



Universiteit
Leiden
The Netherlands

CalcAMP: a new machine learning model for the accurate prediction of antimicrobial activity of peptides

Bournez, C.; Riool, M.; Boer, L. de; Cordfunke, R.A.; Best, L. de; Leeuwen, R. van; ... ;
Westen, G.J.P. van

Citation

Bournez, C., Riool, M., Boer, L. de, Cordfunke, R. A., Best, L. de, Leeuwen, R. van, ...
Westen, G. J. P. van. (2023). CalcAMP: a new machine learning model for the accurate
prediction of antimicrobial activity of peptides. *Antibiotics*, 12(4).
doi:10.3390/antibiotics12040725

Version: Publisher's Version







License: [Creative Commons CC BY 4.0 license](#)

Downloaded from: <https://hdl.handle.net/1887/3618249>

Note: To cite this publication please use the final published version (if applicable).

Article

CalcAMP: A New Machine Learning Model for the Accurate Prediction of Antimicrobial Activity of Peptides

Colin Bournez ¹, Martijn Riool ^{2,†}, Leonie de Boer ², Robert A. Cordfunke ³, Leonie de Best ⁴, Remko van Leeuwen ⁴, Jan Wouter Drijfhout ³, Sebastian A. J. Zaat ², and Gerard J. P. van Westen ^{1,*}

¹ Computational Drug Discovery, Drug Discovery and Safety, Leiden Academic Centre for Drug Research, Leiden University, P.O. Box 9502, 2300 RA Leiden, The Netherlands

² Department of Medical Microbiology and Infection Prevention, Amsterdam Institute for Infection and Immunity, Amsterdam UMC, University of Amsterdam, 1105 AZ Amsterdam, The Netherlands

³ Department Immunology, Leiden University Medical Center, 2300 RC Leiden, The Netherlands

⁴ Madam Therapeutics B.V., Pivot Park Life Sciences Community, Kloosterstraat 9, 5349 AB Oss, The Netherlands

* Correspondence: gerard@lacdr.leidenuniv.nl

† Current address: Laboratory for Experimental Trauma Surgery, Department of Trauma Surgery, University Hospital Regensburg, Am Biopark 9, 93053 Regensburg, Germany.

Abstract: To combat infection by microorganisms host organisms possess a primary arsenal via the innate immune system. Among them are defense peptides with the ability to target a wide range of pathogenic organisms, including bacteria, viruses, parasites, and fungi. Here, we present the development of a novel machine learning model capable of predicting the activity of antimicrobial peptides (AMPs), CalcAMP. AMPs, in particular short ones (<35 amino acids), can become an effective solution to face the multi-drug resistance issue arising worldwide. Whereas finding potent AMPs through classical wet-lab techniques is still a long and expensive process, a machine learning model can be useful to help researchers to rapidly identify whether peptides present potential or not. Our prediction model is based on a new data set constructed from the available public data on AMPs and experimental antimicrobial activities. CalcAMP can predict activity against both Gram-positive and Gram-negative bacteria. Different features either concerning general physicochemical properties or sequence composition have been assessed to retrieve higher prediction accuracy. CalcAMP can be used as an promising prediction asset to identify short AMPs among given peptide sequences.

Keywords: antimicrobial peptides; artificial intelligence; bacteria; drug discovery; machine learning; antimicrobial resistance



Citation: Bournez, C.; Riool, M.; de Boer, L.; Cordfunke, R.A.; de Best, L.; van Leeuwen, R.; Drijfhout, J.W.; Zaat, S.A.J.; van Westen, G.J.P.

CalcAMP: A New Machine Learning Model for the Accurate Prediction of Antimicrobial Activity of Peptides.

Antibiotics **2023**, *12*, 725. <https://doi.org/10.3390/antibiotics12040725>

Academic Editor: Fernando Albericio

Received: 26 February 2023

Revised: 24 March 2023

Accepted: 31 March 2023

Published: 7 April 2023



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

1. Introduction

It is now recognized that an increase in bacterial resistance to conventional antibiotics can lead us to a “post antibiotic era” [1]. Conventional therapeutic strategies often no longer work; hence, there is an urgent need to find new drugs to fight pathogens. Despite a few promising compounds entering the different clinical phases, only two new classes (lipopeptides and oxazolidinones) were discovered in the last 20 years [2]. Moreover, both of them target only Gram-positive (Gram+) bacteria and their representatives already face serious resistance drawbacks [3,4]. Therefore, there is a clear priority to find new antimicrobial compounds, especially against a selection of critical strains published by the World Health Organization (WHO) [5]. Among alternatives to small molecules drugs, antimicrobial peptides (AMPs) are considered as interesting and promising candidates [6]. These peptides, which are already present in the innate immune system of plants, animals, and humans, possess both antimicrobial activity and immunomodulatory properties [7]. AMPs are an essential component of the body’s first line of defense against pathogens, even before the adaptive immune system is activated. Moreover, they exhibit diverse

structural and functional profiles that can be optimized and fine-tuned to enhance their activity further [8,9]. As a result, they offer tremendous potential as novel therapeutic agents for combating a wide range of pathogens.

Several general properties are shared among AMPs such as a number of amino acids (AAs) between 5 and 60, typically a global net positive charge (>3), and amphipathic structures [10]. Still, even if the majority are cationic, several anionic AMPs exist [11,12]. Concerning their conformational characteristics, they show a great diversity of possible 3D structures from linear α -helices to β -sheets or random coils. They can also be cyclic or present with one or several disulfide bridges [10]. Figure 1 represents an overview of this variety in 3D structures among several AMPs.

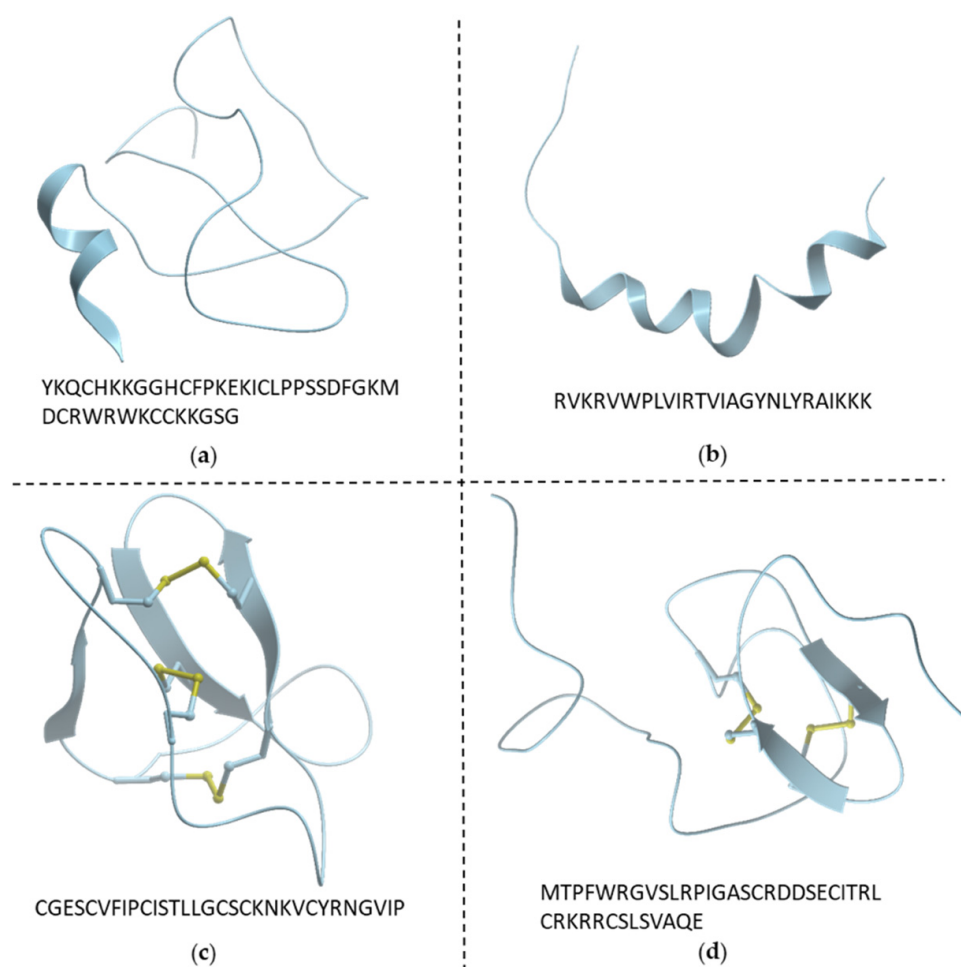


Figure 1. Overview of the variety of 3D AMP structures. (a) Crotamine from PDB code 1Z99; (b) fowlicidin from PDB code 2AMN; (c) circullin B from PDB code 2ERI; (d) LEAP-2 from PDB code 2L1Q. Yellow bonds represent disulfide bridges.

In contrast to conventional antibiotics that inhibit specific intracellular targets, most AMPs act directly on the bacterial cell membrane or crucial cytoplasmic components [13]. Their interaction with the membrane provoke its disruption leading to the death of the bacteria [14]. Therefore, the threshold to develop resistance is higher since it involves a great modification of the membrane [15]. Moreover, since eukaryotic and prokaryotic membranes present different specifications, AMPs can be very selective against bacteria by accumulating at their negatively charged membrane surface [16]. In addition to their antibacterial effects, AMPs may also present antifungal, antiparasitic, antiviral, or even anticancer properties thus strengthening their potential and importance as new therapeutics [17–20]. However, despite their ability, numerous interesting peptides never passed

preclinical stages for various reasons. The most important one is their possible toxicity against human cells, in particular red blood cells, leading to their lysis similar to that of bacteria [21]. Peptides may also present some stability issues, low oral bioavailability, or high cost of production [22–24]. Those limitations are not unsurmountable with peptide modification and engineering, e.g., D-amino acids or N-terminal modification can be considered to improve their characteristics. Nevertheless, these challenges limited large pharmaceutical companies from fully supporting the development of AMPs [25]. As of today, only a few new AMPs are approved by the Food and Drug Administration (FDA) or are in clinical trials [26]. Therefore, in order to help in the discovery of new ones and reduce their cost, several computational approaches were developed.

In silico predictive models typically rely on the primary sequence of proven AMPs from which different compositional and physicochemical descriptors are calculated and used for predictions. Since the beginning of the century, several AMP prediction tools were conceived based on different features and various algorithms. Experimentally validated AMPs can be retrieved from different public databases such as DBAASP, DRAMP, or CAMP [27–29]. In general, such databases also include a predictive model on their website accessible via a webserver. In addition to these ones, several standalone tools have also been developed for this purpose. For instance, iAMP-2 utilizes a fuzzy k-nearest neighbor algorithm and pseudo amino acid composition (PseAAC) to predict not only antimicrobial activity but also other types of activities such as anticancer or antiviral activities [30]. Another tool, iAMPpred, employs a support vector machine (SVM) algorithm and three different types of features (compositional, structural, and physicochemical) [31]. More recently, Bahdra et al. presented AmPEP, which uses the distribution patterns of amino acid properties and a Random Forest (RF) classifier [32]. Subsequently, an updated version focused on shorter peptides (<30 AAs), named DEEP-AmPEP30, was introduced. DEEP-AmPEP30 is based on pseudo k-tuple reduced amino acids composition (PseKRAAC) and a convolutional neural network (CNN) algorithm [33]. However, the first deep learning-based AMP prediction model was conceived by Veltri et al. in 2018. It relies on the peptide's primary sequence composition, converted to a numerical vector, for its prediction [34].

