, two outliers in the prediction of positive class samples, and a significant difference between the median and average in AUC and MCC, the average and median values in 10-fold cross-validation remain higher across all evaluation metrics compared to the other parameters. Therefore, we ultimately selected ESM-2 with the 8 M parameters.

Typically, larger ESM-2 parameters indicate an increase in model size and complexity, and the features extracted using these parameters are expected to be more comprehensive, potentially better representing the 3D structure related information of proteins, which is crucial for peptides with properties generated through binding to receptors. However, in our experiments, the results indicated that the 8 M parameters version outperformed the 650 M parameters version. Considering the quantity of our data and the sequence lengths, we hypothesized that the limited data and relatively short sequences may have resulted in the 650 M version of ESM-2 extracting features that contain more noise and other detrimental factors. Additionally, during selection, we used the default parameters for each classifier, which may have led to suboptimal performance.

The RFECV plot is presented in Figure 3(B) This plot demonstrates that the final dimensionality of the features extracted using the traditional descriptors was 372. Detailed changes in the dimensions of each descriptor are listed in Table 4. We observed that the features extracted using the MWC descriptor were completely removed after RFECV. By contrast, features extracted with only one dimension (HC and IPC) and those extracted with two dimensions (PC) were not removed. This indicates that RFECV does not exclude descriptors based solely on the number of features they extract, although some descriptors extract a large number of features. This also suggests that molecular weight does not significantly contribute to distinguishing bitter peptides from hydrophobicity, polarity, or isoelectric point. In addition, the AACC features were not removed even in one dimension. The AADC descriptor initially had 60 dimensions, but only 2 were removed after RFECV, indicating that the frequency of amino acids and the relative frequency of each dipeptide combination in the protein sequence may influence bitterness. The TFC method extracted the largest number of dimensions (400) and removed the largest number (110). This is likely due to the varying lengths of proteins and the presence of specific amino acids that occur infrequently, resulting in dimensions with values that are predominantly zero, and thus not useful.

To further validate the effect of our integrated features compared to non-integrated features on model performance and to demonstrate the necessity of RFECV, we employed the integrated model to evaluate four feature sets: features extracted by ESM-2 (Esm), features extracted using traditional descriptors (TD), features extracted using traditional descriptors after RFECV (TDRFECV), and our integrated features (ETD). The results are shown in Figure 3(C). The four feature extraction methods did not significantly affect our model's performance in terms of AUC. However, in terms of ACC, our integrated features (ETD) improved by 4.06%, 4.06%, and 4.87% compared to TD, TDRFECV, and Esm, respectively. For Sn, ETD showed improvements of 6.35%, 7.99%, and 5.27% over TD, TDRFECV, and Esm. In Sp, ETD demonstrated enhancements of 1.66%, 0.00%, and 5.00% compared to TD, TDRFECV, and Esm, respectively. As for MCC, ETD improved by 8.57%, 8.53%, and 10.17% compared to TD, TDRFECV, and Esm. These results indicate that our integrated model achieved significant improvements across almost all of the selected evaluation metrics. We also observed that the features extracted using traditional descriptors after RFECV were better at identifying negative samples, while the features extracted by ESM-2 were more adept at recognizing positive samples. By integrating both feature sets, the new features became proficient in identifying both positive and negative samples. This ability to balance between recognizing both types of samples may be a key reason why the integrated features performed better overall.

Evaluation of robust classifiers for stacking model

In this section, we explore the performances of different classifiers using the same set of features and analyze their impact on the model when combined in a stacking architecture.

We first used 13 classifiers, LR, ridge classifier (RC), DT, RF, ETree, GB, Gaussian naive Bayes (GNB), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), SVM, linear support vector (LSVC), K-nearest neighbors (KNN), and nearest centroid (NC), and the features were extracted using the ESM-2 8 M parameter version. All classifiers used the default parameters, except for LR using max-iter = 500, RC using max-iter = 500, KNN using n_neighbors = 5, and RN using radius = 10.0, to allow them to work properly. Considering that some classifiers did not support the AUC calculation, we only used ACC, MCC, Sn, and Sp as evaluation metrics. The results are shown in Figure 4(A) We observed that the differences in the ACC, Sn, and Sp among these classifiers

Table 4 Summary of changes in the dimensions of features extracted using seven descriptors before and after RFECV.

Order	Descriptors	Original dimensions	Removed dimensions	Final dimensions
1	MWC	1	1	0
2	HC	1	0	1
3	PC	2	0	2
4	IPC	1	0	1
5	AACC	20	0	20
6	TFC	400	110	290
7	AADC	60	2	58

