Journal of Cheminformatics

METHODOLOGY Open Access

xBitterT5: an explainable transformer-based framework with multimodal inputs for identifying bitter-taste peptides

Nguyen Doan Hieu Nguyen[1†], Nhat Truong Pham[1†], Duong Thanh Tran[1], Leyi Wei[2], Adeel Malik[3*] and Balachandran Manavalan[1*]

Abstract

Bitter peptides (BPs), derived from the hydrolysis of proteins in food, play a crucial role in both food science and biomedicine by influencing taste perception and participating in various physiological processes. Accurate identification of BPs is essential for understanding food quality and potential health impacts. Traditional machine learning approaches for BP identification have relied on conventional feature descriptors, achieving moderate success but struggling with the complexities of biological sequence data. Recent advances utilizing protein language model embedding and meta-learning approaches have improved the accuracy, but frequently neglect the molecular representations of peptides and lack interpretability. In this study, we propose xBitterT5, a novel multimodal and interpretable framework for BP identification that integrates pretrained transformer-based embeddings from BioT5+ with the combination of peptide sequence and its SELFIES molecular representation. Specifically, incorporating both peptide sequences and their molecular strings, xBitterT5 demonstrates superior performance compared to previous methods on the same benchmark datasets. Importantly, the model provides residue-level interpretability, highlighting chemically meaningful substructures that significantly contribute to its bitterness, thus offering mechanistic insights beyond black-box predictions. A user-friendly web server (https://balalab-skku.org/xBitterT5/) and a standalone version (https://github.com/cbbl-skku-org/xBitterT5/) are freely available to support both computational biologists and experimental researchers in peptide-based food and biomedicine.

Scientific contribution

We propose xBitterT5, a novel multimodal transformer-based framework for the identification of BPs. By utilizing the pretrained BioT5+ model, xBitterT5 effectively extracts high-level representations from both the peptide sequences and their corresponding SELFIES molecular representation. This dual-modality approach enables a more comprehensive understanding of the peptide sequence by leveraging its molecular string, leading to substantial improvements in performance across two benchmark datasets. Additionally, xBitterT5 offers interpretability

†Nguyen Doan Hieu Nguyen and Nhat Truong Pham have contributed equally to this work.

*Correspondence:
Adeel Malik
adeel@procarb.org
Balachandran Manavalan
bala2022@skku.edu
Full list of author information is available at the end of the article

by identifying key molecular substructures that contribute to bitterness, thereby providing mechanistic insights essential for peptide-based food and drug applications.

Keywords Bitter taste peptides, Molecular representations, Text-to-text transfer transformer, Multimodal analysis, Model interpretation

Introduction

In vertebrates, taste perception is a fundamental physiological sense that aids organisms in selecting palatable foods and identifying toxic and nutrient-rich substances. Besides the five classical taste categories, sweet, salty, sour, umami, and the recently recognized ammonium chloride taste [1], bitterness plays a crucial role as an evolutionary defense mechanism against the ingestion of harmful compounds. Bitter peptides (BPs) refer to flavor-active peptides produced during the enzymatic hydrolysis of proteins, which elicit a bitter taste by binding to bitter taste receptors in the oral cavity [2]. In food science, BPs are used as an indicator of food quality and characteristics or quality of foods. Beyond their sensory impact, BPs are biologically active molecules that influence various physiological functions, including modulation of gastrointestinal, respiratory, and immune functions. Notably, BPs have also been implicated in disease contexts such as cancer, with expression patterns observed in pancreatic and ovarian tissues [3, 4]. Given their dual relevance in food and biomedical research, accurate identification of BPs is essential for applications ranging from flavor enhancement to therapeutic design.

Advancements in bioinformatics, particularly machine learning (ML), have enabled the development of computational models to predict BPs from sequence-derived features. Early methods have relied on conventional feature descriptors, including amino acid composition (AAC), dipeptide composition (DPC), pseudo amino acid composition (PAAC), amphiphilic pseudo amino acid composition (APAAC), and amino acid index (AAI), combined with various classical ML algorithms. For example, iBitter-SCM [5] applied a scoring card method based on AAC and DPC. While this approach provided a straightforward solution for classification tasks, it struggled to capture the complexities of biological sequence data. To address these limitations and enhance model performance, the same research group proposed iBitter-Fuse [6], by incorporating a support vector machine (SVM) algorithm and additional descriptors containing a combination of AAC, DPC, PAAC, APAAC, and AAI features, resulting in significant improvements. Subsequently, Bitter-RF [7] demonstrated that the random forest algorithm could improve predictive performance by leveraging a more comprehensive feature set. The recent integration of deep learning (DL) and natural language processing (NLP) into peptide bioinformatics has marked a paradigm shift. BERT4Bitter [8] was among the first to apply bidirectional encoder representations from transformers (BERT) [9], combined with long short-term memory (LSTM) [10] networks, for BP identification. This approach introduced automated feature learning and capturing complex sequence patterns more effectively than handcrafted features, resulting in deeper insights and improved performance. A notable example of this trend is MIMML [11], which combined meta-learning with TextCNN [12]. This approach allowed the model to optimize its learning strategies and fine-tune parameters more effectively, resulting in a more robust and adaptable framework. Hybrid strategies have also emerged, such as iBitter-DRLF [13], which fused representations from sequence-to-sequence attention (SSA), universal representations (UniRep), and bidirectional LSTM (BiLSTM), as inputs for the light gradient boosting machine (LGBM) [14] model. Similarly, the CPM-BP [15] framework employs LGBM with a focus on incorporating features such as the average hydrophobicity of peptides and the percentage of bitter-contributing amino acids in peptides, demonstrating the benefit of incorporating biochemical knowledge into model design.

While the existing methods for BP prediction achieved impressive performances, they often overlook the key aspects of the molecular representations of peptides and their function. Most approaches typically rely on properties or features extracted from peptide sequences, neglecting the valuable insights embedded within the molecular representations of the peptides. This limitation hinders their ability to fully capture the complex relationship between molecular representation of peptide and bitterness. Moreover, many current models lack interpretability. While they can predict whether a peptide is bitter, they often fail to identify the specific molecular substructures or motifs responsible for bitterness. This lack of transparency hinders a deeper understanding of bitterness perception and poses challenges for designing peptides with tailored taste profiles. Finally, most methods are trained from scratch using conventional descriptors, which ignore the advantages of transfer learning. By leveraging pretrained models, which have learned general representations from vast amounts of data, we can potentially improve the efficiency and accuracy of BP identification.

To overcome the abovementioned limitations, we introduce xBitterT5, a novel multimodal, explainable framework for BP identification (Fig. 1). xBitterT5 leverages the capabilities of a pretrained transformer-based model combined with a customized classification head specifically for BP identification. Unlike other models, xBitterT5 employs a unique multimodal approach, integrating both peptide sequence data and their corresponding molecular representations. It harnesses the versatility of transformer models to process heterogeneous data types and utilize pretrained tokenizers with various string representations of BPs.

The key contributions of our study are as follows:

- xBitterT5 is the first model that combines peptide sequences and their self-referencing embedded strings (SELFIES)-based molecular representation for BP identification.
- By utilizing the pretrained BioT5+, xBitterT5 demonstrates a deeper understanding of molecular structures, outperforming existing methods on both training and testing datasets.
- xBitterT5 offers interpretability, enabling the identification of specific molecular substructures within a peptide that contribute to its bitterness. This provides valuable insights into the structure–activity relationship of BPs.
- We have developed a user-friendly web server and made the source code and trained models publicly available via GitHub and Hugging Face, enabling both experimental researchers and computational biologists to utilize xBitterT5 for BP identification.

Methods

Dataset collection and refinement

To ensure a fair comparison with the existing method, we used two benchmark datasets from prior research. The first dataset, BTP640, was originally introduced in the iBitter-SCM [5] and comprises a balanced set of 320 BPs and 320 non-BPs. It has been partitioned into training (designated as BTP-CV) and independent testing sets (designated as BTP-TS) in a ratio of 80:20, providing an equitable condition for other researchers to compare their results.