Still, the majority of current AMP prediction tools suffer from two main drawbacks. First, they do not account for differences in bacterial species or membrane structure differences, whereas the potency of AMPs can vary significantly depending on the target and the type of bacterial envelope [35]. Second, these tools employ randomly selected sequences without experimentally determined antimicrobial activity as the negative data set, rather than using confirmed inactive peptides. Nowadays, it remains difficult to develop a model specific to a bacterial species or a precise strain since little to no experimental data are available to do so. A few methods were conceived to more precisely target bacteria using their Gram classification and a threshold on activity to discriminate AMPs from other peptides (i.e., non-AMPs), but it is only specific to either Gram+ or Gram-negative (Gram−) bacteria [36,37]. Finally, even more recently, a new deep learning-based approach only specific to *Escherichia coli* has been published focusing on small AMPs (<20 AAs) without cysteine residues [38].

Here, we created a new data set composed exclusively of experimentally proven AMPs and Non-AMPs by setting an activity threshold to discriminate them. The experiments contained in our data set were focused on short AMPs (<35 AAs) since they can present potent activity coupled with low toxicity [15,39]. Furthermore, it is much more convenient to work with such peptides due to their small size, and they are simpler to synthesize, optimize and produce on a bigger scale, implying lower costs. Based on this data set, we introduced several novel predictive machine learning (ML) models separated according to the Gram classification. Hence, a specific prediction model was conceived for each class. In addition, an antifungal activity prediction model was created. The prediction method relies on the calculation of several sequence composition and physicochemical descriptors and several different ML algorithms assessed with cross validation (CV) and a holdout test data set.

2. Results

2.1. Exploration of the Data Set

2.1.1. Global Overview

After classification of all bacterial species by their Gram staining (positive or negative), the exploration began with the search of the most tested species. Figure 2a shows a detailed overview of the five most retrieved species per category. For each category, the large difference observed between the first and second most tested microorganisms reveals a clear preference for certain species when running experimental tests. As suspected, *Staphylococcus aureus* and *E. coli* were the most tested Gram+ and Gram− bacterial species, respectively, and for fungi it was *Candida albicans*. While important species from the WHO priority list of antibiotic-resistant bacteria were present, namely *Acinetobacter baumannii* and *Pseudomonas aeruginosa*, other important species were not retrieved in the top five of tested microorganisms, such as *Helicobacter pylori*, *Neisseria gonorrhoeae*, or *Streptococcus pneumoniae* [2,5]. Figure 2b shows a Venn diagram illustrating the distribution of peptides tested against the different categories. As numerous peptides present several activities against different targets, their number is much lower than the number of activities. Moreover, the majority of them are common in both the Gram+ and Gram− data sets. However, there was also a significant number of peptides that are specific to each category (1243, 1600, and 576 for the Gram+, Gram−, and fungi, respectively). Concerning antifungal peptides, much less data are available than for bacteria and the majority of the peptides were retrieved within the Gram category. Still, a significant number of the antifungal peptides was also specific to this category. It appears that one global model for AMP prediction would not sufficiently represent the data since, as shown in Figure 2b, a great number of peptides were uniquely tested to one of the categories: Gram+, Gram−, or fungi.

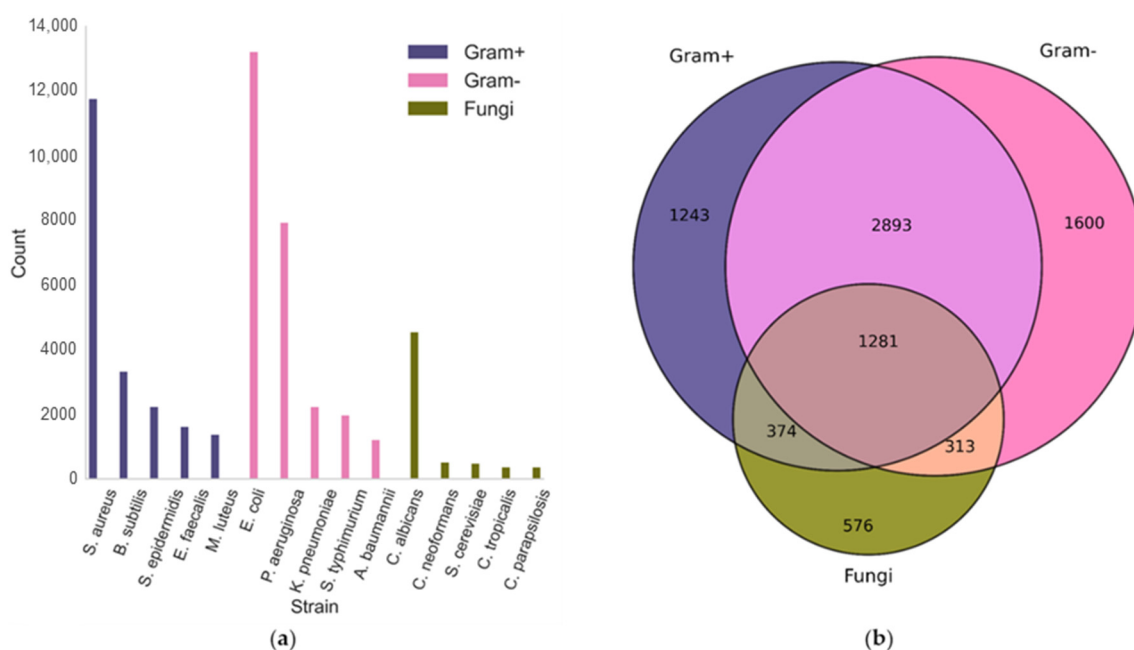


Figure 2. Number of experimental assays retrieved for the top five species by category (a). Venn diagram showing the distribution of peptides per category (b).

2.1.2. AMP/Non-AMP Peptides Analysis

The analysis of the overall amino acid composition (AAC) between AMPs and Non-AMPs is represented in Figure 3a. For both categories, the five most represented AAs were lysine (K), leucine (L), alanine (A), arginine (R), and glycine (G). In general, AMPs are known to be cationic, and they indeed exhibited a higher frequency of positive residues such as lysine or arginine. Additionally, tryptophan (W) was overrepresented in AMPs.

Conversely, the proportion of histidine (H) was slightly higher for Non-AMP peptides, which were also much more enriched in negative residues such as aspartic acid (D) and glutamic acid (E), despite these being in low abundance overall. Finally, a global similarity was observed for non-polar residues such as phenylalanine (F), isoleucine (I), leucine (L), and cysteine (C) participating in the amphiphilic properties of these peptides. These results were in correlation with the global charge difference between the two categories (Figure 3b). As mentioned above, AMPs were significantly more positively charged than Non-AMPs (Mann–Whitney U test, $p < 0.05$). They had an average positive charge of +4.8 (standard deviation SD: 2.70) vs. a charge of +2.8 (SD: 2.56) for Non-AMPs. AMPs were also significantly heavier (Mann–Whitney U test, $p < 0.05$) with a mean of 2241 g·mol⁻¹ (SD: 703) vs. 1911 g·mol⁻¹ (SD: 715) for Non-AMPs (see Figure 3c). Both a PCA and t-SNE analysis were performed on the overall physicochemical descriptors (Supplementary Materials, Figure S1) and AAC (Supplementary Materials, Figure S2). Such projections allow us to quickly see if one can perceive a separation between AMPs and Non-AMPs. A significant overlap existed between the two categories in the PCA space. For t-SNE, the projections were very sparse and small clusters appeared to be quite discriminative but overlap was visible. Therefore, we hypothesize that the specific AA sequence is more important for the biological activity than the overall composition.

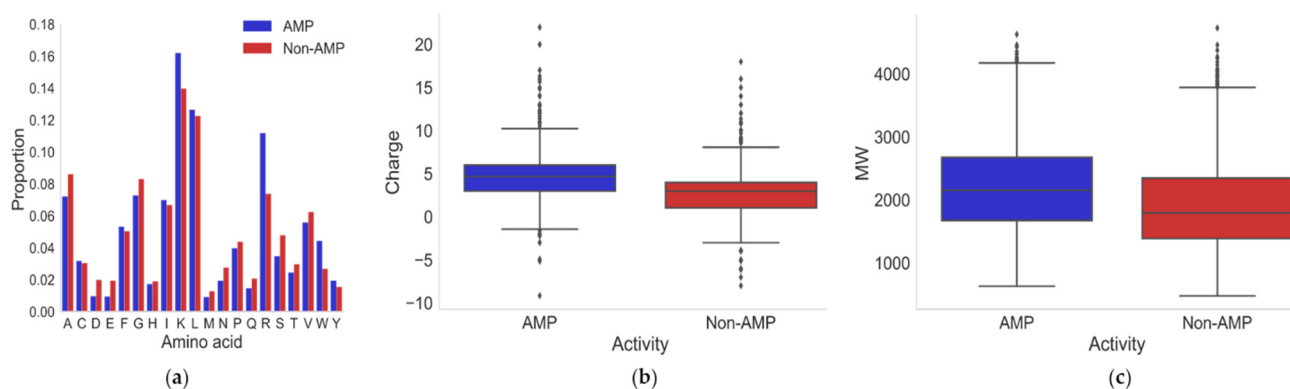


Figure 3. Comparison of amino acid composition (a), global net charge (b), and molecular weight (c) between AMPs and Non-AMPs.

A deeper analysis on the 4174 common peptides tested against both Gram+ and Gram− bacteria and the distribution of AMPs/Non-AMPs is presented within the matrix in Figure 4 (see *Data labelling* section for details on the classification). The majority, 89%, acted either as an AMP or as a Non-AMP in both Gram+ and Gram− categories. Still, a non-negligible part of Gram− bacteria AMPs (7%) were not active against Gram+ bacteria and vice versa (4%). These results reinforce the importance to have a specific model for Gram+ and Gram− bacteria rather than a global one.

		Gram-	
		AMP	Non-AMP
Gram+	AMP	1708 (41%)	301 (7%)
	Non-AMP	174 (4%)	1991 (48%)

Figure 4. Matrix of labels for the common peptides between Gram− and Gram+ categories.

