Application of Machine Learning models as a strategy to screen peptides with antibiofilm capacity

Guilherme S.Lobo, Anália Lourenço, and Maria Olívia Pereira

Universidade do Minho, Campus de Gualtar, 4710-057 Braga, Portugal

Abstract. Biofilm resistance has increased, posing a serious threat to public health, making it imperative to develop effective control measures for biofilm-associated infections. In this study, we aim to explore Machine Learning models for predicting peptides with anti-Biofilm capacity. Two experiments were conducted: one focused on peptides associated with the human immune system and the other more general. For Experiment 1, the Support Vector Machine model ($C = 0.1$, $\gamma = 0.1$, $kernel = Linear$) and for Experiment 2, the Logistic Regression model ($C = 1$, $max_iter = 10$, $Penalty = l1$, $solver = liblinear$) achieved the best performance, demonstrating high accuracy and precision (both at 1.000). This work showcases an in silico approach to predict the antibiofilm capacity of peptides and opens up new avenues for studying the repurposing of peptides used in the treatment of other pathologies.

Keywords: Biofilm · Resistance · Immune System · Machine Learning.

1 Introduction

1.1 Context and motivation

The discovery of antibiotics played a very important role in medicine, which led the scientific community to call this period the "golden age" [1]. However, in recent decades, with the rapid global emergence of antimicrobial resistance, we are now facing one of the greatest threats to medicine [2]. Furthermore, the common idea of single-species bacteria is now outdated, as microorganisms (MO) mostly reside as organized multicellular aggregates called Biofilms [3].

The big generation of biological data associated with the problem mentioned above and the development of effective computational strategies, offer a unique opportunity [4], [5]. The principal strategies used to combat this type of resistance are i) the discovery of peptides and/or chemical molecules capable of damaging Biofilms and ii) the maximum prevention of the appearance of this problem with the use of an optimized therapeutic plan [6]. However, in recent years an innovative strategy has been emerging where it is intended to predict the immune response to a given multiresistant Biofilm. In short, this last strategy would use the ability of the immune system to fight these resistant MO embedded with Biofilms.

Contextualizing the severity of this problem, there are estimates that predict around 10 million deaths by 2050 [2], with immunocompromised patients being the most common targets, but noting that the entire population is at risk.

1.2 Objectives

The main objective is to explore the immune system in the fight against Biofilms and, for this purpose, it is necessary, through computational tools such as Machine Learning (ML), to screen several peptides and their properties. For this, supervised learning algorithms will be used in order to predict the antibiofilm potential of the peptides. In short, the aim is to find an ML model capable of predicting whether or not a given peptide will have antibiofilm capacity.

2 Background

2.1 Biofilm

Biofilms are surface-attached well organized microbial communities able to adhere to biotic or abiotic surfaces. The complexity of the mechanisms by which MO existing in a Biofilm avoids antimicrobials has led to the creation of various terminologies and underlying concepts, such as resistance, heteroresistance, tolerance, and persistence [7].

Resistance can be classified as intrinsic, acquired, or adaptive [8]. Intrinsic resistance is the innate ability of MO to resist a specific antimicrobial encoded in their genome that provides protection [9]. Acquired resistance is when an originally sensitive MO becomes resistant with the acquisition of new genetic material [9], [10]. In adaptive resistance, MO in response to an antimicrobial or environmental factor (pH, temperature, etc) have the ability to change their genetic expression and become increasingly resistant [8], [10], [11].

Heteroresistance, quite common in Biofilms, is defined by the presence of one or more microbial subpopulations with different "levels" of resistance and, as such, with different susceptibility to antibiotics [12]. Tolerance, in turn, refers to the ability of each MO, present in the Biofilm to survive exposure to increased concentrations of antibiotics and, is therefore classified as temporary [8].

In biofilms, in addition to the existence of different microbial populations, there is also a set of distinct microhabitats caused by the establishment of an unequal gradient of oxygen and nutrients [8], [13]. These aspects are directly related to the appearance of persistent cells that reduce their metabolic activity almost in relation to a state of hibernation or "false-mortem" [14]. This type of cells are one of the main reasons why antimicrobials are ineffective, as these persistent cells can regrow after antimicrobial action and maintain the infection [7], [8], [15].

Biofilm cells are protected by a self-produced polymeric matrix that provides their mechanical stability and is the first barrier to the entry and diffusion of

antimicrobial substances [7], [8]. However, impeding antimicrobial penetration does not extensively explain the observed resistance phenomena[4], [16].

Quorum Sensing (QS) is an extremely important cell-cell communication system, which allows to coordinate the mechanisms by which MO within a Biofilm regulate their activities. Thus, QS serves as an intra- and interspecies communication portal, allowing the establishment of intimate relationships of competition or cooperation [17]. QS plays a key role in Biofilm antimicrobial resistance, although the mechanisms behind it are still not fully understood [7], [8].

2.2 Bioinformatic Approach

Faced with the problem presented, it is necessary to use in silico approaches capable of systematically screening known peptides, seeking the desired antimicrobial capabilities. [4], [5]. In this sense, a possible approach would be ML, which is a field of computer science contained in Artificial Intelligence [18]. ML predictions are useful to design lab experiments, namely prioritizing the investigation of certain peptides. Lab results will confirm or deny these predictions and such knowledge can be later used to update the predictive model [18], [19].