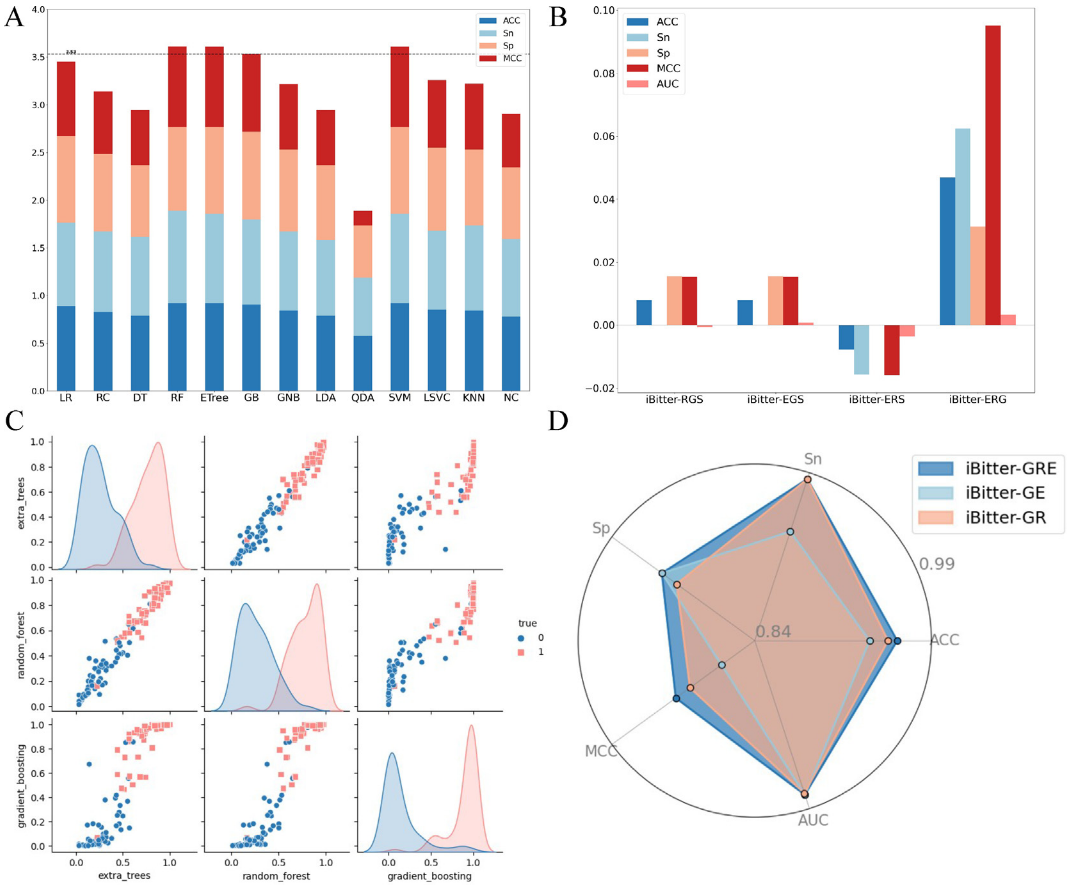

Figure 4. Assessment of classifiers for bitter peptide identification. (A) Evaluation of classifiers. The x-axis represents different models name, and the y-axis represents the scores. (B) The impact of ensemble classifiers on model performance. iBitter-RGS represents the model using RF, GB, and SVM as the base classifiers, iBitter-EGS represents the model using ETree, GB, and SVM as the base classifiers, iBitter-ERS represents the model using ETree, RF, and SVM as the base classifiers, iBitter-ERG represents the model using ETree, RF, and GB as the base classifiers. (C) Scatter plot matrix shows the relationships between the three base classifiers. (D) Radar chart shows the impact of two classifiers on model performance. iBitter-RGE represents the model using RF, GB, and ETree as the base classifiers, iBitter-GE represents the model using GB, and ETree as the base classifiers, iBitter-GR represents the model using GB, and RF as the base classifiers.

were minimal. However, when aggregating these metrics along with the MCC values, it was clear that RF, ETree, GB, and SVM were the top performers among the 13 classifiers. Although GB had the lowest aggregate score (3.53), it significantly outperformed the other nine classifiers. Therefore, considering that stacking aims to combine the strengths of the best-performing models, we used these four models as the base classifiers.

To further investigate the influence of these four classifiers on the model's prediction results, we take the output when all four classifiers (RF, ETree, GB, and SVM) are included as the baseline, and then remove each classifier one by one, leaving the remaining three. The results are compared to the baseline, and the differences are presented in Figure 4(B). When RF, ETree, GB, and SVM are used as the base classifiers, a baseline is established for comparison. When GB was removed, all model metrics, except for specificity, decreased. Conversely, when RF, ETree, or SVM were individually removed, most model metrics improved (ACC increased by 0.0469, Sn by 0.0624, Sp by 0.0313, MCC by 0.908, and AUC by 0.0034). Notably, removing SVM resulted in a significant improvement in the metrics. This indicates that GB is an essential component of the original model, while the presence of SVM negatively impacts model performance. Since the parameters used in these models have not been further optimized, some factors negatively affecting overall model performance can be improved through methods such as grid search. However, although significant performance declines can be optimized, this may also indicate the occurrence of overfitting. We consider that while ESM-2 is excellent at feature extraction, the extracted features may still contain noisy data, and SVM struggles to handle this noise effectively. Therefore, we ultimately selected RF, ETree, and GB as the base classifiers because this combination provided the most robust and reliable performance.

The primary objective of the stacking model is to enhance overall predictive performance and robustness by combining the prediction results of multiple diverse base learners. Selecting base classifiers requires thorough screening across a wide range of models. However, the final selection of RF, ETree, and GB as base classifiers, while effective, seems to contradict this principle, particularly with regard to RF and ETree. To further understand the relationships between these three base classifiers, we visualized their interactions using a scatter matrix plot, as shown in Figure 4(C) As expected, there was a significant correlation between RF and ETree, while GB showed no significant correlation with either RF or ETree. To investigate this phenomenon further, we conducted experiments under the same conditions, systematically removing RF and ETree from the base classifiers to observe the model's performance across various metrics (ACC, AUC, MCC, Sn, and Sp) compared to the model with all three classifiers retained. The results in Figure 4 (D) show that retaining all three models resulted in the best performance across all metrics (with ACC at 0.9609, Sn at 0.9844, Sp at 0.9375, MCC at 0.9229, and AUC at 0.9776). Removing RF led to a significant drop in Sn and MCC (with Sn decreasing by 4.76% and MCC decreasing by 5.19%), while removing ETree resulted in declines in Sp, MCC, and ACC (with Sp decreasing by 1.66%, MCC by 1.61%, and ACC by 0.81%). This indicates that despite the correlation between RF and ETree, they complement each other and contribute significantly to the final predictions of the model. Therefore, RF and ETree are indispensable components of the ensemble model.

Performance analysis

To better evaluate the existing model, we conducted a grid search on the ensemble model and introduced a test set that underwent the same feature-extraction method as the training set. We analyzed the model in detail from several perspectives: the prediction probabilities of different categories of data in the test set, the ROC curve, and the interactions and roles of the three base classifiers in the ensemble.