A PCA and a t-SNE projection were applied based on several general physicochemical descriptors and were subsequently colored by class in Figure 4 matrix. Based on the PCA projection, Figure 5a, a slight separation appeared between AMPs (right) from Non-AMPs (left). However, a significant overlap remained between them. Peptides labelled as AMP for Gram+ and Non-AMP for Gram− (in green) tended to cluster with the common Non-AMP ones on the left side. Whereas peptides labeled AMP for Gram− and Non-AMP for Gram+ (in black) tended to be projected with the common AMPs on the right side. The correlation circle associated with this PCA (Supplementary Materials, Figure S3) shows that component 1 is mostly “Charge” and component 2 is “MW” (molecular weight) and “Length”. The plot confirms that AMPs have on average a higher molecular weight and are more positively charged. The same separation was retrieved in the t-SNE projection, Figure 5b, with an interesting cluster on the bottom of practically only the Non-AMPs projected there. Still, in most cases, AMPs and Non-AMPs are near coincident in the plots, meaning they present similar physicochemical characteristics. The same projections were also produced using AAC (Supplementary Materials, Figure S4). These projections show an horizontal separation between AMPs and Non-AMPs with AMPs being much more dispersed than Non-AMPs. The separation was realized for arginine (R), lysine (K), tryptophan (W), and leucine (L) for the top part (AMPs) and the other AAs for the bottom part (Non-AMPs).

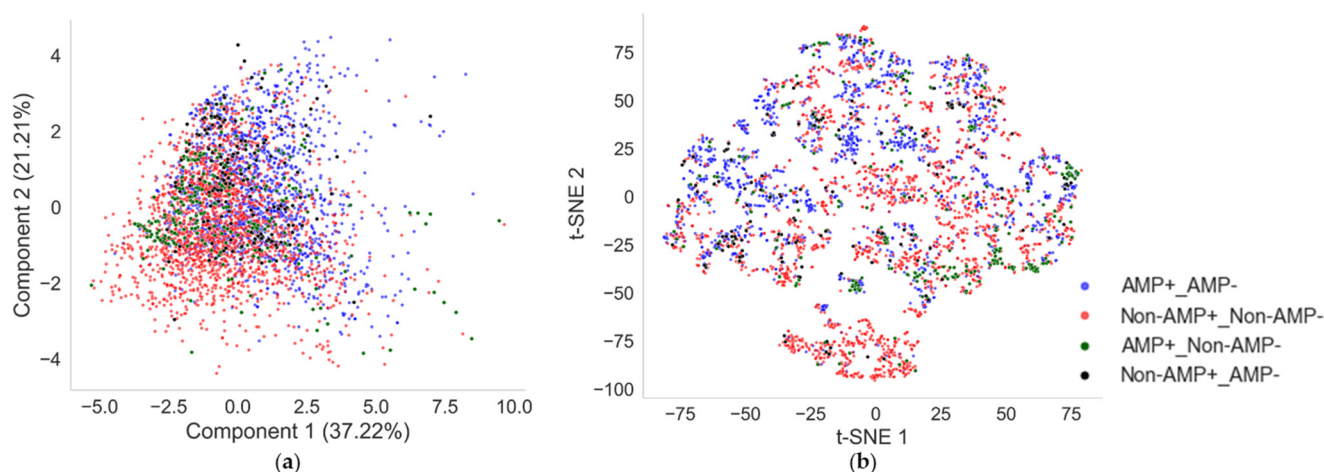


Figure 5. PCA (a) and t-SNE (b) projections of physicochemical descriptors between common peptides of Gram+ and Gram− categories. In blue are peptides are labelled as AMP in both categories, in red are peptides are labelled as Non-AMP in both categories, in green are peptides labelled AMP for Gram+ and Non-AMP for Gram−, and in black is the opposite.

2.2. Antimicrobial Activity Prediction

2.2.1. Feature Selection

As a basis for the prediction model, it is crucial to extract and select features from the peptide’s primary sequence. The features from the sequences can be divided into two main categories: based on physicochemical descriptors or based on AAC. For this study, both types were used and evaluated individually to retain only the most interesting ones. Feature selection was performed using the Random Forest classifier (RF) [40], a ML algorithm that has an extensive track record in both drug discovery and AMP prediction and can be interpreted. Hence, this algorithm was selected in order to obtain features that produce the most accurate peptide classification. This first preliminary assessment, based on a classical RF-classifier, was performed with 10-fold cross validation (CV) experiments each time. The CV process consists of the data set being split into k-folds and k − 1 folds being used as training data while the final fold is retained for evaluation. Therefore, the model is assessed k times where each of the k-folds serve once as the validation data. After that, the mean score and standard deviation for each metric can be calculated.

For sequence-based features, the AAC, the dipeptide composition (DPC), the pseudo amino acid composition (PseAAC) [41], and the composition-transition-distribution (CTD) descriptors [42] were retained. AAC is the frequency of each AA in the sequence and DPC is the same for dipeptides. Thus, they are made of 20 and 400 descriptors, respectively. Contrarily to those two, PseAAC allows us to retain all the sequence-order information. Depending on the initial parameters, it contains a different number of descriptors, derived from the primary sequence and incorporating some sequence-order knowledge. Finally, CTD descriptors are divided in three components: composition (C), the number of AAs with a particular property divided by the length of the sequence; transition (T), the frequency where AAs with a particular property is followed by AAs of another property; and distribution (D), the measure of different lengths of the sequence of the distribution of each property. In total, CTD descriptors are composed of 147 features (21 for C and T and 105 for D) and describes seven properties: charge, hydrophobicity, normalized van der Waals volume, polarity, polarizability, secondary structure, and solvent accessibility. Concerning the global physicochemical (GPC) descriptors, they are composed of the 10 following descriptors: length, MW, global charge, charge density, isoelectric point, instability index, aromaticity, aliphatic index, Boman index, and hydrophobic ratio. Of note, AA scale-based descriptors such as T-scales or Z-scales as well as descriptors derived from the AA index were dropped because of insufficient prediction results even though they were previously shown to perform well on peptide bioactivity modelling [43,44].

The performances were evaluated according to several metrics including accuracy, sensitivity, specificity, area under the receiver operator characteristic (ROC) curve (AUC), and Matthew's correlation coefficient (MCC). The results are shown in Table 1. For all sets of features, individually and aggregated, the predictions between categories were similar and no major difference appeared. For both categories, CTD descriptors were the ones with the best prediction according to all metrics tested. It makes sense that CTD are the descriptors producing the best predictions since they are the most complete ones, including sequence composition and physicochemical criteria. The higher scores in all metrics was obtained using all these set of descriptors together. The choice of these features allowed our basic RF model to achieve an accuracy of 81% in both categories.

Table 1. Comparison of different feature sets for Gram+ (white) and Gram− (grey) AMP prediction. The values in brackets represent the standard deviation obtained via 10-fold cross validation. Bold values indicate the best value per column.

Feature Set (#)	Accuracy	Sensitivity	Specificity	AUC-ROC	MCC
AAC (20)	0.77 (0.02) 0.78 (0.02)	0.77 (0.02) 0.76 (0.01)	0.78 (0.04) 0.81 (0.03)	0.85 (0.01) 0.86 (0.02)	0.55 (0.04) 0.56 (0.03)
CTD (147)	0.79 (0.02) 0.80 (0.01)	0.77 (0.03) 0.79 (0.02)	0.82 (0.03) 0.82 (0.02)	0.87 (0.02) 0.88 (0.01)	0.59 (0.04) 0.61 (0.03)
DPC (400)	0.77 (0.02) 0.77 (0.02)	0.78 (0.03) 0.77 (0.02)	0.76 (0.03) 0.78 (0.03)	0.85 (0.02) 0.86 (0.01)	0.53 (0.04) 0.55 (0.04)
PseAAC (24)	0.77 (0.02) 0.78 (0.02)	0.76 (0.03) 0.75 (0.03)	0.79 (0.02) 0.81 (0.02)	0.85 (0.02) 0.86 (0.01)	0.55 (0.04) 0.55 (0.04)
GPC (10)	0.78 (0.01) 0.78 (0.01)	0.78 (0.02) 0.78 (0.03)	0.79 (0.03) 0.79 (0.02)	0.85 (0.01) 0.86 (0.01)	0.57 (0.02) 0.57 (0.02)
All (601)	0.81 (0.02) 0.81 (0.02)	0.80 (0.03) 0.80 (0.04)	0.83 (0.03) 0.82 (0.03)	0.89 (0.02) 0.89 (0.02)	0.62 (0.05) 0.62 (0.05)

2.2.2. Algorithm Choice

To assess which algorithm best suits our prediction purpose, several different models from classical classification algorithms were tested. Thus, the performance of 14 ML algorithms and one Multi-Layer Perceptron (MLP) was evaluated with 10-fold CV each

time (Tables S1 and S2 in Supplementary Materials). For both categories, i.e., Gram+ and Gram−, it was clearly observed that ensemble tree-based algorithms outperformed, in all metrics, all the other types of ML models as well as the MLP. They achieved their prediction with a mean accuracy of 80% compared to 73% for k-nearest neighbors, for example. Therefore, the decision was made to build and tune a model for each top five tested algorithms in each category: Catboost [45], LightGBM [46], XGBoost [47], Random Forest, and Extra Trees [48].

2.2.3. Performance and Interpretation

For each algorithm selected, two models were created and tuned, one with all the features (601) and one with a supplementary feature selection process. Indeed, feature selection can be an important step to accelerate the learning and training but also improve the performance of any model [49]. In order to identify the best discriminative and useful features, their importance weights were used followed by a recursive elimination with 3-fold CV each time, reducing their numbers to 75. The overall results of each AMP prediction model is presented in Supplementary Materials, Tables S3 and S4. For both categories, no created model stood out, as they all presented similar results within their standard deviation ranges. Moreover, no significant performance losses were observed after our feature selection step. The selection of the best model was therefore made using the external data set. For the Gram+ model (CalcAMP+), the best classifier was the Extra Trees one with all features, as it achieved a prediction accuracy on the external test set of 79% and an MCC of 0.58. For the Gram− model (CalcAMP-), the best one was obtained with the LightGBM algorithm using all of the features. This model obtained an accuracy of 80% and an MCC of 0.61. More results can be found in Table 2 in the *Comparison with Other Prediction Tools* Section 2.2.4.

A “SHapley Additive exPlanations” (SHAP) values [50] analysis was performed to globally interpret the predictions on our test data set. SHAP values allow us to visualize which features are important for the prediction and their contribution. For the two models, the top 20 variable importance plot is shown in Figure 6, with their impacts on the prediction. The features are ranked in descending order and the horizontal scatterplot for each illustrates whether the effect of that feature was associated with a positive or a negative prediction output. For example, for CalcAMP+, a high MW (dots in red) had a strong and positive impact on AMP prediction. Of these top 20 features, only three were common to both models: MW, Charge, and pI. The majority of the top features were from CTD descriptors. They are identifiable by their names beginning with an underscore character, followed by the property and finally the component characteristics: composition (C), transition (T), and distribution (D). From the AAC, only the proportion of tryptophan (W) was represented in CalcAMP+. No features from DPC descriptors and only one from PseAAC were retrieved in each model. For the global physicochemical descriptors, four out of ten were part of the top features in both models, meaning that their importance was high, as those of the CTD descriptors. For CalcAMP+ (A), the impact of the 20 features were quite important as illustrated on the horizontal distribution, whereas for the CalcAMP- (B), except for charge and MW, the impact of the other features were less important.