ML models are meaningful representations of information extracted from raw data using algorithmic mathematical transformations. The model can then be applied to new related data to infer unknown information or gain a better understanding of the data distribution [18], [20]. The main difference between ML and DL is that DL uses a layered model structure, where successive layers represent more and more meaningful representations, and therefore tries to perform automatic learning of the representation without requiring much effort in feature extraction. The choice between traditional ML and DL algorithms depends on the amount and complexity of the data and available computational resources [19]. Taking into account that the main strategies employed in antimicrobial resistance are based on the application of ML, we will summarize its workflow in 3 main steps: i) data processing ii) model selection, iii) model evaluation.

Data processing: The data is composed of samples (rows), and their extracted features (columns) [18], [21]. Each feature can be categorical, ordinal, or numerical. The chosen dataset may, however, include undesirable attributes such as missing values and outliers, which must be addressed before creating models. The quality of the features chosen for the creation of the model is fundamental to obtain a good performance. This includes selecting relevant features and removing irrelevant ones, a fundamental process called feature selection [22]. Its importance can be determined by observing a high correlation with the dependent variable, indicating important features that can predict the desired outcome, and those that do not [21].

Model selection: After processing the data, it´s necessary to select a model based on a certain algorithm. In this way, ML can be classified into two categories: unsupervised ML and supervised ML [18]. Unsupervised ML use unlabeled data, where the result variable is not specified, to discover data structures that have some common characteristics that can be manually labeled with an

appropriate label. In this particular strategy, the models essentially employ Dimensionality reduction and Clustering. Supervised ML use labeled data, where the outcome variable is specified, to predict outcomes from future new data. In this category, the most common models are Probabilistic models, Kernel-based models, K-nearest neighbors, Decision trees, and Ensemble models [18], [23], [24].

Model evaluation: To understand how well the created ML models can infer information from the provided data, the model must be evaluated using appropriate methods [18]. It is critical to verify that the quality of model performance depends only on the training data provided, as the model may fail to generalize unseen data (where poor performance is observed on a given test dataset), a problem known as overfitting. However, the algorithms used may not be ideal for extracting information from the raw data provided, or the created model may not have been well trained (due to scarcity of data, short training time, and inadequate hyperparameters, among other factors), leading to the underfit [18], [23]. These two types of issues are the ones to consider when evaluating your chosen ML model.

2.3 Related work

Using the existing literature, it was possible to verify several articles that have already used ML models. A summary of the main strategies analyzed is described in Table 1, and the different associated colors represent different strategies to combat multiresistant Biofilms. The green color is associated with the prediction of peptides and/or chemical molecules with anti-Biofilm capacity, the orange color is associated with a more personalized therapy with the intention of delaying the emergence of antimicrobial resistance as much as possible, and the yellow color where the articles explore the host immune system as the main weapon in the fight against these multiresistant pathogens.

Table 1: Bibliographic review of the main strategies that employ computational tools, such as ML and DL, in the combat against multiresistant Biofilms.

Authors	Year	Prediction Model	Resume	Ref.
Gupta et al.	2016	SVM, RF	Biofilm inhibiting peptides.	[25]
Srivastava et al.	2020	RF, SVM, KNN	Biofilm inhibiting molecules.	[26]
Sharma et al.	2016	SVM	Predict and design anti-biofilm peptides.	[27]
Bose et al.	2022	SVM, RF	Predict the antibiofilm potencial of peptides.	[28]
Tacconelli et al.	2020	RF	Determining the best antibiotic to avoid resistance.	[29]
Shaban et al.	2022	RF, Logistic Regression, Decision Tree	Predict the ability of an antibiotic to prevent the emergence of resistance.	[24]
Charoenkuan et al.	2022	SVM, RF	Identify defensins, essential molecules for the immune system.	[30]
Zhang et al.	2017	RF, SVM, ANN	Associating a different bacterial infection with a specific immune fingerprint.	[31]
Pavlovic et al.	2021	immuneML	open-source ecosystem for Adpatative Imunne Receptores Repertoires.	[32]

The focus of this project is the exploration of the immune system responses to tackle biofilm-associated infections. in the fight against Biofilms. Hereupon, and taking into account all the literature referred to in Table 1, the article by Bose et.al [28] and Pavlovic et al. [32] are the ones we want to highlight.

When a pathogenic agent, such as MO, enters the body, the immune system is promptly activated to fight this infectious agent [33]. In short, this type of strategy explores the link between antigens and adaptive immune receptor repertoires (AIRR). With the aim of accelerating this process, ML and DL strategies make it possible to predict patterns in antigen binding to AIRR and, in this way, to identify a possible stage of infection as well as to know which antigens are most likely to bind to this type of receptors at that particular stage [32], [33].

By using platforms such as immuneML [32], the goal is to identify these binding patterns and, from there, predict which antigens are most likely to bind to AIRR in a given stage of infection. Thus, by having a certain range of antigens as a result, it is possible to think of a possible stimulation of the immune system based on this set of molecules.

However, there are many challenges, namely the fact that the immune system response is specific to each individual, the fact that Biofilms can even be made up of the same bacteria but in different proportions, leading to each case is particular, making it difficult to predict patterns, requiring a wide variety of complete datasets [34]. Furthermore, it is known that exaggerated stimulation of the immune system is toxic to other cells, that is, there would have to be a subsequent analysis of cell viability to determine a safe dose of the selected antigens [35].

