Machine Learning Modeling and Insights into the Structural Foundations of Polymyxin-like Antimicrobials

Inês Machado,[†] João Inácio,[‡,¶,§] Paula Jorge,[‡,¶,§] and Filipe Teixeira[*,‡]

[†]*Institute for Polymers and Composites, University of Minho, 4800-058 Guimarães, Portugal*

[‡]*Centre of Chemistry, University of Minho, 4710-057 Braga, Portugal*

[¶]*Centre of Biological Engineering, University of Minho, 4710-057 Braga, Portugal*

[§]*LABBELS – Associated Laboratory, Braga/Guimarães, Portugal*

E-mail: fteixeira@quimica.uminho.pt

Abstract

Antimicrobial resistance (AMR) is a silent pandemic that represents an urgent threat to human health. Unfortunately, the antibiotic development pipeline is slow even though AMR has been escalating uncontrollably fast, namely amongst Gram-negative pathogens. Although out of use until recently due to their toxic side effects, polymyxins have been revived as a last-line therapeutic option since all other antibiotics are currently failing. In an attempt to ameliorate their toxicity and improve antimicrobial activity, many studies have been generating polymyxin analogues through different strategies, mostly empirical. As such, there is still a lack of faster and more reliable approaches to make analog design efficient in order to tackle AMR in a timely fashion. The solution to accelerate the discovery of new drugs probably lies in the use of *in silico* approaches, such as machine learning, due to their faster pace and time and

cost efficiency. In this work, machine learning was applied to Quantitative Structure-Activity Relationship (QSAR) modeling with the objective of providing a working semi-quantitative model capable of predicting the activity of polymyxin-like molecules for a given species. For this, we applied four different learning algorithms and ten different families of molecular descriptors to our dataset of 408 molecule/microorganism pairs retrieved from PubChem. The AdaBoost model devised using the CKP set of descriptors was the best performer, with good accuracies and very low false negative and positive predictions. Preliminary exploration of the model's response to systematic changes in the structure of polymyxin B reveals a trend towards increased antimicrobial activity when exchanging some of its constituent amino acids for more lipophilic ones. Experimental studies are already underway based on this model's application and we believe it will become a crucial tool for drug development.

1 Introduction

Antimicrobial resistance represents one of the current biggest health-threats worldwide, whose impact has been heightened by the escalation of multidrug resistant (MDR) Gram-negative bacteria. *Pseudomonas aeruginosa*, *Acinetobacter baumannii*, and *Klebsiella pneumoniae* head the WHO list of priority pathogens not responding to front-line antibiotics[1], and their ability to grow as biofilms further heightens their role as troublesome pathogens highly associated with lower respiratory infections and with high mortality/morbidity rates[2].

Polymyxins (PMs) B and E are the two most studied and utilized variants of the antimicrobial peptide PM group and are currently used in last resort treatments for Gram-negative bacterial infections[3]. PMs were first put into clinical use in the 1950s, but were subsequently replaced by other drugs due to their nephrotoxic and neurotoxic side effects. Nevertheless, improved dosing regimens and the rise of Gram-negative MDR strains led to a renaissance of their clinical use when everything else was failing[4]. Sadly, PM resistance has also emerged, mostly due to outer membrane lower permeability[7], but also through nonspecific mechanisms (e.g. capsules, efflux pumps)[6]. This, along with PM's poor bioavailability, nephro-/neuro-toxicity, and narrow-spectrum activity, compromises what is already the last available treatment option[11].

PMs possess two key structural domains, polar (L-α,γ-diaminobutyric acid, Dab, residues) and hydrophobic (N-terminal fatty acyl chain; position 6-7 residues), which are relevant for PM-lipopolysaccharide (LPS) and outer membrane interaction[5]. Specifically, PMs act against Gram-negative bacteria by outer membrane destabilization/disruption, PM uptake, cell content leakage, and bacterial death[6–8]. Additionally, studies suggest that PMs have an ability to inhibit NADH-quinone oxidoreductase type II (NDH-2). This finding opens up the possibility that a secondary mode of action of polymyxin B (PM-B) against Gram-negative bacteria may involve inhibition of vital respiratory enzymes in the bacterial inner membrane[9]. There is also rising evidence indicating that PM-B nephrotoxicity is associated with DNA damage, leading to chromosome missegregation and genome instability. This

novel mechanistic information may lead to new strategies to overcome the nephrotoxicity of this important last-line class of antibiotics[10].

With no new antibiotics entering the market, repurposing and enhancing what is already available could be the solution. Therefore, PM structural modification aiming at improving its activity against MDR Gram-negative bacteria and diminish its toxicity has gained great interest. Modifications occurring at the Nα-terminal fatty acyl[12], Dab side chains[13], D-Phe6-L-Leu7 motif[14,15], cyclic heptapeptide ring[16], and tripeptide (Dab1-Thr2-Dab3) segment[15,17] have met variable success. Many of these studies are empirically-based[18] (with structure-activity relationships - SARs - often missing) and output analogues inactive, or less active than PMs[6,7,19]. Moreover, claims of PM toxicity reduction are often misleading or undocumented. Additionally, most PM analogue designs are not supported by knowledge of LPS-binding/outer membrane-disrupting mechanisms and, therefore, do not specifically target PM resistance[15]. To this date, more than 2000 PM analogues have been identified, but only a few have proceeded to preclinical studies, clinical trials, FDA approval, or market[11]. As such, an increased understanding of the SARs of this important class of compounds is required to further the development of the next generation of PM-related antibiotics.

