Computational and Structural Biotechnology Journal 23 (2024) 2442–2452

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Classification of bioactive peptides: A systematic benchmark of models and encodings

Edoardo Bizzotto [a], Guido Zampieri [a,*], Laura Treu [a], Pasquale Filannino [b], Raffaella Di Cagno [c], Stefano Campanaro [a]

[a] Department of Biology, University of Padua, Via U. Bassi 58/b, Padova 35131, Italy
[b] Department of Soil, Plant and Food Science, University of Bari Aldo Moro, Via G. Amendola 165/a, Bari 70126, Italy
[c] Faculty of Agricultural, Environmental and Food Sciences, Free University of Bolzano, Piazza Universita, 5, Bolzano 39100, Italy

ARTICLE INFO

Keywords:
Bioactive peptide
Functional classification
Machine learning
Sequence encoding
Systematic evaluation

ABSTRACT

Bioactive peptides are short amino acid chains possessing biological activity and exerting physiological effects relevant to human health. Despite their therapeutic value, their identification remains a major problem, as it mainly relies on time-consuming in vitro tests. While bioinformatic tools for the identification of bioactive peptides are available, they are focused on specific functional classes and have not been systematically tested on realistic settings. To tackle this problem, bioactive peptide sequences and functions were here gathered from a variety of databases to generate a unified collection of bioactive peptides from microbial fermentation. This collection was organized into nine functional classes including some previously studied and some unexplored such as immunomodulatory, opioid and cardiovascular peptides. Upon assessing their sequence properties, four alternative encoding methods were tested in combination with a multitude of machine learning algorithms, from basic classifiers like logistic regression to advanced algorithms like BERT. Tests on a total of 171 models showed that, while some functions are intrinsically easier to detect, no single combination of classifiers and encoders worked universally well for all classes. For this reason, we unified all the best individual models for each class and generated CICERON (Classification of bIoaCtive pEptides fRom micrObial fermeNtation), a classification tool for the functional classification of peptides. State-of-the-art classifiers were found to underperform on our realistic benchmark dataset compared to the models included in CICERON. Altogether, our work provides a tool for real-world peptide classification and can serve as a benchmark for future model development.

1. Introduction

Bioactive peptides (BPs) are short chains of 2 to 50 amino acids with a molecular weight of less than 10 kDa exerting biological effects on unicellular and multicellular organisms [1,2]. BPs can have several beneficial functions and act as anti-inflammatory, antihypertensive and antidiabetic molecules [1]. Additionally, they can possess antibacterial, antiviral, and antifungal properties that are comparable to, or even surpass, those of antibiotics in terms of efficacy [3,4]. Their biological properties make them useful to complement or even replace conventional medication in the treatment of pathologies. At present, they are used in many applications to treat cardiovascular diseases, cancer, obesity, and neurodegenerative disorders [5–7]. Due to their importance in many therapeutic areas, they have received great recognition for their specialized and precise activities on target tissues and their limited bioavailability compared to traditional drugs [8]. Moreover, due to their structure, they can be easily modified to fine-tune their therapeutic potential. In the food industry, they have been studied as toxicity-free additives and can increase the nutritional value of both human and animal food products [9].

There are three main methods to obtain bioactive peptides: chemical and physical hydrolysis, microbial fermentation and enzymatic hydrolysis [8]. In the former, proteins are placed in acidic or basic environments at different temperatures or are subjected to microwave irradiation or ultrasonic treatment to break down amino acid chains [10]. In the latter, single or multiple peptidases from animal or plant sources are added to the substrate to obtain BPs [11,12]. In substrates fermented by bacteria and fungi, microbial peptidases hydrolyze amino acid chains [13,14]. After converting the protein matrix into smaller peptides, the sequence of the BPs can be determined through

* Corresponding author.
E-mail address: guido.zampieri@unipd.it (G. Zampieri).

https://doi.org/10.1016/j.csbj.2024.05.040
Received 19 March 2024; Received in revised form 10 May 2024; Accepted 22 May 2024
Available online 23 May 2024

chromatographic purification and mass spectrometry. BPs can derive from proteolysis where the primary sources are proteins derived from plants, animals, and even marine species [15]. The most extensively studied sources of precursor proteins are those used for human nutrition, including milk and other dairy products, soybean and green leafy vegetables [16,17]. By-products derived from agrifood processes are being considered as an additional source of BPs; their use is of particular interest because extraction of high-value biomolecules can open up a new way for the recycling of waste products, such as animal skin and feathers, or fruit peel [18,19]. Subsequently, the functional activity of BPs must be verified and quantified. A series of steps, including purification, synthesis of the peptide and in vitro testing, must be performed. These procedures are time consuming and expensive, presenting significant challenges that can hinder progress in the drug discovery process. Therefore, there is a need for the development of more efficient and cost-effective methods for these processes.

Given the importance of BPs, there have been several attempts to create in-silico approaches to perform a preliminary assignment of the potential functional properties and facilitate the subsequent discovery and testing process in vitro [20–25]. These methods rely on several databases where peptides from various experiments have been collected and classified according to the BPs functional classes. For instance, BIOPEP-UWM currently lists 62 different functional classes, which can include a single peptide up to hundreds of individual molecules [26]. Using the sequence properties of the peptides, such as amino acid composition, or the presence of sequence patterns of interest, peptides can be assigned to a functional class depending on the type of classifier used. Tools include similarity-based classification using available sequences from databases [26,27] and prediction of physicochemical properties [28]. There have also been several attempts to use machine learning techniques to aid in the detection of BPs and their functional classification [20–25]. The proposed methods used Logistic Regression [29], Support Vector Machines [30] and Random Forests [31] to predict BPs' functional role. In recent years, neural networks [23,32,33] and algorithms based on Natural Language Processing (NLP) have also been employed for the same purpose. These studies tend to focus solely on a single method for classification and in most cases they do not compare the performance of the model with other algorithms in settings representative of real use cases. While the intrinsic characteristics of bioactive peptides require the use of different techniques for classification, there exist no universally accepted benchmark settings for testing this type of problem. Moreover, in order to use these different classification methods, peptide sequences need to be transformed into appropriate signals that can be processed by the algorithm of choice. Various encoding techniques have been developed to take into account different properties [34,35]. For instance, amino acids can be encoded by assigning them a number based on the order of the 20 conventional amino acids, or by calculating their frequency in one sequence. Other methods use the physicochemical parameters or secondary structures of BPs. While these techniques have been proven to be useful in transferring information from the sequence to the machine learning algorithm, they sometimes fail to convey important details, such as physical or chemical properties or relationships occurring among amino acids. This leads to predictors that are highly condition-specific and that cannot be generalized to problems other than the ones they were developed for [36].

The aim of this study was to bridge the existing gap between classification method development and the practical applicability of developed tools. To this goal, we generated a set of classifiers for the identification of peptide functions, with a particular focus on BPs derived from microbial fermentation. There are several reasons why microbial fermentation is preferable to enzymatic or physicochemical methods of extraction. First and foremost, utilizing microbial species is far cheaper than traditional methods and does not require solvents or exogenous enzymes, making it a green and sustainable alternative. Moreover, the high number of species and the variability in the enzyme portfolio produce a higher number of bioactive peptides of different sizes. As far as the authors are aware, there are no studies focusing specifically on BPs from microbial sources with the exception of the formation of specific databases focused on fermented food peptides. For this reason, we propose CICERON, a tool to classify the functions of BPs specifically derived from microbial fermentation.

Different machine learning and encoding methods were systematically evaluated across nine functional classes, for a total of 171 distinct classifiers. This approach allowed us to highlight the differences in the techniques employed and their performance in evaluating the individual classes. The various encodings and machine learning techniques were tested to provide a benchmark for future studies focusing on peptides derived from microbial fermentation. For each considered functional class, the most accurate predictor has been selected to suit the intrinsic characteristics of each function. The final result, CICERON, consists of nine different binary classifiers capable of identifying the products of microbial fermentation-derived BPs. The database of microbial peptides used in this study and the best model for each functional class can be found at https://github.com/BizzoTL/CICERON/.