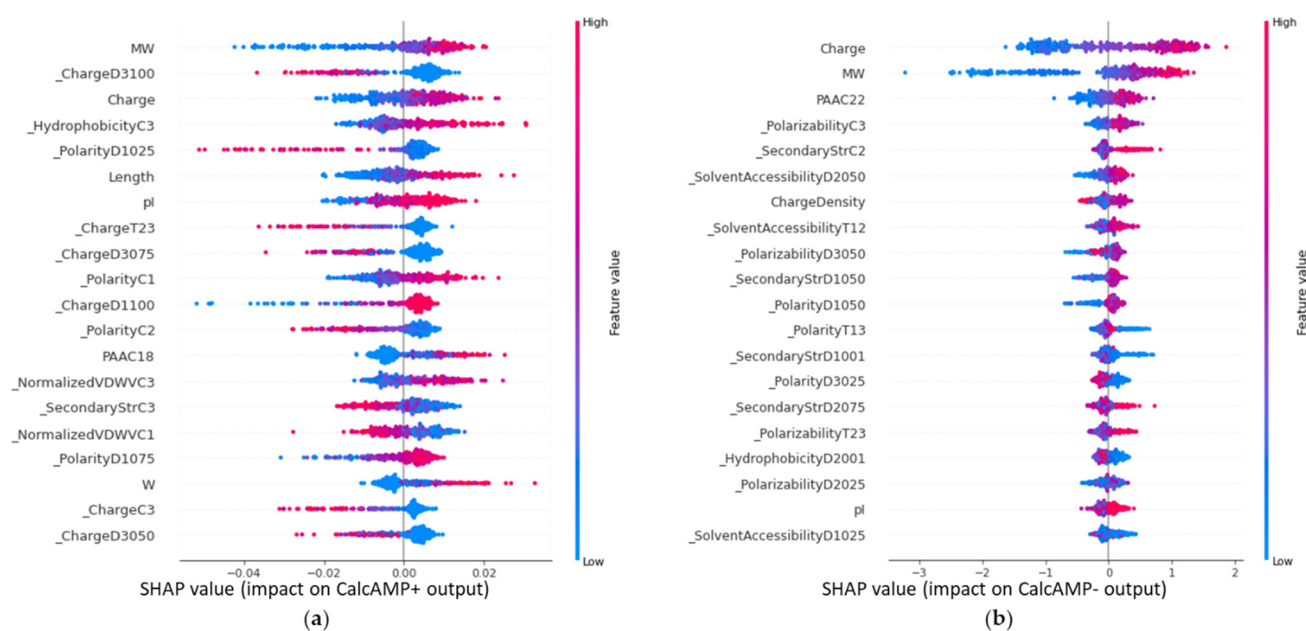


Figure 6. Top 20 features importance plot and their impact on the external test set prediction for CalcAMP+ (a) and CalcAMP- (b). Shown are physicochemical properties such as molecular weight (MW), Charge, or Length. However, the majority of the top features were from CTD descriptors. They are identifiable by their names beginning with an underscore character, followed by the property and finally the component characteristics: composition (C), transition (T), and distribution (D).

A deeper look into the prediction results, especially at the confusion matrix associated with the predictions (Figure 7a), reveals that the great majority of true positives (TP) and true negatives (TN) was similarly predicted in both models. However, the image was less clear for false positives (FP) or false negatives (FN), where a little more than half were predicted differently even if the overall metrics of both models were equivalent. This confusion matrix shows that our models returned different prediction results and the distinction between Gram+ and Gram− bacteria remains important. Finally, the analysis of the probabilities associated with the prediction and not directly the binary output (Figure 7b) displayed a scoring difference between our two models. CalcAMP+ returned lower scores in general for any category, while CalcAMP- had an average close to 1 or 0 for TP and TN predictions. However, the important observation here is that there was a significant scoring difference between TP and FP and between TN and FN for both models. TP scores were significantly higher than FP scores (Mann–Whitney U test $p < 0.05$) and TN scores were significantly lower than FN scores (Mann–Whitney U test $p < 0.05$). Therefore, in order to increase the sensitivity or the specificity of the prediction, one should increase or decrease the classification threshold differently for the CalcAMP+ and CalcAMP- models since they had different scoring scales.

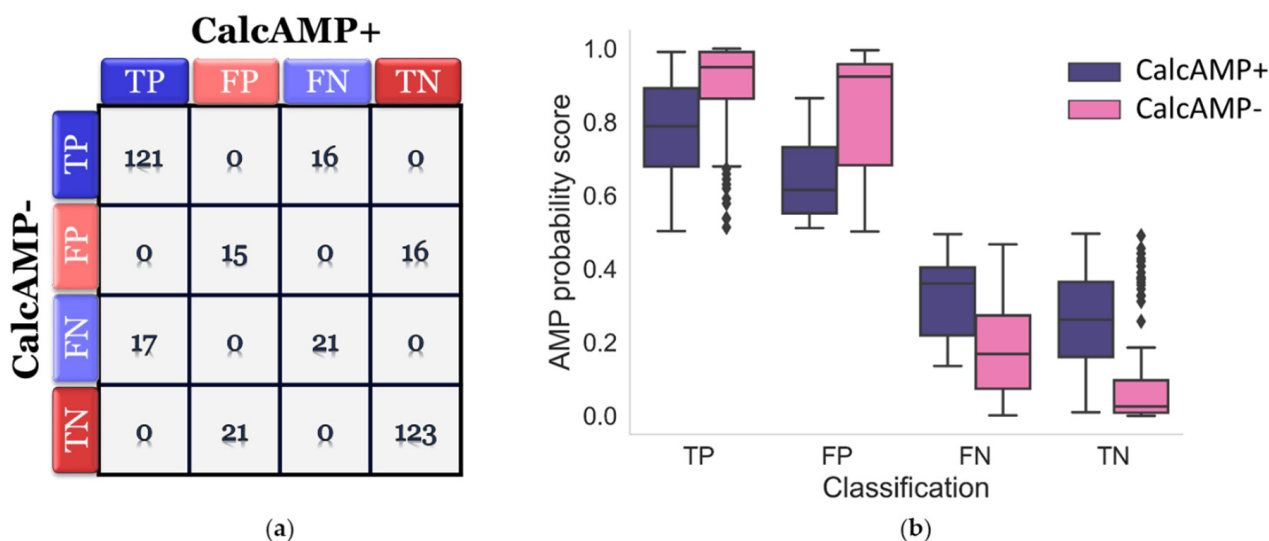


Figure 7. Confusion matrix for CalcAMP+ model prediction versus CalcAMP- (a) and AMP probability score by predicted class (b).

2.2.4. Comparison with Other Prediction Tools

Previously published AMP prediction tools were assessed in order to compare them to our models using the external benchmark data set, composed of AMPs and non-AMPs common to both Gram+ and Gram− bacteria. As already mentioned, most AMP predictive approaches that have been developed are also based on a training set composed of AMPs found in public databases. However, their negative data set is made without an activity threshold but based on random peptide sequences from the UniProt database tagged as Non-AMP. In our case, the method is significantly different since the classification of peptides as AMP or Non-AMP was based on their measured activities (see *Data Labelling*, Section 4.1.2). Nevertheless, our best classifiers were compared to five existing prediction models: iAMPpred, DBAASP, RF-AmPEP30, Deep-AmPEP30, and AMP Scanner Vr.2. These five tools are freely available as webservers, which were used for this comparison.

The comparison demonstrates the superiority of the CalcAMP models over all other tested tools in the global prediction of activity (Table 2). The CalcAMP accuracy was 79% and 80% for Gram+ and Gram−, respectively, versus 67% or less for the others, and CalcAMP demonstrated an MCC of at least 0.58 versus 0.35 or less. However, except for the DBAASP model, all the others had a higher sensitivity (>90%), meaning that they were more prone to predict the peptides as an AMP than CalcAMP. These results can be explained by their different training data sets. For models using randomized negatives, any peptide with a recorded antimicrobial activity is an AMP so in our external data set most of them will be predicted as an AMP. However, the drawback is their lower specificity ($\leq 30\%$), implying a difficulty in discriminating and predicting peptides as Non-AMP if they have a weak activity on their target. Both CalcAMP models presented a balanced high specificity and sensitivity. They were able to efficiently differentiate between peptides with high activity and those with lower ones. Figure 8 shows the different ROC curves (except for DBAASP model) and confirm that our models have high accuracy at various thresholds and are superior to the other models.

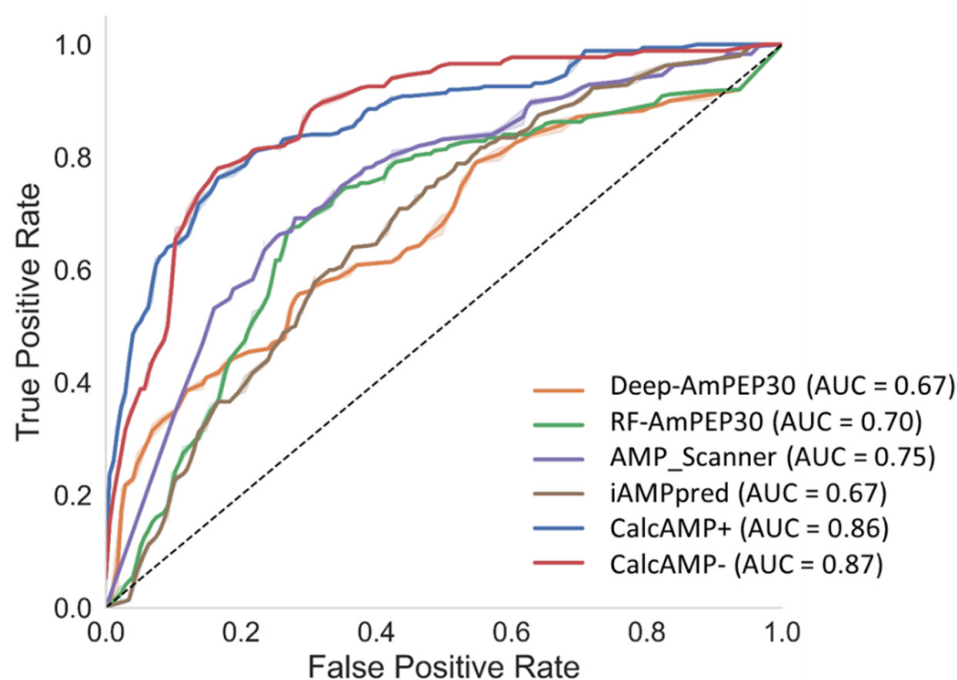


Figure 8. Receiver operator characteristic (ROC) curves of the different AMP classifiers and their area under the curve score.

Table 2. Comparison of different AMP prediction classifiers using the external data set. Bold values indicate the best value per column.

Model	Accuracy	Sensitivity	Specificity	AUC-ROC	MCC
Deep-AmPEP30 ¹	0.60	0.92	0.29	0.71	0.27
RF-AmPEP30 ¹	0.59	0.94	0.25	0.74	0.26
AMP_Scanner	0.61	0.93	0.30	0.75	0.29
iAMPpred	0.60	0.91	0.29	0.67	0.26
DBAASP	0.67	0.74	0.61	- ²	0.35
<i>Average</i>	0.61	0.89	0.35	0.72	0.29
CalcAMP+	0.79	0.79	0.79	0.86	0.58
CalcAMP-	0.80	0.78	0.82	0.87	0.61

¹ For Deep-AmPEP30 and RF-AmPEP30 models, only peptides with a length between 5 and 30 AAs were used since it does not predict using longer ones. ² For DBAASP, AUC-ROC cannot be calculated and the ROC curve could not be displayed since it only returns binary results and the probabilities associated are not accessible.