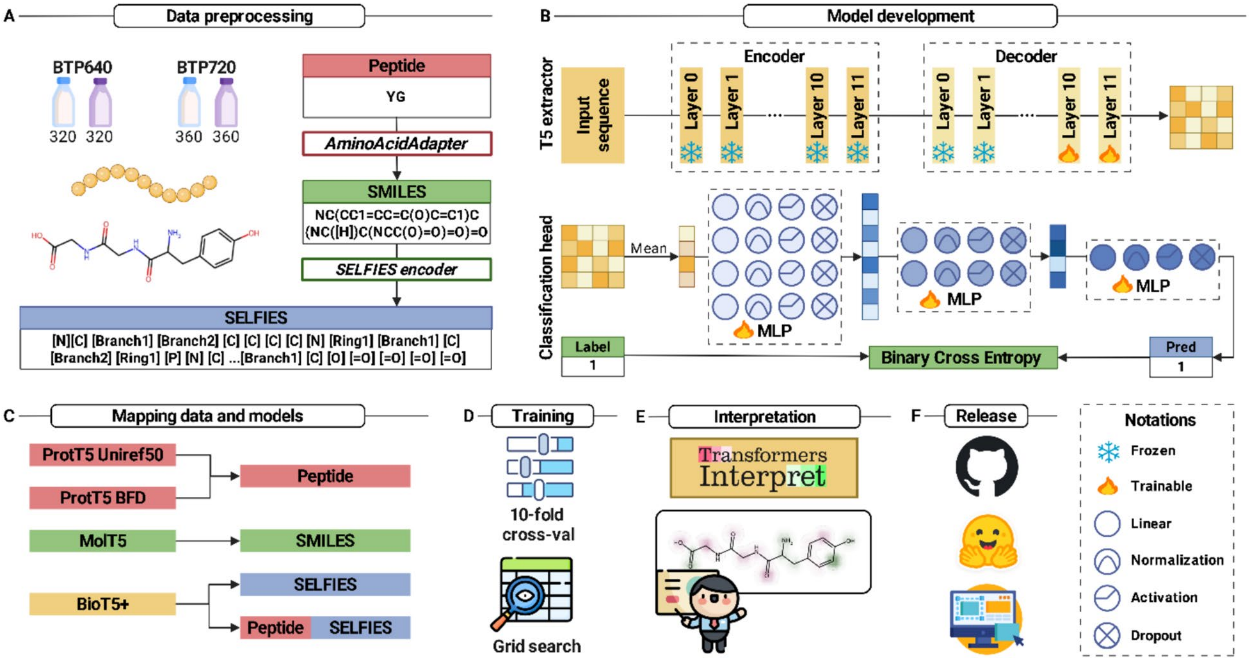


Fig. 1 The overall workflow for our proposed xBitterT5 framework. **A** The dataset preprocessing step, where the original peptide sequences in each dataset are converted into their corresponding molecular representations with SMILES and SELFIES format. **B** The proposed architecture of xBitterT5 consists of a pretrained transformer-based model and a classification head. **C** The data are mapped with the corresponding pretrained transformer models. **D** The proposed architecture is trained on two datasets, BTP640 and BTP720, with 10-fold cross-validation and grid search to find the optimal hyperparameters. **E** We carry on the model interpretation for xBitterT5 to discover which part of the molecular representations contributes to the bitterness of the peptide. **F** We provide the web server, code implementation, and trained weights for two variants of xBitterT5, namely xBitterT5-640 and xBitterT5-720. (Created with BioRender.com)

The second dataset, BTP720 [15], was curated by expanding BTP640 with additional BPs collected from diverse sources. To enhance data quality and label specificity, BTP720 includes only peptides with a unique bitter taste, while sequences exhibiting multiple taste attributes were excluded. This dataset was also split into an 80:20 ratio for 10-fold cross-validation and an independent test. It should be noted that we used the same criteria for model development using 10-fold cross-validation and validated the model transferability with the independent test.

Data preprocessing

The peptides in both BTP640 and BTP720 are relatively short, with most sequences being less than or equal to 15 amino acids. The distribution of peptide sequence lengths within each dataset is depicted in Figure S1. These relatively short sequence lengths can limit the model's ability to learn effectively from the conventional descriptors. To enhance the representation of the peptide sequences and provide the model with a broader dataset, we utilize the simplified molecular input line entry system (SMILES) [16] format to convert these sequences into molecular representations. SMILES constitutes a formalized notation system that enables the representation of chemical structures through concise ASCII string formats. This specification facilitates the encoding of chemical species in a manner that is both compact and interpretable, thereby serving as an essential tool for computational chemistry and cheminformatics. The conversion is carried out using the *chemistry-adapters*[1] package, which offers robust tools for converting amino acid sequences to SMILES [17]. In addition to this initial transformation, we further refine the molecular representation by converting the SMILES strings into SELFIES [18] format. SELFIES represents a novel and robust string-based encoding of molecular graphs. This methodology provides a completely reliable molecular string representation, ensuring accuracy and consistency in depicting molecular structures. This step allows us to achieve a more detailed and nuanced depiction of the molecules, enriching the dataset for our analysis. The advantages of utilizing the SELFIES format in comparison to the SMILES format are illustrated in Figure S2. We implement this transformation step by using the *selfies*[2] package.

Overall architecture

With the rapid advancements in language models, various pretrained models tailored to biological data, such as molecular strings and peptide sequences, have emerged. These models enhance understanding of biological concepts and facilitate a bridge between natural language text and biological data. Among the most notable models developed in this field are ProtTrans [19], MolT5 [20], BioT5 [21], and BioT5+ [22], which are fine-tuned adaptations of the T5 (Text-To-Text Transfer Transformer) [23] model. In this study, we leverage the capabilities of these models to create an effective classifier specifically aimed at identifying BPs.

Our proposed architecture has two primary modules: the T5 extractor and the classification head, shown in Fig. 1b. The T5 extractor exploits the capabilities of a T5 model, effectively leveraging its pretrained weights to facilitate faster convergence and improve overall performance. The T5 model includes two distinct components, the encoder and the decoder, structured sequentially. We freeze the entire encoder component to retain the valuable knowledge acquired from the pretraining phase, ensuring its learned representations remain intact. In contrast, we allow some final blocks of the decoder to be trainable. Once the T5 model processes the input data, it generates a tensor of hidden states, which will be channeled through the classification head module. This module is specifically tasked with identifying BPs based on the representations provided by the T5 extractor.

T5 extractor module

ProtTrans: This research focuses on training two auto-regressive models and four auto-encoder models using two key datasets, UniRef [24] and BFD (Big Fantastic Database) [25, 26]. Both datasets are valuable resources for protein research, though they differ in some aspects. UniRef is a well-established database emphasizing curated protein sequences sourced from UniProt, while BFD is a more recent and larger database encompassing a wider variety of sequences, making it particularly beneficial for training protein language models. To enhance our proposed architecture, we utilize the pretrained weights from the ProtT5 model, a variant of T5 from ProtTrans. We examine the performance of both above-mentioned pretrained models on each of the datasets to determine which pretrained model is more effective for fine-tuning the downstream task of BP classification. The model's input consists of peptide sequences represented with white space between all amino acids. For instance, the peptide sequence "*PA*" will be input into the model as "*P A*", with each constituent amino acid represented separately.

[1] https://pypi.org/project/chemistry-adapters/

[2] https://pypi.org/project/selfies/

MolT5: MolT5 is a variant of the T5 model that was pretrained through a self-supervised mechanism on a substantial amount of unlabeled natural language texts and SMILES molecular strings. This model was specifically designed to facilitate translation between molecules and natural language by introducing two tasks: molecule captioning and text-guided de novo molecule generation. Consequently, it is adept at comprehending both the architecture and functions of molecules. We utilize the SMILES format of molecules as input for this model to extract meaningful representations.

BioT5+: BioT5 is a specialized variant of the T5 model developed as a comprehensive pretraining framework that enhances cross-modal integration in biology by incorporating chemical knowledge and natural language. BioT5 utilizes SELFIES for molecules to ensure entirely robust representations of molecules. Additionally, BioT5 is trained on protein sequences, offering valuable insights into the interactions and properties of biological entities. Building upon BioT5's success, BioT5+ has undergone extensive pretraining and fine-tuning through numerous experiments, addressing three types of problems: classification, regression, and generation, which are divided into 15 different tasks and utilize a total of 21 benchmark datasets. This development has resulted in outstanding performance, achieving state-of-the-art results in most areas. Using BioT5/BioT5+, many models have achieved notable results in downstream tasks [27–30]

Leveraging the cross-modal capabilities of BioT5+, we propose three distinct types of input data for the T5 extractor when utilizing BioT5+ as the pretrained model. First, we provide peptide sequences prefixed by the token *"$<p>$"* to assist with tokenization. To indicate that the input string represents a peptide sequence, we incorporate two special tokens included in the token dictionary of BioT5+: *"$<bop>$"* and *"$<eop>$"*, which denote the beginning and the end of the peptide sequence, respectively. For instance, the peptide sequence *"PA"* will be preprocessed into *"$<bop><p>P<p>A<eop>$"* before being fed into the T5 extractor. Second, we utilize the SELFIES representation of the peptide sequence as the input of the T5 extractor. This string will be preprocessed by adding two special tokens, which are also available in the token dictionary of BioT5+, *"$<bom>$"* and *"$<eom>$"*, to mark the beginning and the end of the molecule. Lastly, we integrate both the peptide sequences and their corresponding SELFIES representations to furnish the model with more comprehensive information from both modalities. Each representation will be preprocessed separately, as outlined earlier, and then concatenated in the format *"[SEQUENCE][SELFIES]"*, where *"[SEQUENCE]"* is the preprocessed peptide sequence and *"[SELFIES]"* is the preprocessed SELFIES input string.