2. Methods

2.1. Dataset preparation

Peptide data from BIOPEP-UWM [26], the Milk Bioactive Peptide Database [37], BioPepDB [38] and FermFoodDB [27] were downloaded and merged into a unified database. Data collected from these databases are represented only by food proteins derived from microbial fermentation. BPs that were present more than once due to overlaps in the database content were clustered together to remove redundancy. Peptides with identical sequences but different functional class assignments were removed to avoid introducing potential biases in classifier training. Additionally, sequences that had more than 90 % similarity were also clustered together if the corresponding peptide belonged to the same functional class, otherwise they were excluded from the analysis. Peptides longer than 100 amino acids, shorter than three and those containing unconventional amino acids or symbols were also removed.

Functional classes were then homogenized by grouping together those with the same or overlapping biological functions, as follows. "Antihypertensive", "ACE-inhibitory" and "Renin-inhibitory" were set as "Antihypertensive" as angiotensin-converting enzymes and renin are the two main regulators of blood pressure in humans [39,40]. "DPP-IV inhibitors" and "alpha-glucosidase inhibitors" were grouped into "Antidiabetic" due to their inhibiting action against type 2 diabetes by reducing blood glucose [41,42]. "Antimicrobial", "antifungal", "antibacterial" and "anticancer" were all defined as "Antimicrobial", incorporating all the peptides with disruptive properties against either bacteria or fungi [43,44]. Anticancer peptides were also included according to their ability to kill bacterial cells [45,46]. "Antithrombotic", and "CaMKII Inhibitor" were set as "Cardiovascular" as antithrombotic peptides have positive effects on vascular circulation [47], while CaMKII inhibitors prevent cardiomyopathies and arrhythmogenesis [48]. "Antiamnestic", "anxiolytic-like", "AChE inhibitors", "PEP-inhibitory" and "neuropeptides" were grouped into "Neuropeptides". While having effects on different neural pathways, these peptides all positively contribute to neurological processes [49–52].

Among all the considered functional classes, BPs associated with celiac disease are the only ones that determine a toxicity-inducing effect due to their allergenic properties; on the contrary, all the others have a positive effect on human health. Such a distinction is expected not to be linked to classification performance differences as only BP sequence is taken into consideration for the analysis and every functional class is treated independently of the others. Finally, classes with fewer than 100 peptides were removed.

2.2. Functional class characterization

Amino acid, dipeptide and tripeptide frequencies were calculated to infer possible relationships between functional classes and amino acid composition. MERCI [53] was used with default settings to discover sequence motifs unique for each class. Binomial tests with false discovery rate correction from the Scipy python package[54] were carried out to infer the statistical significance of the results, in order to ascertain whether the most frequent dipeptide or tripeptide sequences were more prevalent in one class compared to others. For each sequence of interest, the probability of finding any specific motif was calculated across all samples and compared to the distribution of the same motif within the class of interest. Amino acid, dipeptide, and tripeptide motifs that were statistically significant (adjusted p-value lower than 0.05) are reported in Supplementary Table 2.

2.3. Encoding methods

BPs were encoded into appropriate inputs for different machine-learning algorithms (Fig. 1). Four different encoding methods were devised for Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbor (KNN) and Logistic Regression (LR) models. In the "sparse" encoding method, each peptide was transformed into a vector of length 100, representing the maximum possible sequence length in the database. Using a list of the 20 conventional amino acids and an additional element representing empty positions, each amino acid in the sequence was encoded as a 21-element vector. The position of the amino acid in the list was assigned a value of "1," while the remaining 20 elements were set to "0." Thus, for each position in the vector of length 100, there was a corresponding vector of length 21. The "dense" encoding method also utilized a vector of length 100. However, instead of one-hot encoding for each position, the indices from 1 to 20 representing the 20 conventional amino acids were used, along with an element indicating an empty position. The third encoding method, "BLOMAP" [55], used the BLOSUM62 substitution matrix to encode each amino acid as the vector of the substitution probability with other amino acids. The additional character indicating a missing amino acid was given a − 100 % probability of substitution with the other 20 amino acids. The last encoding method, "threemers", uses a sliding window of three amino acids to count all the three-mers in the peptide. These counts were then stored in a dictionary encompassing all possible combinations of the 20 common amino acids. The Neural Network (NN) models used "sparse" and "threemers" encodings to map the input for the first convolutional layer. The encoding for the BERT protein learning model, ProtBERT [56], consisted in separating each sequence into single amino acids and treating them as a single element, as follows. The maximum length of the encoded vector was set to 102 elements. The first element in the vector was a special character that indicates the start of the peptide, followed by the tokenized amino acids, and ending with the special character indicating the end of the sequence. Any remaining positions were padded to ensure uniform vector lengths of 102. The entire peptide sequence is treated as an entire phrase in the language model. All the BPs were encoded with the methods described above and a t-SNE [57] representation was used to infer information on the sequence

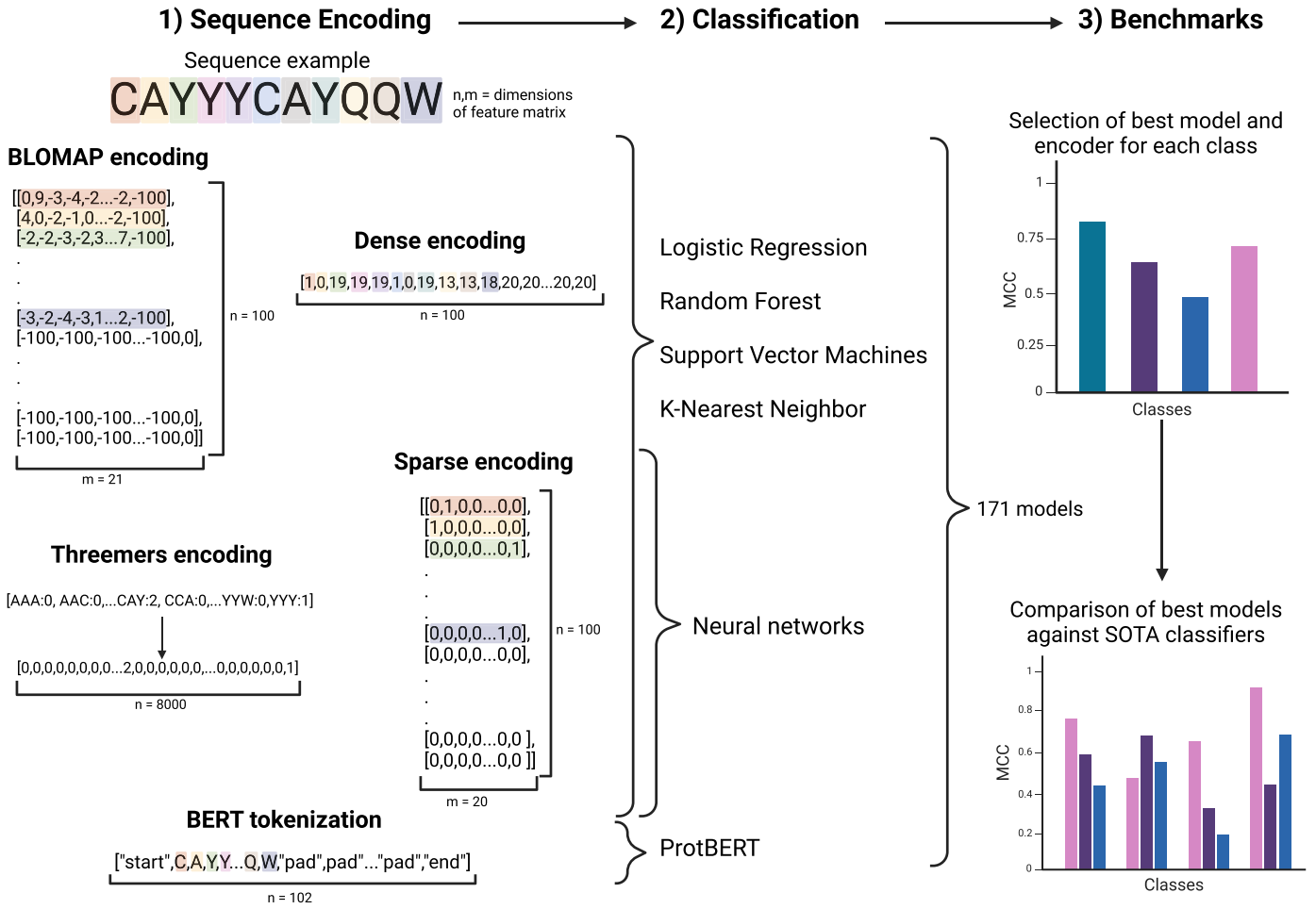

Fig. 1. Example of classification of a bioactive peptide sequence. 1) The sequence was first transformed into the appropriate vectors according to the encoding. 2) The encoded sequence was then used in the classification by different algorithms. 3) The best generated model for each class is selected to be compared against SOTA classifiers.

distribution of functional groups (Supplementary Figure 1).