To further compare CalcAMP with the other tools, an assessment on their own respective external data set was also performed using the external benchmark for AmPEP and a shortened version of the Antimicrobial Peptide Scanner vr.2 validation data set (only peptides with a length between 5 and 30 AAs). More details and results can be found in Supplementary Materials, section *Comparison with other datasets*. As expected, CalcAMP did not perform as well on those data sets as on our own external data set but still displayed an accuracy of around 70% and an MCC > 0.4. Such a loss of performance can once again be explained by the initial labelling difference. Indeed, most peptides labelled as AMPs in those data sets would be labelled Non-AMP in our case, thus accounting for this decrease in prediction power. However, in doing these extra comparisons, our main interest was the Non-AMPs predictions and therefore the sensitivity. Indeed, most or all of their Non-AMPs have never been seen by any of our models. For both external data sets, the sensitivities of CalcAMP+ and CalcAMP- were higher than all the other models, reaching 100% for CalcAMP+ on the Antimicrobial Peptide Scanner vr.2 validation data set (Supplementary Materials, Tables S5 and S6 and Figures S5–S8).

2.3. Antifungal Activity Prediction

Fungal infections still remain a serious threat for humans and, similar to antibiotics, antifungal drugs present some limitations [51,52]. Therefore, antifungal peptides (AFPs) have emerged as new potential treatments to prevent or treat such infections [53], similar to AMPs for bacteria. Even though the initial and main focus here was AMPs, quite a few reported antifungal activities caught our attention. However, given the inferiority of input data (1301 Non-AFPs and 887 AFPs), a lighter protocol was proposed and only models based on RF and ET algorithms were employed. Concerning the initial exploration of the data set with global physicochemical descriptors, no clear separation between AFPs and Non-AFPs was visible with either PCA or t-SNE projections. Similarly, when looking at AAC, no AAs were unbalanced between the categories, with two exceptions. Arginine was enriched in AFPs over Non-AFPs and for alanine the inverse was true. Additionally, like with AMPs, the five most retrieved amino acids were lysine, leucine, glycine, alanine, and arginine. Of note, this could be the result of a bias as these data were retrieved from AMP public databases. Therefore, most of the peptides were synthesized with the aim to target bacteria so they present the characteristics of AMPs.

2.3.1. Performance and Interpretation

Following the same protocol as for the AMPs, two models per algorithm were created and tuned: one with all of the features (601) and one with a feature selection step ending with 75 features. Again, no created model stood out as they all presented similar results within their standard deviation ranges (Supplementary Materials, Table S7). No significant performance losses were observed after our feature selection step either. On the external test set consisting of 30 AFPs and 30 Non-AFPs, the best classifier was the Random Forest one with all features (CalcAFP) with an accuracy of 77% and a great specificity of 90% but a lower sensitivity (63%). The prediction results are presented in Table 3. Such a difference might be explained by the small input imbalance, where roughly 60% were Non-AFPs versus 40% AFPs (due to the small amount of data, the choice was made to leave it as is). Therefore, in our model it is best to discard Non-AFPs from selection rather than identifying the AFPs.

Table 3. Comparison of different AFP prediction classifiers using the external data set. Bold values indicate the best value per column.

Model	Accuracy	Sensitivity	Specificity	AUC-ROC	MCC
iAMPpred	0.52	0.77	0.27	0.56	0.04
ClassAMP	0.48	0.33	0.63	- ¹	−0.03
Antifp	0.50	0.30	0.70	- ¹	0.00
<i>Average</i>	<i>0.50</i>	<i>0.47</i>	<i>0.53</i>	-	<i>0.00</i>
CalcAFP	0.77	0.63	0.90	0.86	0.55

¹ For ClassAMP and Antifp, AUC ROC could not be calculated since we do not have access to the probabilities associated.

Analysis of the model with SHAP values and the top 20 variable importance plot are shown in Figure 9. Similar to the AMP prediction, charge and MW were important features; the higher they were, the higher the positive impact on AFP prediction. However, the descriptor pI was not retrieved in the top 20 features. In correlation with the AAC comparison, the proportion of arginine (R) and also the dipeptide RR were retrieved, and both were highly correlated with AFP prediction. Finally, the presence of three PseAAC descriptors showed that they might be more important and interesting for AFP prediction than for AMP predictions.

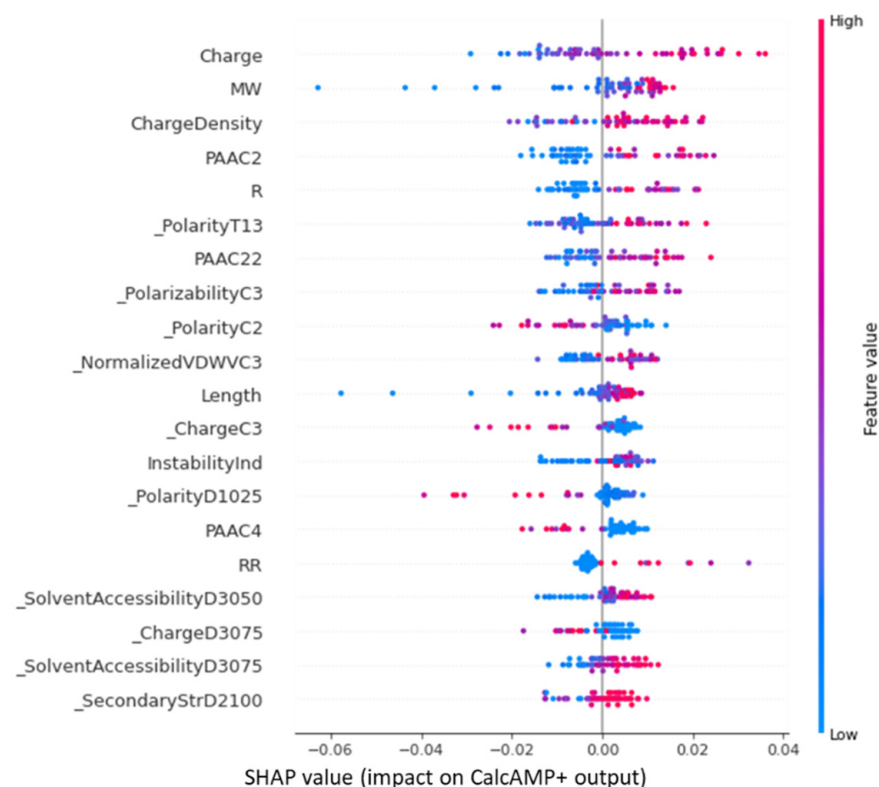


Figure 9. Top 20 feature importance plot and their impact on the external test set prediction for CalcAFP.

2.3.2. Comparison with Other Prediction Tools

The comparison results with other AFP prediction tools is provided in Table 3. Three available tools were evaluated: iAMPpred, ClassAMP [54], and AntiFP [55]. To note, iAMPpred and ClassAMP are not specific AFP prediction tools but multilabel ones proposing AFP prediction. For ClassAMP, the model based on the SVM algorithm was selected. Sequences predicted as antifungal were labeled as AFP while the ones not predicted as antifungal but as other classes were considered Non-AFPs. Similar to the previously developed AMP prediction tools, these tools were developed based on a training set composed of AFPs from public databases without an activity threshold and random peptide sequences from UniProt or Swiss-Prot tagged as non-AFP. These three tools are freely available as a webserver, which we used for this comparison. CalcAFP achieved an accuracy of 77% in contrast with 48% to 52% for the other predictors. It clearly outperformed them on all metrics, except sensitivity where iAMPpred was higher (63% vs. 77%).

Another data set was also evaluated, the one from the Antifp tool. The Antifp_Main validation data set was modified to keep only peptides with a length between 5 and 35 AAs. The details and results can be found in Supplementary Materials, section *Comparison with other datasets*. Unfortunately, with this data set, CalcAFP performance suffered heavy losses in all metrics with an MCC of -0.12 and was not able to perform better than a random model. Still, it maintained a good specificity of 79% (Supplementary Materials, Table S8). Hence, the performance difference can be explained by the different methodologies and initial classification of AFPs/Non-AFPs (Supplementary Materials, Figure S9). Even though our results for AFP prediction were lower than for the AMPs predictions, to the best of our knowledge, CalcAFP is the only model that works exclusively with peptides having experimentally measured antifungal activity and classified AFP/Non-AFP using a threshold.

3. Discussion

The increase in antibiotic resistance urges the discovery of new therapeutics to tackle this issue. AMPs represent a very interesting alternative to small molecules against bacteria. They can have a large spectrum of activity against bacteria (either Gram+, Gram−, or both classes) and fungi but also against viruses, parasites, or even cancer cells. Unlike eukaryotic cells, prokaryotic cells, and particularly their cytoplasmic membranes, are negatively charged and thus it is more convenient for the action of small cationic peptides. Therefore, AMPs tend to be more enriched in arginine or lysine residues compared to peptides showing no antimicrobial activity. However, an overall positive charge alone and the proportion of certain AAs are not enough to correctly discriminate AMPs from Non-AMPs. Other descriptors are important and the combination of several physicochemical and compositional ones is the key to build an efficient prediction model. Here, the creation of a novel public data set allowed us to construct ML models for antimicrobial activity prediction. For each class of bacteria, Gram+ or Gram−, a general accuracy of around 80% was achieved. CalcAMP outperformed existing AMP classifiers and was also able to correctly classify Non-AMP data set from these other tools.

The main limitation of the models comes from the input data gathered from different sources that present quite high heterogeneity. Moreover, it can be difficult to classify a peptide as a generic AMP since there is no clear experimental test to define it. Indeed, several factors can influence the outcome, such as the bacterial strain tested, the growth medium, and the type of activity measured. Since prediction models rely on input data, small changes in the method and choices to discriminate AMPs from Non-AMPs can have great consequences on the output. In an ideal world, models would be species or even strain specific, but as discussed in the introduction a lack of data makes this quite difficult to achieve. Future work will also focus on toxicity prediction with the creation of a new model coupled to the current one allowing one to have both activity and toxicity predictions returned. Indeed, toxicity remains a major issue in AMP design and development; therefore, an effective toxicity prediction tool would be a significant help for the design of potential clinical AMPs. We hope our method can be of great help and thus accelerate the R&D process of finding new AMPs as a potential alternative to antibiotics. Moreover, we have made our current curated data set available for use which could serve as a basis for other experiments and development of tools.

4. Materials and Methods

4.1. Data Preparation

4.1.1. Data Mining and Preprocessing

The data set of peptides serving as input for the different prediction models was built with publicly available data from different databases. At first, seven databases were selected and manually mined: ADAM [56], BaAMPs [57], CAMP [29], DBAASP [27], DRAMP [28], LAMP2 [58], and YADAMP [59]. From these, ADAM, having no precise experimental data on activity, and BaAMPs, containing no relevant data (only activities on biofilms), were rejected. Table 4 lists the databases examined with their URL and number of corresponding peptides.

