Artificial Intelligence in the Life Sciences 2 (2022) 100034

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

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

Research Article

Coupled encoding methods for antimicrobial peptide prediction: How sensitive is a highly accurate model?

Ivan Erjavac [a], Daniela Kalafatovic [b,c], Goran Mauša [a,c,*]

[a] Faculty of Engineering, University of Rijeka, Vukovarska 58, Rijeka 51000, Croatia
[b] Department of Biotechnology, University of Rijeka, Radmile Matejčić 2, Rijeka 51000, Croatia
[c] Center for Artificial Intelligence and Cybersecurity, University of Rijeka, Radmile Matejčić 2, Rijeka 51000, Croatia

ARTICLE INFO

Keywords:
Peptide encoding
Machine learning
Antimicrobial peptide prediction
False positive
Model sensitivity

ABSTRACT

Current application of machine learning in the process of antimicrobial peptide discovery call for the reduction of the false positive predictions that are produced by the classification models. Considering that the positive predictions of high confidence drive modern experimental design, the model's sensitivity is crucial to reduce the number of unnecessary *in vitro* tests. Furthermore, taking into account the expert-based design approaches that employ random mutations on confirmed sequences, the machine learning models are required to distinguish between subtle differences among shuffled sequences. With the goal of reducing the false positive rate and improving sensitivity, we propose a hybrid approach to antimicrobial peptide prediction that utilizes combined encoding models. To this end, we implement models that employ both the physico-chemical features and sequence ordering information to stress the importance of using both representations. We also investigate the usage of binary encoding for peptide representation purposes, a method that is insufficiently represented in related research, which proved to act as a viable low dimensional alternative to the one-hot encoding. Our results, supported by Cochran and McNemar statistical tests and Spearman correlation analysis, indicate that the sequence-based encodings complement the physico-chemical features and their synergic effect yields improvement in terms of every evaluation metric. Finally, the proposed hybrid approach that combines physico-chemical features and binary encoding using logical conjunction was shown to be superior to other single models by a factor of 2.96 in terms of fall-out and up to 6.1% in terms of precision.

1. Introduction

Antimicrobial peptides (AMPs), also known as host defense peptides, are short chains of amino acids produced by the immune systems of living organisms. Their main purpose is to protect the host organism against external threats that are of viral, bacterial, fungal, and parasitic origin [3,15]. Furthermore, some AMPs have displayed anticancer activity [30], which indicates that these naturally occurring defense apparatus can be used to combat one of modern medicine's biggest adversaries. Their origin reaches back to 1939 [3,29], when René Dubos isolated a bacillus capable of attacking live gram-positive bacteria [12]. This bacillus was later named gramicidin [14], and is now considered to be the first discovered antimicrobial peptide [3,29].

In recent years, the research on antimicrobial peptides has gained more attention than ever [19]. This interest can be attributed to the alarming accumulation of drug-resistant microbes causing an increasingly large number of infections [34]. The extensive usage of traditional antibiotics, combined with the highly adaptive and mutable nature of microbes, has led to the emergence of a new global health crisis [22,26]. According to the WHO, the antimicrobial resistance problem is leading the world towards a post-antibiotic era [26], and is set to become one of the leading causes of deaths [22]. Considering the potency and broad activity spectrum of AMPs, the scientific community is becoming progressively interested in turning these naturally occurring amino acid sequences into modern-day antibiotics. However, *in vitro* testing of peptides to potentially discover their antimicrobial properties is a time-consuming operation which also requires considerable resources [4]. Moreover, should there be a need to test hundreds, or even thousands of candidate peptides, the entire process of AMP discovery becomes very slow and inflexible.

With the goal of faster discovery of new antimicrobial peptides, researchers have been combining the contemporary *in silico* testing methodologies with the traditional *in vitro* peptide evaluation to augment their drug discovery workflow [35,41]. The general idea is to create machine learning models capable of discerning which peptides are more likely to present antimicrobial activity from the ones which are

* Corresponding author at: Faculty of Engineering, University of Rijeka, Vukovarska 58, Rijeka 51000, Croatia.
E-mail address: gmausa@riteh.hr (G. Mauša).

https://doi.org/10.1016/j.ailsci.2022.100034
Received 31 December 2021; Received in revised form 5 March 2022; Accepted 13 April 2022
Available online 22 April 2022

less likely to do so. Using this newfound knowledge, garnered during the *in silico* testing process, researchers can then focus their *in vitro* testing efforts onto the promising peptide candidates with high potential of having antimicrobial activity. When successful, this combined testing methodology leads to less *in vitro* testing, and more AMP discoveries. Moreover, this approach can be used to identify general antimicrobial activity, but also to further specify the exact type of activity for an AMP [40]. Lastly, the usage of evolutionary algorithms in this drug discovery process enables the *in silico* generation, and eventually *in vitro* synthesis of previously non-existent peptides with antimicrobial activities [42].

This paper compares and evaluates methodologies of peptide representation in the form of numerical vectors, with the aim to construct an augmented methodology that will be used to reduce the number of falsely identified antimicrobial peptides. The first approach uses the physico-chemical descriptors of peptides computed from their amino acid sequences (referred to as the *phys_chem* method). The second approach is a sequence-based approach which uses one-hot encoding and binary encoding of categorical variables to represent amino acids within the amino acid sequence (referred to as the *one-hot_encoding and binary_encoding* methods). Furthermore, we introduce a third approach that uses both feature generation methods simultaneously to account for the physico-chemical descriptors of peptides and the exact ordering of amino acids within the amino acid sequence (referred to as the *PC_one-hot* and *PC_binary* methods). The methodologies are assessed and compared to stress out the importance of using both the physico-chemical features and the sequence order information when making AMP predictions.

We propose an approach for reducing the number of non-antimicrobial peptides that make it into the *in vitro* testing phase. The rationale for our aim is that the expert will benefit more from a machine learning model that is able to accurately produce a few promising solutions than a model with higher overall accuracy that yields false positives with high confidence. The operation of logical conjunction is conducted between the predictions of the *phys_chem* model and encoding-based models in an effort to minimize the fall-out metric, also known as the false positive rate (FPR). Our results indicate that this approach of combining the prediction of divergent models can indeed be used for the reduction of the false positive rate and act as a strong candidate for future applications in peptide discovery.

2. Background

The usage of machine learning for the purposes of predicting AMPs is a known interdisciplinary endeavour, however new caveats are constantly being uncovered, meaning there is always room for improvement. For instance, a common representation scheme in Artificial Neural Network and Support Vector Machine (SVM) prediction models is to use physico-chemical properties of peptides [20,37,39]. Another approach to create machine learning models that predict antimicrobial peptides is to utilize the compositional features of peptides in the form of PseAAC [7]. While both of these approaches represent valid resolutions to the problem of antimicrobial activity prediction, it has been argued that the usage of both physico-chemical and sequence order descriptors leads to the construction of models able to achieve even higher performance [32].

To this end, many researchers have used physico-chemical features in combination with various compositional descriptors to consider more peptide information while developing machine learning models. Compositional features, along with structural and physico-chemical features, were used to develop an SVM-based classifier for identifying AMPs and their functional types [21,40]. Considering that using only amino acid composition (AAC) does not contribute with sufficient amount of information, this particular model has also utilized the pseudo amino acid composition (PseAAC) descriptor. The difference between the two is that AAC loses all of the sequence order information, while PseAAC keeps the sequence order information, although only partially [9]. The pseudo amino acid composition descriptor is a commonly used compositional descriptor, and has been used numerous times for AMP predictions [21,23,40,41].