A possible solution would be to add to this immunotherapy a potential peptide that somehow weakens the Biofilm so that it no longer needs an exacerbated immune response. Therefore, the work by Bose et al. [28] is a reference where several ML models were explored in order to choose the one with the greatest predictive capacity. This ability would be that from a set of features associated with the sequence of a given peptide, it is possible to identify the presence of anti-biofilm capacity.

Taking all the literature consulted into account, it is possible to conclude that combating resistant Biofilms is not an easy task, however, the appearance of computational tools, such as ML, and a large number of biological data increases the expectation of finding possible solutions. The solution that we particularly want to explore is based on personalized medicine with the exploration, in the first stage, ML models capable of predicting the antibiofilm capacity of a certain set of peptides and, in a future phase, synergizing this stage with the exploration of the immuneML platform.

3 Methodology

For this work we had two different approaches: i) “Exp 1”- antibiofilm peptides associated only with the human immune system; ii) “Exp 2” - a more general set of antibiofilm peptides. The general workflow is schematized in Fig. 1.

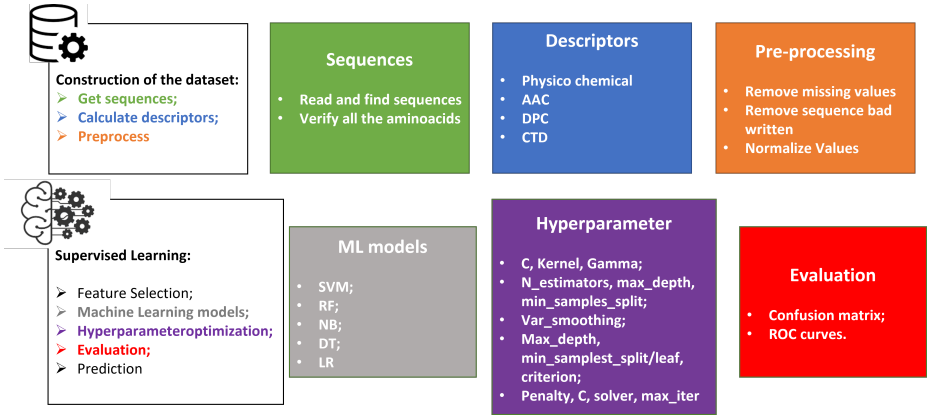

Fig. 1: A schematic representation of the workflow used for all the approaches.

3.1 Data loading and preparation

3.1.1 Negative Class dataset: For all approaches performed and mentioned above, the dataset belonging to the negative class (without antibiofilm capacity) was the same. The dataset in question was selected with the help of different databases such as UniProt [36], Quorum Sensing Peptide Prediction Server (QSPProd) [37], and NCBI [38]. In this way, the peptides that were associated with some essential process in the formation, preservation, or development of the biofilm were taken into account, since, as they are essential for the MO under study, it did not make sense to associate these peptides with an antibiofilm capacity.

3.1.2 Positive Class dataset: For **Experiment 1**, the objective would be to have in this dataset of positive class (with antibiofilm capacity) peptides that somehow had already been associated with the human immune system. That said, we chose to collect all peptides linked to antibiofilm capacity whose origin is Homo sapiens from the CAMPr4 database. Regarding **Experiment 2**, which is a more generic exploration of antimicrobial peptides, all those cataloged as antibacterial in the DRAMP database were collected. These positive class datasets were later added, for each experiment in particular, to the negative class dataset, the final datasets are taken as references for the rest of this work.

In addition to the positive class datasets already mentioned, a new dataset was also explored that gathered synthetic and stapled antimicrobial peptides taken from the DRAMP database. This dataset was also connected with the negative class dataset in order to serve as a "verification dataset" to allow observing whether the ML models that were subsequently chosen really have a high predictive capacity even in a dataset where they were not trained.

3.1.3 Case study-peptide repurposing : The DRAMP database allows access to numerous peptides that are used as antivirals, anticancer, antiparasitics, and insecticides. Thus, what was intended was to gather all the desired characteristics of these peptides in a dataset and perform a repurposing analysis. That

is, to explore peptides that currently have another purpose and could also have antibiofilm capacity.

3.1.4 Feature Extraction: Then, it was necessary to transform the peptide sequence into a mathematical vector. For this, the software packages protParam ("ProteinAnalysis" module) and Biopython were used [39]. The features obtained were 571, of which 4 were general features such as sequence size, molecular weight, aromaticity, and isoelectric point, 20 were associated with the composition of each amino acid, 400 with the dipeptide composition, 147 represented different physicochemical properties such as "hydrophobicity", "normalized van der Waals volume", "polarity", "polarizability", "charge", "secondary structure", "solvent accessibility", among others and.

3.1.5 Pre-processing: This step was characterized by "cleaning" the dataset, that is, removing peptides whose features are associated with "NaN" values, incorrect characters in the sequences, duplicate peptides, non-relevant columns (with only zero and repeated values), and in the data results after this filtering, apply a normalization process.

3.2 Supervised Learning

To carry out this work, in the first place we decided to reserve 20% of the data for the final test and with the remaining dataset (80% of the data) a 10-fold cross-validation process was taken into account, in which a part of the dataset data, called the validation set, was used for testing and the other nine were used for training. This process was iterated more than ten times, using each of the ten parts in turn as the validation set. Because our dataset is unbalanced, we use stratified sampling to ensure that each fold receives an equal percentage of positive and negative peptides during cross-validation.