Table 4. List of public AMP database used and their corresponding number of unique peptides.

Database	Number of Unique Sequence ²
ADAM 1 (A Database of Anti-Microbial Peptides) http://bioinformatics.cs.ntou.edu.tw/ADAM/index.html	7007
BaAMPs 1 (Biofilm-Active AMPs Database) http://www.baamps.it/	225
CAMP (Collection of Anti-Microbial Peptides) http://www.camp.bicnirrh.res.in/	8177
DBAASP (Database of Antimicrobial Activity and Structure of Peptides) https://dbaasp.org/	17,783
DRAMP (Data Repository of Antimicrobial Peptides) http://dramp.cpu-bioinfor.org/	22,259
LAMP2 (Linking Antimicrobial Peptides) http://biotechlab.fudan.edu.cn/database/lamp/index.php	23,253
YADAMP (Yet Another Database of Antimicrobial Peptides) http://yadamp.unisa.it/	2525

¹ In grey are the databases that were rejected. ² accessed in March 2021.

A primary filter was applied to the selected databases to keep only activity against either bacteria or fungi; all other activities (virus, parasites, cells, etc.) were discarded. Then, a second filter was implemented to retain only interesting activity types such as minimal inhibitory or minimal bactericidal concentrations (MIC/MBC) or 99.9% lethal concentration (LC_{99.9}). Therefore, nonstandard or unclear activity types were rejected. In order to easily and accurately compare experimental activity values, they were all converted to μM , the most predominantly used, using Equation (1). Finally, since the study focuses on the prediction of natural short AMPs, only peptides having a length between 5 and 35 AAs were considered. Among them, repetitive sequences of single AAs (e.g., RRRRRRR, AAAAAAAAA) and sequences containing unnatural AAs were left out. The full data set is available in the Supplementary Materials.

$$(\mu\text{M}) = \frac{C(\mu\text{g}\cdot\text{mL}^{-1})}{MW(\text{Da})} \times 1000, \quad (1)$$

4.1.2. Data Labelling

Each activity recorded was grouped by Gram classification using the bacteria species tested. Unlike existing tools, with a negative data set constructed by selecting random sequences from UniProt, [32] our strategy was to work exclusively with peptides that were experimentally tested. Depending on their activities, peptides were classified either as AMP (1) or Non-AMP (0) based on reported MIC/MBC or equivalent LC_{99.9} values for antimicrobially active AMPs [60]. The activity threshold was set at $\leq 15 \mu\text{M}$ for a peptide to be considered as active (strong) and above $25 \mu\text{M}$ for inactive ones (or weaker AMPs that need higher concentrations to show antimicrobial properties). The values between these thresholds were discarded as they are considered in an area where labelling was not certain enough. The majority of peptides have been tested against multiple species of bacteria and hence present several activity values. Therefore, to take into account this heterogeneity, a specific workflow was set up for each peptide. If all activity values belonged to the active category (Gram+, Gram−, or fungi; $\leq 15 \mu\text{M}$) or if the majority was in this category while none of the values are above $25 \mu\text{M}$, then the peptide was labelled as an AMP (1). Conversely, if all activity values are higher than $25 \mu\text{M}$ or if the majority is and none are below $15 \mu\text{M}$, then the peptide was labelled as Non-AMP (0). In every other case, the antimicrobial activity was considered as unsure, and the peptide could not be labelled as either active or inactive and was discarded for the creation of the prediction models. This method is a means for us to tackle the heterogeneity of the experimental data and to take into consideration the uncertainty of antimicrobial activity caused by different

experimental conditions and/or experimental errors. In Figure 10, different examples of labelling depending on the situation of the peptide and its related activities are displayed. The categories were balanced between AMPs and Non-AMPs; there were slightly fewer AMPs for the Fungi category but no major imbalance was present. Our final data set is composed as follows:

- Gram+: 5791 peptides; 2849 Non-AMP (49%) and 2942 AMP (51%)
- Gram–: 6087 peptides; 3163 Non-AMP (52%) and 2924 AMP (48%)
- Fungi: 2544 peptides; 1475 Non-AMP (58%) and 1069 AMP (42%)

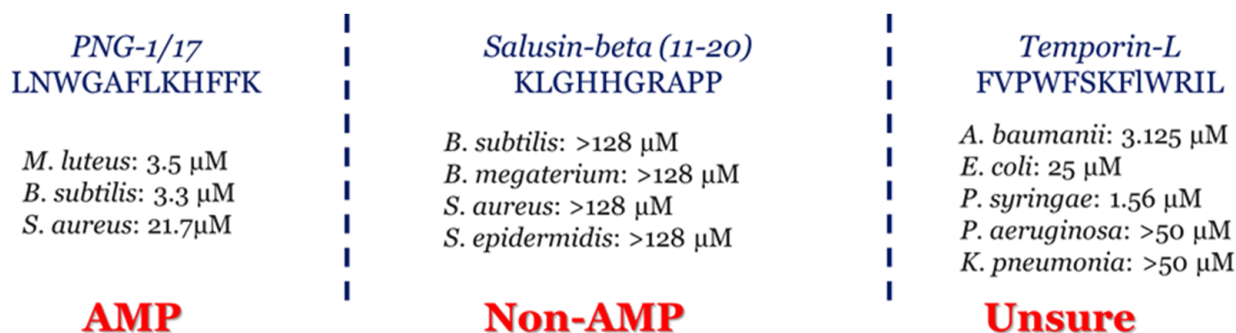


Figure 10. Different examples of peptides with their reported experimental activities and their label.

4.1.3. Creation of the Data Sets

To build training data sets and in order to avoid redundancy, the software CD-HIT [61] was used with a cutoff of 95% to remove highly similar sequences for both AMPs and Non-AMPs distinctly before aggregating them. For the Gram– category, since there was a small imbalance between AMPs and Non-AMPs, 350 Non-AMPs sequences were removed randomly before the model construction. The external data set test was composed of 350 peptides (175 AMPs and 175 Non-AMPs). It was used for the evaluation and comparison of our models to the previously developed models. For its creation, all the common AMPs and Non-AMPs of the Gram+ and Gram– data set were assembled. Then, a clustering was performed to select 300 representative peptides of each category. It was followed by filtering using CD-HIT with a cutoff of 80%. Finally, this was followed with a random selection on the remaining peptides to reach 175 AMPs and 175 Non-AMPs. Figure 11 shows how the training set was representative of the entire data set via a PCA projection based on physicochemical descriptors.

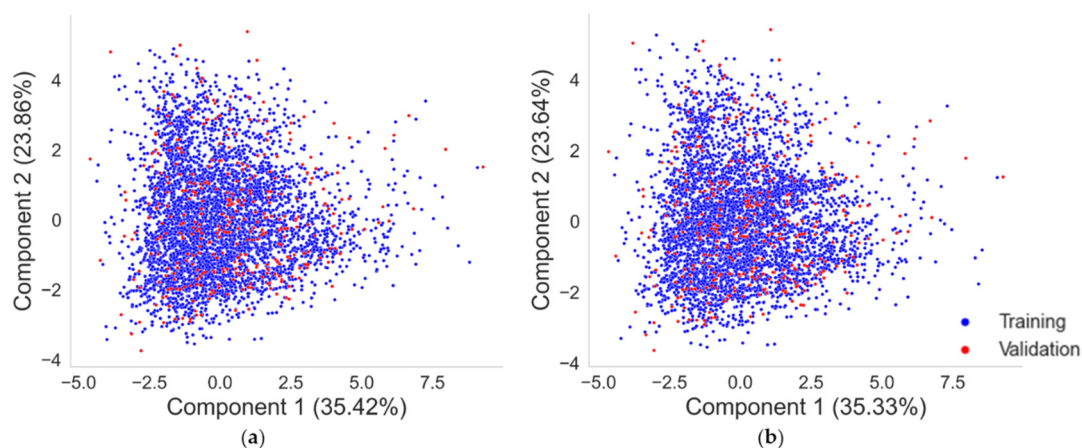


Figure 11. PCA projections of the training set (blue) and external test set (red) for Gram+ (a) and Gram– (b) categories.

Concerning the AFP prediction external data set, a simple clustering on our training data set to select 30 AFPs and 30 Non-AFPs representatives was performed.

4.2. Machine Learning Experiments

4.2.1. Feature Calculation

A preliminary exploration and assessment of the different existing types of descriptors (single, double, tri-peptide composition, Moran, Geary, Moreau-Broto, etc.) and co-variance-encoding methods (auto, cross, and auto-cross using different AA descriptor scales) was achieved. The ones that showed insufficient performance with a simple RF model were discarded and only the most promising were kept for the rest of the study. Thus, only AAC, DPC, CTD, GPC, and PseAAC (see section “Feature selection” for their brief description) were retained. All the calculations were made using Python 3.7 and the packages modLAMP 4.3.0 and PyBioMed 1.0.

4.2.2. Model Comparison

The comparison of different ML algorithms to develop a classification model between AMP and non-AMP were made using the package PyCaret 2.3.6. PyCaret is an open-source, low-code library allowing users to quickly compare several different ML algorithms. The algorithms included were Random Forest (RF), Extra Trees (ET), Extreme Gradient Boosting (xgboost), Light Gradient Boosting Machine (lightgbm), Ada Boost (ada), Gradient Boosting (gbc), CatBoost (catboost), Logistic Regression (lr), SVM linear kernel (svm), Naive Bayes (nb), Decision Tree (DT), Ridge (ridge), K-Nearest Neighbor (knn), Quadratic Discriminant Analysis (qda), Linear Discriminant Analysis (lda), and a Dummy Classifier (dummy). In addition, a Multi-layer Perceptron model created with Scikit-Learn 0.23.2 was added for the comparison step. The comparison was made using the “Classification” modules, without changing the parameters, and for each category (Gram+, Gram−), the top five models were kept for further analysis.

4.2.3. Model Creation and Tuning

For each retained algorithm, two models were created and tuned, one with all the features and one with a set of 75 features selected. The packages lightgbm 3.1.1, xgboost 1.5.0, and catboost 0.26.20 were used for the creation of our LightGBM, XGBoost, and CatBoost classifiers, respectively. For our RF and ET classifiers we used Scikit-Learn 0.23.2. The tuning and optimization of our created models were performed with the Scikit-Learn API and the “RandomizedSearchCV” function (3-fold CV per change of parameter) on the whole training data set. Once the hyperparameters were selected, the final model was established with them and the performance was evaluated first via a 10-fold CV and then with the external test set.

4.2.4. Feature Selection

Once the model calculated with all features was optimized, the process of feature selection started with the help of feature importance weights. First, Scikit-Learn API with the “SelectFromModel” function was used to remove unimportant features. Then, the recursive feature elimination (RFE) module was used, with 3-fold CV, to select 75 features from those remaining. RFE selects best features by recursively removing the least important features until the desired number of features is reached. The choice of 75 features was determined by plotting accuracy vs. number of features and the appearance of a plateau from this number.