To preserve as much sequence order information as possible, other encoding methodologies have been trialed, such as one-hot encoding that retains only the information about the order of amino acids [25]. A hybrid approach that used one-hot encoding paired with physico-chemical features to create numerical representations of peptides was successfully applied for the prediction of anticancer peptide activity [6]. Recent comparative studies [32,33] were conducted with the goal of summarizing and comparing known peptide encoding methods that are commonly used when approaching the task of biomedical classification. To the best of our knowledge, binary encoding remains an unexplored and unused encoding scheme in the context of AMP prediction, which is why we chose to study its potential benefits on the predictive performance.

Finally, recent research reflects on the need to minimize the amount of false positives that the classification models yield. Most recent studies tried improving the sensitivity of the predictive performance by fusing features using a logistic regression equation [18] and by applying rigorous classification threshold on a Convolutional Neural Network model to select only the most promising peptides for the *in vitro* testing phase [41]. With this in mind, we propose a method of combining predictions of models based on different types of descriptors (physico-chemical or compositional) in an effort to reduce the false positive rate and improve the sensitivity of the model.

3. Methodology

This section introduces our methods and Fig. 1 visualizes the implemented workflow. We present our data sources and pre-processing, explain the feature extraction process, describe the model used for prediction, and outline which metrics were used for evaluation of model performance. Furthermore, Fig. 1 also shows how the feature generation methods were utilized to construct various datasets and which models were examined.

3.1. Datasets

The antimicrobial dataset was constructed by combining the non-patent positive instances from the *Data Repository of Antimicrobial Peptides* database (DRAMP) [31] and the negative instances from the *UniProt* database [11]. To ensure that the negative instances are not longer than 75 amino acids and are not fragments of longer sequences, the following query has been run in the UniProt database: "*NOT keyword:antimicrobial length:[0 TO 75] fragment:no AND reviewed:yes*". This query also ensures that the negative dataset contains reviewed and verified peptides ("*reviewed:yes*") with no antimicrobial potency ("*NOT keyword:antimicrobial*"). UniProt closely monitors AMPs, so any peptides with detected antimicrobial activity would not be returned by this query.

Data cleaning consisted of filtering duplicated sequences, sequences that contained other than the 20 natural amino acids, and sequences whose length is considered an outlier. The outlier detection is set to preserve the range between the 5th and the 95th percentile of the positive instances, which corresponds to sequences with length in the range between 10 and 75 amino acids with inclusive boundaries. The positive set was reduced to 5043 instances out of the initial 5819, while the filtered negative set contained 16,910 out of 24,751 original instances.

The consistency of the sequence length distributions between the classes has been recognized as an important data filtration step [38,41]. Fig. 2a and 2b illustrate the distribution of peptide lengths for the positive and negative instances, respectively. The 16,910 negative instances are skewed towards the upper limit of the sequence lengths, while the 5043 positive instances are more present in the lower limit of the sequence lengths. This could lead to the model implicitly associating the

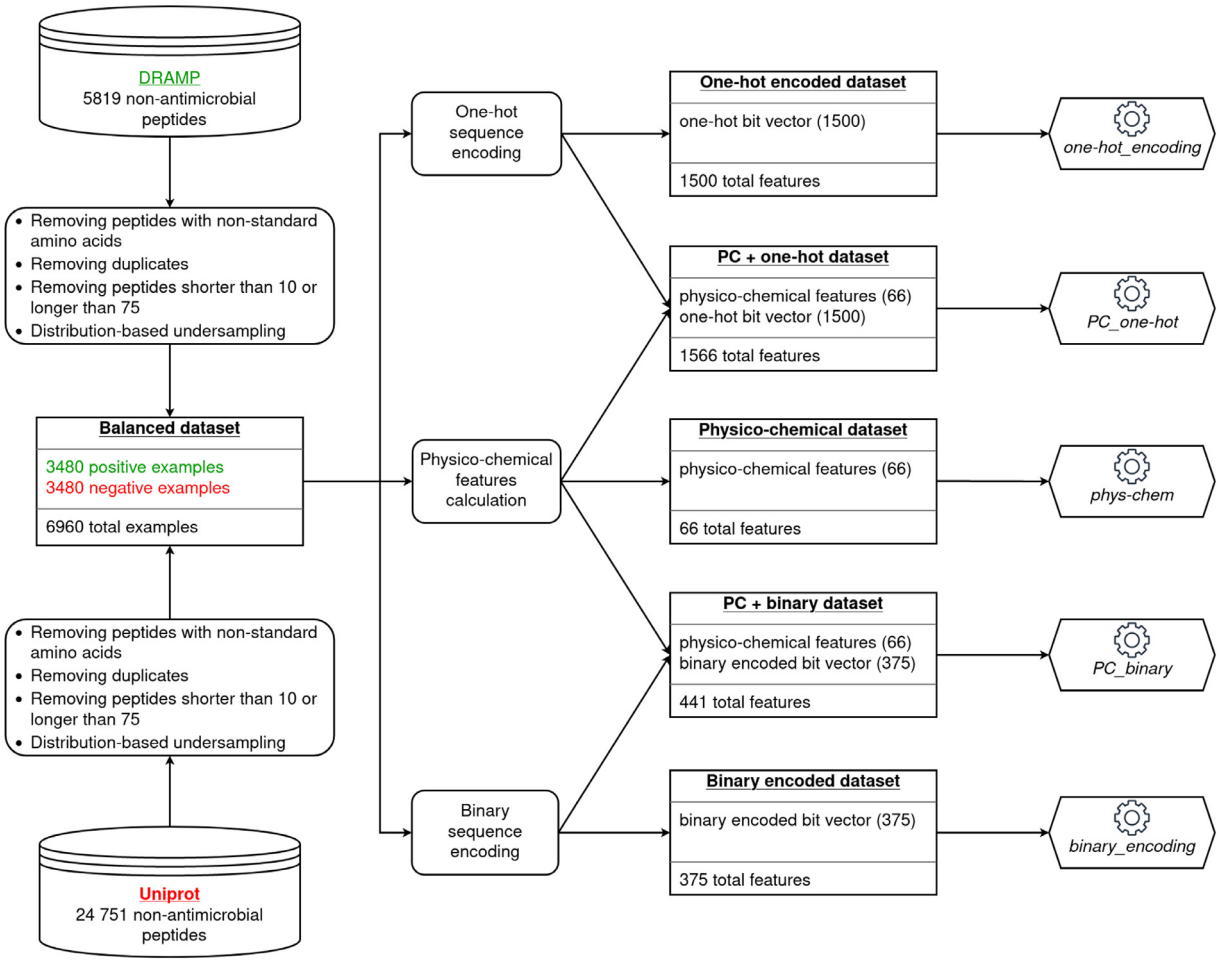


Fig. 1. The dataset construction process including two sources of data (DRAMP and Uniprot databases), data cleaning and sampling, three feature generation methods and combined encoding models.

longer peptides with the negative class, and the shorter peptides with the positive class. Therefore, data pre-processing also involved distribution-based undersampling to reduce the effect of divergent sequence length distributions of the positive and negative classes. The reduction of the instances from both classes resulted in a fully balanced dataset that contains 3480 positive and 3480 negative instances, which are equally distributed with respect to sequence length as presented in Fig. 2c and 2d. To analytically check the distributions similarity between classes, we applied the nonparametric Mann-Whitney U statistical test, which confirmed that their distributions indeed come from the same population, producing *p*-value of 0.84. The process of blending positive and negative instances into an operational dataset is depicted in Fig. 1.