Many researchers began to study the option of mimicking the physicochemical properties of PMs. Frecer *et al.*[20] reported a series of cyclic amphipathic peptides consisting of alternating cationic (Lys) and nonpolar (Phe) residues, loosely based on the amphipathic properties of the PM-B structure. The compounds exhibited potent antimicrobial activity against bacteria of the genera *Escherichia*, *Salmonella*, *Pseudomonas*, *Klebsiella*, and *Shigella*, and a high affinity for LPS. Velkov *et al.*[7] addressed the modelling of pharmacophores based on the collective two-dimensional SAR data of PMs in the literature combined with the three-dimensional model of the PM-B-LPS complex, confirming that the positive charge of the Dab side chains represents key characteristic. Hydrophobic properties are also key features in the Nα fatty acyl chain and positions 6 and 7 on the cyclic heptapeptide ring. The pharmacophore model showed that the PM-B molecule can be divided into a set of polar

and hydrophobic domains, namely the polar residue segments Dab and Thr residues, the hydrophobic Nα fatty acyl chain, and the D-Phe6-L-Leu7 motif. The model further highlighted the integral scaffolding function of the linear tripeptide and the cyclic heptapeptide to maintain the optimal distance between each domain, giving the structure its amphiphilicity, an essential property for antimicrobial activity.

The prospects for discovering a novel antibiotic are actually great when considering the vast possibilities among the existing 10^{30}-10^{60} drug-like chemicals, with $20n$ variants per n-length canonical amino acid. But individual experiments cannot be conducted on every candidate molecule both in terms of time and money. The conventional process of antibiotic development is not only slow, tedious, and expensive, but also has a high failure rate, contrasting with the fast and continuous process that is bacterial evolution[21,22]. This incapacity to keep up with AMR development is illustrated by the small amount (only 14) of new approved antibiotics between 2014 and 2019[23].

Computational approaches are key to overcome the antibiotic crisis and surpass conventional development pipelines. In fact, several antibiotics or drug candidates with putative antimicrobial activity and minimal toxicity have been identified through machine learning (ML) and quantitative structure–activity relationships (QSAR)[21]. ML is a branch of artificial intelligence that deals with the development of algorithms and models that can automatically learn patterns from data and perform tasks without explicit instructions. In the recent decade, with the systematic generation and management of data on an unprecedented scale and increases in computational power, ML has begun to explore new frontiers in many fields, including biology and chemistry. ML is particularly suited to exploratory tasks with combinatorially or exponentially complex solutions. Thus, ML is an excellent approach to many challenges in antibiotic science because it can generalize from training data to explore new solutions, speeding up the identification of physiological processes involved in drug–target interactions (e.g. mechanisms of action, cytotoxicity pathways, resistance mechanisms)[24–27].

With the various existing approaches of ML, QSAR emerged as the most frequent appli-

cation area. In view of the large libraries of compounds now being treated by combinatorial chemistry and high-throughput screening, the use of computational techniques such as QSAR modeling is highly advisable. QSAR serves as lead optimization in early drug discovery, before they are subjected to more intensive studies, such as receptor docking and empirical determination of *in vitro* and *in vivo* activity/toxicity. QSAR methodology consists in the representation of the chemical structure using molecular descriptors, which serve as useful physicochemical information to determine the correlation between the chemical structure and the biological activity. Nowadays, there are thousands of molecular descriptors with the potential to be applied in drug design[22,27,28].

The successful development of new antimicrobials based on the PM scaffold would greatly benefit from the development of QSAR models to aid navigating this vast chemical space. In this work, we endeavor to explore several approaches to model the antimicrobial activity (measured by its Minimum Inhibitory Concentration) of PM-like molecules towards an assortment of microbial species using different ML strategies. The best performing model is further explored in terms of its response under systematic mutations of the PM-B structure, in order to gain new insights onto the most preponderant features of highly active PM derivatives. The main goal is to provide a working model capable of discerning the most promising molecules towards a given species and thus aid in the quick development of much needed new antimicrobial agents.

2 Methodology

2.1 Polymyxin Activity Dataset

To the best of our knowledge, there are no previous datasets devoted to the antimicrobial activity of PMs and PM-like molecules. Thus, a large dataset containing the Minimum Inhibitory Concentration (MIC) for 408 molecule/microorganism pairs was collected from PubChem[29]. Data collection procedures encompassed Polymixin B nonapeptide (CID 123978),

as well as the 1000 most similar structures (based on the 2-dimensional Tanimoto fingerprint), which were first filtered by the availability of antimicrobial assay data (Biological Assays), and then by the availability of a defined value for MIC, MIC_{50}, or MIC_{95}.

The data was then curated for the removal of duplicates, entries without description of the targeted microorganism or pertaining to drug-combination studies. During the curation process, the information regarding the targeted microorganism in each assay was condensed into two variables: one containing the taxonomic genus of the target (T_xG), and another one concerning a broader classification of the type of microorganism (M_{Typ}), which can take one of three values: Gram-negative bacteria, Gram-positive bacteria, or fungi.

Preliminary calculations aiming for a regression model of the MIC failed, prompting a semi-quantitative approach targeting the quartile (among the full data) of the MIC reported for a given compound/target pair. This strategy not only allowed for some semi-quantitative assessment of the inhibitory activity of novel compounds, but also ensured that the different categories of the target are equally represented in the data. The curated dataset of 399 entries is provided in the Electronic Supplementary Information (ESI).