As shown in Supplementary Figure S1, the peptide sequences across the datasets have variable lengths. When converted to molecular representations, longer peptides typically generate longer molecular representations, further increasing the disparity in input lengths. To normalize the sequence lengths for the training process, we applied sequence padding, extending all representations to match the maximum length. These padding tokens are included solely to facilitate the uniform batch processing and are explicitly masked during the attention computation within the T5 extractor. Consequently, the output embeddings maintain a fixed dimensionality, independent of the original input sequence lengths or the number of padding tokens. To obtain the final representation for classification, we applied mean pooling across the token embeddings before passing them to the classification head.

Classification head module

The classification head module contains three main groups of layers, each consisting of a fully connected layer, a normalization layer, an activation layer, and a dropout layer (Fig. 1b). The number of nodes in the linear layers is determined by the T5 output hidden states' shape, where the number of nodes in the latter is equal to the previous divided by 4. The specific configurations of fully connected layers for each pretrained model are shown in Supplementary Table S1.

To ensure compatibility between the pretrained T5 extractor and the classification head module while utilizing the available components from the original T5 model, we incorporate the "T5LayerNorm" and "NewGELUActivation" submodules from the T5 implementation. Normalization is a crucial technique in transformer models, as it helps stabilize and expedite training by normalizing the inputs to each layer. The "T5LayerNorm", also known as Root Mean Square Layer Normalization (RMSNorm) [31], is specifically designed for the T5 model and employs a layer normalization that scales without shifting. Consequently, variance is calculated without accounting for the mean, resulting in no bias. The formula for the RMSNorm layer is described below:

$$RMS(a) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} a_i^2} \tag{1}$$

$$\overline{a}_i = \gamma \frac{a_i}{RMS(a) + \epsilon} \tag{2}$$

where a is an input vector of size n, ϵ is a small constant added for numerical stability, with the default value to be

10^{-6}, γ is a learnable parameter vector, and $\overline{a}_i$ is the final output of the normalization step at the index i.

The Gaussian Error Linear Unit (GELU) [32] is an activation function commonly used in transformer architectures, including the T5 model. It synthesizes the advantageous properties of the Rectified Linear Unit (ReLU) [33] and Sigmoid functions by facilitating a smooth activation of inputs grounded in a probabilistic approximation of the Gaussian cumulative distribution function. This characteristic of GELU promotes the generation of smoother gradients, which in turn enhances the convergence behavior of deep neural networks. Consequently, GELU has been demonstrated to improve the overall performance and efficiency of model training within various deep-learning architectures. The formula for GELU activation is described below:

$$GELU(x) = x\Phi(\mathrm{x}) \tag{3}$$

$$\Phi(\mathrm{x}) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right] \tag{4}$$

where x is the input of the activation function, $x\Phi(\mathrm{x})$ represents the Gaussian cumulative distribution function and erf is the error function. The GELU formula can serve as an activation function within a deep learning model. However, an approximate version of GELU is commonly employed to enhance computational efficiency due to its balance of performance and computational cost. The built-in "NewGELUActivation" layer in the T5 model exemplifies this with the following formula:

$$GELU(x) \approx 0.5x\left(1 + \tanh\left[\sqrt{\frac{2}{\pi}}\left(x + 0.044715x^3\right)\right]\right) \tag{5}$$

Finally, we incorporate dropout layers into our model architecture to address the risk of overfitting. Given the constraints of our relatively small dataset, we explore a diverse set of dropout rates to assess their impact on model performance. This allows us to identify the optimal dropout rate based on the evaluation of the training dataset.

Training mechanisms

To identify the most generalized configuration for each pretrained model and data type while facilitating a fair comparison with other existing methods, we implement the stratified 10-fold cross-validation methodology, ensuring that each fold maintains the same class distribution as the original dataset. We apply the grid search algorithm to determine the optimal configuration, which we then use to train the models on the entire training dataset. This approach ensures the generalization of the model and avoids data leakage issues. To mitigate overfitting, we adopted a two-pronged strategy. First, we designed a compact model architecture to limit the model's capacity to memorize training data. Second, we incorporated regularization techniques, including dropout, which randomly deactivates neurons during training, and weight decay, which penalizes large weights. Collectively, these strategies reduce model complexity and promote better generalization to unseen data.

Model interpretation

Having the final models, which have been trained on the whole training set of each dataset, we conduct model interpretation to understand which aspects of the input contribute to model predictions. We carry out this step by utilizing *transformers-interpret*,[3] which is a tool specifically designed for model explainability in conjunction with the transformers[4] package. Its operation mechanism is based on the Integrated Gradients algorithm, which calculates the integral of gradients concerning inputs along the path from a specified baseline to the input. It is important to highlight that the Integrated Gradients algorithm is classified as a primary attribution algorithm, assessing the contribution of each input feature to the model's output. Since this research focuses on sequence classification, particularly BP classification, we employ the *"SequenceClassificationExplainer"* class from the package to acquire the attribution score of each token in the input sequence, indicating its importance in the prediction. Upon obtaining the attribution score, we employed RDKit[5] to generate a molecular visualization for a better representation. This approach allows us to illustrate the regions of the molecule that exert the most significant contribution to the predictive outcomes.

Performance evaluation metrics

In this research, we implemented seven popular metrics, which are often used in binary classification problems [34–37] to evaluate the performance of the proposed models. The formulas of the metrics are presented as follows:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$SP = \frac{TN}{TN + FP} \tag{7}$$

[3] https://pypi.org/project/transformers-interpret/

[4] https://pypi.org/project/transformers/

[5] https://pypi.org/project/rdkit/

$$PRE = \frac{TP}{TP + FP} \tag{8}$$

$$SN = \frac{TP}{TP + FN} \tag{9}$$

$$F1 = \frac{2 \times TP}{2 \times TP + FN + FP} \tag{10}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{11}$$

$$TPR = \frac{TP_i}{TP_i + FN_i} \tag{12}$$

$$FPR = \frac{FP_i}{FP_i + TN_i} \tag{13}$$

$$AUC = \sum_{i=1}^{n-1} (TPR_{i+1} + TPR_i) \cdot \frac{FPR_{i+1} - FPR_i}{2} \tag{14}$$

where ACC refers to accuracy, SP denotes specificity, PRE signifies precision, and SN represents sensitivity. The Matthews correlation coefficient (MCC) is a balanced measure of the classification model's performance. The F1 score is derived as the harmonic mean of precision and recall. The terms TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative, respectively. Furthermore, we evaluate the performance of our proposed model using the area under the receiver operating characteristic (ROC) curve (AUC). We selected the MCC score as the primary metric for optimizing the model's hyperparameters during model development and optimization.

Results and discussion

Assessment of different configurations on the BTP640 dataset

To identify the optimal model architecture for BP prediction, we systematically evaluated the performance of various configurations using the BT640 dataset. Models were trained with different inputs: peptide sequences, molecular representations, such as SMILES and SELFIES, and a multimodal combination of peptide sequences and SELFIES. These inputs were encoded using several pretrained models, including ProtT5-Uniref50, ProtT5-BFD, MolT5, and BioT5+.