Table 1
The physico-chemical feature space of peptides.

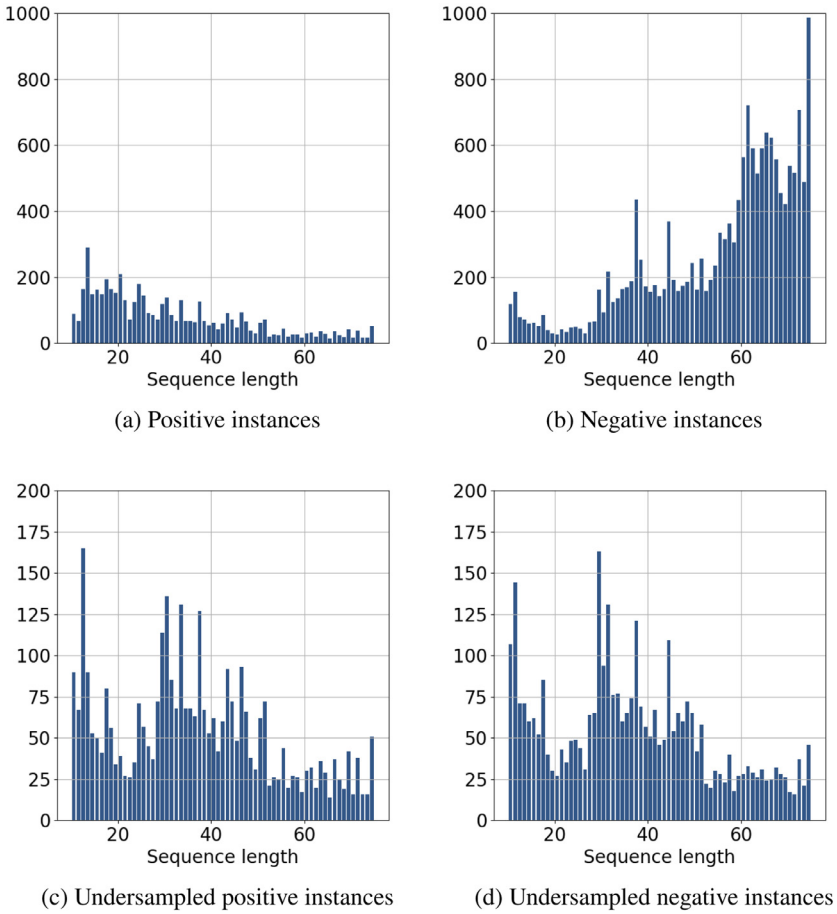
Descriptor category	Number of Components
BLOSUM indices	10
Cruciani properties	3
FASGAI vectors	6
Kidera factors	10
MS-WHIM scores	3
ProtFP	8
ST-scales	8
T-scale	5
VHSE scales	8
Z-scales	5
Total	66

3.2. Feature representation

To establish the numerical representation of the peptides in the constructed dataset, three feature generation methods were used. The first method leverages the physico-chemical descriptors of peptides, while the other two methods are based on the ordering of amino acids in a sequence.

3.2.1. Physico-chemical features

The physico-chemical feature generation method is based on the calculation of peptide's properties from the amino acid sequence, in a format similar to FASTA. The categories of peptide descriptors and the number of their components are listed in Table 1. In total, each peptide is described using 66 features that encompass hydrophobic, bulky, steric, electronic, topological, structural, alignment and interaction properties, alpha and turn propensities, compositional characteristics, etc. To calculate these physico-chemical features of peptides, we have employed the *Peptides* package [27], which is available for the R programming language. An important property of computed features is that each descriptor is calculated as the average value over the entire sequence. For example, in the case of *Z-scales*, each of the 20 amino acids has it's own 5 specific *Z-scale* values. To calculate the *Z-scale* values for a sequence such as "LMCTHPLDCSN", we compute the average of each of the 5 *Z-scale* values over the entire sequence. The expression used by the *Peptides* package to calculate the *Z-scales* descriptor is shown in Eq. (1), where N represents the number of amino acids in the sequence, while



(a) Positive instances (b) Negative instances

(c) Undersampled positive instances (d) Undersampled negative instances

Fig. 2. Distribution of sequence lengths for positive and negative instances before and after undersampling.

$Z_j^{(i)}$ denotes the j^{th} *Z-scale* value of the i^{th} amino acid.

$$Z_j = \frac{\sum_{i=1}^{N} Z_j^{(i)}}{N}, \forall j \in [1,5] \quad (1)$$

Although these features describe a peptide sequence from compositional, electrostatic, structural, and other aspects, the very nature of their calculation signifies that the physico-chemical descriptors are not always sufficient for representing a sequence. If we were to take the aforementioned sequence ("LMCTHPLDCSN") and simply switch the places of the first two amino acids ("MLCTHPLDCSN"), the *Z-scale* values for these two sequences would be identical. Taking this into consideration, we can conclude that additional features are to be used if we were to effectively capture the ordering of the amino acids within a sequence. To help and reflect the ordering of the amino acids, this feature generation method is used in combination with the other two feature generation methods to create datasets for combined encoding models (*PC_one-hot dataset* and *PC_binary dataset*). This method is also used to construct a single encoding dataset named the *physico-chemical dataset*.

3.2.2. One-hot encoding

The one-hot encoding is used to represent a categorical value as a vector of bits. Each bit represents one possible value, meaning the vector length is equal to the number of possible categories that we are dealing with. In the context of this paper, we are dealing with peptides comprised of the 20 natural amino acids and the bit vector for a single amino acid has a fixed length of 20 bits. Each of the 20 amino acids is positionally assigned to one bit and to encode an amino acid in the bit vector, the assigned bit is set to 1, while all the others are set to 0. The amino acids were assigned to the bits in an alphabetical order: the amino acid that is alphabetically first (A - alanine) is assigned to the last, rightmost bit, while the amino acid that is alphabetically last (Y - tyrosine) is assigned to the first, leftmost bit. The full encoding of each amino acid is available in Table 2. To construct the feature vector for each of the sequences, the bit vectors of all of the amino acids in the sequence are concatenated.

This encoding scheme requires the amino acid sequences to be of fixed length. Given that we have chosen a maximum length of 75 amino

Table 2
The one-hot and binary encoding of amino acids, which are ordered alphabetically and assigned rank for interpretability.