The results are presented in Figure 5(A) Violin plots were used to visualize the prediction probabilities of the model for different categories in the test set. The prediction probabilities for Class 0 (non-bitter peptides) were primarily concentrated at 0–0.4, with a median of approximately 0.1. For Class 1 (bitter peptides), the prediction probabilities primarily ranged 0.6–1, with a median of approximately 0.9. This distribution indicates that the model has a high degree of separation between the two categories. Additionally, we observed that the distribution for Class 1 was more concentrated than that for Class 0, suggesting that the model's predictions for Class 1 were more accurate and reliable. Conversely, the prediction probabilities for Class 0 are relatively dispersed. This implied that our model was better at identifying potentially bitter peptides than non-bitter peptides. Figure 5(B) illustrates the ROC curves for each fold in the 10-fold cross-validation, the average ROC curve from the 10-fold cross-validation, and the ROC curve for the test set. The results showed that the AUC value for fold 1 was 0.984, which was the highest among all folds, indicating excellent model performance. The AUC value for fold 7 was 0.897, the lowest among all folds, but still close to 0.9, indicating good model performance across all folds. The AUC values for the other folds ranged 0.902–0.937, demonstrating consistent model performance across the different folds. The AUC value on the test set is 0.978, indicating very good model performance on the test set as well. We conclude that the differences in AUC values between the cross-validation folds are minimal, indicating stable model performance and the minimal impact of different data splits on the model. The model performed well in both the cross-validation and test sets, demonstrating high classification capability.

Compare with existing methods

In this section, we use the same training and independent datasets as in the previous models to assess and compare the predictive performance of iBitter-GRE with the existing models.[19–24] The performance comparison results for the test set are presented in Table 5 and Figure 6. The results of the 10-fold cross-validation are presented in Supplementary Table S1 and Supplementary Figure S1.

Compared to the other six models, our model achieved significant improvements in five metrics: ACC, AUC, Sn, and MCC. The ACC reached 0.961 for the first time, which is an improvement of 1.7% over the best previous result. The Sn and MCC improved by 4.9% and 3.5%, respectively. Considering that previous models had already achieved commendable performance, these improvements, although relatively small, were particularly remarkable.

The significant improvement in our model can be attributed to three major factors. First, we utilize the powerful LLM ESM-2 for feature extraction. Second, in terms of features, we followed the existing research literature and exclusively used descriptors capable of extracting structural, physical, and biochemical information on bitter peptides. Finally, we employ the stacking

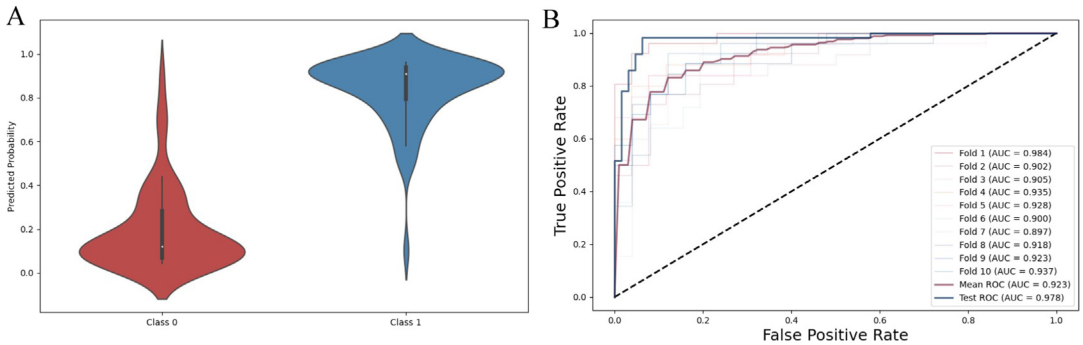

Figure 5. Performance of Ensembled model. (A) Violin plot of the prediction results of the model for different classes on the test set. Class 0 represents non-bitter peptides, and Class 1 represents actual bitter peptides. (B) ROC curves for the performance of the model on 10-fold cross-validation, 10-fold average, and test set. Mean ROC represents the average ROC curve from 10-fold cross-validation, Test ROC represents the ROC curve on the independent test set.

Table 5 Performance comparison of iBitter-GRE with existing methods.

Model	ACC	AUC	Sn	Sp	MCC
iBitter-SCM	0.844	0.904	0.844	0.844	0.688
iBitter-Fuse	0.930	0.933	0.938	0.922	0.859
Bert4Bitter	0.922	0.964	0.938	0.906	0.844
iBitter-DRLF	0.945	0.977	0.922	**0.969**	0.892
Bitter-RF	0.938	**0.978**	0.938	0.938	0.875
CPM-BP (BTP640)	0.836	0.836	0.773	0.903	0.680
iBitter-GRE	**0.961**	**0.978**	**0.984**	0.938	**0.923**

ensemble learning method and integrate multiple-base classifiers to allow the model to learn features from as many perspectives as possible.

Interpreting feature importance using SHAP analysis

In this section, we employ SHAP[30] analysis to assess the impact of the extracted features. The analysis is conducted from two perspectives: the most significant SHAP value for each individual fea-