Starting from the simplified molecular-input line-entry system (SMILES) representation of each molecule in the dataset, several families of molecular descriptors were calculated using the RDKit software package[30]. The naming scheme for these sets of descriptors, their composition, and the number of features present in each set are provided in Table 1.

2.2 Machine Learning Models

In order to explore the potential of ML methods modeling the antimicrobial activity of PM-like molecules, four supervised learning algorithms were considered: logistic regression[34], decision tree[35], random forest[36], and AdaBoost[37], as implemented in the Scikit-learn package[38], version 1.0.2. The variables T_xG and M_{Typ} were added to each set of molecular descriptors in order to form the feature set used by each model. Each algorithm/descriptor set pair was trained targeting a multi-class prediction of the MIC quartile, using a 65:35

Table 1: Labeling convention used for the sets of molecular descriptors used in this work, as well as their general description and the number of features ($n_{feat.}$) within each set.

Acronym	Description	$n_{feat.}$
Gen.	BalabanJ, BertzCT, Ipc, HallKierAlpha, MolLogP, MolWt, HeavyAtomCount, NumHeteroatoms, NumRotatableBonds, NumValenceElectrons, RingCount, FractionCSP3, TPSA, LabuteASA[31]	14
Hb	Descriptors related to H-Bond formation: NHOHCount, NOCount, NumHAcceptors, NumHDonors	4
CKP	κ-form Kier and Hall indices[32]	15
PEOE_VSA	MOE-type descriptors using partial charges and surface area contributions	14
SMR_VSA	MOE-type descriptors using Molar Refractivity and surface area contributions	10
SLopP_VSA	MOE-type descriptors using LogP and surface area contributions	12
Estate_VSA	MOE-type descriptors using Kier and Hall's Estate indices and surface area contributions	11
AC2D	2-dimensional autocorrelation functions[31]	192
BCUT2D	Perlman's BCUT metrics[33]	8
FG	Counting of functional group fragments	85

split between the training and the testing data.

All models were created as a data pipeline, where all numerical fields were first scaled to zero mean and unit standard deviation and all non-numerical variables were codified using one-hot encoding, prior to being fed onto the main algorithm. The logistic regression and decision tree models were trained on the transformed data using the default hyper-parameters defined in their implementation. On the other hand, the random forest and AdaBoost models required optimization of some of their hyper-parameters. For the random forest models, the number of estimators (trees) (n_{est}), as well as the fraction of samples (n_s) and features (n_f) considered by each estimator, were optimized using a 5-fold cross-validation strategy, scanning 100 random combinations of n_{est}, n_s, and n_f. A similar cross-validation scheme was also used in the case of the AdaBoost models, optimizing the number of estimators (trees) (n_{est}), the depth of the base estimators (d_{est}), and the learning rate (r_L). Further analysis of each model was carried out using in-house developed Python scripts for accessing the importance of individual features, partial dependence of each model's response to the most important features, as well as response of the best-performing models to systematic mutations of the PM-B structure. These scripts are also provided in the ESI.

3 Results and Discussion

3.1 Characterization of the Dataset

Of the 399 data points collected, 366 were related to antibacterial activity, of which 287 were for Gram-negative bacteria and 79 for Gram-positive bacteria. In addition, 33 entries were related to antifungal activity. Among bacteria, the most represented genera were *Escherichia* (109 entries), *Pseudomonas* (81 entries), *Salmonella* (58 entries), and *Staphylococcus* (54 entries). These four genera make up about 82.5% of the bacterial data. The least represented genera of bacteria were *Priestia*, *Yersinia*, *Enterobacter*, *Shigella*, *Vibrio*, and *Proteus*, with only one entry per genus. With regard to fungi, about 76% of the reported values concerned

the genus *Candida*, with *Cryptococcus* being the least studied genus among fungi.

Regarding the compounds, 86 entries (about 21% of the data) concerned PM-B1 in the neutral form, followed by PM-B1 sulfate, with 56 entries (13.5% of the data).

As for the collected MIC values, these ranged between 0.006 μM and 256 μM, with an average of 26 ± 43 μM. These data are quite asymmetrical (as can be seen from the density distribution shown in Figure 1), with the boundary of the first quartile (Q1) located at 1.25 μM and the upper limit of the third quartile (Q3) at 32.0 μM, with a median of 4.0 μM. As shown in Figure 1, MIC data are distributed in a rather asymmetric and multi-modal way, with the main modal peak close to the median, but a considerable distance between the median and Q3. This can be partially attributed to the MIC determined in assays with fungi. The MIC for Gram-positive bacteria appears to be spread over a wider region of the MIC spectrum (up to about 150 μM). The outlier values above 120 μM appear to be common to multiple assays and are likely to reflect the maximum concentration threshold used in various assays.

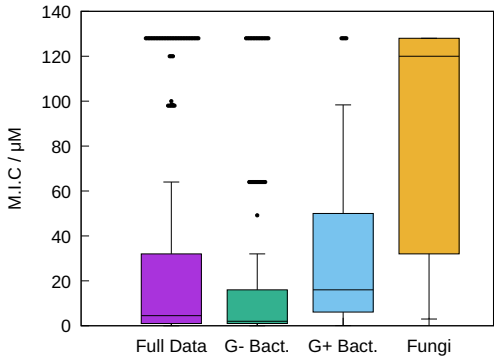

Figure 1: Boxplot representation of the collected MIC values in the full data, as well as for the sub-categories regarding the type of microorganism. Outlier values above 150 μM (3 entries, all regarding Gram-negative bacteria) were omitted for clarity.