Amino acid	Rank	One-hot	Binary
A	1	00000000000000000001	00001
C	2	00000000000000000010	00010
D	3	00000000000000000100	00011
E	4	00000000000000001000	00100
F	5	00000000000000010000	00101
G	6	00000000000000100000	00110
H	7	00000000000001000000	00111
I	8	00000000000010000000	01000
K	9	00000000000100000000	01001
L	10	00000000001000000000	01010
M	11	00000000010000000000	01011
N	12	00000000100000000000	01100
P	13	00000001000000000000	01101
Q	14	00000010000000000000	01110
R	15	00000100000000000000	01111
S	16	00001000000000000000	10000
T	17	00010000000000000000	10001
V	18	00100000000000000000	10010
W	19	01000000000000000000	10011
Y	20	10000000000000000000	10100

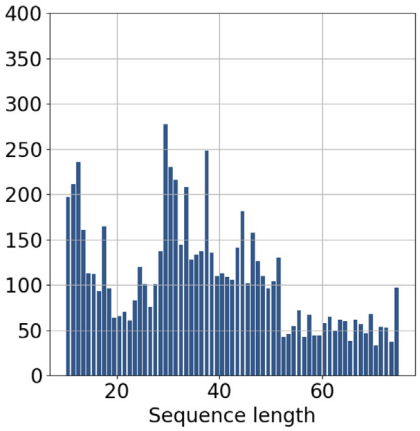


Fig. 3. The distribution of the sequence lengths in the final dataset.

acids, the feature vector of a one-hot encoded peptide will always have a total length of 1500, no matter what the actual length of the peptide is. Naturally, most of the peptides in the dataset are shorter than 75 amino acids, as shown in Fig. 3. To account for this, after the bit vectors of the amino acids have been concatenated, all remaining elements of the feature vector are set to -1. For example, a peptide composed of 40 amino acids will be represented by an 800 bits long feature vector and the remaining 700 features set to the defined value of -1. This method is used to construct the *one-hot encoded dataset*, which is dedicated solely to the one-hot encoding scheme. Also, to help mitigate the shortcomings of the physico-chemical approach, this method is also used in combination with the physico-chemical feature generation method to construct the *PC_one-hot dataset*.

3.2.3. Binary encoding

Similarly to one-hot encoding, the binary encoding turns categorical values into vectors of bits, which correspond to the assigned rank number of each category. In the context of this paper, these numbers range from 1 to 20, and are assigned using the alphabetical criterion of the single letter code of natural amino acids, as presented in Table 2. Using the binary encoding scheme, each amino acid can be represented using exactly 5 bits. The binary encoding method approach differs from the one-hot encoding method in that it uses a feature space that is 4 times smaller. Considering the sequence limit of 75 amino acids, the feature vector will have a length of 375 features. Just like with one-hot encoding, the bit vectors of all amino acids are concatenated to form the feature vector, with any remaining features being padded with the value of -1. Should binary encoding be used on a peptide consisting of 40 amino acids, the first 200 elements of the feature vector would have a value of either 0 or 1, while the remaining 175 elements would be set to -1.

Along with the construction of the *binary encoded dataset*, this encoding scheme is also used to create another dataset with combined encodings named the *PC_binary dataset*.

3.3. Prediction model

The prediction model that was used in this research is the *Support Vector Machine* (SVM) with the RBF kernel. The SVM models are capable of solving binary classification tasks, and do so by fitting a hyperplane that separates the two classes. The main advantage of the SVM model is that it implicitly maps the data into higher dimensions using the kernel trick, which enables the separation of classes in higher dimensions when the separation cannot be done in lower dimensions. The motivation for the use of SVM lies in the fact that it has been used numerous times for AMP prediction purposes [16,21], and is widely regarded as an advanced and well-performing model, even outside the field of chemoinformatics [2,13,36]. To construct the optimal hyperplane that separates the two classes, the SVM model relies on distances between the data instances. Taking this into account, it is important to scale the features down to similar scales. The one-hot encoded features and the binary encoded features already exist on similar scales because they can assume one of three values: -1, 0, or 1. On the other hand, the physico-chemical features have varying ranges. In the context of this research, the second component of the *MSWHIM* descriptor ($MSWHIM^{(2)}$) has a range of [-0.18, 0.78], while the first component of *T-scales* ($T^{(1)}$) exists on a scale of [-8.78, 2.98]. Taking this into consideration, the *physico-chemical dataset* was scaled using the feature scaling technique of standardization, while the *one-hot encoded dataset* and *binary encoded dataset* were left unaltered. The *PC_one-hot dataset* and *PC_binary dataset* were only partially standardized: the physico-chemical features of the combined datasets were included in the standardization process, but the one-hot encoded and the binary encoded features were left in their original form. The z-score standardization function was implemented using the expression denoted in Eq. (2).

$$X_j'^{(i)} = \frac{X_j^{(i)} - \mu^{(i)}}{\sigma^{(i)}}, \forall j \in [1, m], i \in [1, 66] \tag{2}$$

The $X^{(i)}$ represents the ith raw, unstandardized feature, while $X'^{(i)}$ denotes the new, standardized version of the feature. As denoted in Table 1, the total number of the physico-chemical features is 66, which is why i spans in the range from 1 to 66. The $\mu^{(i)}$ and $\sigma^{(i)}$ are the mean and standard deviation of the ith feature, respectively. This calculation will standardize each of the features to have a mean of 0 and a unit variance, enabling SVM to perform better [1].

The SVM classification models were implemented using the *scikit-learn* Python library [28]. The hyperparameters of the models were optimized using grid search during the nested cross-validation evaluation, which is why a range of optimal values may be identified, as for the amount of regularization (hyperparameter C). The search has been conducted on hyperparameters $C \in [0.1, 30]$, *kernel* $\in$ {rbf, sigmoid}, $\gamma \in$ {auto, scale}, *shrinking* $\in$ {True, False}, and *probability* $\in$ {True, False} to find the best performing combination for each model of single and combined encodings. The hyperparameters that were identified as optimal for each encoding method are shown in Table 3.

3.4. Evaluation

This section introduces all of the metrics that were used to evaluate the performance of the binary classification models. We provide each metric's description, as well as the calculation which defines the metric. We will be using the commonly used abbreviations for the confusion matrix elements: true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

As presented in Table 4, we have utilized some of the commonly used metrics like accuracy, precision, and recall. We have also employed the F_1 score, a metric that mitigates the precision/recall trade-off [5], indicating that high values for both metrics are needed to prove that a model is truly performing well. These metrics can be intuitively interpreted, and can provide a reasonably good insight into general model performance. However, a more comprehensive metric is needed to objectively evaluate a model's performance [8]. To this end, we have used the Matthews Correlation Coefficient (MCC), which is an often used metric for measuring the performance of binary classifiers. The MCC calculation in Table 4 gives a number in the range of [-1, 1]. In a machine learning context, a value of -1 represents extremely poor model performance, while 1 signifies that the model performs perfectly. Due to their mathematical nature, other metrics run the risk of giving excessively optimistic classification results. MCC tends to avoid this type of exaggeration [8], which is why it was included as an indicator of model performance in this research. Finally, since recent research reflects on the need for lowering the amount of false positives produced by AMP prediction models [18,21], we have also included the fall-out metric,

Table 3
Grid search results for hyperparameter optimization.

	Model				
Hyperparameter	phys-chem	one-hot_encoding	binary_encoding	PC_one-hot	PC_binary
C	10–27	10–24	12–27	10–18	6–15
kernel	rbf	rbf	rbf	rbf	rbf
γ	scale	scale	scale	scale	scale
shrinking	True	True	True	True	True
probability	True	True	True	True	True

Table 4
Evaluation metrics used for estimating the models' predictive power.

Metric	Description	Calculation expression
Accuracy	Proportion of correctly classified instances, out of all testing instances.	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	Proportion of instances correctly identified as positive, out of all testing instances that the model has classified as positive.	$\frac{TP}{TP+FP}$
Recall	Proportion of instances correctly identified as positive, out of all testing instances that are labeled as positive.	$\frac{TP}{TP+FN}$
F_1 score	The harmonic mean of precision and recall.	$2 \cdot \frac{Precision \cdot Recall}{Precision+Recall}$
Matthew's Correlation Coefficient	Measures the correlation between the observed labels and the predicted outcomes.	$\frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP) \cdot (TP+FN) \cdot (TN+FP) \cdot (TN+FN)}}$
Fall-out	Proportion of instances falsely identified as positive, out of all testing instances that are labeled as negative.	$\frac{FP}{FP+TN}$

also known as the false positive rate (FPR). Unlike the other metrics used in this research, lower values of the FPR are indicative of a better model performance. In the context of this research, a low FPR means that a model produces less *in vitro* testing candidates which do not actually have antimicrobial properties.