3.2.1 Baseline ML models: A prediction model was developed using various ML algorithms, including Support Vector Machines (SVM), Random Forest (RF), Naive Bayes (NB), Decision Tree (DT), and Logistic Regression (LR) classifiers. Our aim was to select the algorithm that provides the best predictive performance for antibiofilm activity in out-of-sample data. We used the "Scikit-learn" [40] package to, at first, implement the referred algorithms and then train and test.

3.2.2 Hyperparameters optimization: Next, we decided to carry out a grid search, the 'Grid-SearchCV' function from the scikit-learn package to identify the best parameters for each model. The grid search generated numerous models due to the high number of possible combinations. Ultimately, the grid search enabled us to identify the model with the optimal parameters, thereby achieving the highest possible scores. It should be noted that we've specified that the cross-validation scores should be based on the f1-score. The hyperparameters taken into account for this optimization are referred in Fig. 1 in the same order as the mentioned models.

3.2.3 Performance Evaluation: To evaluate the predictions made by each particular model, extensive metrics are available. Metrics include accuracy, pre-

cision, recall, F1-score and area under the receiver operating characteristic curve (ROC-AUC). Most of these metrics are calculated based on the confusion matrix, taking into account the values TP = True Positive, TN = True Negative, FP = False Positive, and FN = False Negative.

4 Results and Discussion

The two distinct experiences used throughout this work had as their main purposes the exploration of peptides directly associated with the immune system and a more generalized exploration where a greater range of natural peptides were taken into account, respectively. Having said that, we will then demonstrate and highlight the results that we find most interesting for each particular experiment.

In **Experience 1**, the dataset used had the already described 571 features, being composed of 2838 peptides, of which 2420 were associated with a negative class and 418 with a positive class, that is, without and with antibiofilm capacity, respectively. Regarding **Experience 2**, the dataset only differentiated in the peptides associated with the positive class, and the dataset used had 6490 peptides and the same number of features.

For both datasets grid search was conducted for hyperparameters optimization, considering SVM, RF, NB, DT, and LR models, using a 5-fold cross-validation on the training set. The best configurations for each experience were then used to build ML models evaluated on the test set. Table 2 and Table 3 show the best hyperparameters obtained for each of the ML models, as well as the values of the respective metrics after being evaluated on the test set.

Table 2: Metrics resulting from testing the ML models for the Experiment 1 dataset with the respectively optimized hyperparameters.

ML Models	Hyperparameters	F1-Score	Accuracy	Precision
SVM	• C = 0.1 • gamma = 0.1 • Kernel = Linear	1.000	1.000	1.000
Random Forest	• max_depth = None • min_samples_split = 5 • n_estimators = 100	0.956	0.988	1.000
Naive Bayes	• varsmoothing = 1e-12	0.573	0.829	0.454
Decision Tree	• Criterion = gini • max_depth = None • min_samples_leaf = 4 • min_samples_split = 2	0.894	0.970	0.935
Logistic Regressor	• C = 1 • max_iter = 10 • Penalty = l1 • solver = liblinear	0.988	0.996	0.988

As be seen from Table 2, there are 3 models that have very similar and high metrics. However, as the objective at this stage would be to choose the best

model against a set of optimized hyperparameters, it was decided to choose the one that really stood out and, therefore, for **Experience 1**, the SVM model with *C* of 0.1, *gamma* of 0.1, and linear *kernel* with values of 1.000 for all evaluated metrics was chosen. SVM is one of the most commonly used classifiers for peptide prediction [41]. SVM works particularly well for binary classification problems. The model works by separating samples into different classes using a hyperplane, which can be expressed in a high dimensional space through kernel transformations. Is a robust model that can be used for both classification and regression.

Table 3: Metrics resulting from testing the ML models for the Experiment 2 dataset with the respectively optimized hyperparameters.

ML Models	**Hyperparameters**	**F1-Score**	**Accuracy**	**Precision**
SVM	• C = 0.1 • gamma = 0.1 • Kernel = Linear	0.998	0.997	0.995
Random Forest	• max_depth = None • min_samples_split = 2 • n_estimators = 500	0.990	0.988	0.984
Naive Bayes	• varsmoothing = 1e-12	0.879	0.855	0.920
Decision Tree	• Criterion = entropy • max_depth = None • min_samples_leaf = 1 • min_samples_split = 2	0.980	0.975	0.985
Logistic Regressor	• C = 1 • max_iter = 10 • Penalty = l1 • solver = liblinear	1.000	1.000	1.000

With regard to choosing the best model described in Table 3, it is possible to observe that for **Experience 2** there were four ML models with very high-performance evaluation metrics. Thus, it was decided to choose the LR with *C* of 10, *max_iterations* of 200, *Penalty* is l1, and *solver* is liblinear, with values of 1,000 for all evaluated metrics. LR is widely used for the prediction of [41] peptides, being suitable for binary classification problems. This model depicts the relationship between independent variables and the probability of belonging to a given class using a logistic function. In the context of peptide prediction, the independent variables can be the characteristics of the peptides, such as the features used.

As the values of the evaluated metrics were quite high for both chosen models, we decided to carry out a deeper analysis of the respective confusion matrix, represented in Fig. 2A and Fig. 2B, respectively.