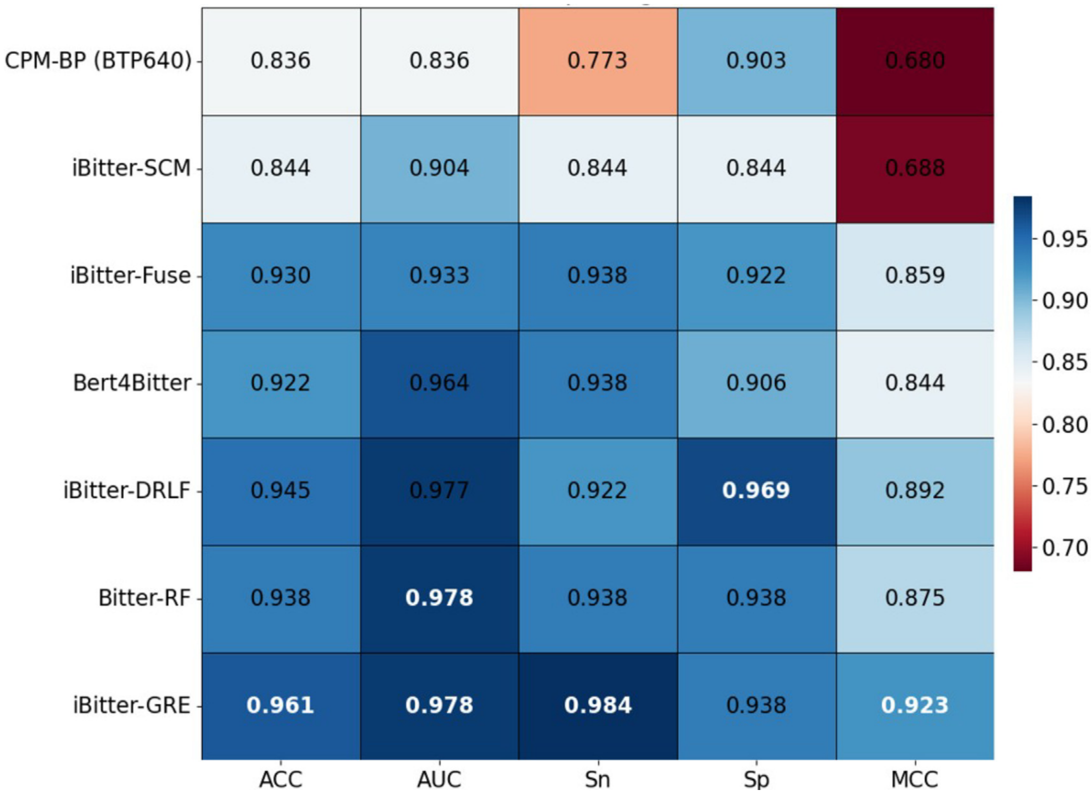

Figure 6. Heatmap of comparison with the existing model under different metrics.

ture and the most significant SHAP values representing the average importance of features. By analyzing the top 30 most significant features, we can identify those that have the greatest influence on the model's predictions, providing insights into their contribution to the overall performance of the model. Despite ESM-2 being a pre-trained large language model, the specific meanings of each extracted feature dimension remain unclear. However, given that these features constitute a significant component of our model, we have incorporated them into our SHAP analysis. Additionally, by integrating ESM-2 features, we aim to gain a deeper understanding of whether the features extracted by the large language model exert a substantial and positive influence on our model's predictive performance and whether they significantly impact each individual sample. These efforts allow us to intuitively appreciate the improvements brought by the introduction of ESM-2 to our model, rather than relying solely on experimental data.

Figure 7(A) displays the top 30 most significant SHAP values for the average importance of features. To facilitate the analysis, the samples have been arranged in a sequence from positive to negative. This ordering allows for a clearer examination of the SHAP value distribution and the relative impact of each feature across the sample categories, offering insights into how these features differentiate between positive and negative samples. The most impactful features, as indicated by the prominent red and blue regions, provide insights into the model's prediction dynamics: red indicates a positive contribution (i.e., an increase in the probability of the predicted outcome), while blue represents a negative contribution (i.e., a decrease in the predicted probability). The majority of the top 30 most significant features are derived from ESM-2 embeddings, with esm_feature_11, esm_feature_235, esm_feature_177, and esm_feature_342 emerging as particularly influential. These features exhibit consistent contributions across a wide range of samples, indicating their substantial role in guiding model predictions. Additionally, biochemical properties such as hydrophobicity, isoelectric point, and transition-based features like transition_L_P and distribution_L_middle_50% are also highly impactful. These features, grounded in biochemical and structural properties, contribute valuable domain-specific information that enhances the model's ability to differentiate between classes. Their presence among the top features highlights the interplay between sequence-derived embeddings and fundamental biochemical characteristics in driving the model's overall predictive performance. We also visualized the combined SHAP values of the remaining features for each sample. Although the cumulative SHAP values of these features are significantly higher than those of the top features, when examined individually, each feature's contribution is far less impactful compared to the top 30 features. This highlights that while the aggregated effect of the remaining features may be substantial, their individual influence on the model's predictions is considerably weaker than that of the top-ranked features.

Further analysis reveals that while ESM-2 is a powerful protein feature extraction tool and dominates the top 30 most important features, not all ESM-2 features play a critical role in the output for every sample. For instance, features such as esm_feature_183, esm_feature_17, and esm_feature_110, although showing some contribution (indicated by red or blue) in a few samples, have negligible or invisible contributions for most samples (represented by white). Additionally, some ESM-2 features exhibit substantial contributions only for specific samples. For example, esm_feature_179 plays a crucial role in predicting a few negative samples, while esm_feature_120 is significant for predicting a few positive samples. However, it is undeniable that certain top ESM-2 features, such as esm_feature_11, esm_feature_235, and esm_feature_177, consistently make significant contributions across the overall sample set. Moreover, the impact of these key features is largely aligned with the actual outcomes, further underscoring their importance in model predictions. The features extracted using traditional biochemical properties, such as transition_L_P, hydrophobicity, distribution_L_middle_50%, and isoelectric_point, exhibit varying contributions to sample predictions. transition_L_P represents the transition frequency from leucine to proline, and although it generally leads to accurate classifications, the visualizations show a tendency to predict negative outcomes. This could suggest that bitterness in peptides might be related to the transition frequency between leucine and proline. Hydrophobicity, calculated using the Kyte-Doolittle scale[59] for each amino acid in a protein sequence, shows a tendency to predict positive outcomes, which may indicate that the hydrophobic index plays a significant role in the bitterness of peptides. distribution_L_middle_50% reflects the frequency of leucine occurring in the middle 50% of the protein sequence. Despite having little influence on the predictions for most samples—likely due to the absence of leucine in the middle 50% of shorter sequences—this feature tends to predict negative outcomes. Combined with the data from transition_L_P, we can speculate that leucine may be closely linked to the mechanism behind bitterness in peptides. Finally, isoelectric_point, which calculates the isoelectric point of the protein sequence, tends to predict positive outcomes. Historically, the isoelectric point has been used as a research criterion