The assessment of differences in prediction probabilities between analyzed models was achieved by the non-parametric Spearman's rank-order correlation, which measures the strength and direction of a monotonic relationship between paired data [17]. The yielded correlation coefficient r_s takes on a value within the range [-1, 1], with the following interpretation of its absolute value:

- $0.0 \leq |r_s| < 0.2$ - very low correlation,
- $0.2 \leq |r_s| < 0.4$ - low correlation,
- $0.4 \leq |r_s| < 0.6$ - moderate correlation,
- $0.6 \leq |r_s| < 0.8$ - strong correlation,
- $0.8 \leq |r_s| \leq 1.0$ - very strong correlation.

The advantage of Spearman's correlation is that it does not require the data to be normally distributed nor does it assume linear relationship between the observations [24]. To investigate whether the binary predictions of the analyzed models come from the same distribution, we have utilized the Cochran's Q statistical test. In a machine learning context, this test can be used to check for statistically significant differences between the predictions of three or more models, with a confidence level of 95% [10]. Upon rejecting the null-hypothesis of Cochran's Q test, the McNemar test with Bonferroni correction is applied for pairwise comparison to further specify models which display significantly differing predictions.

4. Results

This paper demonstrates how various peptide representation techniques have been scrutinized, compared, and combined in an effort to produce prediction models capable of identifying antimicrobial peptides with low tendency for false positives. Firstly, we set out to investigate whether the AMP classification task may be solved by simple process of clustering the peptides into two categories using their length and amino acid composition as dependent variables. The K-means algorithm was applied in the attempt to cluster the peptides into two groups which were then compared with the two categories of positive and negative antimicrobial instances. The second approach was to cluster the data using the physico-chemical features from the dataset that was constructed as presented in the workflow from Fig. 1. Both attempts resulted with Rand index of $R \approx 0.5$, proving this task not to be a trivial one and the necessity of using prediction models of higher complexity. In addition, we confirmed that the under-sampling technique yielded datasets that are not biased by peptide length.

The classification models were evaluated using a balanced dataset containing matching amounts of antimicrobial and non-antimicrobial peptides. To allow for a fair and objective comparison of the chosen methodologies, SVM was the prediction model of choice, being the dominant approach in related work, and hyperparameter tuning based on the grid search optimization process has preceded the testing of the models. The evaluation approach of choice was *nested 10-fold cross-validation*. The dataset is randomly split into 10 folds, after which the model is trained, tuned and tested 10 times, with each fold being used for testing once. Before model evaluation, the model's hyperparameters are optimized on the 9 training folds, ensuring that the test fold is completely left out from the hyperparameter tuning and model training processes.

To compare the approaches as objectively as possible, the fold splitting was done identically for each of the methods, meaning each of the folds across all of the trialed datasets contain the same sets of peptides. Finally, each metric is calculated as an average of the metric values that were produced in each of the 10 runs. Besides the average, the standard deviation is calculated for each of the metrics as well. The results that have been acquired through the evaluation of the models developed using the described approaches are presented in Table 5.

4.1. Combining model predictions

Considering the fact that both the *phys-chem* model and the sequence-based models are known to yield highly accurate predictions, we have set out to jointly employ them in order to improve the sensitivity, i.e. fall-out of the prediction model. The predictions of the *phys-chem* and the *one-hot_encoding* models have been put through the operations of logical conjunction (*multi_comb_1*) and logical disjunction (*add_comb_1*). The *multi_comb_1* approach uses the logical *AND* operation to classify a peptide as antimicrobial only if both the *phys-chem* and the *one-hot_encoding* models classify the peptide as antimicrobial. Similarly, the *add_comb_1* method uses the logical *OR* operation to classify a peptide as antimi-

Table 5
The results of 10-fold cross-validation. The table is horizontally separated into three segments, representing single encoding models, combined encoding models and hybrid models, respectively. The best performing model for each evaluation metric is marked in bold.

Method	Accuracy	Precision	Recall	F_1 score	MCC	Fall-out
phys-chem	87.7% ± 1.0%	85.7% ± 1.2%	90.4% ± 1.0%	88.0% ± 0.9%	0.755 ± 0.019	15.1% ± 1.4%
one-hot_encoding	91.2% ± 0.8%	90.8% ± 0.9%	91.6% ± 1.3%	91.2% ± 0.8%	0.824 ± 0.016	9.3% ± 1.0%
binary_encoding	87.7% ± 1.1%	88.0% ± 1.3%	87.4% ± 1.5%	87.7% ± 1.1%	0.755 ± 0.021	11.9% ± 1.5%
PC_one-hot	**92.1% ±0.8%**	91.8% ± 1.0%	92.6% ± 1.4%	**92.2% ±0.8%**	**0.843 ±0.016**	8.3% ± 1.1%
PC_binary	91.1% ± 1.2%	90.8% ± 1.4%	91.5% ± 1.8%	91.1% ± 1.2%	0.822 ± 0.024	9.3% ± 1.5%
multi_comb_1	90.0% ± 0.8%	94.1% ± 1.2%	85.3% ± 0.8%	89.5% ± 0.8%	0.803 ± 0.016	5.4% ± 1.1%
add_comb_1	88.9% ± 1.0%	83.6% ± 1.1%	**96.8% ±0.9%**	89.7% ± 0.9%	0.788 ± 0.019	19.0% ± 1.4%
multi_comb_2	88.0% ± 0.7%	**94.1% ±0.8%**	81.2% ± 1.3%	87.1% ± 0.8%	0.768 ± 0.013	**5.1% ±0.7%**
add_comb_2	87.4% ± 0.7%	81.6% ± 0.9%	96.6% ± 0.4%	88.5% ± 0.5%	0.761 ± 0.012	21.9% ± 1.2%

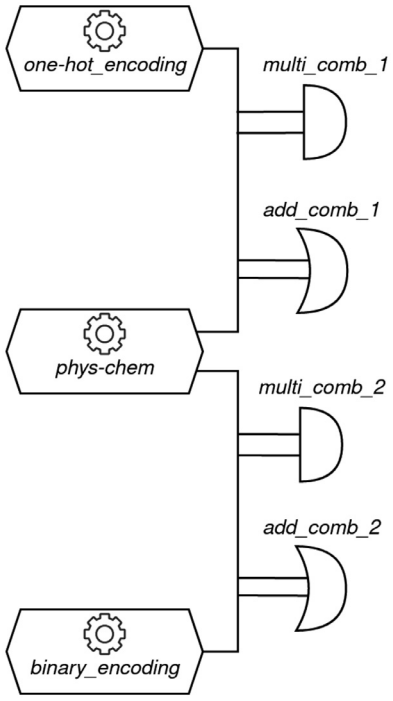


Fig. 4. Combining the single model predictions to create the hybrid models.

crobial if at least one of the two models classifies it as antimicrobial. This combined workflow is depicted in Fig. 4, and the results that the proposed methodologies have produced are presented in Table 5.

The usage of these approaches introduces a precision-recall trade-off. The *multi_comb_1* approach is the most rigorous one. Less peptides are classified as antimicrobial, but those that are classified as antimicrobial are more likely to be true positives. This has the effect of increasing the precision, but lowering the recall. Conversely, the *add_comb_1* is a more liberal approach. More peptides are classified as antimicrobial, however at the cost of producing more false positives. This approach will decrease the precision, while increasing the recall metric. Analogously, the *phys-chem* model is also combined with the *binary_encoding* model to construct the *multi_comb_2* and the *add_comb_2* methods.