2.4. Classifiers

For every functional prediction, the samples belonging to the specific functional group under investigation were labeled as "0", while the remaining samples were labeled as "1". In this classification scheme, the positive samples for each class are exclusively those that pertain to the specific function under consideration. Conversely, the negative samples comprise all samples from the remaining classes, excluding the one currently under examination. This strategic selection of the negative dataset aims to capture a representative distribution of microbial fermentation peptides, enabling the models to more effectively distinguish the positive class from the rest of experimentally obtainable examples. LR, SVM, RF and KNN were implemented using the Scikit-learn Python Package version 1.1.3 [58]. Convolutional NNs were implemented using the Keras framework version 2.11.0 [59]. The architecture consists of two convolutional layers, each followed by two max pooling operations. This was then followed by a dropout layer and a flattening operation. Two dense layers were added, with the last one being the output layer. A callback function was implemented to stop the model training if the loss increases, in order to avoid overfitting.

2.5. Model evaluation and parameter optimization

Each of the functional classes in the database was modeled using a binary classifier for each machine learning algorithm previously described. Due to the significant imbalance in the number of BPs present in the different classes, accuracy alone cannot adequately describe the performance of the models. The metrics used in the classification were therefore the Area Under the Receiver Operating Characteristic curve (AUROC) and Matthew's Correlation Coefficient (MCC). For both traditional classifiers and NN the dataset was split into training, validation, and test sets in a ratio of 70:20:10. The number of epochs was set to 50, and the batch size was set to 20. The encoding methods tested for the CNN models were "sparse" and "threemers". The protein learning model, based on the ProtBERT pre-trained model, was implemented using the HugginFace Transformers framework version 4.26.0 [60]. Each binary classifier was trained using an 80:20 training test split for 20 epochs. After each epoch, the trained classifier was tested on the validation set and the resulting metrics were recorded. Among all the machine learning methods tested, the model with the best MCC obtained from the test set was selected as the representative for each class.

The following parameters were selected in order to optimize the performance of the model: "C", "penalty", "solver" for LR; "criterion", "max_features", "n_estimators" for RF; "algorithm", "leaf_size", "n_neighbors", "weights" for KNN; "C", "class_weight", "gamma", "kernel" for SVM. A grid search was performed to select the best parameters for each model, followed by cross-validation, using the GridSearchCV module from scikit-learn. The best parameters obtained for each model, based on the highest MCC value obtained, are described in Supplementary Table 1. For NN models, a genetic algorithm was developed using the DEAP Python library [61]. The weights of the positive and negative classes, learning rates, and kernel size were encoded as a vector and used as individuals in the genetic algorithm. The MCC obtained from the test set served as the fitness value for the individuals. After each generation, the top 5 individuals with the highest fitness were selected as parents for producing offspring in the next generation. A randomly selected individual from the pool of the best individuals had a 25 % chance of changing each hyperparameter using a random value. The allowed intervals for the class weights ranged from 1 to 10 (float values), for the learning rates it ranged from 0.0001 to 0.1, and for the kernel size it ranged from 5 to 500 (integer values). For the first 10 generations, the number of offspring was set to 15, while the remaining 5 individuals were randomly generated within the set of possible intervals, in order to explore more the search space and avoid local optima. After the 10th generation, all 20 individuals were generated as offspring from the best individuals of the previous generations. In order to select the best hyperparameters for the singular classes, the genetic algorithm consisted of 20 generations of 20 individuals each for every functional class. The vector used to obtain the best neural network model for each class is described in Supplementary Table 1. The metrics obtained from the best individual of the 20 generations were further evaluated using five-fold cross-validation to obtain the final metric values.

2.6. CICERON implementation

CICERON consists of a Python script that takes one or more FASTA files as input and returns one or more functional predictions for every peptide reported in the file. The first step consists of checking that the input sequence is present in the database of bioactive peptides and if the match is identical then the associated function is reported in the final table. The second step consists of the search for motifs: if the sequence contains one of the class-specific motifs previously found using MERCI, the corresponding functional class is reported. In each step, if multiple associations are confirmed, they are all included. Finally, the best classifier model for each functional class predicts the probability of the peptide belonging to that group or not. The results are collected in a tab-separated file for each file in input.

3. Results and discussion

3.1. Exploration of sequence characteristics across functional classes

In our starting dataset, a total of 13,123 peptides divided into 56 functional groups were obtained from different online databases. After filtering the sequences, and merging functionally overlapping or related classes, the final database consisted of 3990 BPs divided into nine different functional groups (Table 1). Such classes present substantial differences both in terms of number of peptides and sequence characteristics. The number of BPs per class ranges between 107 for "immunomodulatory" to 1386 for "antihypertensive". The difference in the number and the length of BPs for each class reflects the difficulty in performing certain essays for the identification of functions in vitro and the higher interest in certain functions over others. It is also possible that some biological functions can be performed only by a very limited number of AA sequences, thus intrinsically limiting the number of BP in the class. Moreover, after the production of the digested substrates, peptides are filtered from other molecules based on the molecular weight, prioritizing shorter peptides over longer ones and thus reducing the number of BPs that are usually longer, such as antimicrobial BPs. While the average peptide length is 9.5 amino acids, the length distribution is quite different between functional classes. For antidiabetic and immunomodulatory classes the peptide length is shorter than the global average, while for antimicrobial the peptides are significantly longer

Table 1
Average sequence length, number of peptides and range of peptide length from the shortest to the longest sequence for each functional class.

Peptide class	Average peptide length	Number of peptides	Peptide length variation
Antidiabetic	6.94 ± 3.16	263	3–21
Antihypertensive	7.03 ± 4.11	1386	3–54
Antimicrobial	22.19 ± 14.75	540	3–94
Antioxidant	6.69 ± 5.18	1001	3–57
Cardiovascular	11.58 ± 11.75	148	3–84
Celiac disease	11.74 ± 4.39	240	4–34
Immunomodulatory	7.56 ± 4.39	107	3–20
Neuropeptides	9.67 ± 4.09	165	3–47
Opioid	7.22 ± 5.93	140	3–31
All classes	9.57 ± 3.88	3990	3–94

than the mean, more than double in comparison to other functional groups (Supplementary Figure 2). In the t-SNE representations [57] (Supplementary Figure 1), the different classes are not clustered together, with the exception of celiac disease peptides in “sparse” and “threemers” encodings, where the majority of these samples are clustered in a unique group.

To compare the amino acid usage in the functional classes, single amino acid, dipeptide, and tripeptide frequencies were plotted (Fig. 2). The amino acid frequency plot in Fig. 2A reveals that some BP classes have distinct characteristics. For example, celiac disease BPs have the highest frequency of proline and glutamine, opioid peptides are enriched in tyrosine and glycine, while cardiovascular BPs have slightly higher frequencies of alanine and the highest frequency of the negatively charged amino acids aspartic acid and glutamic acid.