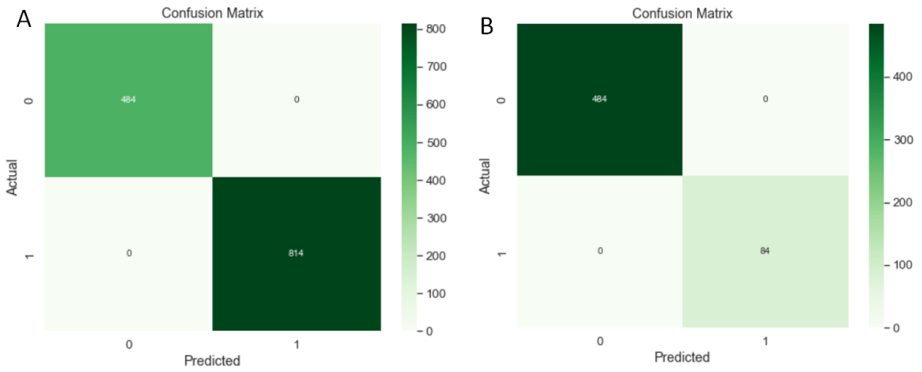

Fig. 2: Predictive capability, for de dataset test, illustrated in the confusion matrix. A) Experience 1- SVM model; B) Experiment 2- LR model.

The analysis of these figures is justified by the fact that, as the dataset does not have an equal distribution of the analyzed classes, we could be camouflaging the true performance of the models. However, by observing the confusion matrix, we quickly observe that the selected models only have TP and TN, allowing us to eliminate our concern and confirm the good performance obtained.

As the results obtained so far, in terms of predictive capacity, was quite good for both experiments, we decided to check how these models would behave in a new dataset never trained before. For this, we used synthetic and stamped antibiofilm peptides. Therefore, this new dataset with 2420 peptides of negative class and 1552 of positive class and with the same number of features was used as a test dataset to evaluate the selected models in the experiments under study. The results obtained for SVM and LR models are represented by the confusion matrix in Fig. 3 A and B, respectively.

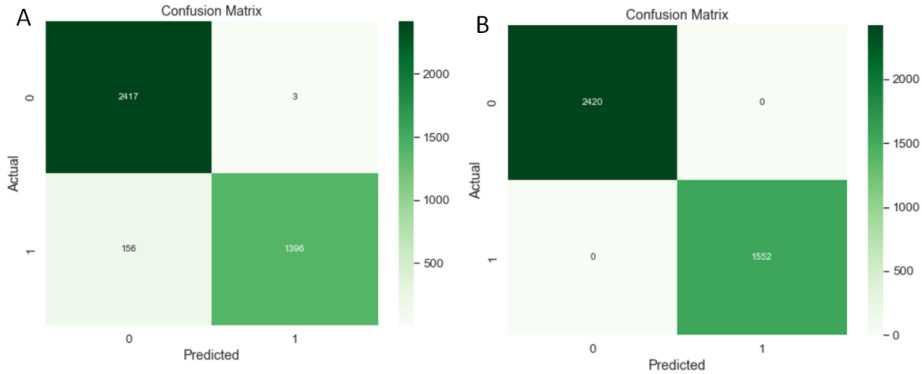

Fig. 3: Resulting confusion matrix for previously selected models: A) SVM B)LR
.

From the brief interpretation of the confusion matrix, it can be immediately verified that the predictive capacity of both models remains quite high, with the LR model (chosen from Exp-2) obtaining better metrics with maximum values of 1.000. However, it would be wrong to conclude that, as a result of the data obtained, we should only work with LR, because it depends on the objective of what we are looking for. That is, if perhaps the interest is to explore peptides that, in a way, have already been associated with the Homo sapiens organism and, more specifically, with its immune system, it is indeed more appropriate to use the SVM model. Thus, in our opinion, and taking into account the main objective of relating the fight against biofilms with the immune system, we would choose the model obtained in Exp 1-SVM. This model, as it is associated with the human immune system, provides greater guarantees that a possible peptide may, at a much more advanced stage of the work, be considered as a possible adjuvant therapeutic option.

Another approach explored was taken towards the question: **Can repurposing be possible?**. In this way, a new dataset was created, which contained antiviral, anti-cancer, antiparasitic, and insecticidal peptides. In this new dataset, the class of these peptides was not known. So the objective was to determine whether the peptides that are already associated with other pathologies could also be directed to the problem addressed in this work. That said, we used the SVM model selected in the first experiment to predict which class the peptides under analysis would belong to, with the results shown in Fig. 4.

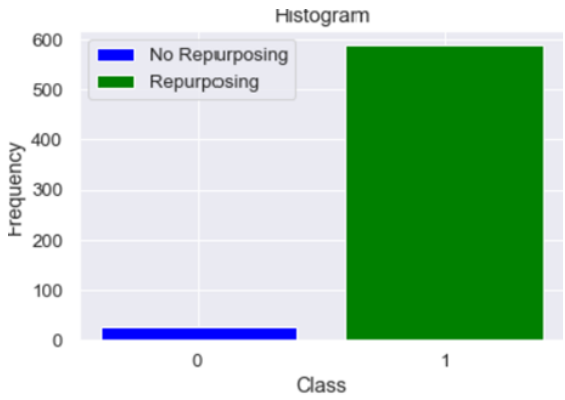

Fig. 4: Frequency of peptides that could be associated with a negative class and a positive class.