3.2 Variable Selection and Model Refinement

All models were trained targeting the quartile position of each entry in the dataset, with the ultimate aim of understanding which modifications to the PM scaffold would yield more

antimicrobial activity. For reference, the MIC collected for PM-B against *P. aeruginosa* is 4.0 μM. Hence, a classification of Q1 or Q2 for a novel structure would signal such structure as a promising candidate for future synthesis and testing, with the most promising ones being those classified as Q1.

Because of this, the evaluation of each model took into account not only the accuracy scores for the train and test sets, but also the true positive rate for Q1 (i.e. the fraction of cases were a compound was correctly predicted as Q1, or $f(Q1|Q1)$). Moreover, our evaluation also took into account the undesirable metrics $f(Q1|Q4)$ and $f(Q4|Q1)$, which translate to promoting a particularly inactive molecule (at least for the selected microbial target) and wasting a good proposed structure, respectively[39].

Upon optimization of the hyper-parameters, most random forest models required relatively small forests (n_{est} between 10 and 25), which is adequate considering the number of points in the data set. Most models favoured the use of all available features for each tree ($n_f = 1.0$), with the exception of the model using the Hb set, which showed the best cross-validation accuracy score for $n_f = 0.55$. On the other hand, most random forest algorithms opted for each tree to consider only a fraction of the presented data, n_s, between 0.16 and 0.26.

The AdaBoost models exhibited a similar preference for small values of n_{est}, with the exception of the model using the SLopP_VSA set of descriptors, which required $n_{est} = 50$. Each of the trees in the AdaBoost model were usually limited to a maximum depth (d_{est}) of 10, with the exception of the AdaBoost models using the FG set ($d_{est} = 2$), as well as that of the models using the SLopP_VSA and AC2D sets, which attained maximum accuracy for $d_{est} = 100$. The optimal hyper-parameters of the random forest and AdaBoost models are provided in the ESI.

The results depicted in Figure 2 show the behaviour of the 40 models considered in this work through the metrics detailed above. It is noteworthy that $f(Q4|Q1)$ is always very low for all models, and was thus excluded from further considerations. Overall, the logistic

regression models performed the worst (Figure 2a), with accuracies in the train set never exceeding 60% and considerable values of $f(Q1|Q1)$. The decision tree models, despite performing better than the logistic regression ones, showed some considerable over-fitting behaviour, as well as relatively low scores of $f(Q1|Q1)$ in the test data (Figure 2b). Indeed, the combination of multiple decision trees in either a bagging (random forest) or boosting (AdaBoost) configuration yielded models with some interesting characteristics.

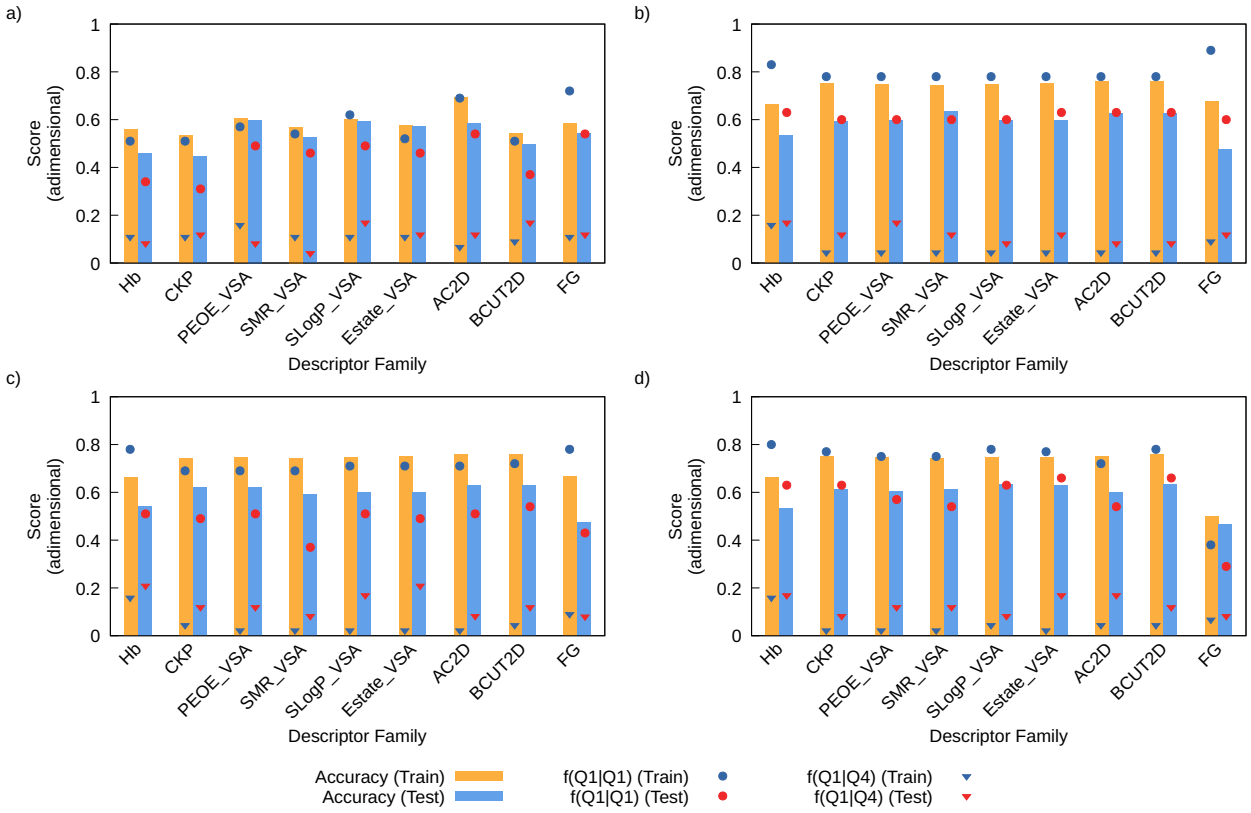

Figure 2: Values of different scores (overall accuracy, $f(Q1|Q1)$ and $f(Q1|Q4)$) for each family of descriptors (*see* Table 1) and algorithms: a) logistic regression; b) decision tree; c) random forest, and d) AdaBoost.