Distinct patterns are also visible in the frequency of dipeptides (Fig. 2B), with high-frequency dipeptides mainly localized in celiac disease-associated and opioid BPs, immediately followed by antidiabetics. This trend is further confirmed by the tripeptide distribution (Fig. 2C), where celiac disease and opioid peptides have nine and five amino acid triplets, respectively, occurring at higher frequencies. Other highly frequent tripeptides belong to the neuropeptides class (FLR, LRF and NFL) and the immunomodulatory class (RKP). PIP and PGP are tripeptides that have a high frequency in multiple classes. A binomial test also identified that many other sequences were present with a higher

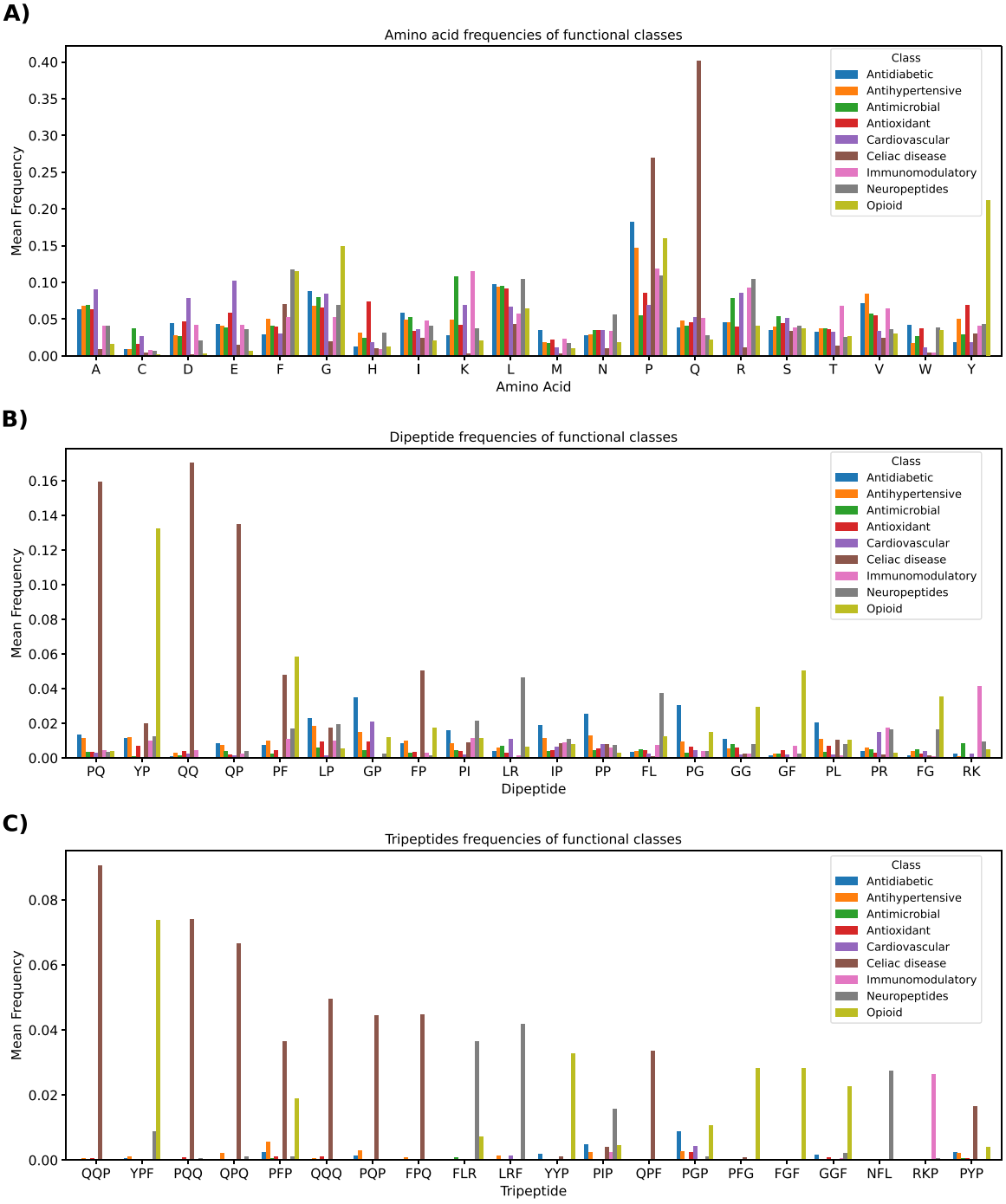

Fig. 2. Amino acid composition of peptides for each functional class: A) Mean amino acid frequencies for single amino acids, B) Mean dipeptide frequencies of the 20 most abundant dipeptides, C) Mean tripeptide frequencies for the 20 most abundant tripeptides.

frequency in certain classes, providing additional insights and potential patterns that can be exploited for classification (Supplementary Table 2).

Further analysis using MERCI [53] revealed unique motifs for some functional classes. For celiac disease, the QPF triplet was found 67 times in the 240 samples and 9 other motifs were found in at least 10 % of the sequences. Glutamine and proline were prominent in these motifs, occurring at least once per motif. All the discovered motifs from MERCI are distinct for each class and are not found in samples that belong to a different function. Although unique motifs can be used to positively identify classes, they should not be solely relied upon for functional classification. For example, in the work of Tomer et al. [62], the QPQ triplet was described as highly conserved in celiac disease; however, in the database produced in this work, this motif was found in 21 non-celiac disease peptides. Thus, the presence of a motif alone does not unequivocally determine the functional class of a peptide. Another possible option is that these 21 peptides have multiple functions, including also celiac toxicity. The distribution of amino acid frequencies, along with specific substructures, can provide insights and aid in the identification and classification of bioactive peptides, but they should not be the sole method of peptide functional classification. All the motifs found for each class are reported in Supplementary Table 2.

3.2. Benchmark model evaluation across functional classes

In order to establish a benchmark for the classification of BPs, combinations of several sequence encoding and machine learning techniques were tested across the nine functional classes (Fig. 1, see 2.3). To address the class imbalance in the dataset, MCC was used as the evaluation metric for comparing the performance of different classifiers, as shown in Fig. 3. The obtained AUROC was also reported in Supplementary Figure 2, and the performance metrics obtained during training and testing are reported in Supplementary Table 1.

The Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbor (KNN) and Logistic Regression (LR) algorithms were assessed in conjunction with “sparse”, “dense”, “BLOMAP” and “threemers” encoding methods. In general, each classifier exhibited varying levels of performance when coupled with the different encoding methods and no single encoding method outperformed the others, as evidenced by the findings presented in Fig. 3. For example, the “threemers” encoding exhibited exceptional effectiveness in the classification of cardiovascular, immunomodulatory, and antioxidant classes, likely due to the presence of specific tripeptides that augmented the classification power. On the other hand, the "dense" and "BLOMAP" encodings yielded the most favorable MCC values for neuropeptides and antihypertensives, respectively. Despite such a heterogeneity, the "sparse" and "threemers" encoding methods yielded the highest number of classifiers with MCC values exceeding 0.5, and consequently they were selected as the encoding methods for NN models to reduce the training time. Again, a highly variable performance was observed depending on the considered functional class. Overall, the variability in the optimal encoding method across all classes underscores the necessity for a robust benchmark encompassing a wide range of encoding strategies to ensure the comprehensive capture of relevant features in BPs classification.

When comparing classification model types, while no single machine learning method consistently excelled across all classes, RF demonstrated superior models for five out of nine classes, indicating its effectiveness in discovering relevant features within the dataset. KNN, on the other hand, displayed less favorable overall performance, with the exception of the antidiabetic class. However, the performance for this class, the one only under 0.5 MCC, is the lowest among all classes, indicating that the considered encoding methods are not sufficient to capture enough information to correctly classify these peptides. Notably, LR emerged as the optimal classification method for the celiac disease class, achieving the highest MCC value among all the classifiers at 0.923 with the “threemers” encoding. This excellent performance can be attributed to the high fraction of distinctive motifs in the class, which facilitates effective separation based on the presence of specific tripeptides, a characteristic more prevalent within this class (Fig. 2 and Supplementary Table 2). Along with celiac disease BPs, opioids and antimicrobials display the highest scores on average, with MCC values of 0.783 and 0.724, respectively, achieved by NN in conjunction with the “sparse” feature encoding. This enhanced performance for antimicrobials was likely due to the significantly longer peptide sequences in antimicrobial peptides, averaging over 22 amino acids, four times the overall average length of all BPs. For the opioid class, the high performance is likely due to the presence of specific motifs, as it is the second class with the most frequent tripeptide motifs, after celiac disease, as seen in Fig. 2. Moreover, we verified that the NLP-based ProtBERT model, pre-trained on a vast corpus of sequences, does not provide improved performances in the present tasks except for antimicrobial peptides, where ProtBERT exhibited a slight performance improvement over NN, achieving an MCC of 0.731. The length of antimicrobial peptides makes them more likely to be classified using the BERT-derived algorithm. However, it is worth noting that the use of the native ProtBERT tokenizer as an encoding method for such a model might have limited its performance in certain cases.