4.2. Comparing prediction probabilities

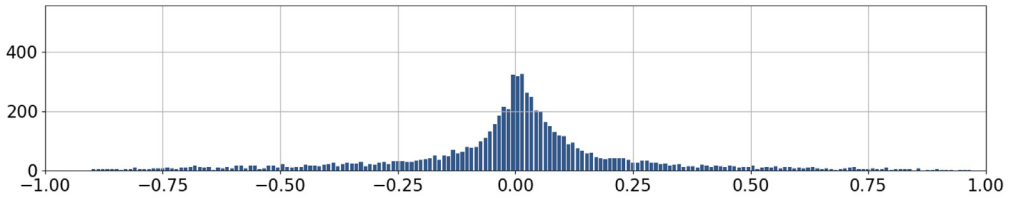
To compare the output probabilities that each of the individual, non-combined models have yielded, we have subtracted them. As presented in Fig. 5, the differences for all compared models are centered around 0, indicating that the prediction probabilities are similar. This was confirmed by the Spearman's correlation analysis, which yielded correlation coefficients r_s of 0.732, 0.782, and 0.859 for models compared in Fig. 5a, Fig. 5b, and Fig. 5c, respectively, indicating their strong and very strong relationship.

Taking the variances and Spearman's correlation coefficients into account, the relationship between sequence-based approaches is stronger than between them and the *phys-chem* model. The Cochran's Q statistical test compared the classification predictions of the three models, and with the *p*-value of 1.28e-09 it revealed there exsists a statistically significant difference between at least one pair of models. The McNemar statistical test with the Bonferroni correction was used to identify that this result is due to the differences between the *phys-chem* model and the sequence based models (Fig. 5a,b), while the difference between the *one-hot_encoding* and the *binary_encoding* models was not identified as significant (Fig. 5c).

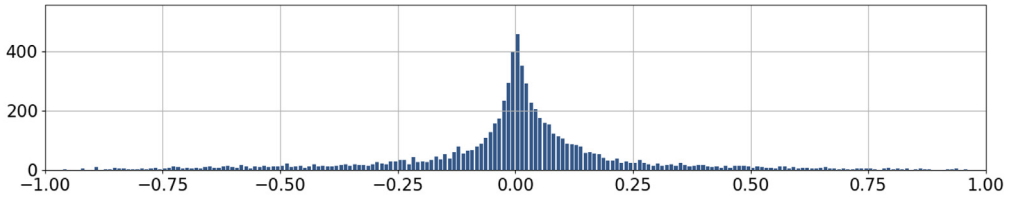
4.3. Permuting highly probable sequences

An additional round of model evaluation on *de novo* generated sequences has been conducted to verify the sensitivity of the models. Upon evaluation of the *phys-chem* model, peptides that were predicted as antimicrobial with high certainty (probability 90% or higher) in any of the testing folds were singled out, and used as the basis for creating the *permutation set*. Among the 1737 peptides that satisfied this criterion, 66 of them were chosen, randomly selecting one peptide for each of the viable peptide lengths ($length \in [10, 75]$). Their amino acid sequences were randomly shuffled 10 times to create 10 new peptides, for a total of 660 new peptides in the *permutation set*. We additionally confirmed that none of the 660 newly generated sequences appear in the training dataset.

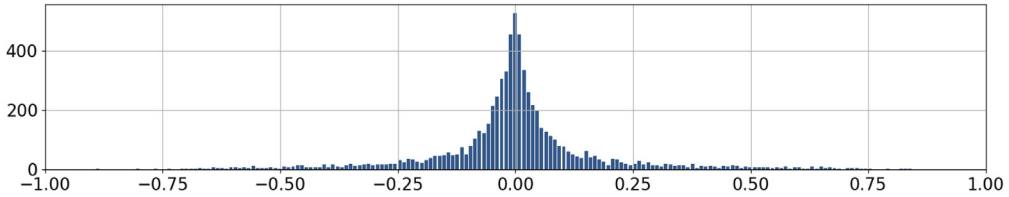
Considering that the 66 chosen sequences come from various testing folds, we made sure that the permutations are given to the same model as it's original peptide. In a sense, we have created a new testing dataset with 660 peptides of unknown antimicrobial activity for which we wanted to test the models' sensitivity. The output prediction probabilities have been compared as presented in Fig. 6. The overlaid histograms in Fig. 6a,b reveal that the predictions of the sequence-based models are scattered across the entire probability range, while the predictions of the *phys-chem* model are grouped in the $P \in [0.9, 1.0]$ range. This is due to the *phys-chem* model's insensitivity to amino acid order in the sequence, which is why the model outputs identical probabilities for all permutations, emphasizing the importance of sequence order information once again. Fig. 6c shows the differences in prediction probabilities for the *binary_encoding* and *one-hot_encoding* models, and exhibits a grouping around 0, with a variance of $\sigma^2 = 0.086$. The predictions of the three models have also been compared using the Cochran's Q statistical test which has confirmed that a significant difference exists between at least one pair of the examined models (p-value = 5.18e-55). Subsequently, the McNemar statistical test with the Bonferroni correction has revealed that a significant difference exists between all models. Finally, the Spearman's correlation coefficients are 0.06, -0.02, and 0.43 for models compared in Figs. 6a, 6b, and 6c, respectively. This confirms that although there is a certain degree of similarity (moderate corre-



(a) Comparison of *phys_chem* and *binary_encoding* ($\mu = -0.0013$, $\sigma^2 = 0.071$, $r_s = 0.732$, McNemar p-value = 4.09e-06).

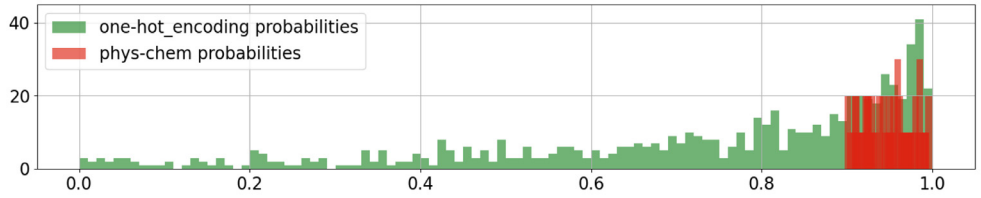


(b) Comparison of *phys_chem* and *one-hot_encoding* ($\mu = -0.0015$, $\sigma^2 = 0.059$, $r_s = 0.782$, McNemar p-value = 5.39e-08).

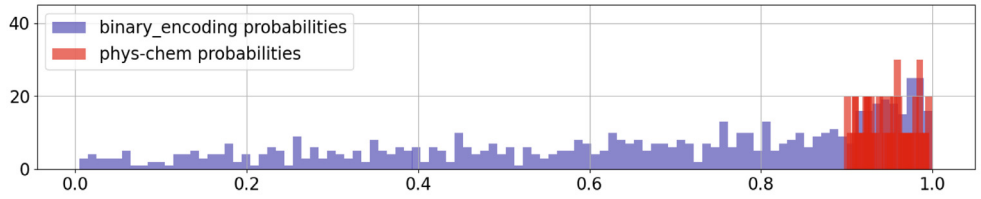


(c) Comparison of *binary_encoding* and *one-hot_encoding* ($\mu = -0.0002$, $\sigma^2 = 0.036$, $r_s = 0.859$, McNemar p-value = 0.078).