The overall performance of the random forest models showed some of the highest accuracy scores in the training data. Unfortunately, the over-fitting issues in these models were particularly pronounced when looking at the $f(Q1|Q1)$ and $f(Q1|Q4)$ scores (Figure 2c). These problems appear to be somewhat mitigated by the AdaBoost models, specially with respect to $f(Q1|Q4)$ (Figure 2d).

Regarding the performance of each set of molecular descriptors, the fraction of functional groups (FG) standed out negatively. Despite its modest performance when in combination with the logistic regression algorithm, its use in decision tree or random forest models showed significant over-fitting, affecting the important $f(Q1|Q1)$ score. Additionally, FG was the worst performing set of descriptors for the AdaBoost models. Likewise, the Hb set also yielded some of the weakest models, usually by increasing $f(Q1|Q4)$ in both the train and test sets, irrespective of the algorithm used.

The performance of the descriptor sets derived from surface area contributions (VSA-based descriptors) varied significantly between algorithms. They yielded logistic regression models with neglectable over-fitting, but with relatively low accuracy and $f(Q1|Q1)$ scores, and always with relatively high $f(Q1|Q4)$ scores, specially in the case of SLopP_VSA (Figure 2a). In combination with the decision tree algorithm, the accuracy of the resulting models appears to be linked to the $f(Q1|Q1)$ score (Figure 2b). In general, these families of descriptors performed well, albeit with an overall tendency to significantly increase the $f(Q1|Q4)$ score in the test phase.

In turn, the topological-rooted sets of descriptors (CPK, AD2D, and BCUT2D) typically performed well, with overall accuracy and $f(Q1|Q1)$ scores in par with those found when using the VSA-based descriptors. However, with the exception of the logistic regression models, these descriptor sets usually presented lower $f(Q1|Q4)$ scores in the test data than models trained using the VSA-based descriptors. Two noteworthy cases are the random forest model trained using the AC2D set (Figure 2c) and the AdaBoost model trained using the CKP set. Both models show an acceptable overall accuracy (approximately 80% and 65% in the train and test sets, respectively), high $f(Q1|Q1)$ scores, and very low $f(Q1|Q4)$ scores. This prompted the AdaBoost model devised using the CKP set of descriptors to be selected for further studying.

3.3 Analysis of the Model's Response

The Kier and Hall descriptors forming the CPK set describe the structural and geometric properties of a molecule, including its flexibility, polarity, and hydrophobicity. They are widely used for the analysis of the biological activity of compounds, mainly with respect to their lipophilic and hydrophilic affinity, proving to be efficient for the discrimination of compounds with different levels of affinity[32,40]. The good performance of this set of descriptors in the particular case of predicting the antimicrobial activity of PMs and their analogues is consistent with the established mode of action of these compounds.

The relative importance of each feature considered in the selected AdaBoost model was evaluated using the Permutation Importance (PI) method, in which the weight of each feature is related to the change in the model's outcome when said feature is replaced with randomly generated data[41]. The relative weights of each feature are represented in Figure 3 and suggest that the model is particularly sensitive to 5 features: the two features describing the microorganism (T_XG and M_{Typ}) as well as three molecular descriptors ($^1\chi$, $^0\chi$, and $^3\kappa$). These five features make up 71% of the total PI.

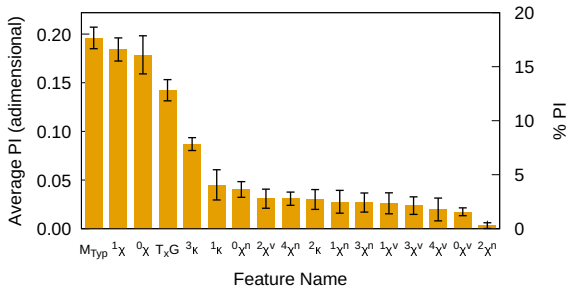

Figure 3: Average Permutation Importance (PI) of the features in the AdaBoost model using the CPK set of molecular descriptors calculated using 10 noisy replicas of the data set for each feature. The error bars represent the standard deviation of the PI over the 10 replicas and the right y axis indicates the value of the average PI normalized to percentage.

3.3.1 Influence of the Biological Target

The most important feature was M_{Typ}, with a PI weight of 17.7%. This result reflects the pre-modelling assessment of the data, in which the MIC values of the Gram-positive bacteria and (even more so) of the fungi was significantly larger than that of the Gram-negative bacteria (Cf. Figure 1). This is well illustrated in Figure 4a, which shows the model's partial dependence with this feature. Indeed, assays using Gram-negative bacteria were more likely to be classified as Q1, whereas those targeting either Gram-positive bacteria or fungi were more likely to be classified as Q3 or Q4, respectively. This partial dependence behaviour reflects the distribution of the MIC shown in Figure 1, suggesting that this pattern was learned by the AdaBoost model.