4.2.5. Evaluation Metrics

The developed classifiers were systematically assessed using a 10-fold CV and five metrics: Accuracy (Acc), Sensitivity (Sen), Specificity (Spe), area under the ROC curve (AUC), and Matthew’s correlation coefficient (MCC). The external test set, with data

unseen by our models, was not included into the CV process and serves for supplementary evaluation and comparison. The metrics are defined as follows:

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \quad (2)$$

$$Sen = \frac{TP}{TP + FN} \quad (3)$$

$$Spe = \frac{TN}{TN + FP} \quad (4)$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad (5)$$

where TP (True positive) is the number of correctly predicted AMPs, TN (True Negative) is the number of correctly predicted Non-AMPs, FP (False Positive) is the number of Non-AMPs incorrectly predicted as AMPs, and FN (False Negative) is the number of AMPs incorrectly predicted as Non-AMPs. AUC is the area under the ROC curve: the plot of the true positive rate (sensitivity) as a function of the false positive rate (1—specificity). Accuracy (Equation (2)) is a global metric representing the sum of true positives and true negatives divided by the total number of the data; it indicates the proportion of correct predictions. Sensitivity (Equation (3)) and specificity (Equation (4)) focus on how well the classifier predicts AMPs and non-AMPs, respectively. AUC measures the ability to correctly distinguish between classes. All of them are between 0 and 1, and the higher value the better the performance of the model. Matthew's correlation coefficient (Equation (5)) also measures the overall quality of a binary classifier and is widely used in the field of ML. The MCC value is between -1 and 1 . Again, the closer to 1 , the better the performance, and a value of 0 indicates that the model is no better than a random prediction and -1 is a total disagreement between prediction and reality.

4.3. Implementation

All ML experiments were performed and implemented using Python 3.7. All the figures are made using Matplotlib and Seaborn packages. Input, generated, or analyzed data used in this study are included in this article's supplementary data sets or uploaded in Zenodo: <https://doi.org/10.5281/zenodo.7588702>. The code is available at: <https://github.com/CDDLeiden/CalcAMP>.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/antibiotics12040725/s1>, Figure S1: PCA (a) and t-SNE (b) projections of physicochemical descriptors between AMPs and Non-AMPs; Figure S2: PCA (a) and t-SNE (b) projections of amino acid composition between AMPs and Non-AMPs; Figure S3: Correlation circle associated to PCA shown in Figure 5 in the main text; Figure S4: PCA (a) and t-SNE (b) projections of amino acid composition between common peptides of Gram+ and Gram− categories; Table S1: Comparison of ML classifiers tested for Gram+ AMP prediction (10 Times 10-Fold Cross-Validation); Table S2: Comparison of ML classifiers tested for Gram− AMP prediction (10 Times 10-Fold Cross-Validation); Table S3: Results of ML classifiers created for Gram+ AMP prediction; Table S4: Results of ML classifiers created for Gram− AMP prediction; Figure S5: Matrix of labels for the common peptides between training dataset of CalcAMP+ and CalcAMP- models and AmPEP external benchmark dataset; Table S5: Comparison of different AMP prediction classifiers using AmPEP benchmark dataset; Figure S6: Receiver operator characteristic (ROC) curves of the different AMP classifiers and their area under the curve score obtained using AmPEP external benchmark dataset; Figure S7: Matrix of labels for the common peptides between training dataset of CalcAMP+ and CalcAMP- models and adapted Antimicrobial Peptide Scanner vr.2 validation dataset; Table S6: Comparison of different AMP prediction classifiers using adapted Antimicrobial Peptide Scanner vr.2 validation dataset; Figure S8: Receiver operator characteristic (ROC) curves of the

different AMP classifiers and their area under the curve score obtained using adapted Antimicrobial Peptide Scanner v2 validation dataset; Table S7: Results of ML classifiers created for AFP prediction; Figure S9: Matrix of labels for the common peptides between training dataset of CalcAFP model and adapted Antifp main validation dataset; Table S8: Comparison of different AFP prediction classifiers using adapted Antifp main validation dataset.

Author Contributions: Conceptualization, C.B., M.R., L.d.B. (Leonie de Boer), R.A.C., L.d.B. (Leonie de Best), R.v.L., J.W.D., S.A.J.Z. and G.J.P.v.W.; methodology, C.B. and G.J.P.v.W.; writing—original draft preparation, C.B. and G.J.P.v.W.; supervision, S.A.J.Z. and G.J.P.v.W.; data curation, C.B., L.d.B. (Leonie de Boer) and M.R.; funding acquisition, R.v.L., J.W.D., S.A.J.Z. and G.J.P.v.W. All authors have read and agreed to the published version of the manuscript.

Funding: This project received funding from the Dutch Scientific Council GDST-NWO science industry cooperation programme Chemistry of Advanced Materials, project number 729.001.024.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data used in this study were uploaded in Zenodo: <https://doi.org/10.5281/zenodo.7588702>, code: <https://github.com/CDDLeiden/CalcAMP>.

Acknowledgments: We would like to thank Roelof van der Kleij for his help using the university IT infrastructure.

Conflicts of Interest: Madam Therapeutics is a commercial company that aims to put these type of AMPs on the market.

References

1. Prestinaci, F.; Pezzotti, P.; Pantosti, A. Antimicrobial Resistance: A Global Multifaceted Phenomenon. *Pathog. Glob. Health* **2015**, *109*, 309. [CrossRef]
2. Tacconelli, E.; Carrara, E.; Savoldi, A.; Harbarth, S.; Mendelson, M.; Monnet, D.L.; Pulcini, C.; Kahlmeter, G.; Kluytmans, J.; Carmeli, Y.; et al. Discovery, Research, and Development of New Antibiotics: The WHO Priority List of Antibiotic-Resistant Bacteria and Tuberculosis. *Lancet Infect. Dis.* **2018**, *18*, 318–327. [CrossRef]
3. Eliopoulos, G.M.; Meka, V.G.; Gold, H.S. Antimicrobial Resistance to Linezolid. *Clin. Infect. Dis.* **2004**, *39*, 1010–1015. [CrossRef]
4. Munoz-Price, L.S.; Lolans, K.; Quinn, J.P. Emergence of Resistance to Daptomycin during Treatment of Vancomycin-Resistant Enterococcus Faecalis Infection. *Clin. Infect. Dis.* **2005**, *41*, 565–566. [CrossRef] [PubMed]
5. WHO Publishes List of Bacteria for Which New Antibiotics Are Urgently Needed. Available online: <https://www.who.int/news/item/27-02-2017-who-publishes-list-of-bacteria-for-which-new-antibiotics-are-urgently-needed> (accessed on 9 November 2022).
6. Mahlapuu, M.; Håkansson, J.; Ringstad, L.; Björn, C. Antimicrobial Peptides: An Emerging Category of Therapeutic Agents. *Front. Cell. Infect. Microbiol.* **2016**, *6*, 194. [CrossRef]
7. Nijnik, A.; Hancock, R. Host Defence Peptides: Antimicrobial and Immunomodulatory Activity and Potential Applications for Tackling Antibiotic-Resistant Infections. *Emerg. Health Threat. J.* **2009**, *2*, 7078. [CrossRef]
8. Fjell, C.D.; Hiss, J.A.; Hancock, R.E.W.; Schneider, G. Designing Antimicrobial Peptides: Form Follows Function. *Nat. Rev. Drug Discov.* **2012**, *11*, 37–51. [CrossRef]
9. Finlay, B.B.; Hancock, R.E.W. Can Innate Immunity Be Enhanced to Treat Microbial Infections? *Nat. Rev. Microbiol.* **2004**, *2*, 497–504. [CrossRef]
10. Huan, Y.; Kong, Q.; Mou, H.; Yi, H. Antimicrobial Peptides: Classification, Design, Application and Research Progress in Multiple Fields. *Front. Microbiol.* **2020**, *11*, 2559. [CrossRef] [PubMed]
11. Lai, Y.; Villaruz, A.E.; Li, M.; Cha, D.J.; Sturdevant, D.E.; Otto, M. The Human Anionic Antimicrobial Peptide Dermcidin Induces Proteolytic Defence Mechanisms in Staphylococci. *Mol. Microbiol.* **2007**, *63*, 497–506. [CrossRef]
12. Schitteck, B.; Hipfel, R.; Sauer, B.; Bauer, J.; Kalbacher, H.; Stevanovic, S.; Schirle, M.; Schroeder, K.; Blin, N.; Meier, F.; et al. Dermcidin: A Novel Human Antibiotic Peptide Secreted by Sweat Glands. *Nat. Immunol.* **2001**, *2*, 1133–1137. [CrossRef]
13. Magana, M.; Pushpanathan, M.; Santos, A.L.; Leanse, L.; Fernandez, M.; Ioannidis, A.; Giulianotti, M.A.; Apidianakis, Y.; Bradfute, S.; Ferguson, A.L.; et al. The Value of Antimicrobial Peptides in the Age of Resistance. *Lancet Infect. Dis.* **2020**, *20*, e216–e230. [CrossRef] [PubMed]
14. Benfield, A.H.; Henriques, S.T. Mode-of-Action of Antimicrobial Peptides: Membrane Disruption vs. Intracellular Mechanisms. *Front. Med. Technol.* **2020**, *2*, 20. [CrossRef]
15. Kim, H.; Jang, J.H.; Kim, S.C.; Cho, J.H. De Novo Generation of Short Antimicrobial Peptides with Enhanced Stability and Cell Specificity. *J. Antimicrob. Chemother.* **2014**, *69*, 121–132. [CrossRef] [PubMed]
16. Zasloff, M. Antimicrobial Peptides of Multicellular Organisms. *Nature* **2002**, *415*, 389–395. [CrossRef] [PubMed]