Our result shows that the MolT5 model employing SMILES as input exhibited the lowest performance across all evaluated metrics (Fig. 2a), with MCC, ACC, F1, and AUC values of 0.675, 0.832, 0.820, and 0.829, respectively. Conversely, models utilizing ProtT5-Uniref50,

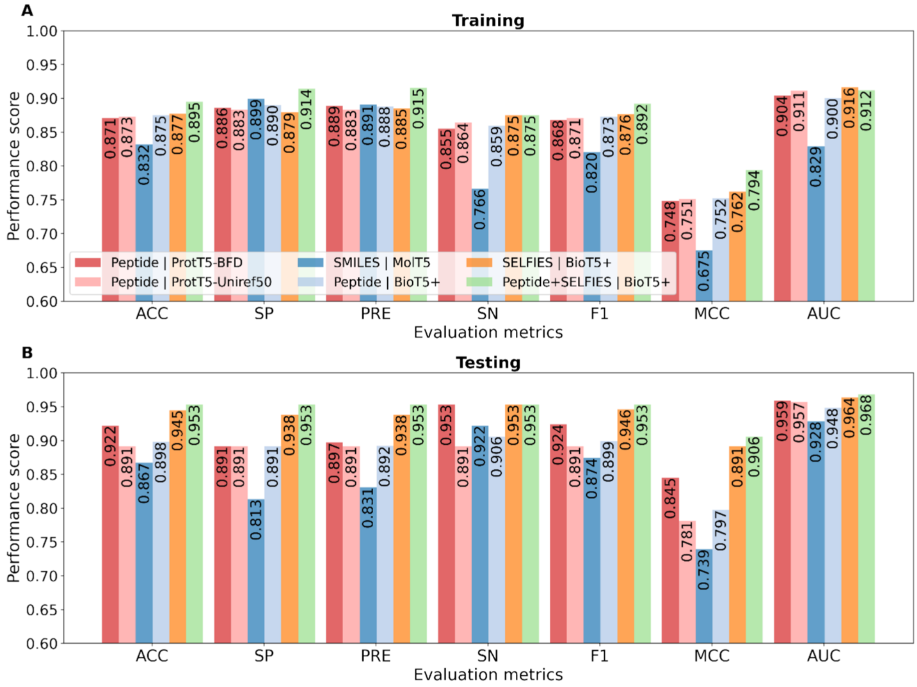


Fig. 2 Comparison of the proposed architecture using different data types and pretrained models on BTP640. **A** Performance of 10-fold cross-validation on the training set. **B** Performance on the independent dataset

ProtT5-BFD, and BioT5+ with peptide sequences as input demonstrated superior performance, with MCC values of 0.748, 0.751, and 0.752, respectively. Notably, when using the SELFIES representation with BioT5+, significantly improved performance was observed, with MCC gains of 1.00–8.70% compared to other single-input (unimodal) models. Furthermore, the multimodal BioT5+ model, combining peptide sequences with SELFIES, achieved the highest performance during training, with MCC, ACC, F1, and AUC values of 0.794, 0.895, 0.892, and 0.912, respectively. Compared to unimodal models, the multimodal model achieved significant improvements ranging from 3.20% to 11.90% in MCC, 1.80–6.30% in ACC, 1.60–7.20% in F1, and 0.10–8.30% in AUC.

To assess the transferability of our trained models, we evaluated them using an independent dataset (Fig. 2b). Consistent with the training results, the MolT5 model with SMILES input consistently underperformed (ACC < 87%). Among the models trained solely on peptide sequences, ProtT5-Uniref50 achieved the highest performance, exhibiting MCC, ACC, F1, and AUC values of 0.845, 0.922, 0.924, and 0.959, respectively. However, the BioT5+ model using SELFIES representations surpassed all single-modality models, achieving MCC, ACC, F1, and AUC values of 0.891, 0.945, 0.946, and 0.964, respectively. The multimodal BioT5+ model again demonstrated the highest performance with MCC, ACC, F1, and AUC values of 0.906, 0.953, 0.953, and 0.968, respectively. These results represent significant improvements over all other models, with increases ranging from 1.50% to 16.70% in MCC, 0.80–8.60% in ACC, 0.70–7.90% in F1, and 0.40–4.0% in AUC.

Overall, these results suggest that BioT5+ did not effectively capture meaningful information from peptide sequences due to its training on a relatively limited number of sequences. Conversely, ProtT5-Uniref50 and ProtT5-BFD, while trained on a larger database, appeared to be hindered by inherent challenges associated with processing short peptides. Interestingly, leveraging the SELFIES representation, derived from peptide leads to a slight improvement, suggesting that SELFIES may capture information complementary to the raw peptide sequence. The multimodal approach, integrating both peptide sequences and molecular representations, achieved the most accurate prediction, demonstrating the effectiveness of integrating diverse data modalities for BP prediction. Consequently, we selected the multimodal BioT5+ model as the optimal solution for the BTP640 dataset, designated as xBitterT5-640.

Assessment of different configurations on the BTP720 dataset

A recent study by Yu et al. [15] refined and updated the BP dataset, resulting in a more comprehensive version comprising 720 sequences (BTP720). To assess the generalizability of our previously optimized model, xBitterT5-640, we evaluated its performance on this updated dataset (Table S2). To ensure fair evaluation, we excluded all overlapping sequences present in the BTP640 training set from the BTP720 dataset. As shown in Table S2, xBitterT5-640 showed limited transferability to the BTP720. This outcome aligns with previous research [15] that highlights significant differences between the BTP640 and the BTP720 datasets. These findings underscore a well-known challenge in ML research: models trained on a smaller or less diverse dataset may struggle to generalize effectively to newer, unseen data.

To address this, we developed a new model specifically for this dataset, following the same procedure outlined in the previous section. Figure 3a demonstrates the performance of various configurations on the BTP720 training dataset. The model utilizing MolT5 with SMILES representation performed poorly, achieving MCC, ACC, F1, and AUC values of 0.752, 0.873, 0.869, and 0.901, respectively. While the models using peptide sequences as input showed improved performance, they still fell short of the BioT5+ model with SELFIES representations. This model achieved MCC, ACC, F1, and AUC values of 0.790, 0.892, 0.888, and 0.919. Notably, the model employing BioT5+, incorporating both peptide and SELFIES as input, achieved the best performance, with MCC, ACC, F1, and AUC values of 0.814, 0.903, 0.895, and 0.915, respectively. Interestingly, the multimodal approach improvements range from 2.40% to 6.20% in MCC, 1.10–3.00% in ACC, 0.70–2.60% in F1, and 0.20–1.40% in AUC compared to the other models.

We also evaluated different models on the independent dataset to validate the transferability (Fig. 3b). Interestingly, while the BioT5+ model with peptide sequences achieved the same ACC value as the MolT5 with SMILES, it exhibited the lowest overall performance on the independent dataset, with MCC, F1, and AUC values of 0.778, 0.887, and 0.937, respectively. Conversely, the BioT5+ model with SELFIES representations outperformed all other single unimodal models. Once again, the multimodal BioT5+ model demonstrated the best performance across all metrics, achieving MCC, ACC, F1, and AUC values of 0.879, 0.938, 0.934, and 0.980, respectively. These results represent significant improvements compared to all unimodal

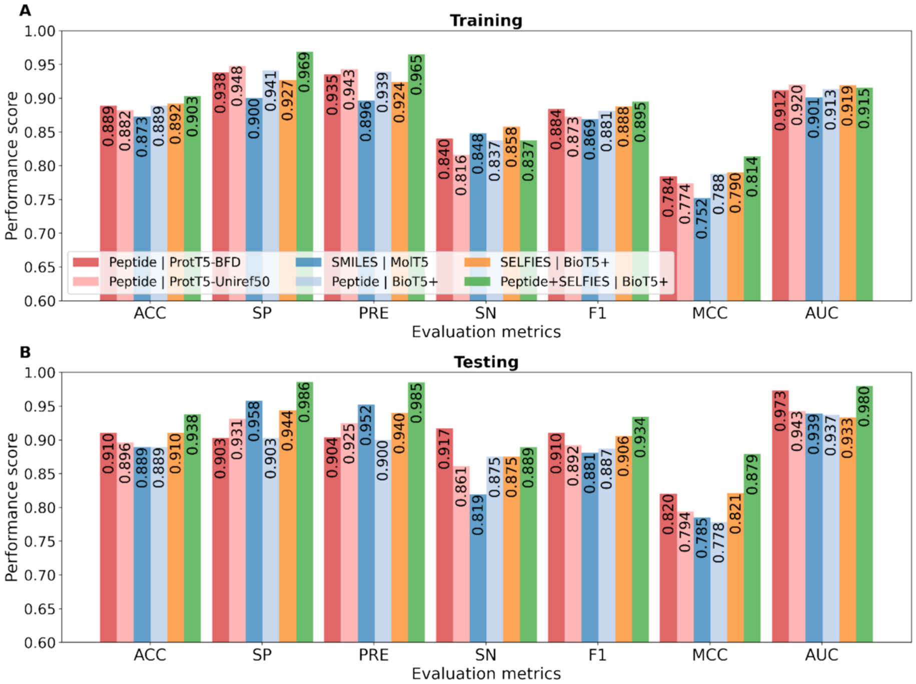


Fig. 3 Comparison of the proposed architecture using different data types and pretrained models on BTP720. **A** Performance of 10-fold cross-validation on the training set. **B** Performance on the independent dataset

models, with gains of 5.80–10.10% in MCC, 2.80–4.90% in ACC, 2.40–5.30% in F1, and 0.70–4.70% in AUC.