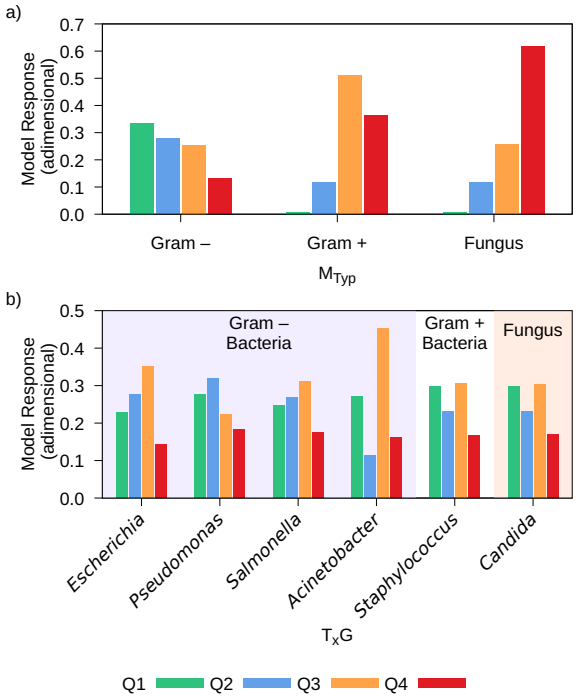

Figure 4: Partial dependence graphs of the AdaBoost model using CPK molecular descriptors: a) by M_{Typ}, and b) by T_xG, for the most common genera in the dataset (i.e. with more than 10 entries in the data).

The other feature relating to the microorganism used in each entry, T_xG, ranked in forth place, with a PI of 12.8%. According to these results, the model appears to be using M_{Typ}

as a first sieve to classify the incoming data, and T_xG as a secondary sieve. This reflects the zeitgeist that PMs are particularly active towards all genera of Gram-negative bacteria.

Indeed, the partial dependence data plotted in Figure 4a suggests that, despite the model's tendency to classify assays carried out using Gram-negative bacteria as Q1, the ones targeting the *Escherichia*, *Salmonella*, and specially the *Acinetobacter* genus have a boost towards a Q3 classification, hinting at these genera being more resistant to the PM analogues found in the data.

3.3.2 Influence of the Molecular Descriptors

From the point of view of the molecular descriptors, the model's response is dominated by ${}^1\chi$, ${}^0\chi$, and, to a lesser extent, ${}^3\kappa$. These three molecular descriptors gather about 40% of the PI. When combining these weights with those describing the target, the five most preponderant features collected 71% of the PI. The model's partial dependence plots of these three features are depicted in Figure 5, showing the likelihood of a given classification outcome (Q1, Q2, Q3, or Q4) with varying values of the feature at hand, when all other features are random. In all cases, the partial dependence plots appear to be divided into a "low-value" regime and a "high-value" one, with a somewhat chaotic partial response when transitioning between the two states.

Both ${}^0\chi$ and ${}^1\chi$ are atomic connectivity indexes derived from the molecule's connectivity matrix, with

$$ {}^0\chi = \sum_i \frac{1}{\sqrt{d_i}} \tag{1} $$

and

$$ {}^1\chi = \sum_i \sum_{j<i} \frac{1}{\sqrt{d_i d_j}} \tag{2} $$

where di_i is the number of heavy (non-hydrogen) atoms connected to atom i and the sums cover all heavy atoms[31]. Thus, an increase in the number of heavy atoms connected to few heavy atoms (small ${}^0\chi$) favours a classification as Q1, whereas an increase of ${}^0\chi$ (either by

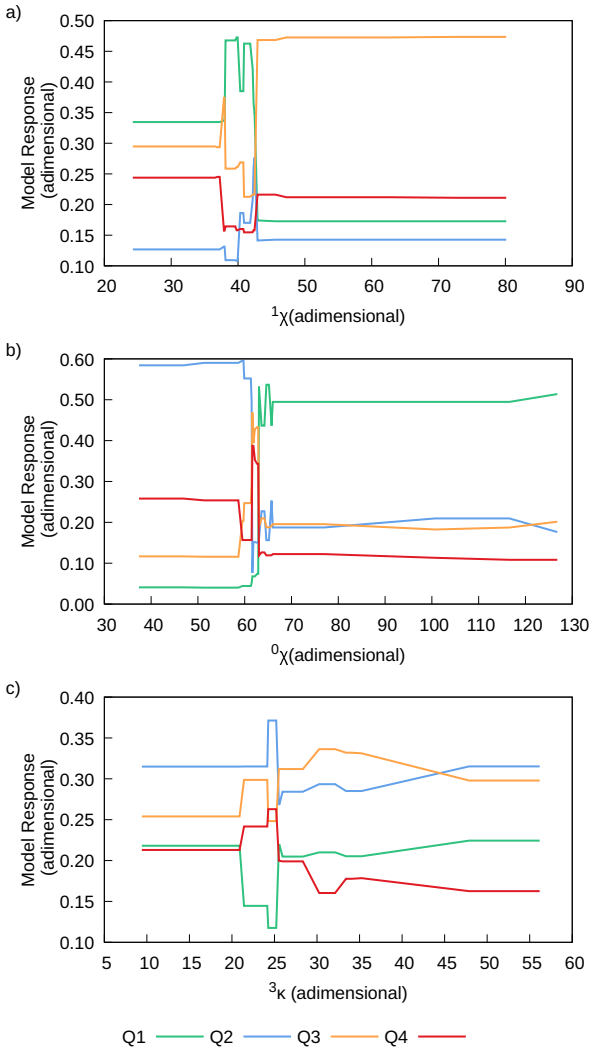

Figure 5: Partial dependence graphs of the AdaBoost model using CPK molecular descriptors: a) with respect to $^1\chi$, b) with respect to $^0\chi$, and c) with respect to $^3\kappa$.

removal of heavy atoms or by greatly increasing ramification) makes a classification as Q2 more likely, as shown in Figure 5b.