By interpreting the histogram, it is quickly observable that most of the peptides in this dataset, due to the same features analyzed, could be considered for use against resistant MO, leading to an affirmative answer to the previously posed hypothesis.

Finally, when we compare the results of our global approach with the different approaches referenced in green in Tab. 1, it is evident that our models compared

to the dataset in question also have a high predictive capacity, being therefore considered adequate in the identification of peptides with antibiofilm cacity.

5 Conclusion and future work

5.1 Conclusion

In this work, we developed a ML pipeline for the classification of antibiofilm peptides, followed by the analysis of possible repurposing of peptides applied in other pathologies.

We used different datasets with the same set of 571 features, varying only in the selected peptides. This approach allowed us to obtain the best ML model for both experiments. For Experiment 1, which explores peptides related to the human immune system, the most suitable ML model is the SVM (C = 0.1, $gamma$ = 0.1; $kernel$ = Linear), and for Experiment 2, with a more general approach in terms of peptide selection, it is LR(C =1, max_iter =10; $Penalty$ =l1; $solver$ =liblinear).

Additionally, with the help of the SVM model, we found that there is a large set of peptides used in other pathologies that could also be taken into account as a possible solution to tackle biofilm-associated infections encompassing resistant MO.

5.2 Future work

This work provides valuable information that can serve as a basis for future research. However, the next step would be to explore the models obtained and, with the aid of a regression strategy, explore possible values of the Biofilm Minimum Inhibitory Concentration (MBIC), with the aim of actually being able to prove in practical terms the possible predictions obtained in silico for a given set of MO.

Furthermore, it would be extremely interesting to explore the immuneML platform with the aim of selecting a set of antigens that, given a given state of infection, would contribute to an increase in the immune response.

Thus, in an ideal situation, it would be possible to present a set of antigens and peptides, obtained with the aid of ML models given a certain set of characteristics, which would be related through a future adjuvant process where the main combat weapon is the immune system.

6 Code availability

The pipelines implemented in this work were entirely developed in Python, and are freely available at https://github.com/GuilhermeLoboSousa/antibiofilm_projecto .

References

[1] E. Martens and A. L. Demain, "The antibiotic resistance crisis, with a focus on the United States," *Journal of Antibiotics*, vol. 70, no. 5, pp. 520–526, 2017, ISSN: 18811469. DOI: 10.1038/ja.2017.30.

[2] P. Shankar, "Book review: Tackling drug-resistant infections globally," *Archives of Pharmacy Practice*, vol. 7, no. 3, p. 110, 2016, ISSN: 2045-080X. DOI: 10.4103/2045-080x.186181.

[3] H. C. Flemming and S. Wuertz, "Bacteria and archaea on Earth and their abundance in biofilms," *Nature Reviews Microbiology*, vol. 17, no. 4, pp. 247–260, 2019, ISSN: 17401534. DOI: 10.1038/s41579-019-0158-9. [Online]. Available: http://dx.doi.org/10.1038/s41579-019-0158-9.

[4] P. Jorge, A. Lourenço, and M. O. Pereira, "New trends in peptide-based anti-biofilm strategies: a review of recent achievements and bioinformatic approaches," *Biofouling*, vol. 28, no. 10, pp. 1033–1061, 2012, ISSN: 08927014. DOI: 10.1080/08927014.2012.728210.

[5] N. F. Azevedo, S. P. Lopes, C. W. Keevil, M. O. Pereira, and M. J. Vieira, "Time to "go large" on biofilm research: Advantages of an omics approach," *Biotechnology Letters*, vol. 31, no. 4, pp. 477–485, 2009, ISSN: 01415492. DOI: 10.1007/s10529-008-9901-4.

[6] B. Parrino, D. Schillaci, I. Carnevale, *et al.*, "Synthetic small molecules as anti-biofilm agents in the struggle against antibiotic resistance," *European Journal of Medicinal Chemistry*, vol. 161, pp. 154–178, 2019, ISSN: 17683254. DOI: 10.1016/j.ejmech.2018.10.036. [Online]. Available: https://doi.org/10.1016/j.ejmech.2018.10.036.

[7] S. P. Lopes, P. Jorge, A. M. Sousa, and M. O. Pereira, "Discerning the role of polymicrobial biofilms in the ascent, prevalence, and extent of heteroresistance in clinical practice," *Critical Reviews in Microbiology*, vol. 47, no. 2, pp. 162–191, 2021, ISSN: 15497828. DOI: 10.1080/1040841X.2020.1863329. [Online]. Available: https://doi.org/10.1080/1040841X.2020.1863329.

[8] P. Jorge, A. P. Magalhães, T. Grainha, *et al.*, "Antimicrobial resistance three ways: Healthcare crisis, major concepts and the relevance of biofilms," *FEMS Microbiology Ecology*, vol. 95, no. 8, pp. 1–17, 2019, ISSN: 15746941. DOI: 10.1093/femsec/fiz115.

[9] J. M. Blair, M. A. Webber, A. J. Baylay, D. O. Ogbolu, and L. J. Piddock, "Molecular mechanisms of antibiotic resistance," *Nature Reviews Microbiology*, vol. 13, no. 1, pp. 42–51, 2015, ISSN: 17401534. DOI: 10.1038/nrmicro3380. [Online]. Available: http://dx.doi.org/10.1038/nrmicro3380.