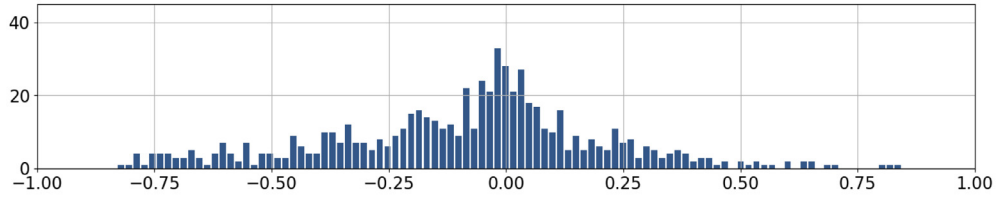
Fig. 5. Differences in prediction probabilities.



(a) Prediction probabilities of the *phys-chem* and *one-hot_encoding* models ($r_s = 0.06$, McNemar p-value = 1.5e-36).



(b) Prediction probabilities of the *phys-chem* and *binary_encoding* models ($r_s = -0.02$, McNemar p-value = 7.70e-34).



(c) Differences in prediction probabilities for the *binary_encoding* and *one-hot_encoding* models ($\mu = -0.09$, $\sigma^2 = 0.086$, $r_s = 0.43$, McNemar p-value = 3.38e-10).

Fig. 6. Comparing prediction probabilities for the *permutation set*.

lation) between the *one-hot_encoding* and the *binary_encoding* models, a statistically significant difference exists between their predictions.

5. Discussion

The focus of this paper is the ability of AMP prediction models to discern between sequence permutations, and their ability to maintain a low false positive rate. The means of achieving this aim are the numerical representations of peptides used to develop SVM models capable of distinguishing peptides with antimicrobial properties. We have analyzed the traditional approach based on physico-chemical features and two sequence-based approaches that use categorical variable encoding techniques to encode amino acids into bit vectors. Although all of the single models have produced fair results according to Table 5, the *one-hot_encoding* shown better overall performance over the *phys_chem* and *binary_encoding* models by a margin of 3.5%, referring to the accuracy metric. In the context of this research, we argue that a trade-off is introduced regarding the chosen approach and the results which the adopted approach outputs. The sequence-based methods produce better results, however they are limited by the fact that the sequence lengths must be limited with an upper bound. On the other hand, the calculation of physico-chemical features does not require such limitations, though it does indeed produce inferior results.

In addition to this, the results convey that the *one-hot_encoding* model has outperformed the *binary_encoding* model by 3.5%, with respect to accuracy. Another essential factor that needs to be taken into account when interpreting these results is data dimensionality. The feature vector of the *one-hot_encoding* model is four times larger than that of the *binary_encoding* model (1500 against 375). The data dimensionality issue may prove to be crucial should we analyze longer proteins instead of shorter peptides. The one-hot encoding scheme might also prove to be infeasible for long sequences and small datasets, as the number of features might exceed the number of observations. Furthermore, taking into consideration that Fig. 5 suggests that the *binary_encoding* model produces probabilities similar to the ones of the *one-hot_encoding* model, we conclude that the binary encoding scheme can function as a viable, sustainable, and well-performing alternative to the one-hot encoding technique in the context of AMP prediction.

Another noticeable trend is that the combination of the physico-chemical and sequence encoding features undeniably leads to better results. The *PC_one-hot* model displayed better performance than the individualistic *phys-chem* and *one-hot_encoding* models by margins of 4.4% and 0.9%. Analogously, the *PC_binary* model has outperformed *phys-chem* and *binary_encoding* models by 3.4%. This leads us to deduce that the peptide's physico-chemical descriptors and the amino acid ordering information both have an important role in the classification of peptides.

Finally, we have made an effort to combine the predictions of the contrasting methods to guide the models towards the improvement of the recall, precision, and fall-out metrics. The logical disjunctions of the *phys-chem* model and the sequence-based models create the *add_comb* approaches which produce a larger set of potentially antimicrobial peptides. More *in vitro* testing candidates are produced, inevitably leading to a larger number of false positives as well. On the other hand, the *multi_comb* approach is based on the logical conjunction of the models. This method produces a smaller set of candidates that make it into the *in vitro* testing phase, but the ones that are to be tested *in vitro* have a higher chance of actually having antimicrobial properties. This method of prediction aggregation can be used to reduce the number of false positives, which enables the faster discovery of new antimicrobial peptides during *in vitro* testing.

6. Conclusion

In this paper, we have explored the usage of different methods for the numerical representation of peptides with the goal of developing machine learning models that are able to differentiate between antimicrobial and non-antimicrobial peptides. We point out the characteristics, advantages, and disadvantages of each of the representation methodologies, namely the physico-chemical descriptors, one-hot and binary encoding and their combinations. We emphasize the importance of using the calculated physico-chemical descriptors of the peptides together with the encoded amino acid sequences to preserve both the intrinsic chemical properties, as well as the amino acid ordering information, through the implementation of combined encoding strategies that utilize both of the aforesaid feature generation techniques. Furthermore, with the goal of reducing the number of false positives generated by the prediction models, we propose workflows that combine the predicted outputs of the diverging methods. Our case study showed that the proposed combined encoding models and hybrid models outperform the single encoding models in every evaluation metric. The performance of single models on the *permutation set* has demonstrated that the physico-chemical features are insensitive to shifting amino acids within a sequence, which justifies the usage of combined and hybrid models. Finally, comparing the one-hot encoding and binary encoding methods, we argue that binary encoding acts as an effective alternative to one-hot encoding, while introducing a performance-dimensionality trade-off. The binary encoding method produces results which are inferior to the ones of the one-hot encoding method, but with the advantage of reducing the data dimensionality by 75%.

This paper also advocates the importance of analyzing the performance of predictive models from the expert's point of view, in the sense that false positive predictions of high confidence are more costly than the false negative ones because the former ones may lead to unnecessary experimental expense. To achieve the aim of minimizing the false positive rate, we propose hybrid models named *multi_comb_1* and *multi_comb_2* that require both the physico-chemical descriptors-based model and the one-hot encoding-based model to agree on declaring a peptide as positive. The results have shown that this methodology is superior to other single models by a factor of 1.82 to 2.96 in terms of fall-out. It also exhibits the improvement of precision between 3.3 and 6.1%.

Declaration of Competing Interest

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

Acknowledgments

This research was funded by the Croatian Science Foundation/Hrvatska zaklada za znanost [UIP-2019-04-7999]; and by the University of Rijeka [uniri-pr-tehnic-19-10].

References

[1] Ahsan MM, Mahmud M, Saha PK, Gupta KD, Siddique Z. Effect of data scaling methods on machine learning algorithms and model performance. Technologies 2021;9(3):52.
[2] Bagwari A, Joshi P. Approaches of sentiment analysis: a review. Des Eng 2021:1219–32.
[3] Bahar AA, Ren D. Antimicrobial peptides. Pharmaceuticals 2013;6(12):1543–75. doi:10.3390/ph6121543.
[4] Balouiri M, Sadiki M, Ibnsouda SK. Methods for *in vitro* evaluating antimicrobial activity: a review. J Pharm Anal 2016;6(2):71–9. doi:10.1016/j.jpha.2015.11.005.
[5] Buckland M, Gey F. The relationship between recall and precision. J Am Soc Inf Sci 1994;45(1):12–19.
[6] Chen J, Cheong HH, Siu SW. Xdeep-acpep: deep learning method for anticancer peptide activity prediction based on convolutional neural network and multitask learning. J Chem Inf Model 2021;61(8):3789–803.
[7] Chen W, Ding H, Feng P, Lin H, Chou K-C. iACP: a sequence-based tool for identifying anticancer peptides. Oncotarget 2016;7(13):16895.
[8] Chicco D, Jurman G. The advantages of the matthews correlation coefficient (MCC) over f1 score and accuracy in binary classification evaluation. BMC Genom 2020;21(1):1–13.