Overall, these findings are consistent with our observations from the BTP640 dataset, demonstrating the effectiveness and robustness of the multimodal approach in BP identification. Therefore, we selected the BioT5+ model integrating both peptide and SELFIES inputs as the final model for the BTP720 dataset, designated it as xBitterT5-720.

Performance comparison between the proposed method and the existing methods on the same datasets

We compared the performance of our proposed model, xBitterT5-640, with existing methods on the BTP640 training dataset. As shown in Fig. 4a, while the recent iBitter-DRLF method outperformed the first BP predictor, iBitter-SCM, xBitterT5-640 significantly outperformed the best existing method (iBitter-DRLF), with notable improvements ranging from 1.70% to 9.40% in MCC and 0.60–4.50% in ACC. These results highlight the advantage of integrating multimodal inputs, combining peptide sequences with their corresponding SELFIES representations. Further evaluation on an independent BTP640 dataset showed that all existing methods performed better than the first BP predictor, iBitter-SCM, but xBitterT5-640 again demonstrated superior performance, with improvements ranging from 1.70% to 21.80% in MCC and 0.90–10.90% in ACC. Importantly, xBitterT5-640 achieved a balanced performance between SN and SP, both reaching 0.953, demonstrating that our model is not biased towards with particular class, a common limitation in several existing methods. These findings suggest that xBitterT5-640 effectively captures biologically meaningful information through multimodal integration, resulting in significant performance improvements on both training and independent datasets.

For the BTP720 dataset, we compared our model, xBitterT5-720, with the CPM-BP, the only available method trained and evaluated on the same dataset. Other approaches developed using the BTP640 dataset were excluded from comparison, as their performance does not generalize well to BTP720 due to broader sample diversity. This limitation was demonstrated in the previous section, where we observed a notable performance drop when evaluating BTP720 using the xBitterT5-640 model. Figure 5a illustrates that xBitterT5-720 consistently outperformed CPM-BP across all global metrics during training. Specifically, our model demonstrated significant improvements of 11.60% in MCC, 5.40% in ACC, 4.70% in F1, and 4.70% in AUC. Furthermore, xBitterT5-720 consistently surpassed CPM-BP on the independent dataset, with improvements of MCC by 6.30%, ACC by

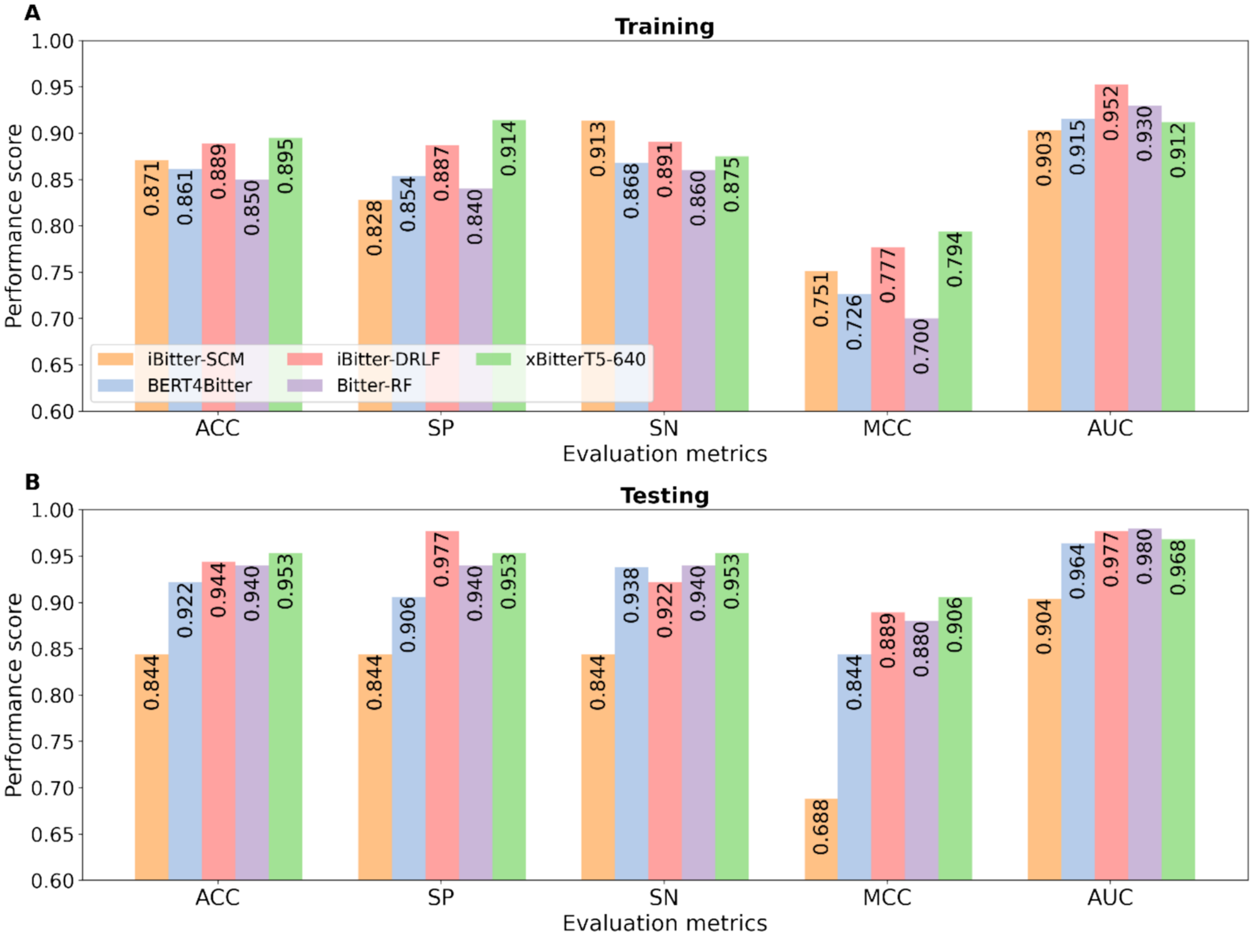


Fig. 4 Performance comparison of xBitterT5-640 and existing methods on the BTP640 dataset. **A** Performance of 10-fold cross-validation on the training set. **B** Performance on the independent dataset

3.50%, F1 by 4.10%, and AUC by 7.50%. These consistent improvements across both datasets demonstrate the effectiveness of integrating the BioT5+ model with multimodal inputs, enabling the model to extract and leverage meaningful information from both peptide and molecule sequences. Overall, these findings confirm the robust generalizability and superior accuracy of our proposed xBitterT5 framework in bitter peptide identification.

Visualization of model interpretation

To enhance model interpretability of xBitterT5, we analyzed how the model evaluates tokens within a sequence, indicating their significance to prediction outcomes (see Methods section). When model-identified important substructures align with known biological or physicochemical properties, this offers valuable insight into the molecular basis of bitter taste perception. Among two input modalities, peptide sequences and molecular representations, we excluded the peptide modality from interpretation due to the peptides in our dataset being relatively short, limiting the ability to extract meaningful physicochemical characteristics or amino acid composition for attribution. In contrast, molecular representations encode more detailed structural information, enabling the model to capture relevant substructures that influence bitterness. This distinction is empirically supported by our modality-wise performance analysis using the same pretrained model (see Assessment section). Accordingly, all interpretation analyses were conducted solely on the SELFIES representations.

We first analyzed xBitterT5-640 on selected samples from the BTP640 independent dataset (Table 1). For BPs (samples 1 and 2), the model correctly identified substructures corresponding to Proline (P) and Tyrosine (Y), both of which are known contributors to bitterness [38, 39]. Notably, the unique conformation of P facilitates the folding of the peptide backbone, facilitating receptor interaction and enhancing bitter perception. Similarly,