[10] M. Arzanlou, W. C. Chai, and H. Venter, "Intrinsic, adaptive and acquired antimicrobial resistance in Gram-negative bacteria," *Essays in Biochemistry*, vol. 61, no. 1, pp. 49–59, 2017, ISSN: 00711365. DOI: 10.1042/EBC20160063.

[11] Z. Pang, R. Raudonis, B. R. Glick, T. J. Lin, and Z. Cheng, "Antibiotic resistance in Pseudomonas aeruginosa: mechanisms and alternative ther-

apeutic strategies," *Biotechnology Advances*, vol. 37, no. 1, pp. 177–192, 2019, ISSN: 07349750. DOI: 10.1016/j.biotechadv.2018.11.013. [Online]. Available: https://doi.org/10.1016/j.biotechadv.2018.11.013.

[12] S. N. Saravolatz, H. Martin, J. Pawlak, L. B. Johnson, and L. D. Saravolatz, "Ceftaroline-heteroresistant staphylococcus aureus," *Antimicrobial Agents and Chemotherapy*, vol. 58, no. 6, pp. 3133–3136, 2014, ISSN: 10986596. DOI: 10.1128/AAC.02685-13.

[13] L. Dewachter, M. Fauvart, and J. Michiels, "Bacterial Heterogeneity and Antibiotic Survival: Understanding and Combatting Persistence and Heteroresistance," *Molecular Cell*, vol. 76, no. 2, pp. 255–267, 2019, ISSN: 10974164. DOI: 10.1016/j.molcel.2019.09.028. [Online]. Available: https://doi.org/10.1016/j.molcel.2019.09.028.

[14] C. W. Hall and T. F. Mah, "Molecular mechanisms of biofilm-based antibiotic resistance and tolerance in pathogenic bacteria," *FEMS Microbiology Reviews*, vol. 41, no. 3, pp. 276–301, 2017, ISSN: 15746976. DOI: 10.1093/femsre/fux010.

[15] B. P. Conlon, "Staphylococcus aureus chronic and relapsing infections: Evidence of a role for persister cells: An investigation of persister cells, their formation and their role in S. aureus disease," *BioEssays*, vol. 36, no. 10, pp. 991–996, 2014, ISSN: 15211878. DOI: 10.1002/bies.201400080.

[16] T. Grainha, A. P. Magalhães, L. D. Melo, and M. O. Pereira, "Pitfalls associated with discriminating mixed-species biofilms by flow cytometry," *Antibiotics*, vol. 9, no. 11, pp. 1–17, 2020, ISSN: 20796382. DOI: 10.3390/antibiotics9110741.

[17] C. Grandclément, M. Tannières, S. Moréra, Y. Dessaux, and D. Faure, "Quorum quenching: Role in nature and applied developments," *FEMS Microbiology Reviews*, vol. 40, no. 1, pp. 86–116, 2015, ISSN: 15746976. DOI: 10.1093/femsre/fuv038.

[18] S. Badillo, B. Banfai, F. Birzele, *et al.*, "An Introduction to Machine Learning," *Clinical Pharmacology and Therapeutics*, vol. 107, no. 4, pp. 871–885, 2020, ISSN: 15326535. DOI: 10.1002/cpt.1796. arXiv: 2209.00939.

[19] A. Muller and S. Guido, "Introduction to machine learning with python—a guide for data scientists. o'reillymedia," *Inc., Sebastopol, CA*, 2016.

[20] J. Alzubi, A. Nayyar, and A. Kumar, "Machine Learning from Theory to Algorithms: An Overview," *Journal of Physics: Conference Series*, vol. 1142, no. 1, 2018, ISSN: 17426596. DOI: 10.1088/1742-6596/1142/1/012012.

[21] C. V. Gonzalez Zelaya, "Towards explaining the effects of data preprocessing on machine learning," *Proceedings - International Conference on Data Engineering*, vol. 2019-April, pp. 2086–2090, 2019, ISSN: 10844627. DOI: 10.1109/ICDE.2019.00245.

[22] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing*, vol. 300, pp. 70–79, 2018, ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2017.11.

077. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231218302911.

[23] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, vol. 415, pp. 295–316, 2020, ISSN: 18728286. DOI: 10.1016/j.neucom.2020.07.061. arXiv: 2007.15745. [Online]. Available: https://doi.org/10.1016/j.neucom.2020.07.061.

[24] T. F. Shaban and M. Y. Alkawareek, "Prediction of qualitative antibiofilm activity of antibiotics using supervised machine learning techniques," *Computers in Biology and Medicine*, vol. 140, p. 105 065, 2022, ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2021.105065. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482521008593.

[25] S. Gupta, A. K. Sharma, S. K. Jaiswal, and V. K. Sharma, "Prediction of biofilm inhibiting peptides: An In silico Approach," *Frontiers in Microbiology*, vol. 7, no. JUN, pp. 1–11, 2016, ISSN: 1664302X. DOI: 10.3389/fmicb.2016.00949.