The effect of ramification can also be observed when considering the partial dependence with respect to ${}^1\chi$, shown in Figure 5a. Highly ramified structures usually obtain a lower value of ${}^1\chi$, and have an increased probability of being labelled as Q2, while less ramified ones hold a greater chance of being classified as Q3. This aspect of the model's response appears to contradict the above discussion centred in ${}^0\chi$. Because both features hold approximately the same weights (Cf. Figure 3), one would argue that the penalty shown in Figure 5b for ${}^0\chi > 65$ identifies compounds with lower molecular weight (more properly, with less heavy atoms), with the sensitivity towards ramification being handled mostly by ${}^1\chi$. That being said, it is worthy to highlight that the model's partial dependence with respect to ${}^1\chi$ points towards an optimal region that maximizes the probability of achieving a Q1 classification at about ${}^1\chi \approx 40$, suggesting an optimal value for the number of ramification motifs.

The third Kier's kappa shape index (${}^3\kappa$) is defined as

$$ {}^3\kappa = \begin{cases} \frac{(n-1)(n-3)^2}{p_3^2} & \text{if } n \text{ is odd,} \\ \frac{(n-3)(n-2)^2}{p_3^2} & \text{if } n \text{ is even.} \end{cases} \tag{3} $$

where n is the number of non-hydrogen atoms, and p_3 is the number of paths of length 3 (i.e. groups of atoms connected using three bonds). Contrary to ${}^0\chi$ and ${}^1\chi$, ${}^3\kappa$ increases with increasing number of heavy atoms, but decreases with increasing number of possible length 3 paths allowed by the molecular topology, which can be achieved either by increasing ramification (specially when occurring in the middle of longer chains), or via introduction of cyclic groups[32]. As shown in Figure 5, the model's response with respect to ${}^3\kappa$ suggests that this feature is mostly used to distinguish between Q1 and Q2 classifications, with the probability of the former being at a minimum when the probably of the latter is at its maximum.

3.3.3 Systematic Mutations of the Polymixin B Scaffold

In order to have a more immediate sense of the model's response, the structure of PM-B was systematically mutated in positions 1 to 3 and 5 to 10 using glycine (Gly), leucine (Leu), lysine (Lys), and glutamic acid (Glu). These mutations reflect the change in the model's outcome when varying the steric hindrance at a particular position (Gly *versus* Leu), or upon introduction of a basic or acid amino acid residue (Lys *versus* Glu). In all predictive runs of the model, the microbial target was fixed at the Gram-negative bacterial species *P. aeruginosa*. The results from these so-called "mutations" on the PM-B structure are shown in Figure 6. Along these systematic mutations, the distance (in feature space) of the proposed structures to the centre (average) of the available data was monitored in order to estimate whether the new molecular structures would generate a set of descriptors within the range for which the model was trained. The largest average distance from the data average was observed in the case of Gly substitution. In this set of molecules, the average distance to the data centre was 2.16 ± 0.17 in the adimensional feature space, which compares very favourably with the average distance to the centre of 3.1 ± 2.8 found in the data set. All other series of PM-B mutations fell even closer to the data centre, with average distances of 1.55 ± 0.15, 1.46 ± 0.15, and 1.60 ± 0.16 for exchanges with Leu, Lys, and Glu, respectively.

Regarding the systematic exchange of each of the constituent amino acids by Gly, the general trend is for conserving the Q2 classification (the combination PM/*P. aeruginosa* itself being ranked Q2 in the data). Nevertheless, substitution of Leu7 by Gly appears to improve the antimicrobial activityi, as shown in Figure 6a. On the other hand, the model suggests a negative impact on the predicted antimicrobial activity when replacing Phe6 by Gly.

The previous observations are in sharp contrast to what is observed when replacing each aminoacid residue by Leu, which usually results in a more optimistic prediction of antimicrobial activity, as shown in Figure 6b. Again, the major exception is the introduction of Leu in position 6 for which the model predicts a value for the MIC between 1.5 μM and