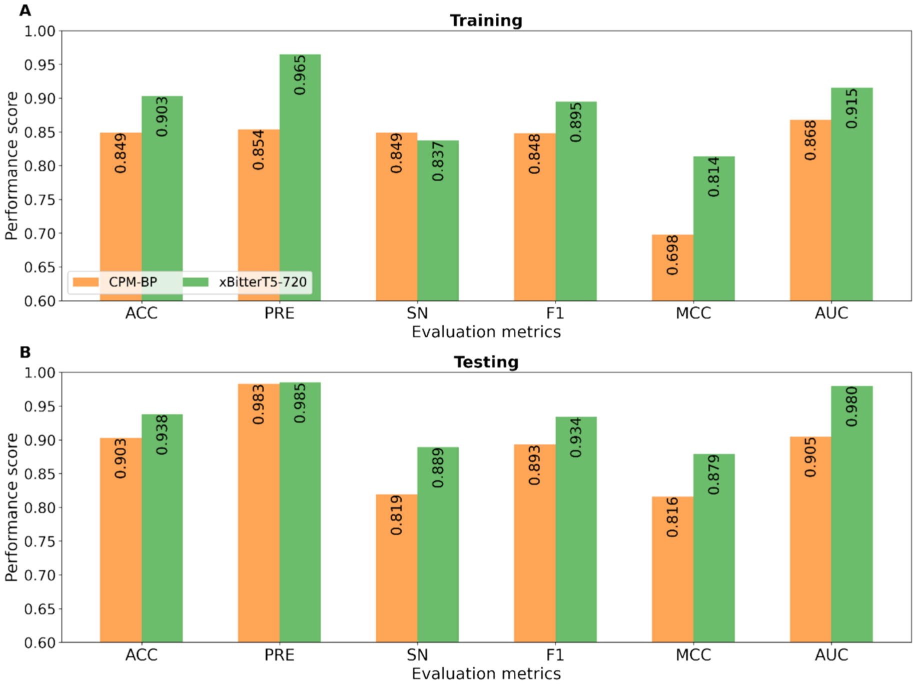
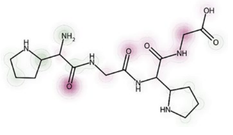
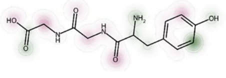
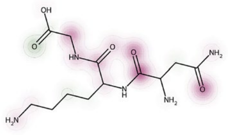
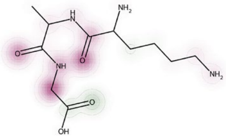


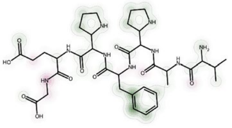
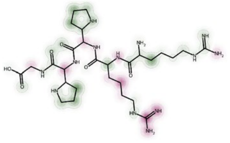
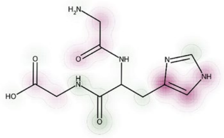
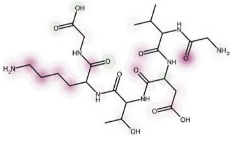
Fig. 5 Performance comparison of xBitterT5-720 and existing methods on the BTP720 dataset. **A** Performance of 10-fold cross-validation on the training set. **B** Performance on the independent dataset

Table 1 Model interpretation of identifying BPs and non-BPs using the xBitterT5-640 model on the selected samples of the BTP640 independent dataset

	
Sample 1. PGP	Sample 2. YG
	
Sample 3. NK	Sample 4. KA

The green color highlights the molecular substructures that contribute positively to the prediction of a peptide being classified as bitter. In contrast, the red color signifies the molecular substructures that contribute to the prediction of the non-bitter classification

Table 2 Model interpretation of identifying BPs and non-BPs using the xBitterT5-720 model on the selected samples of the BTP720 independent dataset

	
Sample 5. VAPFPE	Sample 6. RRPP
	
Sample 7. GH	Sample 8. GVDTK

The green color highlights the molecular substructures that contribute positively to the prediction of a peptide being classified as bitter. In contrast, the red color signifies the molecular substructures that contribute to the prediction of the non-bitter classification

Y aromatic side chain has been previously associated with bitter taste modulation. In contrast, for non-BPs (samples 3 and 4), the model highlighted the substructures of Asparagine (N) and Alanine (A) amino acid groups, respectively. These findings are consistent with prior studies, where L-asparagine is reported as tasteless, unlike its D-form, which exhibits a sweet flavor [40], with neither form contributing to bitterness. Furthermore, A is known for its mild, sweet flavor profile, often described as having a flat and food-like taste [41]. The attribution results demonstrate the model's ability to distinguish structurally relevant features that influence bitter perception.

Further analysis on the BTP720 independent dataset revealed more refined attribution patterns Table 2. For instance, in sample 5 (a BP), the model correctly focused on the substructure derived from P and Phenylalanine (F), both of which exhibit high attribution scores, consistent with prior research indicating their role in enhancing bitterness [38, 39]. Similarly, sample 6, where the model highlights two distinct P substructures. Conversely, in samples 7 and 8 (non-BPs), the model correctly focused on substructures associated with Histidine (H), Lysine (K), and Glycine (G). These amino acids are known to be tasteless or slightly sweet (H), salty (K) [42], and sweetness (G) [41].

Overall, molecular attribution maps generated for the BTP720 independent dataset demonstrated greater accuracy and precision compared to the BTP640 independent dataset. This improvement is likely attributable to the more stringent construction criteria applied during the construction of the BTP720 dataset, which excludes peptides with multiple or ambiguous taste attributes [15].

Table 3 The comparison of the predicted labels between xBitterT5 models and BERT4Bitter on the known BPs in Group B

Sequence	xBitterT5-720	xBitterT5-640	BERT4Bitter
FYPELF*	BP	NA	NA
VEVFAPPF	BP	BP	BP
FFVAPFPQVFGK	BP	Non-BP	Non-BP
RGPPFFIV	BP	BP	BP
RGPPFIV*	BP	NA	NA
LRF	Non-BP	Non-BP	BP
PQVF	Non-BP	Non-BP	Non-BP
LRL*	BP	NA	NA
GPFPIIV*	BP	NA	NA
AQTQSLVYPFPGPIPNSLPQNIPPLTQ	BP	BP	BP

(* denotes the peptide present in the BTP640 training set; NA denotes not applicable.)

As a result, the BTP720 dataset provides a clearer classification boundary, allowing the model to learn more discriminative patterns. Consequently, the model trained on the BTP720 exhibits enhanced transferability and interpretability when applied to the independent dataset.

Case study

To further evaluate the performance of our proposed method, we conducted an additional evaluation based on BPs from recent literature. Specifically, we employed the eleven reported BPs and three experimentally validated BPs (referred to as Group A) utilized in a previous study [15], and supplemented this dataset with ten additional BPs (Group B) reported in a more recent study [43], to enable a more comprehensive assessment. Predictions from CPM-BP, iBitter-SCM, and iBitter-Fuse for Group B were unavailable due to the absence of publicly accessible web servers and standalone programs. It should be noted that none of the BPs from these groups were included in the refined BTP720 training dataset, although five BPs from Group A and four from Group B were present in the original BTP640 training dataset.

Table 3 compared the performance of our xBitterT5 models with BERT4Bitter on the reported BPs, while a comprehensive comparison between our proposed method and all existing methods on the reported BPs used in the previous study is shown in Table S3. xBitterT5-720 and CPM-BP accurately predicted all reported BPs, while the other tools performed poorly on Group A. Notably, xBitterT5-640, iBitter-SCM, iBitter-Fuse, and BERT4Bitter consistently predicted incorrect labels for four BPs: "FALPQYLK," "LHLPLPLL," "LPLPLLQSW," and "FALPQYL.". This discrepancy may be attributed to the differing distributions of the BTP720 and BTP640 datasets. In contrast, our xBitterT5-640 performed better than the other three models. For Group B, xBitterT5-720 accurately predicted eight out of ten BPs, compared to only three by xBitterT5-640 and four by BERT4Bitter. Collectively, these results demonstrate that xBitterT5-720 achieves the highest predictive accuracy across both groups, with xBitterT5-640 consistently outperforming earlier sequence-based models. The strong performance of xBitterT5 on unseen, literature-reported BPs underscores its ability to capture generalizable and discriminative features for BP identification, even in challenging, real-world scenarios.

Web server implementation

To ensure accessibility for both experimentalists and computational researchers, we have developed a web server for xBitterT5, freely available at https://balalab-skku.org/xBitterT5/. This platform provides a user-friendly interface that allows for peptide sequence submission in FASTA format via both file upload and direct input. xBitterT5 provides two distinct models: xBitterT5-640 and xBitterT5-720, with the latter set as the default option. The input peptide sequences undergo standardized preprocessing steps to generate a combined feature representation, which is subsequently utilized for prediction. The resulting predictions can be downloaded in CSV format. To enhance transparency and reproducibility, the complete source code is publicly available at https://github.com/cbbl-skku-org/xBitterT5/, and pretrained weights for both model variants can be accessed via Hugging Face: (xBitterT5-640: https://huggingface.co/cbbl-skku-org/xBitterT5-640/ and xBitterT5-720:

https://huggingface.co/cbbl-skku-org/xBitterT5-720/). We recommend using xBitterT5-720 for most applications, given its enhanced training data quality and superior performance generalizability.