As for the classes with lower performance, several reasons can be attributed to the cases of erroneous classification. For the antidiabetic and immunomodulatory classes, the low number of available peptide sequences likely impacts the classification performance. Increasing the number of sequences in the positive class could help the detection of patterns associated with these functional classes. Antioxidant BPs are among the classes with the most sequences, but a consistent fraction is shorter than six AA residues. This might prevent the identification of meaningful patterns but rather act as a confounding factor for this class. Moreover, antioxidants generally constitute a functionally broad biological class, and they could be divided into subclasses, for example taking into account their oxidative action, to carefully capture intra-class diversity. While these reasons could contribute to the low classification performance for these classes, the implementation of different encoding or machine learning methods could partially compensate for the limitations in the data.

Such a systematic evaluation unveiled a high heterogeneity across functional classes, both in terms of distinctive sequence characteristics and model power. No single classification algorithm or encoding method is capable of systematically outperforming others, highlighting the need for robust benchmark testing with multiple techniques to better elucidate the differences in class-specific features. Ultimately, the best-performing encoding and classification method for each class was selected and subsequently incorporated into the CICERON tool.

3.3. Comparison between CICERON and state-of-the-art classifiers

To evaluate the practical effectiveness of current BP classification methods, CICERON was tested against state-of-the-art (SOTA) classifiers specific to the various peptide functional classes, which were selected upon careful literature inspection [62–67]. The main characteristics of SOTA models and their development dataset are summarized in Table 2. It is worth noting that certain functional classes, such as opioid, immunomodulatory, and cardiovascular peptides, could not be included in our comparative analysis due to the absence of available classifiers for these specific categories. Additionally, for the antihypertensive and neuropeptides classes, the tools described in relevant literature were not accessible for utilization. For the antimicrobial class, it is important to mention that the web server version of the tool allowed the submission of only one sequence at a time, while the source code version lacked comprehensive installation and usage guidance for the package. For the other three classes, namely antidiabetic, antioxidative, and celiac disease, the database of peptides generated in this study was used to compare the performance of these purpose-specific tools against CICERON.

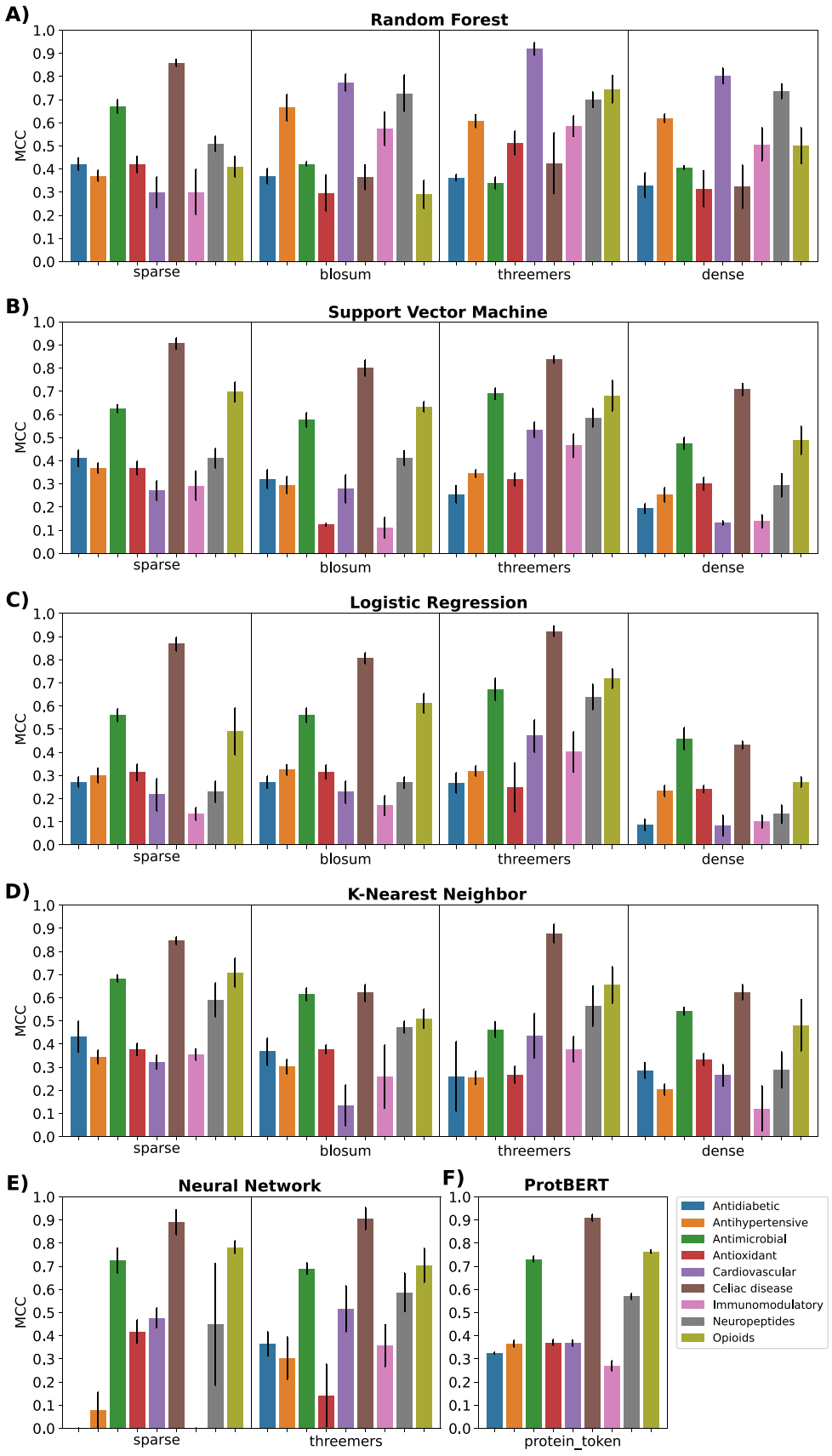

Fig. 3. Matthews Correlation Coefficient value on the test set for every single classifier and encoding method. Each plot describes the performance of a different machine learning method: A) Random Forest, B) Support Vector Machine, C)Logistic Regression, D) K-Nearest Neighbor, E) Neural Network, F) ProtBERT.

Table 2
Comparison of state-of-the-art classifier and CICERON for each functional class. Opioid, immunomodulatory and cardiovascular classifiers were included despite the absence of previously available classifiers, to provide a comprehensive overview of the information produced in this work. The characteristics that were considered are the following: model, encoding, type of tool and dataset used; MCC on test dataset, limitations of the tool.

Functional Class	Tools	Model	Encoding	Type of tool	Dataset	MCC on test set	Limitations	Ref
Antidiabetic	AntiDMPpred	Random Forest	Multiple encodings were merged	Web tool	236 antidiabetic and 236 non-antidiabetic peptides	0.133	Validated on peptides between 5 and 50 AA	[63]
	CICERON	KNN	Sparse (One hot encoding)	Python package	263 antidiabetic and 3727 non-antidiabetic peptides	0.432	No limitations	This work
Celiac disease	CDpred	ExtraTree classifier	Amino Acid Composition (AAC).	Web tool	503 celiac disease and 503 non-celiac disease peptides	0.47	No limitations	[62]
	CICERON	Logistic Regression	Threemers (Tripeptide composition)	Python package	240 celiac disease and 3750 non-celiac disease peptides	0.923	No limitations	This work
Antioxidant	AnOxPP	BiLSTM Neural Network	22 Amino Acids Descriptors (AADs).	Web tool	1060 antioxidant and 1060 non-antioxidant peptides	-0.005	Applicable on peptide sequences between 2 and 19 AA	[64]
	CICERON	Random Forest	Threemers (Tripeptide composition)	Python package	1001 antioxidant and 2989 non-antioxidant peptides	0.513	No limitations	This work
Neuropeptides	Target-ensC_NP	Ensemble of ETC, LGBM, SVM, XGB, and ADA	One-hot encoding of single AA	Python package	2435 neuropeptides and 2435 non-neuropeptides	MCC not available	Tool not available for use	[65]
	CICERON	Random Forest	Dense (One hot encoding)	Python package	165 neuropeptides and 3825 non-neuropeptides	0.736	No limitations	This work
Antihypertensive	Ensemble-AHTPpred	Ensemble of RF, SVM and XGB	431 numerical features	Web tool	913 antihypertensive and 913 non hypertensive peptides	MCC not available	Tool not available for use	[67]
	CICERON	Random Forest	BLOMAP	Python package	1386 antihypertensive and 2604 non hypertensive peptides	0.665	No limitations	This work
Antimicrobial	Antimicrobial-peptide-generation	SeqGAN and BERT	BERT tokenization	Web tool and Python package	4134 AMPs and 4134 non-AMPs	MCC not available	Validated on peptides between 11 and 30 AA; web tool only accepting one sequence at once; python package lacking a guide for utilization and requiring retraining the model	[66]
	CICERON	ProtBERT	BERT tokenization	Python package	540 AMPs and 3450 non-AMPs	0.731	No limitations	This work
Opioid	CICERON	Random Forest	Threemers (Tripeptide composition)	Python package	140 opioid and 3858 non opioid peptides	0.745	No limitations	This work
Cardiovascular	CICERON	Random Forest	Threemers (Tripeptide composition)	Python package	156 cardiovascular and 3842 non cardiovascular peptides	0.919	No limitations	This work
Immunomodulatory	CICERON	Random Forest	Threemers (Tripeptide composition)	Python package	107 immunomodulatory and 3891 non immunomodulatory peptides	0.586	No limitations	This work