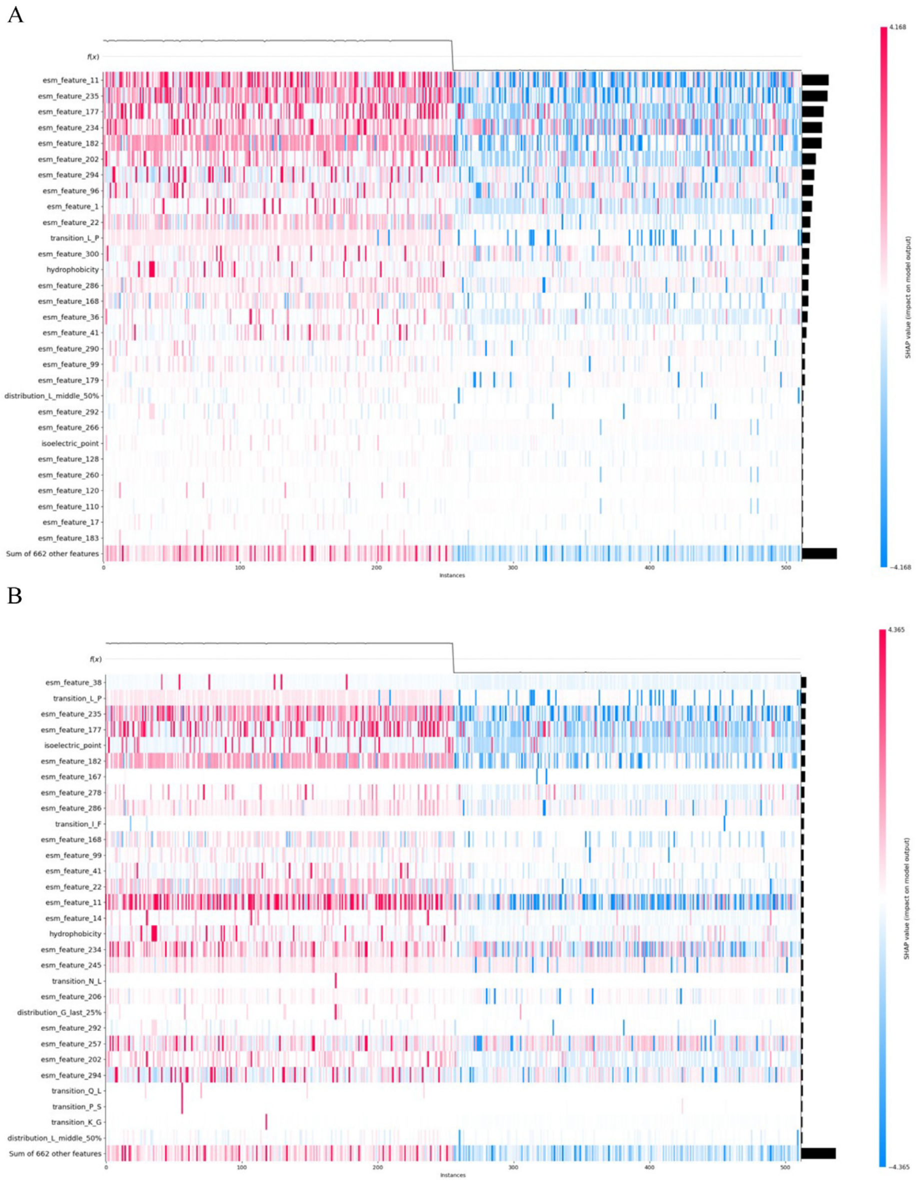

Figure 7. Feature importance analysis by SHAP values. (A) SHAP values heatmap focusing on the most significant 30 SHAP value for average importance of features. (B) SHAP values heatmap focusing on the most significant 30 SHAP value for each feature. I Instances are arranged in the order from positive samples to negative samples both (A) and (B) for better analysis.

for identifying bitter peptides, suggesting its continued relevance in distinguishing between bitter and non-bitter peptides.

Considering that certain features may significantly impact the predictions for specific samples, we generated another SHAP heatmap focusing on the most significant 30 SHAP values for each feature, as shown in Figure 7(B) Compared to the results in Figure 6(A), the importance ranking of features extracted by ESM-2 has changed significantly, while the number of features extracted using biochemical properties has greatly increased. The reason for this phenomenon may be that not every amino acid is present in the sample sequences, leading to limitations when using biochemical properties as feature extraction methods. Nevertheless, it is undeniable that features extracted using biochemical properties tend to be more specific.

For example, transition_I_F, transitionN_L, transition_Q_L, transition_P_S, and transition_K_G represent the transition frequencies from isoleucine to phenylalanine, from asparagine to leucine, from glutamine to leucine, from proline to serine, and from lysine to glycine, respectively. Among them, transition_I_F and transition_L_P are more inclined to predict samples with high values as negative, while transitionN_L, transition_Q_L, transition_P_S, and transition_K_G tend to predict samples as positive. This aligns with the view that the presence of leucine, isoleucine, phenylalanine, lysine, or valine at either the C- or N-terminus increases the likelihood that a peptide will be bitter. The bitterness intensity is magnified if a basic amino acid (such as arginine, lysine, or histidine) is at the N-terminus and a hydrophobic amino acid is simultaneously at the C-terminus.[60] On the other hand, we found that hydrophobicity tends to classify samples with high hydrophobicity values as bitter peptides, which can be corroborated by the earlier use of the Q-rule states in classifying bitter peptides.[61,62] The isoelectric_point tends to predict bitter peptides, indicating that they may have a higher isoelectric point, meaning they are electrically neutral (equal positive and negative charges) at a relatively high pH. Research has shown that cheese produced from milk with a lower pH is more bitter.[27] The biochemical properties we used have been confirmed in real-world studies,[27,59–62] not only reflecting the validity of the features we employed and explaining why these features are feasible but also further verifying the scientific basis of using these properties to judge bitterness.

Conclusions

In the present study, we introduced iBitter-GRE, a novel stacked ensemble learning methodology designed for the precise identification of bitter peptides. The approach commenced with the extraction of features using seven traditional descriptors and was subsequently refined using RFECV. Furthermore, we integrated advanced features derived from the ESM-2 8 M model, which were combined with the initial dataset to enhance the robustness and feature representation of the model. The resultant feature set was input into an ensemble model comprising RF, ETree, and GB as base classifiers, and LR as the meta-classifier. This configuration was systematically evaluated against six established methodologies in the field, iBitter-SCM, Bert4Bitter, iBitter-Fuse, iBitter-DRLF, Bitter-RF, and CPM-BP(BTP640), on a curated test dataset. iBitter-GRE demonstrated superior predictive performance, as evidenced by the following metrics: ACC, 0.961; Sn, 0.984; Sp, 0.938; MCC, 0.923; andAUC, 0.978. These results underscore the efficacy and broad applicability of the method for the identification of bitter peptides, positioning iBitter-GRE as a significant advancement in this field.