Limitations and future work

Although xBitterT5 demonstrates superior performance compared to existing methods on both the BTP640 and BTP720 datasets, several avenues for future improvement remain. Firstly, the current framework focuses solely on bitter peptides; however, the underlying architecture is readily extendable to other flavor-associated peptides, including those responsible for sweet, sour, salty, and umami peptides. Expanding the model's scope to multiple flavors could enhance its ability to discern the distinct molecular substructure associated with each taste. However, this advancement would require the development of an extensive dataset with precisely annotated labels to minimize potential ambiguities during model training and improve generalizability. Second, integrating additional pretrained models trained on diverse molecular data representations is a promising direction. Exploring alternative molecular string representations, including InChI [44], DeepSMILES [45], and Group SELFIES [46], could provide complementary perspectives on peptide molecular representation. This multimodal approach has the potential to offer a more comprehensive understanding of the complex relationship between a peptide's molecular substructure and its flavor perception. Finally, the proposed methodology is not restricted to flavor peptides alone. It holds significant promise for a broad spectrum of peptide therapeutics, including anticancer [47, 48], antimicrobial, tumor-homing [49, 50], and cell-penetrating [51, 52] peptides. By leveraging chemically informed molecular representations, our approach enables systematic exploration of the structural and functional properties of short peptide sequences, opening new avenues for peptide-based drug discovery and design.

Conclusion

The presence of bitter peptides, a byproduct of enzymatic hydrolysis or fermentation processes in many foods, can significantly diminish product quality and impact consumption. Accurately identifying these peptides is crucial for both improving detection efficiency and optimizing food processing strategies. Here, we introduce xBitterT5, a novel multimodal and interpretable DL framework for identifying BPs that integrates peptide sequences and their SELFIES-based molecular representations. This approach leverages the pretrained BioT5+ model, which is specifically adapted to capture both sequential features from the two modalities. To our knowledge, this is the first application of BioT5+ for multimodal peptide-based prediction, demonstrating an advancement in the field. Comprehensive evaluations on two benchmark datasets, BTP640 and BTP720, demonstrate that our model variants (xBitterT5-640 and xBitterT5-720) consistently achieve superior performance. These results highlight the advantage of incorporating molecular-level representations, which overcome the inherent information limitations of short peptide sequences alone. Furthermore, xBitterT5 provides an unprecedented level of interpretability, providing insights into the specific molecular strings that contribute to bitterness classification. For future research, we recommend the BPT720 dataset due to its larger data size and a wider variety of unique tastes. To ensure accessibility, the proposed method is deployed as a user-friendly web server at https://balalab-skku.org/xBitterT5/. All corresponding code implementation is available at http://github.com/cbbl-skku-org/xBitterT5/, and pretrained weights for xBitterT5-640 and xBitterT5-720 can be accessed via https://huggingface.co/cbbl-skku-org/xBitterT5-640/ and https://huggingface.co/cbbl-skku-org/xBitterT5-720/, respectively. We expect that xBitterT5 will not only advance BP identification but also serve as a foundation for broader applications in peptide-centric research, including anticancer, cell-penetrating peptides, and antimicrobial peptides.

Abbreviations

A	Alanine
AAC	Amino Acid Composition
AAI	Amino Acid Index
ACC	Accuracy
APAAC	Amphiphilic Pseudo Amino Acid Composition
AUC	Area Under the Receiver Operating Characteristic Curve
BERT	Bidirectional Encoder Representations from Transformers
BP	Bitter Peptide
BiLSTM	Bidirectional Long Short-Term Memory
DL	Deep Learning
DPC	Dipeptide Composition
F	Phenylalanine
FN	False Negative
FP	False Positive
G	Glycine
GELU	Gaussian Error Linear Unit
H	Histidine
K	Lysine
LGBM	Light Gradient Boosting Machine
LSTM	Long Short-Term Memory
MCC	Matthews Correlation Coefficient
ML	Machine Learning
N	Asparagine
NLP	Natural Language Processing
P	Proline
PAAC	Pseudo Amino Acid Composition

PRE	Precision
RMSNorm	Root Mean Square Layer Normalization
ROC	Receiver Operating Characteristic
ReLU	Rectified Linear Unit
SELFIES	Self-Referencing Embedded Strings
SMILES	Simplified Molecular Input Line Entry System
SN	Sensitivity
SP	Specificity
SSA	Sequence-to-Sequence Attention
SVM	Support Vector Machine
T5	Text-To-Text Transfer Transformer
TN	True Negative
TP	True Positive
UniRep	Universal Representations
Y	Tyrosine

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-025-01078-1.

Additional file 1.

Acknowledgements

The authors would like to thank the Korea Bio Data Station (K-BDS) provided computing resources, including technical support. The authors also thank the Korea Association for ICT Promotion (KAIT) for computing resources and technical support.

Author contributions

Nguyen Doan Hieu Nguyen: Conceptualization, Methodology, Software, Validation, Visualization, Writing—original draft, Writing—review and editing. Nhat Truong Pham: Conceptualization, Methodology, Software, Validation, Writing—original draft, Writing—review and editing. Duong Thanh Tran: Methodology, Review and Editing. Leyi Wei: Validation, Writing—review and editing. Adeel Malik: Conceptualization, Validation, Writing—review and editing. Balachandran Manavalan: Conceptualization, Writing—review and editing, Writing—original draft, Supervision, Investigation, Funding acquisition.

Funding

This work was supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (RS-2024-00344752). This research was supported by the Department of Integrative Biotechnology, Sungkyunkwan University (SKKU) and the BK21 FOUR Project.

Data availability

The web server for xBitterT5, along with the training and independent datasets, is freely accessible at https://balalab-skku.org/xBitterT5/. The implementation of the model has been made publicly available at https://github.com/cbbl-skku-org/xBitterT5/, along with the pretrained weights at https://huggingface.co/cbbl-skku-org/xBitterT5-640/ and https://huggingface.co/cbbl-skku-org/xBitterT5-720/ for xBitterT5-640 and xBitterT5-720, respectively.

Declarations

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Integrative Biotechnology, College of Biotechnology and Bioengineering, Sungkyunkwan University, Suwon, Gyeonggi-do 16419, Republic of Korea. [2]Center for Artificial Intelligence Driven Drug Discovery, Faculty of Applied Science, Macao Polytechnic University, Macau, China. [3]Institute of Intelligence Informatics Technology, Sangmyung University, Seoul 03016, Republic of Korea.

Received: 18 March 2025 Accepted: 8 August 2025
Published online: 20 August 2025

References

1. Liang Z, Wilson CE, Teng B, Kinnamon SC, Liman ER (2023) The proton channel OTOP1 is a sensor for the taste of ammonium chloride. Nat Commun 14(1):6194. https://doi.org/10.1038/s41467-023-41637-4
2. Chandrasekaran S, Luna-Vital D, de Mejia EG (2020) Identification and comparison of peptides from chickpea protein hydrolysates using either bromelain or gastrointestinal enzymes and their relationship with markers of type 2 diabetes and bitterness. Nutrients 12(12):3843
3. Tang W, Wan S, Yang Z, Teschendorff AE, Zou Q (2017) Tumor origin detection with tissue-specific miRNA and DNA methylation markers. Bioinformatics 34(3):398–406. https://doi.org/10.1093/bioinformatics/btx622
4. Liu S, Shi T, Yu J, Li R, Lin H, Deng K (2024) Research on bitter peptides in the field of bioinformatics: a comprehensive review. Int J Mol Sci. https://doi.org/10.3390/ijms25189844
5. Charoenkwan P, Yana J, Schaduangrat N, Nantasenamat C, Hasan MM, 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. https://doi.org/10.1016/j.ygeno.2020.03.019
6. Charoenkwan P, Nantasenamat C, Hasan MM, Moni MA, Lio P, Shoombuatong W (2021) Ibitter-fuse: a novel sequence-based bitter peptide predictor by fusing multi-view features. Int J Mol Sci. https://doi.org/10.3390/ijms22168958
7. Zhang Y-F, Wang Y-H, Gu Z-F, Pan X-R, Li J, Ding H et al (2023) Bitter-RF: a random forest machine model for recognizing bitter peptides. Front Med. https://doi.org/10.3389/fmed.2023.1052923
8. Charoenkwan P, Nantasenamat C, Hasan MM, 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. https://doi.org/10.1093/bioinformatics/btab133
9. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Minneapolis, Minnesota: Association for Computational Linguistics; 2019. p. 4171–86.
10. Graves A, Graves A (2012) Long short-term memory. Supervised sequence labelling with recurrent neural networks 2012:37–45
11. He W, Jiang Y, Jin J, Li Z, Zhao J, Manavalan B et al (2022) Accelerating bioactive peptide discovery via mutual information-based meta-learning. Brief Bioinform. https://doi.org/10.1093/bib/bbab499
12. Kim Y. Convolutional Neural Networks for Sentence Classification.
13. Jiang J, Lin X, Jiang Y, Jiang L, Lv Z (2022) Identify bitter peptides by using deep representation learning features. Int J Mol Sci. https://doi.org/10.3390/ijms23147877
14. Fan J, Ma X, Wu L, Zhang F, Yu X, Zeng W (2019) Light gradient boosting machine: an efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data. Agric Water Manage 225:105758
15. Yu Y, Liu S, Zhang X, Yu W, Pei X, Liu L et al (2024) Identification and prediction of milk-derived bitter taste peptides based on peptidomics technology and machine learning method. Food Chem 433:137288. https://doi.org/10.1016/j.foodchem.2023.137288
16. Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci 28(1):31–36
17. Sul. Sulstice/ChemistryAdapters. 0.0.2.1 ed: Zenodo; 2021.
18. Krenn M, Häse F, Nigam A, Friederich P, Aspuru-Guzik A (2020) Self-referencing embedded strings (SELFIES): a 100% robust molecular string representation. Mach Learn Sci Technol. https://doi.org/10.1088/2632-2153/aba947
19. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L et al (2022) Prottrans: toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell 44(10):7112–7127. https://doi.org/10.1109/TPAMI.2021.3095381
20. Edwards C, Lai T, Ros K, Honke G, Cho K, Ji H. Translation between Molecules and Natural Language. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics; 2022. p. 375–413.
21. Pei Q, Zhang W, Zhu J, Wu K, Gao K, Wu L, et al. BioT5: Enriching Cross-modal Integration in Biology with Chemical Knowledge and Natural Language Associations. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing 2023. p. 1102–23.