[26] G. N. Srivastava, A. S. Malwe, A. K. Sharma, V. Shastri, K. Hibare, and V. K. Sharma, "Molib: A machine learning based classification tool for the prediction of biofilm inhibitory molecules," *Genomics*, vol. 112, no. 4, pp. 2823–2832, 2020, ISSN: 10898646. DOI: 10.1016/j.ygeno.2020.03.020. [Online]. Available: https://doi.org/10.1016/j.ygeno.2020.03.020.

[27] A. Sharma, P. Gupta, R. Kumar, and A. Bhardwaj, "DPABBs: A Novel in silico Approach for Predicting and Designing Anti-biofilm Peptides," *Scientific Reports*, vol. 6, no. January, pp. 1–13, 2016, ISSN: 20452322. DOI: 10.1038/srep21839.

[28] B. Bose, T. Downey, A. K. Ramasubramanian, and D. C. Anastasiu, "Identification of Distinct Characteristics of Antibiofilm Peptides and Prospection of Diverse Sources for Efficacious Sequences," *Frontiers in Microbiology*, vol. 12, no. February, pp. 1–20, 2022, ISSN: 1664302X. DOI: 10.3389/fmicb.2021.783284.

[29] E. Tacconelli, A. Górska, G. De Angelis, *et al.*, "Estimating the association between antibiotic exposure and colonization with extended-spectrum β-lactamase-producing Gram-negative bacteria using machine learning methods: a multicentre, prospective cohort study," *Clinical Microbiology and Infection*, vol. 26, no. 1, pp. 87–94, 2020, ISSN: 14690691. DOI: 10.1016/j.cmi.2019.05.013.

[30] P. Charoenkwan, N. Schaduangrat, S. M. Hasan Mahmud, O. Thinnukool, and W. Shoombuatong, "Recent Development of Machine Learning-Based Methods for the Prediction of Defensin Family and Subfamily," *EXCLI Journal*, vol. 21, pp. 757–771, 2022, ISSN: 16112156. DOI: 10.17179/excli2022-4913.

[31] J. Zhang, I. M. Friberg, A. Kift-Morgan, *et al.*, "Machine-learning algorithms define pathogen-specific local immune fingerprints in peritoneal

dialysis patients with bacterial infections," *Kidney International*, vol. 92, no. 1, pp. 179–191, 2017, ISSN: 15231755. DOI: 10.1016/j.kint.2017.01.017.

[32] M. Pavlović, L. Scheffer, K. Motwani, *et al.*, "immuneML: an ecosystem for machine learning analysis of adaptive immune receptor repertoires," *Nature*, no. i, 2021. DOI: 10.1038/s42256-021-00413-z.

[33] C. Kanduri, M. Pavlović, L. Scheffer, *et al.*, "Profiling the baseline performance and limits of machine learning models for adaptive immune receptor repertoire classification," *GigaScience*, vol. 11, pp. 1–21, 2022, ISSN: 2047217X. DOI: 10.1093/gigascience/giac046.

[34] V. Greiff, G. Yaari, and L. G. Cowell, "Mining adaptive immune receptor repertoires for biological and clinical information using machine learning," *Current Opinion in Systems Biology*, vol. 24, pp. 109–119, 2020, ISSN: 24523100. DOI: 10.1016/j.coisb.2020.10.010. [Online]. Available: https://doi.org/10.1016/j.coisb.2020.10.010.

[35] A. Y. An, K. Y. G. Choi, A. S. Baghela, and R. E. Hancock, "An Overview of Biological and Computational Methods for Designing Mechanism-Informed Anti-biofilm Agents," *Frontiers in Microbiology*, vol. 12, no. April, pp. 1–24, 2021, ISSN: 1664302X. DOI: 10.3389/fmicb.2021.640787.

[36] A. Bateman, M. J. Martin, S. Orchard, *et al.*, "UniProt: the universal protein knowledgebase in 2021," *Nucleic Acids Research*, vol. 49, no. D1, pp. D480–D489, 2021, ISSN: 13624962. DOI: 10.1093/nar/gkaa1100.

[37] A. Rajput, K. T. Bhamare, A. Thakur, and M. Kumar, "Biofilm-i: A Platform for Predicting Biofilm Inhibitors Using Quantitative Structure—Relationship (QSAR) Based Regression Models to Curb Antibiotic Resistance," *Molecules*, vol. 27, no. 15, 2022, ISSN: 14203049. DOI: 10.3390/molecules27154861.

[38] R. Agarwala, T. Barrett, J. Beck, *et al.*, "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Research*, vol. 44, no. D1, pp. D7–D19, 2016, ISSN: 13624962. DOI: 10.1093/nar/gkv1290.

[39] P. J. Cock, T. Antao, J. T. Chang, *et al.*, "Biopython: Freely available Python tools for computational molecular biology and bioinformatics," *Bioinformatics*, vol. 25, no. 11, pp. 1422–1423, 2009, ISSN: 13674803. DOI: 10.1093/bioinformatics/btp163.

[40] A. Baranwal, B. R. Bagwe, and V. M, "Machine Learning in Python," vol. 12, pp. 128–154, 2019. DOI: 10.4018/978-1-5225-9902-9.ch008.

[41] X. Y. Ng, B. A. Rosdi, and S. Shahrudin, "Prediction of antimicrobial peptides based on sequence alignment and support vector machine-pairwise algorithm utilizing LZ-complexity," *BioMed Research International*, vol. 2015, 2015, ISSN: 23146141. DOI: 10.1155/2015/212715.