Despite these promising outcomes, our model still exhibits certain limitations. Specifically, the limited number of training samples may result in the model being adept only at recognizing bitter peptides from specific sources. Although our model employs a comprehensive range of feature extraction methods to obtain features that encompass both structural-related and biochemical information, the ESM-2 parameter versions used do not accurately reconstruct the precise structure-related information of proteins. This limitation implies that increasing the length of protein sequences may lead to deviations in prediction results. Additionally, the descriptors employed for extracting biochemical information may not be entirely comprehensive, potentially causing some non-bitter peptides to be erroneously classified as bitter peptides due to shared biochemical properties with our training samples.

To address these challenges, we plan to collaborate with biological sequencing research teams to acquire a larger and more diverse set of validated bitter peptides through wet lab experiments and extensive literature reviews. Expanding our training dataset in this manner will enable the model to learn a broader and more generalizable range of features associated with bitter peptides, thereby enhancing its generalization capability. Furthermore, with the introduction of ESM-3,[63] which offers more refined and accurate protein structure-related representations using fewer parameters and reduced computational time, we aim to integrate this improved version to further optimize our model's performance. In addition, regarding traditional descriptors, we intend to work closely with bitter peptide research teams to accurately identify which feature extraction methods are both precise and effective for predicting bitter peptides. This collaboration will also help us discern which historically used methods may no longer provide accurate predictions. By implementing these validated methods through computer programming, we aim to make our feature engineering process more scientific and reliable, thereby strengthening the overall robustness and accuracy of our prediction model.

Data availability

The work and source code are available to researchers and developers at https://github.com/EuclidLv/iBitter-GRE. We also established a freely available online web server at http://www.bioai-lab.com/iBitter-GRE.

CRediT authorship contribution statement

Jingwei Lv: Writing – original draft, Visualization, Software, Methodology, Funding acquisition, Data curation. **Aoyun Geng:** Writing – review & editing, Visualization, Validation, Methodology. **Zhuoyu Pan:** Project administration, Methodology, Investigation. **Leyi Wei:** Resources, Methodology, Conceptualization. **Quan Zou:** Validation,

Resources, Methodology. **Zilong Zhang:** Writing – original draft, Visualization, Validation, Resources, Project administration, Funding acquisition, Conceptualization. **Feifei Cui:** Writing – review & editing, Resources, Project administration, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Funding

The work was supported by the National Natural Science Foundation of China (Nos. 62450002, 62262015, 62102064, 62322112) and Science and Technology special fund of Hainan Province (ZDYF2024GXJS018) and Internal Research Grants of Macao Polytechnic University (no. RP/CAI02/2023) and the Science and Technology Development Fund of Macau (no. 0177/2023/RIA3).

DATA AVAILABILITY

The code and data used in the experiment can be accessed through the GitHub link mentioned in the article.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary material to this article can be found online at https://doi.org/10.1016/j.jmb.2025.169005.

Received 15 November 2024;
Accepted 10 February 2025;
Available online 13 February 2025

Keywords:
bitter peptide;
sequence information;
machine learning;
stacking;
classification

References

1. Smail, H.O., (2019). The roles of genes in the bitter taste. *AIMS Genet.* **6** (04), 088–097.
2. Mennella, J.A., Spector, A.C., Reed, D.R., Coldwell, S.E., (2013). The bad taste of medicines: overview of basic research on bitter taste. *Clin. Ther.* **35** (8), 1225–1246.
3. Hsu, P.-K., Pan, F.F., Hsieh, C.-S., (2020). McIRBP-19 of bitter melon peptide effectively regulates diabetes mellitus (DM) patients' blood sugar levels. *Nutrients* **12** (5), 1252.
4. Muhialdin, B.J., Rani, N.F.A., Hussin, A.S.M., (2020). Identification of antioxidant and antibacterial activities for the bioactive peptides generated from bitter beans (Parkia speciosa) via boiling and fermentation processes. *Lwt* **131**, 109776.
5. Hernandez, D.F., Mojica, L., de Mejia, E.G., (2024). Legume-derived bioactive peptides: role in cardiovascular disease prevention and control. *Curr. Opin. Food Sci.*101132.
6. Chu, X., Zhu, W., Li, X., Su, E., Wang, J., (2024). Bitter flavors and bitter compounds in foods: identification, perception, and reduction techniques. *Food Res. Int.*114234.
7. Knudsen, L.J., Lokerse, I., Pedrotti, M., Nielsen, S.D.-H., Dekker, P., Rauh, V., Fogliano, V., Larsen, L.B., (2024). Differences in development of volatiles and bitter peptides between pre-and post-hydrolysed lactose-free UHT milk during storage. *Int. Dairy J.* **157**, 106029.
8. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379** (6637), 1123–1130.
9. Varadi, M., Bertoni, D., Magana, P., Paramval, U., Pidruchna, I., Radhakrishnan, M., Tsenkov, M., Nair, S., Mirdita, M., Yeo, J., (2024). AlphaFold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res.* **52** (D1), D368–D375.
10. Fu, X., Duan, H., Zang, X., Liu, C., Li, X., Zhang, Q., Zhang, Z., Zou, Q., Cui, F., (2024). Hyb_SEnc: an antituberculosis peptide predictor based on a hybrid feature vector and stacked ensemble learning. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 1–17.
11. Duan, H., Zhang, Y., Qiu, H., Fu, X., Liu, C., Zang, X., Xu, A., Wu, Z., Li, X., Zhang, Q., et al., (2024). Machine learning-based prediction model for distant metastasis of breast cancer. *Comput. Biol. Med.* **169**, 107943.
12. Tong, J., Liu, S., Zhou, P., Wu, B., Li, Z., (2008). A novel descriptor of amino acids and its application in peptide QSAR. *J. Theor. Biol.* **253** (1), 90–97.
13. Yin, J., Diao, Y., Wen, Z., Wang, Z., Li, M., (2010). Studying peptides biological activities based on multidimensional descriptors (E) using support vector regression. *Int. J. Pept. Res. Ther.* **16**, 111–121.
14. Liang, G., Yang, L., Kang, L., Mei, H., Li, Z., (2009). Using multidimensional patterns of amino acid attributes for QSAR analysis of peptides. *Amino Acids* **37**, 583–591.
15. Soltani, S., Haghaei, H., Shayanfar, A., Vallipour, J., Asadpour Zeynali, K., Jouyban, A., (2013). QSBR study of bitter taste of peptides: application of GA-PLS in combination with MLR, SVM, and ANN approaches. *Biomed. Res. Int.* **2013**, (1)501310
16. Xu, B., Chung, H.Y., (2019). Quantitative structure–activity relationship study of bitter di-, tri-and tetrapeptides using integrated descriptors. *Molecules* **24** (15), 2846.
17. Kim, H.-O., Li-Chan, E.C., (2006). Quantitative structure–activity relationship study of bitter peptides. *J. Agric. Food Chem.* **54** (26), 10102–10111.
18. Wu, J., Aluko, R.E., (2007). Quantitative structure-activity relationship study of bitter di-and tri-peptides including relationship with angiotensin I-converting enzyme inhibitory activity. *J. Peptide Sci.: Off. Publ. Eur. Peptide Soc.* **13** (1), 63–69.