22. Pei Q, Wu L, Gao K, Liang X, Fang Y, Zhu J, et al. Biot5+: Towards generalized biological understanding with iupac integration and multi-task tuning. arXiv preprint arXiv:240217810. 2024.
23. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M et al (2020) Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res 21(140):1–67
24. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, Consortium tU (2014) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 31(6):926–932. https://doi.org/10.1093/bioinformatics/btu739
25. Steinegger M, Mirdita M, Söding J (2019) Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. Nat Methods 16(7):603–606. https://doi.org/10.1038/s41592-019-0437-4
26. Steinegger M, Söding J (2018) Clustering huge protein sequence sets in linear time. Nat Commun 9(1):2542. https://doi.org/10.1038/s41467-018-04964-5
27. Nguyen N, Pham NT, Tran D, Manavalan B. Lang2Mol-Diff: A Diffusion-Based Generative Model for Language-to-Molecule Translation Leveraging SELFIES Representation. Bangkok, Thailand: Association for Computational Linguistics; 2024. p. 128–34.
28. Tran D, Pham NT, Nguyen N, Manavalan B. Mol2Lang-VLM: Vision- and Text-Guided Generative Pre-trained Language Models for Advancing Molecule Captioning through Multimodal Fusion. Bangkok, Thailand: Association for Computational Linguistics; 2024. p. 97–102.
29. Luo Y, Yang K, Hong M, Liu XY, Nie Z, Zhou H, et al. Learning Multi-view Molecular Representations with Structured and Unstructured Knowledge. Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Barcelona, Spain: Association for Computing Machinery; 2024. p. 2082–93.
30. Pei Q, Wu L, Gao K, Zhu J, Yan R. Enhanced BioT5+ for Molecule-Text Translation: A Three-Stage Approach with Data Distillation, Diverse Training, and Voting Ensemble. Bangkok, Thailand: Association for Computational Linguistics; 2024. p. 48–54.
31. Zhang B, Sennrich R. Root Mean Square Layer Normalization. In: Wallach H, Larochelle H, Beygelzimer A, d\textquotesingle Alché-Buc F, Fox E, Garnett R, editors. Advances in Neural Information Processing Systems: Curran Associates, Inc.; 2019.
32. Hendrycks D, Gimpel K. Gaussian Error Linear Units (GELUs).
33. Agarap A. Deep learning using rectified linear units (relu). arXiv preprint arXiv:180308375. 2018.
34. Wang C, Zou Q (2024) MFPSP: identification of fungal species-specific phosphorylation site using offspring competition-based genetic algorithm. PLoS Comput Biol 20(11):e1012607. https://doi.org/10.1371/journal.pcbi.1012607
35. Wang C, He Z, Jia R, Pan S, Coin LJ, Song J et al (2024) Planner: a multi-scale deep language model for the origins of replication site prediction. IEEE J Biomed Health Inform 28(4):2445–2454. https://doi.org/10.1109/JBHI.2024.3349584
36. Manavalan B, Subramaniyam S, Shin TH, Kim MO, Lee G (2018) Machine-learning-based prediction of cell-penetrating peptides and their uptake efficiency with improved accuracy. J Proteome Res 17(8):2715–2726. https://doi.org/10.1021/acs.jproteome.8b00148
37. Manavalan B, Basith S, Shin TH, Wei L, Lee G (2019) Meta-4mcpred: a sequence-based meta-predictor for accurate DNA 4mc site prediction using effective feature representation. Molecular Therapy Nucleic Acids 16:733–744. https://doi.org/10.1016/j.omtn.2019.04.019
38. Sun X, Zheng J, Liu B, Huang Z, Chen F (2022) Characteristics of the enzyme-induced release of bitter peptides from wheat gluten hydrolysates. Front Nutr 9:1022257
39. Ishibashi N, Sadamori K, Yamamoto O, Kanehisa H, Kouge K, Kikuchi E et al (1987) Bitterness of phenylalanine- and tyrosine-containing peptides. Agric Biol Chem 51(12):3309–3313. https://doi.org/10.1080/00021369.1987.10868574
40. Gal J (2012) The discovery of stereoselectivity at biological receptors: arnaldo Piutti and the taste of the asparagine enantiomers—history and analysis on the 125th anniversary. Chirality 24(12):959–976
41. Schiffman SS, Dackis C (1975) Taste of nutrients: amino acids, vitamins, and fatty acids. Percept Psychophys 17(2):140–146. https://doi.org/10.3758/BF03203878
42. Amit SK, Uddin MM, Rahman R, Islam SMR, Khan MS (2017) A review on mechanisms and commercial aspects of food preservation and processing. Agric Food Secur 6(1):51. https://doi.org/10.1186/s40066-017-0130-8
43. Belitz HD, Wieser H (1985) Bitter compounds: occurrence and structure-activity relationships. Food Rev Int 1(2):271–354. https://doi.org/10.1080/87559128509540773
44. Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D (2015) Inchi, the IUPAC international chemical identifier. J Cheminform 7(1):23. https://doi.org/10.1186/s13321-015-0068-4
45. O'Boyle N, Dalke A. DeepSMILES: an adaptation of SMILES for use in machine-learning of chemical structures. 2018.
46. Cheng AH, Cai A, Miret S, Malkomes G, Phielipp M, Aspuru-Guzik A (2023) Group SELFIES: a robust fragment-based molecular string representation. Digital Discovery 2(3):748–758. https://doi.org/10.1039/d3dd00012e
47. Sangaraju VK, Pham NT, Wei L, Yu X, Manavalan B (2024) MAcppred 2.0: stacked deep learning for anticancer peptide prediction with integrated spatial and probabilistic feature representations. J Mol Biol 436(17):168687. https://doi.org/10.1016/j.jmb.2024.168687
48. Thi Phan L, Woo Park H, Pitti T, Madhavan T, Jeon Y-J, Manavalan B (2022) MLACP 2.0: an updated machine learning tool for anticancer peptide prediction. Computa Struct Biotechnol J. 20:4473–4480. https://doi.org/10.1016/j.csbj.2022.07.043
49. Charoenkwan P, Schaduangrat N, Lio P, Moni MA, Manavalan B, Shoombuatong W (2022) NEPTUNE: a novel computational approach for accurate and large-scale identification of tumor homing peptides. Comput Biol Med 148:105700. https://doi.org/10.1016/j.compbiomed.2022.105700
50. Charoenkwan P, Chiangjong W, Nantasenamat C, Moni MA, Lio' P, Manavalan B et al (2022) SCMTHP: a new approach for identifying and characterizing of tumor-homing peptides using estimated propensity scores of amino acids. Pharmaceutics. https://doi.org/10.3390/pharmaceutics14010122
51. Zhang X, Wei L, Ye X, Zhang K, Teng S, Li Z et al (2023) SiameseCPP: a sequence-based Siamese network to predict cell-penetrating peptides by contrastive learning. Brief Bioinform 24(1):bbac545. https://doi.org/10.1093/bib/bbac545
52. Manavalan B, Patra MC (2022) MLCPP 2.0: an updated cell-penetrating peptides and their uptake efficiency predictor. J Mol Biol 434(11):167604. https://doi.org/10.1016/j.jmb.2022.167604

Publisher's Note