Classification results on the test set reveal that CICERON outperformed SOTA models across all the three comparable classes. Specifically, CICERON exhibited MCC values that were consistently 30–50 % higher than those of the benchmarked SOTA models. The notable difference observed in the antidiabetic classifier's performance can be attributed to the relatively constrained dataset employed, characterized by a limited number of samples for both positive and negative classes, thereby restricting the model's classification capabilities. In contrast, the lower MCC values observed for the celiac disease and antioxidative classes primarily stem from an increased number of false negatives in the predictions. It is worth mentioning that a portion of the positive samples for these classes was collected from the same databases utilized for data generation in this study. Consequently, it is the negative dataset that fundamentally contributes to the different performance of the tools.

Most of the SOTA class-specific classifiers are trained on a large dataset of positive examples and a random dataset of negative examples. While this ratio favors the classification of the positive class, it does not take into account that in a real-life scenario of a microbial fermentation the peptide that is generated that does not belong to the positive class is not random but is more dependent on the substrate utilized. For example, if the substrate to be used is from a protein matrix derived from grains, the amount of celiac disease peptides would be much higher than the rest of the other classes. For this reason, in the case of microbial fermentation, the choice of the positive/negative dataset plays an important role in order to get closer to the real-life scenario, as the resulting products are not limited to only one functional class. CICERON adopts a distinct strategy since it considers peptides from other functional classes as negative datasets, thus enhancing its ability to discern both positive and negative sample characteristics.

An additional distinction lies in the peptide length criteria, where each class has predefined minimum and maximum sequence length constraints (Table 2), thereby restricting the classification to peptides within those specific ranges for SOTA classifiers. If enzymatic proteolysis or physicochemical proteolysis are also compared, the length of the peptides obtained is different than in the case of microbial fermentation,

where only certain microbial enzymes are present and thus the digested proteins produce longer peptides. This highlights the need to have separate datasets for each proposed proteolysis method in order to obtain more accurate predictions similar to a real experiment. Upon a comprehensive review of all SOTA classifiers, it emerges there is not a common encoding or classification methodology shared among these models. If the objective is to obtain a specific class of peptide, i.e. antimicrobial peptides are needed from fermentation, a class-specific classifier can in principle be more suited to the case. However, it cannot be excluded that a function-specific classifier is not able to see if the peptide has more than one function. In one extreme case, if the peptide is antihypertensive but at the same time it is associated with celiac disease, it could cause problems for a patient suffering from the disease. A preliminary screening with CICERON could help identify these cases very early in the screening process, eliminating the need for costly and time-consuming tests.

4. Conclusions

To the best of the authors' knowledge, this is the first systematic benchmark specifically focusing on the functional classification of BPs by utilizing a comprehensive dataset of peptides derived from food fermentation. In the context of generating bioactive molecules from fermentation, CICERON can aid in the discovery of products in the final phases of the analysis, reducing the need for costly and time consuming in vitro experiments for the characterization of the functional role of the peptides generated. In this study, we successfully established new machine learning classifiers for previously unaddressed functional classes, including opioids, cardiovascular peptides, and immunomodulatory peptides. Furthermore, classifiers for antidiabetic, celiac disease and antioxidative peptides demonstrated superior performance on a test close to a real-case scenario compared to previously developed models. Nevertheless, certain classes, such as antidiabetic and immunomodulatory peptides, still present challenges in terms of classification performance, highlighting the importance of expanding the number of sequences in these groups. Our comprehensive analysis, encompassing diverse encoding techniques and classification methodologies, unveiled the unique requirements of each functional class for accurate functional identification. Obviously, there is no one-size-fits-all solution for all classes. In general, combining a multitude of encoding and machine learning techniques within the same task could increase classification performance, while a robust feature selection could help to obtain more powerful classifiers by removing unnecessary features [68]. The most relevant ones can then be analyzed to determine whether they contribute positively or negatively to the classification [68]. Moreover, the examination of amino acid composition revealed distinctive preferences within certain groups, enabling discrimination based on specific sequence motifs. While these attributes do not exclusively determine functional group affiliation, they can significantly enhance the classification efficacy of the classifiers. From the present study, it clearly emerges that conventional methods such as LR can be effectively employed in classification tasks and that more advanced algorithms such as NN do not necessarily provide better results. Another important aspect resulting from the comparison of the models generated in this study with SOTA classifiers is the fact that CICERON is a standalone application comprising nine different classifiers targeting as many functional classes, which do not require separate installation, providing a simplified framework for multi-class analysis. In addition, there are no limitations in terms of minimum or maximum peptide length and total number of peptides that can be analyzed at once. In conclusion, while CICERON provides valuable insights into the classification of peptide functions, the implications arising from the aforementioned limitations, particularly in certain functional classes, warrant careful consideration. However, in the context of a microbial fermentation experiment, where the goal is to characterize the full spectrum of peptides present within the substrate, a versatile tool like CICERON proves to be more efficient than SOTA classifiers in identifying multiple functions associated with one single peptide, making it a valuable tool for exploratory investigations into peptide functions within complex biological systems.

Funding

This study was supported by the MIUR Research Projects of National Relevance (PRIN) "Future-proof bioactive peptides from food by-products: an eco-sustainable bioprocessing for tailored multifunctional foods" (PROACTIVE) – Prot. 2020CNRB84 and by the MIUR FFO-2022 "L'innovazione delle biotecnologie nell'era della medicina di precisione, dei cambiamenti climatici e dell'economia circolare (Una piattaforma biotecnologica per lo sfruttamento dei sottoprodotti agroindustriali nella coltivazione di microrganismi benefici)".

Author statement

The authors declare that all of them have seen and approved the final version of the manuscript being submitted. The manuscript is the authors' original work, has not received prior publication and is not under consideration for publication elsewhere.

CRediT authorship contribution statement

Guido Zampieri: Writing – review & editing, Methodology, Conceptualization. **Edoardo Bizzotto:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Pasquale Filannino:** Writing – review & editing, Funding acquisition. **Laura Treu:** Writing – review & editing, Supervision. **Stefano Campanaro:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization. **Raffaella Di Cagno:** Writing – review & editing, Funding acquisition.

Declaration of Competing Interest

The authors declare no conflict of interest.

Data Availability

The database used in this study, the scripts to reproduce the analyses, and the best model for each functional class can be found at https://github.com/BizzoTL/CICERON/.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.05.040.

References

[1] Bhandari D, Rafiq S, Gat Y, Gat P, Waghmare R, Kumar V. A review on bioactive peptides: physiological functions, bioavailability and safety. Int J Pept Res Ther 2020;26:139–50. https://doi.org/10.1007/s10989-019-09823-5.
[2] Sánchez A, Vázquez A. Bioactive peptides: a review. Food Qual Saf 2017;1:29–46. https://doi.org/10.1093/fqsafe/fyx006.
[3] Da Silva J, Leal EC, Carvalho E. Bioactive antimicrobial peptides as therapeutic agents for infected diabetic foot ulcers. Biomolecules 2021;11:1894. https://doi.org/10.3390/biom11121894.
[4] Haney EF, Mansour SC, Hancock REW. Antimicrobial peptides: an introduction. In: Hansen PR, editor. Antimicrob. Pept. Methods Protoc. New York, NY: Springer; 2017. p. 3–22. https://doi.org/10.1007/978-1-4939-6737-7_1.
[5] Perlikowska R, Silva J, Alves C, Susano P, Pedrosa R. The therapeutic potential of naturally occurring peptides in counteracting SH-SY5Y cells injury. Int J Mol Sci 2022;23:11778. https://doi.org/10.3390/ijms231911778.
[6] da Costa JP, Cova M, Ferreira R, Vitorino R. Antimicrobial peptides: an alternative for innovative medicines? Appl Microbiol Biotechnol 2015;99:2023–40. https://doi.org/10.1007/s00253-015-6375-x.