17. Cesare, G.B.D.; Cristy, S.A.; Garsin, D.A.; Lorenz, M.C. Antimicrobial Peptides: A New Frontier in Antifungal Therapy. *mBio* **2020**, *11*, e02123-20. [[CrossRef](#)]
18. Rivas, L.; Luque-Ortega, J.R.; Andreu, D. Amphibian Antimicrobial Peptides and Protozoa: Lessons from Parasites. *Biochim. Biophys. Acta (BBA)—Biomembr.* **2009**, *1788*, 1570–1581. [[CrossRef](#)]
19. Ahmed, A.; Siman-Tov, G.; Hall, G.; Bhalla, N.; Narayanan, A. Human Antimicrobial Peptides as Therapeutics for Viral Infections. *Viruses* **2019**, *11*, 704. [[CrossRef](#)]
20. Gaspar, D.; Veiga, A.S.; Castanho, M.A.R.B. From Antimicrobial to Anticancer Peptides. A Review. *Front. Microbiol.* **2013**, *4*, 294. [[CrossRef](#)]
21. Hollmann, A.; Martinez, M.; Maturana, P.; Semorile, L.C.; Maffia, P.C. Antimicrobial Peptides: Interaction With Model and Biological Membranes and Synergism With Chemical Antibiotics. *Front. Chem.* **2018**, *6*, 204. [[CrossRef](#)]
22. Marr, A.K.; Gooderham, W.J.; Hancock, R.E. Antibacterial Peptides for Therapeutic Use: Obstacles and Realistic Outlook. *Curr. Opin. Pharm.* **2006**, *6*, 468–472. [[CrossRef](#)]
23. Bruno, B.J.; Miller, G.D.; Lim, C.S. Basics and Recent Advances in Peptide and Protein Drug Delivery. *Ther. Deliv.* **2013**, *4*, 1443. [[CrossRef](#)]
24. Lei, J.; Sun, L.; Huang, S.; Zhu, C.; Li, P.; He, J.; Mackey, V.; Coy, D.H.; He, Q. The Antimicrobial Peptides and Their Potential Clinical Applications. *Am. J. Transl. Res.* **2019**, *11*, 3919.
25. Moretta, A.; Scieuzo, C.; Petrone, A.M.; Salvia, R.; Manniello, M.D.; Franco, A.; Lucchetti, D.; Vassallo, A.; Vogel, H.; Sgambato, A.; et al. Antimicrobial Peptides: A New Hope in Biomedical and Pharmaceutical Fields. *Front. Cell. Infect. Microbiol.* **2021**, *11*, 453. [[CrossRef](#)]
26. Li, S.; Wang, Y.; Xue, Z.; Jia, Y.; Li, R.; He, C.; Chen, H. The Structure-Mechanism Relationship and Mode of Actions of Antimicrobial Peptides: A Review. *Trends Food Sci. Technol.* **2021**, *109*, 103–115. [[CrossRef](#)]
27. Pirtskhalava, M.; Amstrong, A.A.; Grigolava, M.; Chubinidze, M.; Alimbarashvili, E.; Vishnepolsky, B.; Gabrielian, A.; Rosenthal, A.; Hurt, D.E.; Tartakovsky, M. DBAASP v3: Database of Antimicrobial/Cytotoxic Activity and Structure of Peptides as a Resource for Development of New Therapeutics. *Nucleic Acids Res.* **2021**, *49*, D288–D297. [[CrossRef](#)] [[PubMed](#)]
28. Kang, X.; Dong, F.; Shi, C.; Liu, S.; Sun, J.; Chen, J.; Li, H.; Xu, H.; Lao, X.; Zheng, H. DRAMP 2.0, an Updated Data Repository of Antimicrobial Peptides. *Sci. Data* **2019**, *6*, 148. [[CrossRef](#)] [[PubMed](#)]
29. Wagh, F.H.; Gopi, L.; Barai, R.S.; Ramteke, P.; Nizami, B.; Idicula-Thomas, S. CAMP: Collection of Sequences and Structures of Antimicrobial Peptides. *Nucleic Acids Res.* **2014**, *42*, D1154–D1158. [[CrossRef](#)] [[PubMed](#)]
30. 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*, 168–177. [[CrossRef](#)]
31. Meher, P.K.; Sahu, T.K.; Saini, V.; Rao, A.R. Predicting Antimicrobial Peptides with Improved Accuracy by Incorporating the Compositional, Physico-Chemical and Structural Features into Chou's General PseAAC. *Sci. Rep.* **2017**, *7*, 42362. [[CrossRef](#)]
32. Bhadra, P.; Yan, J.; Li, J.; Fong, S.; Siu, S.W.I. AmPEP: Sequence-Based Prediction of Antimicrobial Peptides Using Distribution Patterns of Amino Acid Properties and Random Forest. *Sci. Rep.* **2018**, *8*, 1697. [[CrossRef](#)] [[PubMed](#)]
33. Yan, J.; Bhadra, P.; Li, A.; Sethiya, P.; Qin, L.; Tai, H.K.; Wong, K.H.; Siu, S.W.I. Deep-AmPEP30: Improve Short Antimicrobial Peptides Prediction with Deep Learning. *Mol. Ther.—Nucleic Acids* **2020**, *20*, 882–894. [[CrossRef](#)] [[PubMed](#)]
34. Veltri, D.; Kamath, U.; Shehu, A. Deep Learning Improves Antimicrobial Peptide Recognition. *Bioinformatics* **2018**, *34*, 2740–2747. [[CrossRef](#)] [[PubMed](#)]
35. Lee, T.-H.; Hofferek, V.; Separovic, F.; Reid, G.E.; Aguilar, M.-I. The Role of Bacterial Lipid Diversity and Membrane Properties in Modulating Antimicrobial Peptide Activity and Drug Resistance. *Curr. Opin. Chem. Biol.* **2019**, *52*, 85–92. [[CrossRef](#)] [[PubMed](#)]
36. Speck-Planche, A.; Kleandrova, V.V.; Ruso, J.M.; DS Cordeiro, M.N. First Multitarget Chemo-Bioinformatic Model To Enable the Discovery of Antibacterial Peptides against Multiple Gram-Positive Pathogens. *J. Chem. Inf. Model.* **2016**, *56*, 588–598. [[CrossRef](#)] [[PubMed](#)]
37. Vishnepolsky, B.; Gabrielian, A.; Rosenthal, A.; Hurt, D.E.; Tartakovsky, M.; Managadze, G.; Grigolava, M.; Makhatadze, G.I.; Pirtskhalava, M. Predictive Model of Linear Antimicrobial Peptides Active against Gram-Negative Bacteria. *J. Chem. Inf. Model.* **2018**, *58*, 1141–1151. [[CrossRef](#)] [[PubMed](#)]
38. Wang, C.; Garlick, S.; Zloh, M. Deep Learning for Novel Antimicrobial Peptide Design. *Biomolecules* **2021**, *11*, 471. [[CrossRef](#)]
39. Ramesh, S.; Govender, T.; Kruger, H.G.; Torre, B.G.D.L.; Albericio, F. Short AntiMicrobial Peptides (SAMPs) as a Class of Extraordinary Promising Therapeutic Agents. *J. Pept. Sci.* **2016**, *22*, 438–451. [[CrossRef](#)]
40. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
41. Prediction of Protein Cellular Attributes Using Pseudo-Amino Acid Composition—Chou—2001—Proteins: Structure, Function, and Bioinformatics—Wiley Online Library. Available online: <https://onlinelibrary-wiley-com.ezproxy.leidenuniv.nl/doi/10.1002/prot.1035> (accessed on 5 October 2021).
42. Govindan, G.; Nair, A.S. Composition, Transition and Distribution (CTD)—A Dynamic Feature for Predictions Based on Hierarchical Structure of Cellular Sorting. In Proceedings of the 2011 Annual IEEE India Conference, Hyderabad, India, 16–18 December 2011; pp. 1–6.
43. Van Westen, G.J.; Swier, R.F.; Wegner, J.K.; IJzerman, A.P.; van Vlijmen, H.W.; Bender, A. Benchmarking of Protein Descriptor Sets in Proteochemometric Modeling (Part 1): Comparative Study of 13 Amino Acid Descriptor Sets. *J. Cheminform.* **2013**, *5*, 41. [[CrossRef](#)]

44. Van Westen, G.J.; Swier, R.F.; Cortes-Ciriano, I.; Wegner, J.K.; Overington, J.P.; IJzerman, A.P.; van Vlijmen, H.W.; Bender, A. Benchmarking of Protein Descriptor Sets in Proteochemometric Modeling (Part 2): Modeling Performance of 13 Amino Acid Descriptor Sets. *J. Cheminform.* **2013**, *5*, 42. [[CrossRef](#)]
45. Dorogush, A.V.; Ershov, V.; Gulin, A. CatBoost: Gradient Boosting with Categorical Features Support. *arXiv* **2018**, arXiv:1810.11363.
46. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: New York, NY, USA, 2017; Volume 30.
47. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 14–18 August 2012; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794.
48. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely Randomized Trees. *Mach. Learn.* **2006**, *63*, 3–42. [[CrossRef](#)]
49. Wei, G.; Zhao, J.; Feng, Y.; He, A.; Yu, J. A Novel Hybrid Feature Selection Method Based on Dynamic Feature Importance. *Appl. Soft Comput.* **2020**, *93*, 106337. [[CrossRef](#)]
50. Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *arXiv* **2017**, arXiv:1705.07874.
51. Brown, G.D.; Denning, D.W.; Gow, N.A.R.; Levitz, S.M.; Netea, M.G.; White, T.C. Hidden Killers: Human Fungal Infections. *Sci. Transl. Med.* **2012**, *4*, 165rv13. [[CrossRef](#)]
52. Wiederhold, N.P. Antifungal Resistance: Current Trends and Future Strategies to Combat. *Infect. Drug Resist.* **2017**, *10*, 249. [[CrossRef](#)]
53. Fernández de Ullivarri, M.; Arbulu, S.; Garcia-Gutierrez, E.; Cotter, P.D. Antifungal Peptides as Therapeutic Agents. *Front. Cell. Infect. Microbiol.* **2020**, *10*, 105. [[CrossRef](#)]
54. Joseph, S.; Karnik, S.; Nilawe, P.; Jayaraman, V.K.; Idicula-Thomas, S. ClassAMP: A Prediction Tool for Classification of Antimicrobial Peptides. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2012**, *9*, 1535–1538. [[CrossRef](#)]
55. Agrawal, P.; Bhalla, S.; Chaudhary, K.; Kumar, R.; Sharma, M.; Raghava, G.P.S. In Silico Approach for Prediction of Antifungal Peptides. *Front. Microbiol.* **2018**, *9*, 323. [[CrossRef](#)]
56. Lee, H.-T.; Lee, C.-C.; Yang, J.-R.; Lai, J.Z.C.; Chang, K.Y. A Large-Scale Structural Classification of Antimicrobial Peptides. *BioMed Res. Int.* **2015**, *2015*, 475062. [[CrossRef](#)]
57. Di Luca, M.; Maccari, G.; Maisetta, G.; Batoni, G. BaAMPs: The Database of Biofilm-Active Antimicrobial Peptides. *Biofouling* **2015**, *31*, 193–199. [[CrossRef](#)] [[PubMed](#)]
58. Ye, G.; Wu, H.; Huang, J.; Wang, W.; Ge, K.; Li, G.; Zhong, J.; Huang, Q. LAMP2: A Major Update of the Database Linking Antimicrobial Peptides. *Database* **2020**, *2020*, baaa061. [[CrossRef](#)] [[PubMed](#)]
59. Piotto, S.P.; Sessa, L.; Concilio, S.; Iannelli, P. YADAMP: Yet Another Database of Antimicrobial Peptides. *Int. J. Antimicrob. Agents* **2012**, *39*, 346–351. [[CrossRef](#)]
60. De Breij, A.; Riool, M.; Cordfunke, R.A.; Malanovic, N.; de Boer, L.; Koning, R.I.; Ravensbergen, E.; Franken, M.; van der Heijde, T.; Boekema, B.K.; et al. The Antimicrobial Peptide SAAP-148 Combats Drug-Resistant Bacteria and Biofilms. *Sci. Transl. Med.* **2018**, *10*, eaan4044. [[CrossRef](#)] [[PubMed](#)]
61. Huang, Y.; Niu, B.; Gao, Y.; Fu, L.; Li, W. CD-HIT Suite: A Web Server for Clustering and Comparing Biological Sequences. *Bioinformatics* **2010**, *26*, 680–682. [[CrossRef](#)] [[PubMed](#)]

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