[9] Chou K-C. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Curr Proteom 2009;6(4):262–74.
[10] Conover WJ. Practical nonparametric statistics, vol 350. John Wiley & Sons; 1999.
[11] Consortium U. Uniprot: a worldwide hub of protein knowledge. Nucleic Acids Res 2019;47(D1):D506–15.
[12] Dubos RJ. Studies on a bactericidal agent extracted from a soil bacillus: I. preparation of the agent. its activity *in vitro*. J Exp Med 1939;70(1):1.
[13] Gullapalli AS, Mittal VK. Early detection of Parkinson's disease through speech features and machine learning: a review. In: ICT with intelligent applications. Springer; 2022. p. 203–12.
[14] Hotchkiss RD, Dubos RJ. Fractionation of the bactericidal agent from cultures of a soil bacillus. J Biol Chem 1940;132(2):791–2.
[15] Huan Y, Kong Q, Mou H, Yi H. Antimicrobial peptides: classification, design, application and research progress in multiple fields. Front Microbiol 2020;11:2559. doi:10.3389/fmicb.2020.582779.
[16] Lata S, Sharma B, Raghava GP. Analysis and prediction of antibacterial peptides. BMC Bioinform 2007;8(1):1–10.
[17] Lehman A. JMP for basic univariate and multivariate statistics: a step-by-step guide. SAS Institute; 2005. ISBN 9781590477793.
[18] Lertampaiporn S, Vorapreeda T, Hongsthong A, Thammarongtham C. Ensemble-amppred: robust amp prediction and recognition using the ensemble learning method with a new hybrid feature for differentiating amps. Genes 2021;12(2):137. Basel.
[19] Mahlapuu M, Håkansson J, Ringstad L, Björn C. Antimicrobial peptides: an emerging category of therapeutic agents. Front Cell Infect Microbiol 2016;6:194. doi:10.3389/fcimb.2016.00194.
[20] Manavalan B, Shin TH, Kim MO, Lee G. Aippred: sequence-based prediction of anti-inflammatory peptides using random forest. Front Pharmacol 2018;9:276. doi:10.3389/fphar.2018.00276. https://www.frontiersin.org/article/10.3389/fphar.2018.00276.
[21] Meher PK, Sahu TK, Saini V, Rao AR. Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into chou's general pseaac. Sci Rep 2017;7(1):1–12.
[22] Mercer DK, Torres MDT, Duay SS, Lovie E, Simpson L, von Köckritz-Blickwede M, de la Fuente-Nunez C, O'Neil DA, Angeles-Boza AM. Antimicrobial susceptibility testing of antimicrobial peptides to better predict efficacy. Front Cell Infect Microbiol 2020;10:326. doi:10.3389/fcimb.2020.00326.
[23] Mousavizadegan M, Mohabatkar H. Computational prediction of antifungal peptides via Chou's PseAAC and SVM. J Bioinform Comput Biol 2018;16(04):1850016.
[24] Myers JL, Well AD. Research design and statistical analysis. HarperCollins; 1991. ISBN 9780673464149.
[25] Müller AT, Hiss JA, Schneider G. Recurrent neural network model for constructive peptide design. J Chem Inf Model 2018;58(2):472–9.
[26] Organization WH. Global action plan on antimicrobial resistance. World Health Organization; 2015.
[27] Osorio D, Rondón-Villarreal P, Torres R. Peptides: a package for data mining of antimicrobial peptides. Small 2015;12:44–444.
[28] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: machine learning in python. J Mach Learn Res 2011;12:2825–30.
[29] Phoenix D, Dennison S, Harris F. Antimicrobial peptides: their history, evolution, and functional promiscuity. Antimicrob Pept 2013;8:1–37.
[30] Rozek T, Wegener K, Bowie J, Olver I, Carver J, Wallace J, et al. The antibiotic and anticancer active aurein peptides from the Australian bell frogs litoria aurea and litoria raniformis: the solution structure of aurein 1.2. Eur J Biochem 2000;267:5330–41. doi:10.1046/j.1432-1327.2000.01536.x.
[31] Shi G, Kang X, Dong F, Liu Y, Zhu N, Hu Y, et al. Dramp 3.0: an enhanced comprehensive data repository of antimicrobial peptides. Nucleic Acids Res 2021;1:D488–96.
[32] Spänig S, Heider D. Encodings and models for antimicrobial peptide classification for multi-resistant pathogens. BioData Min 2019;12(1):1–29.
[33] Spänig S, Mohsen S, Hattab G, Hauschild A-C, Heider D. A large-scale comparative study on peptide encodings for biomedical classification. NAR Genom Bioinform 2021;3(2):lqab039.
[34] Spellberg B, Guidos R, Gilbert D, Bradley J, Boucher HW, Scheld WM, Bartlett JG, Edwards John J. The epidemic of antibiotic-resistant infections: a call to action for the medical community from the Infectious Diseases Society of America. Clin Infect Dis 2008;46(2):155–64. doi:10.1086/524891. the Infectious Diseases Society of America.
[35] Tallorin L, Wang J, Kim WE, Sahu S, Kosa NM, Yang P, Thompson M, Gilson MK, Frazier PI, Burkart MD, et al. Discovering de novo peptide substrates for enzymes using machine learning. Nat Commun 2018;9(1):1–10.
[36] Tchagna Kouanou A, Mih Attia T, Feudjio C, Djeumo AF, Ngo Mouelas A, Nzogang MP, Tchito Tchapga C, Tchiotsop D. An overview of supervised machine learning methods and data analysis for covid-19 detection. J Healthc Eng 2021;2021.
[37] Torrent M, Andreu D, Nogués VM, Boix E. Connecting peptide physicochemical and antimicrobial properties by a rational prediction model. PLoS One 2011;6(2):e16968.
[38] Veltri D, Kamath U, Shehu A. Deep learning improves antimicrobial peptide recognition. Bioinformatics 2018;34(16):2740–7.
[39] Wei L, Zhou C, Chen H, Song J, Su R. ACPred-fL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. Bioinformatics 2018;34(23):4007–16.
[40] Xiao X, Wang P, Lin W-Z, Jia J-H, Chou K-C. iAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types. Anal Biochem 2013;436(2):168–77. doi:10.1016/j.ab.2013.01.019.
[41] Yan J, Bhadra P, Li A, Sethiya P, Qin L, Tai HK, Wong KH, Siu SW. Deep-AmPEP30: improve short antimicrobial peptides prediction with deep learning. Mol Ther Nucleic Acids 2020;20:882–94.
[42] Yoshida M, Hinkley T, Tsuda S, Abul-Haija YM, McBurney RT, Kulikov V, Mathieson JS, Galiñanes Reyes S, Castro MD, Cronin L. Using evolutionary algorithms and machine learning to explore sequence space for the discovery of antimicrobial peptides. Chem 2018;4(3):533–43. doi:10.1016/j.chempr.2018.01.005.