[7] Ma W, Li N, Lin L, Wen J, Zhao C, Wang F. Research progress in lipid metabolic regulation of bioactive peptides. Food Prod Process Nutr 2023;5:10. https://doi.org/10.1186/s43014-022-00123-y.

[8] Akbarian M, Khani A, Eghbalpour S, Uversky VN. Bioactive peptides: synthesis, sources, applications, and proposed mechanisms of action. Int J Mol Sci 2022;23:1445. https://doi.org/10.3390/ijms23031445.

[9] Zaky AA, Simal-Gandara J, Eun J-B, Shim J-H, Abd El-Aty AM. Bioactivities, applications, safety, and health benefits of bioactive peptides from food and by-products: a review. Front Nutr 2022;8.

[10] Kadam SU, Tiwari BK, Álvarez C, O'Donnell CP. Ultrasound applications for the extraction, identification and delivery of food proteins and bioactive peptides. Trends Food Sci Technol 2015;46:60–7. https://doi.org/10.1016/j.tifs.2015.07.012.

[11] Cruz-Casas DE, Aguilar CN, Ascacio-Valdés JA, Rodríguez-Herrera R, Chávez-González ML, Flores-Gallegos AC. Enzymatic hydrolysis and microbial fermentation: the most favorable biotechnological methods for the release of bioactive peptides. Food Chem Mol Sci 2021;3:100047. https://doi.org/10.1016/j.fochms.2021.100047.

[12] Najafian L, Babji AS. Production of bioactive peptides using enzymatic hydrolysis and identification antioxidative peptides from patin (Pangasius sutchi) sarcoplasmic protein hydolysate. J Funct Foods 2014;9:280–9. https://doi.org/10.1016/j.jff.2014.05.003.

[13] Sharma P, Sosalagere C, Kehinde BA, Choudhary B. Chapter 15 - Bioactive peptides production using microbial resources. In: Singh J, Sharma D, editors. Microb. Resour. Technol. Sustain. Dev. Elsevier; 2022. p. 299–317. https://doi.org/10.1016/B978-0-323-90590-9.00015-8.

[14] Raveschot C, Cudennec B, Coutte F, Flahaut C, Fremont M, Drider D, et al. Production of bioactive peptides by Lactobacillus species: from gene to application. Front Microbiol 2018;9.

[15] Cheung RCF, Ng TB, Wong JH. Marine peptides: bioactivities and applications. Mar Drugs 2015;13:4006–43. https://doi.org/10.3390/md13074006.

[16] Chakrabarti S, Guha S, Majumder K. Food-derived bioactive peptides in human health: challenges and opportunities. Nutrients 2018;10:1738. https://doi.org/10.3390/nu10111738.

[17] Kitts D.D., Weiler K. Bioactive Proteins and Peptides from Food Sources. Applications of Bioprocesses used in Isolation and Recovery. Curr Pharm Des n.d.; 9:1309–1323.

[18] Costa EM, Oliveira AS, Silva S, Ribeiro AB, Pereira CF, Ferreira C, et al. Spent yeast waste streams as a sustainable source of bioactive peptides for skin applications. Int J Mol Sci 2023;24:2253. https://doi.org/10.3390/ijms24032253.

[19] Harnedy PA, FitzGerald RJ. Bioactive peptides from marine processing waste and shellfish: a review. J Funct Foods 2012;4:6–24. https://doi.org/10.1016/j.jff.2011.09.001.

[20] Du Z, Ding X, Xu Y, Li Y. UniDL4BioPep: a universal deep learning architecture for binary classification in peptide bioactivity. Brief Bioinform 2023:bbad135. https://doi.org/10.1093/bib/bbad135.

[21] Fang Y, Xu F, Wei L, Jiang Y, Chen J, Wei L, et al. AFP-MFL: accurate identification of antifungal peptides using multi-view feature learning. Brief Bioinform 2023;24:bbac606. https://doi.org/10.1093/bib/bbac606.

[22] Sharma R, Shrivastava S, Kumar Singh S, Kumar A, Saxena S, Kumar Singh R. Deep-AFPpred: identifying novel antifungal peptides using pretrained embeddings from seq2vec with 1DCNN-BiLSTM. Brief Bioinform 2022;23:bbab422. https://doi.org/10.1093/bib/bbab422.

[23] Sharma R, Shrivastava S, Kumar Singh S, Kumar A, Saxena S, Kumar Singh R. Deep-ABPpred: identifying antibacterial peptides in protein sequences using bidirectional LSTM with word2vec. Brief Bioinform 2021;22:bbab065. https://doi.org/10.1093/bib/bbab065.

[24] Tang W, Dai R, Yan W, Zhang W, Bin Y, Xia E, et al. Identifying multi-functional bioactive peptide functions using multi-label deep learning. Brief Bioinform 2022;23:bbab414. https://doi.org/10.1093/bib/bbab414.

[25] Zhang Y, Lin J, Zhao L, Zeng X, Liu X. A novel antibacterial peptide recognition algorithm based on BERT. Brief Bioinform 2021;22:bbab200. https://doi.org/10.1093/bib/bbab200.

[26] Minkiewicz P, Iwaniak A, Darewicz M. BIOPEP-UWM database of bioactive peptides: current opportunities. Int J Mol Sci 2019;20:5978. https://doi.org/10.3390/ijms20235978.

[27] Chaudhary A, Bhalla S, Patiyal S, Raghava GPS, Sahni G. FermFooDb: A database of bioactive peptides derived from fermented foods. Heliyon 2021;7:e06668. https://doi.org/10.1016/j.heliyon.2021.e06668.

[28] PepCalc.com - Peptide property calculator 2015. https://pepcalc.com/ (accessed May 5, 2023).

[29] Dai R, Zhang W, Tang W, Wynendaele E, Zhu Q, Bin Y, et al. BBPpred: sequence-based prediction of blood-brain barrier peptides with feature representation learning and logistic regression. J Chem Inf Model 2021;61:525–34. https://doi.org/10.1021/acs.jcim.0c01115.

[30] Meng C, Jin S, Wang L, Guo F, Zou Q. AOPs-SVM: a sequence-based classifier of antioxidant proteins using a support vector machine. Front Bioeng Biotechnol 2019;7.

[31] Zhao D, Teng Z, Li Y, Chen D. iAIPs: identifying anti-inflammatory peptides using random forest. Front Genet 2021;12.

[32] Chen J, Cheong HH, Siu SWI. xDeep-AcPEP: deep learning method for anticancer peptide activity prediction based on convolutional neural network and multitask learning. J Chem Inf Model 2021;61:3789–803. https://doi.org/10.1021/acs.jcim.1c00181.

[33] Lei Y, Li S, Liu Z, Wan F, Tian T, Li S, et al. A deep-learning framework for multi-level peptide–protein interaction prediction. Nat Commun 2021;12:5465. https://doi.org/10.1038/s41467-021-25772-4.

[34] Spänig S, Mohsen S, Hattab G, Hauschild A-C, Heider D. A large-scale comparative study on peptide encodings for biomedical classification. NAR Genom Bioinforma 2021;3:lqab039. https://doi.org/10.1093/nargab/lqab039.

[35] Spänig S, Heider D. Encodings and models for antimicrobial peptide classification for multi-resistant pathogens. BioData Min 2019;12:7. https://doi.org/10.1186/s13040-019-0196-x.

[36] Sidorczuk K, Gagat P, Pietluch F, Kała J, Rafacz D, Bąkała L, et al. Benchmarks in antimicrobial peptide prediction are biased due to the selection of negative data. Brief Bioinform 2022;23:bbac343. https://doi.org/10.1093/bib/bbac343.