19. Charoenkwan, P., Yana, J., Schaduangrat, N., Nantasenamat, C., Hasan, M.M., Shoombuatong, W., (2020). iBitter-SCM: identification and characterization of bitter peptides using a scoring card method with propensity scores of dipeptides. *Genomics* **112** (4), 2813–2822.
20. Charoenkwan, P., Nantasenamat, C., Hasan, M.M., Manavalan, B., Shoombuatong, W., (2021). BERT4Bitter: a bidirectional encoder representations from transformers (BERT)-based model for improving the prediction of bitter peptides. *Bioinformatics* **37** (17), 2556–2562.
21. Charoenkwan, P., Nantasenamat, C., Hasan, M.M., Moni, M.A., Lio', P., Shoombuatong, W., (2021). iBitter-fuse: a novel sequence-based bitter peptide predictor by fusing multi-view features. *Int. J. Mol. Sci.* **22** (16), 8958.
22. Jiang, J., Lin, X., Jiang, Y., Jiang, L., Lv, Z., (2022). Identify bitter peptides by using deep representation learning features. *Int. J. Mol. Sci.* **23** (14), 7877.
23. Zhang, Y.-F., Wang, Y.-H., Gu, Z.-F., Pan, X.-R., Li, J., Ding, H., Zhang, Y., Deng, K.-J., (2023). Bitter-RF: a random forest machine model for recognizing bitter peptides. *Front. Med.* **10**, 1052923.
24. Yu, Y., Liu, S., Zhang, X., Yu, W., Pei, X., Liu, L., Jin, Y., (2024). Identification and prediction of milk-derived bitter taste peptides based on peptidomics technology and machine learning method. *Food Chem.* **433**, 137288.
25. Liu, R., Fu, X., Yan, S., Zhang, Z., Cui, F., (2023). AIPPT: predicts anti-inflammatory peptides using the most characteristic subset of bases and sequences by stacking ensemble learning strategies. *In: 2023 IEEE international conference on bioinformatics and biomedicine (BIBM): 5-8 Dec. 2023*, pp. 23–29.
26. Xu, W., Wu, L., Liu, S., Liu, X., Cao, X., Zhou, C., Zhang, J., Fu, Y., Guo, Y., Wu, Y., (2022). Structural basis for strychnine activation of human bitter taste receptor TAS2R46. *Science* **377** (6612), 1298–1304.
27. Kuhfeld, R., Eshpari, H., Atamer, Z., Dallas, D., (2023). A comprehensive database of cheese-derived bitter peptides and correlation to their physical properties. *Crit. Rev. Food Sci. Nutr.*, 1–15.
28. Xiao, C., Zhou, Z., She, J., Yin, J., Cui, F., Zhang, Z., (2024). PEL-PVP: Application of plant vacuolar protein discriminator based on PEFT ESM-2 and bilayer LSTM in an unbalanced dataset. *Int. J. Biol. Macromol.* **277**, 134317.
29. Song, Y.-Y., Ying, L., (2015). Decision tree methods: applications for classification and prediction. *Shanghai Arch. Psychiatry* **27** (2), 130.
30. Li, Z., (2022). Extracting spatial effects from machine learning model using local interpretation method: an example of SHAP and XGBoost. *Comput. Environ. Urban Syst.* **96**, 101845.
31. Mustaqim, A.Z., Adi, S., Pristyanto, Y., Astuti, Y., (2021). The effect of recursive feature elimination with cross-validation (RFECV) feature selection algorithm toward classifier performance on credit card fraud detection. *In: 2021 International conference on artificial intelligence and computer science technology (ICAICST)*. IEEE, pp. 270–275.
32. Matoba, T., Hata, T., (1972). Relationship between bitterness of peptides and their chemical structures. *Agric. Biol. Chem.* **36** (8), 1423–1431.
33. Ishibashi, N., Kubo, T., Chino, M., Fukui, H., Shinoda, I., Kikuchi, E., Okai, H., Fukui, S., (1988). Taste of proline-containing peptides. *Agric. Biol. Chem.* **52** (1), 95–98.
34. Nosho, Y., Otagiri, K., Shinoda, I., Okai, H., (1985). Studies on a model of bitter peptides including arginine, proline and phenylalanine residues. II. 1) bitterness behavior of a tetrapeptide (Arg-Pro-Phe-Phe) and its derivatives. *Agric. Biol. Chem.* **49** (6), 1829–1837.
35. Vasylenko, T., Liou, Y.-F., Chen, H.-A., Charoenkwan, P., Huang, H.-L., Ho, S.-Y., (2015). SCMPSP: prediction and characterization of photosynthetic proteins based on a scoring card method. *In: BMC bioinformatics*. Springer, pp. 1–16.
36. Huang, H.-L., (2014). Propensity scores for prediction and characterization of bioluminescent proteins from sequences. *PLoS One* **9**, (5)e97158
37. Ishibashi, N., Sadamori, K., Yamamoto, O., Kanehisa, H., Kouge, K., Kikuchi, E., Okai, H., Fukui, S., (1987). Bitterness of phenylalanine-and tyrosine-containing peptides. *Agric. Biol. Chem.* **51** (12), 3309–3313.
38. Ney, K.H., (1979). Bitterness of peptides: amino acid composition and chain length. ACS Publications.
39. Ishibashi, N., Arita, Y., Kanehisa, H., Kouge, K., Okai, H., Fukui, S., (1987). Bitterness of leucine-containing peptides. *Agric. Biol. Chem.* **51** (9), 2389–2394.
40. Lin, Z.-H., Long, H.-X., Bo, Z., Wang, Y.-Q., Wu, Y.-Z., (2008). New descriptors of amino acids and their application to peptide QSAR study. *Peptides* **29** (10), 1798–1805.
41. Tyagi, A., Kapoor, P., Kumar, R., Chaudhary, K., Gautam, A., Raghava, G., (2013). In silico models for designing and discovering novel anticancer peptides. *Sci. Rep.* **3** (1), 2984.
42. Kumar, R., Chaudhary, K., Singh Chauhan, J., Nagpal, G., Kumar, R., Sharma, M., Raghava, G.P., (2015). An in silico platform for predicting, screening and designing of antihypertensive peptides. *Sci. Rep.* **5** (1), 12512.
43. Gautam, A., Chaudhary, K., Kumar, R., Sharma, A., Kapoor, P., Tyagi, A., net Osddcio, Raghava, G.P., (2013). In silico approaches for designing highly effective cell penetrating peptides. *J. Transl. Med.* **11**, 1–12.
44. Minkiewicz, P., Dziuba, J., Iwaniak, A., Dziuba, M., Darewicz, M., (2008). BIOPEP database and other programs for processing bioactive peptide sequences. *J. AOAC Int.* **91** (4), 965–980.
45. Atkins, P.W., Ratcliffe, R.G., de Paula, J., Wormald, M., (2023). Physical chemistry for the life sciences. Oxford University Press.
46. Kyte, J., Doolittle, R.F., (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157** (1), 105–132.
47. Janin, J., (1979). Surface and inside volumes in globular proteins. *Nature* **277** (5696), 491–492.
48. Bjellqvist, B., Hughes, G.J., Pasquali, C., Paquet, N., Ravier, F., Sanchez, J.C., Frutiger, S., Hochstrasser, D., (1993). The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences. *Electrophoresis* **14** (1), 1023–1031.
49. Zhang, Z., Wood, W.I., (2003). A profile hidden Markov model for signal peptides generated by HMMER. *Bioinformatics* **19** (2), 307–308.