[37] Nielsen SD, Beverly RL, Qu Y, Dallas DC. Milk bioactive peptide database: a comprehensive database of milk protein-derived bioactive peptides and novel visualization. Food Chem 2017;232:673–82. https://doi.org/10.1016/j.foodchem.2017.04.056.

[38] Li Q, Zhang C, Chen H, Xue J, Guo X, Liang M, et al. BioPepDB: an integrated data platform for food-derived bioactive peptides. Int J Food Sci Nutr 2018;69(8):963. https://doi.org/10.1080/09637486.2018.1446916.

[39] Moguel-Concha D del R, Borges-Martínez JE, Cid-Gallegos MS, Juárez-Chairez MF, Gómez-Gómez AL, Téllez-Medina DI, et al. Antioxidant and renin inhibitory activities of peptides from food proteins on hypertension: a review. Plant Foods Hum Nutr 2023;78:493–505. https://doi.org/10.1007/s11130-023-01085-3.

[40] Iwaniak A, Minkiewicz P, Darewicz M. Food-originating ACE inhibitors, including antihypertensive peptides, as preventive food components in blood pressure reduction. Compr Rev Food Sci Food Saf 2014;13:114–34. https://doi.org/10.1111/1541-4337.12051.

[41] Ojeda MJ, Cereto-Massagué A, Valls C, Pujadas G, Pujadas G. DPP-IV, an important target for antidiabetic functional food design. In: Martinez-Mayorga K, Medina-Franco JL, editors. Foodinformatics Appl. Chem. Inf. Food Chem. Cham: Springer International Publishing; 2014. p. 177–212. https://doi.org/10.1007/978-3-319-10226-9_7.

[42] van de Laar FA, Lucassen PL, Akkermans RP, van de Lisdonk EH, Rutten GE, van Weel C. α-Glucosidase inhibitors for patients with type 2 diabetes: results from a cochrane systematic review and meta-analysis. Diabetes Care 2005;28:154–63. https://doi.org/10.2337/diacare.28.1.154.

[43] Fernández de Ullivarri M, Arbulu S, Garcia-Gutierrez E, Cotter PD. Antifungal peptides as therapeutic agents. Front Cell Infect Microbiol 2020;10. https://doi.org/10.3389/fcimb.2020.00105.

[44] Seyfi R, Kahaki FA, Ebrahimi T, Montazersaheb S, Eyvazi S, Babaeipour V, et al. Antimicrobial peptides (AMPs): roles, functions and mechanism of action. Int J Pept Res Ther 2020;26:1451–63. https://doi.org/10.1007/s10989-019-09946-9.

[45] Tolos (Vasii) AM, Moisa C, Dochia M, Popa C, Copolovici L, Copolovici DM. Anticancer potential of antimicrobial peptides: focus on buforins. Polymers 2024;16:728. https://doi.org/10.3390/polym16060728.

[46] Qin Y., Qin Z.D., Chen J., Cai C.G., Li L., Feng L.Y., et al. From Antimicrobial to Anticancer Peptides: The Transformation of Peptides. Recent Patents Anticancer Drug Discov n.d.;14:70–84.

[47] Cheng S, Tu M, Liu H, Zhao G, Du M. Food-derived antithrombotic peptides: Preparation, identification, and interactions with thrombin. Crit Rev Food Sci Nutr 2019;59:S81–95. https://doi.org/10.1080/10408398.2018.1524363.

[48] Reyes Gaido OE, Nkashama LJ, Schole KL, Wang Q, Umapathi P, Mesubi OO, et al. CaMKII as a therapeutic target in cardiovascular disease. Annu Rev Pharm Toxicol 2023;63:249–72. https://doi.org/10.1146/annurev-pharmtox-051421-111814.

[49] Čolović MB, Krstić DZ, Lazarević-Pašti TD, Bondžić AM, Vasić VM. Acetylcholinesterase inhibitors: pharmacology and toxicology. Curr Neuropharmacol 2013;11:315–35. https://doi.org/10.2174/1570159×11311030006.

[50] Ganina KK, Dugina YL, Zhavbert ES, Ertuzun IA, Epstein OI, Mukhin VN, et al. Antiamnesic effects divaza and its component model β-amyloid amnesia]. Zh Nevrol Psikhiatr Im S S Korsakova 2016;116:69–74. https://doi.org/10.17116/jnevro20161169169-74.

[51] Mizushige T. Neuromodulatory peptides: orally active anxiolytic-like and antidepressant-like peptides derived from dietary plant proteins. Peptides 2021;142:170569. https://doi.org/10.1016/j.peptides.2021.170569.

[52] Hsieh C-H, Wang T-Y, Hung C-C, Hsieh Y-L, Hsu K-C. Isolation of prolyl endopeptidase inhibitory peptides from a sodium caseinate hydrolysate. Food Funct 2016;7:565–73. https://doi.org/10.1039/C5FO01262G.

[53] Vens C, Rosso M-N, Danchin EGJ. Identifying discriminative classification-based motifs in biological sequences. Bioinformatics 2011;27:1231–8. https://doi.org/10.1093/bioinformatics/btr110.

[54] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020;17:261–72. https://doi.org/10.1038/s41592-019-0686-2.

[55] Maetschke S, Towsey M, Boden M. BLOMAP: an encoding of amino acids which improves signal peptide cleavage site prediction. In: Chen Y, Wong L, editors. Proc. 3rd Asia-Pac. Bioinforma. Conf. Adv. Bioinforma. Comput. Biol. Singapore: Imperial College Press; 2005. p. 141–50.

[56] Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M. ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics 2022;38:2102–10. https://doi.org/10.1093/bioinformatics/btac020.

[57] Maaten L, van der, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9:2579–605.

[58] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. J Mach Learn Res 2011;12:2825–30.

[59] Chollet, F. & others. Keras: Deep Learning for humans 2015.

[60] Wolf T., Debut L., Sanh V., Chaumond J., Delangue C., Moi A., et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing 2020. https://doi.org/10.48550/arXiv.1910.03771.
[61] De Rainville F.-M., Fortin F.-A., Gardner M.-A., Parizeau M., Gagné C. DEAP: a python framework for evolutionary algorithms. Proc. 14th Annu. Conf. Companion Genet. Evol. Comput., Philadelphia Pennsylvania USA: ACM; 2012, p. 85–92. https://doi.org/10.1145/2330784.2330799.
[62] Tomer R, Patiyal S, Dhall A, Raghava GPS. Prediction of celiac disease associated epitopes and motifs in a protein. Front Immunol 2023;14:1056101. https://doi.org/10.3389/fimmu.2023.1056101.
[63] Chen X, Huang J, He B. AntiDMPpred: a web service for identifying anti-diabetic peptides. PeerJ 2022;10:e13581. https://doi.org/10.7717/peerj.13581.
[64] Qin D, Jiao L, Wang R, Zhao Y, Hao Y, Liang G. Prediction of antioxidant peptides using a quantitative structure−activity relationship predictor (AnOxPP) based on bidirectional long short-term memory neural network and interpretable amino acid descriptors. Comput Biol Med 2023;154:106591. https://doi.org/10.1016/j.compbiomed.2023.106591.
[65] Akbar S, Mohamed HG, Ali H, Saeed A, Khan AA, Gul S, et al. Identifying neuropeptides via evolutionary and sequential based multi-perspective descriptors by incorporation with ensemble classification strategy. IEEE Access 2023;11:49024–34. https://doi.org/10.1109/ACCESS.2023.3274601.
[66] Cao Q, Ge C, Wang X, Harvey PJ, Zhang Z, Ma Y, et al. Designing antimicrobial peptides using deep learning and molecular dynamic simulations. Brief Bioinform 2023;24:bbad058. https://doi.org/10.1093/bib/bbad058.
[67] Lertampaiporn S, Hongsthong A, Wattanapornprom W, Thammarongtham C. Ensemble-AHTPpred: a robust ensemble machine learning model integrated with a new composite feature for identifying antihypertensive peptides. Front Genet 2022;13.
[68] Deng H, Ding M, Wang Y, Li W, Liu G, Tang Y. ACP-MLC: a two-level prediction engine for identification of anticancer peptides and multi-label classification of their functional types. Comput Biol Med 2023;158:106844. https://doi.org/10.1016/j.compbiomed.2023.106844.