50. Dubchak, I., Muchnik, I., Holbrook, S.R., Kim, S.-H., (1995). Prediction of protein folding class using global description of amino acid sequence. *Proc. Natl. Acad. Sci.* **92** (19), 8700–8704.
51. Johnson, M.M., Wilke, C.O., (2020). Site-specific amino acid distributions follow a universal shape. *J. Mol. Evol.* **88** (10), 731–741.
52. Pavlyshenko, B., (2018). Using stacking approaches for machine learning models. *In: 2018 IEEE second international conference on data stream mining & processing (DSMP)*. IEEE, pp. 255–258.
53. Liaw, A., Wiener, M., (2002). Classification and regression by randomForest. *R News* **2** (3), 18–22.
54. Natekin, A., Knoll, A., (2013). Gradient boosting machines, a tutorial. *Front. Neurorob.* **7**, 21.
55. Sharaff, A., Gupta, H., (2019). Extra-tree classifier with metaheuristics approach for email classification. *In: Advances in computer communication and computational sciences: proceedings of IC4S 2018*. Springer, pp. 189–197.
56. Hosmer Jr, D.W., Lemeshow, S., Sturdivant, R.X., (2013). Applied logistic regression. John Wiley & Sons.
57. Wang, Y., Zhai, Y., Ding, Y., Zou, Q., (2023). SBSM-Pro: support bio-sequence machine for proteins. *arXiv e-prints*. arXiv:2308.10275.
58. Li, Y., Wei, X., Yang, Q., Xiong, A., Li, X., Zou, Q., Cui, F., Zhang, Z., (2024). msBERT-Promoter: a multi-scale ensemble predictor based on BERT pre-trained model for the two-stage prediction of DNA promoters and their strengths. *BMC Biol.* **22** (1), 126.
59. Kastritis, P.L., Visscher, K.M., van Dijk, A.D., Bonvin, A.M., (2013). Solvated protein–protein docking using Kyte-Doolittle-based water preferences. *Proteins Struct. Funct. Bioinf.* **81** (3), 510–518.
60. Lemieux, L., Simard, R., (1992). Bitter flavour in dairy products. II. A review of bitter peptides from caseins: their formation, isolation and identification, structure masking and inhibition. *Lait* **72** (4), 335–385.
61. Hamilton, J.S., Hill, R., Van Leeuwen, H., (1974). A bitter peptide from Cheddar cheese. *Agric. Biol. Chem.* **38** (2), 375–379.
62. Sebald, K., Dunkel, A., Hofmann, T., (2019). Mapping taste-relevant food peptidomes by means of sequential window acquisition of all theoretical fragment ion–mass spectrometry. *J. Agric. Food Chem.* **68** (38), 10287–10298.
63. Hayes, T., Rao, R., Akin, H., Sofroniew, N.J., Oktay, D., Lin, Z., Verkuil, R., Tran, V.Q., Deaton, J., Wiggert, M., (2024). Simulating 500 million years of evolution with a language model. *bioRxiv*2024.2007.2001.600583.