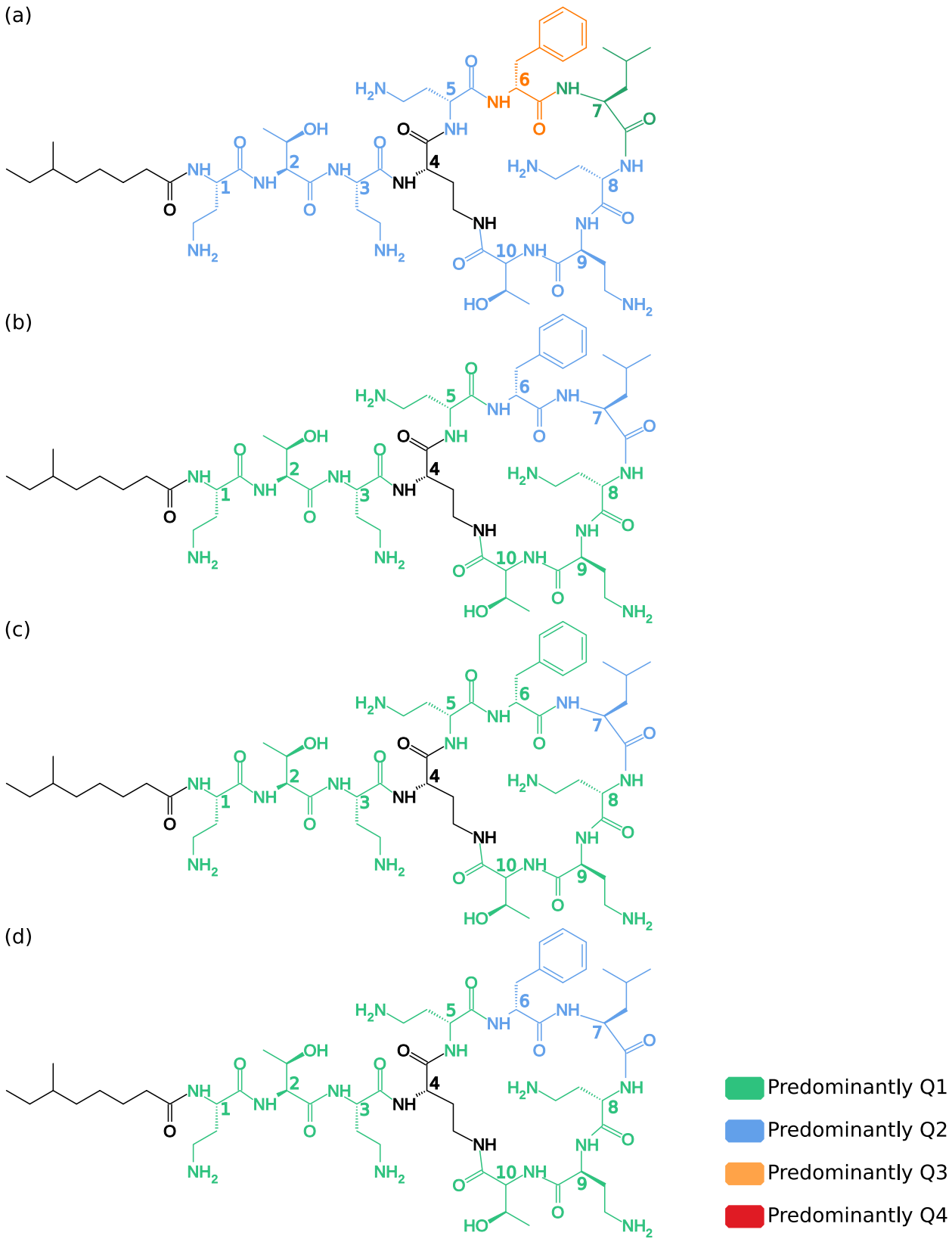

Figure 6: Most probable classification regarding the antimicrobial activity towards *P. aeruginosa* of mutated variants of PM-B upon systematically changing each amino acid residue for: a) Gly, b) Leu, c) Lys, and d) Glu.

4.0 μM (Q2). It should be noted that the position 7 of PM-B is already occupied by Leu, and the corresponding Q2 prediction is coherent with the collected data. It is likely that the more lipophilic character of Leu promped the model to predict an enhanced antimicrobial activity when one (and only one) of the constituent aminoacids is replaced by Leu.

As shown in Figures 6c and 6d, the systematic exchange of each amino acid residue by either Lys or Glu, respectively, resulted in an improvement over the predicted antimicrobial activity of PM-B. The major exception to this trend was, again, Leu7, for which the model's predictions suggested that its substitution for either an acidic or basic amino acid does not bring a distinct advantage over the original PM-B struture. Furthermore, the exchange of Phe6 by Glu also appeared to maintain the predicted antimicrobial activity within the boundaries of Q2. The results depicted in Figure 6c for Lys replacement are particularly interesting, as they suggest that the substitution of Dab by Lys may increase the antimicrobial activity. As in the case of the Leu substitutions, these predictions may be linked to an increase in lipophilicity, perceived by the model by the increase in the amino acid side chains.

4 Conclusions

In this work, we applied the AdaBoost algorithm to generate a semi-quantitative model of the antimicrobial activity of PM-B analogs using well established molecular descriptors. The present model resulted from a systematic exploration of different combinations of ML algorithms and sets of molecular descriptors, and can adequately predict the MIC (by ranges) of a given compound/target combination. This allows the use of the model for rapidly accessing whether a proposed structure can be considered as a viable candidate for novel PM-derived antibiotics. Analysis of this model confirmed insights previously obtained from the available data, such as the greater activity of PM derivatives towards Gram-negative bacteria, and its relatively small antimycotic activity. More interestingly, preliminary exploration of

the model's response to systematic changes to the PM-B structure revealed a trend for increased antimicrobial activity when exchanging some of its constituent amino acids by more lipophilic ones.

Author Contributions

I.M. redacted the first version of the manuscript and revised the results from the calculations. J.I. carried out the data collection and curation, trained the ML models, and performed the analysis of the ML models. P.J. planned the data collection and revised the collected data, advised on critical aspects of the biological assays data, and revised the manuscript. F.T. planned the modelling work, implemented the software routines for the training and analysis of the ML models, and redacted the revised manuscript.

Supporting Information Available

Additional data (collected data set, software code for using the final model, scores of all tested ML models, optimized hyper-parameters for all random forest and AdaBoost models, and partial dependence plots for the features with less than 10% PI) are available in the Electronic Supporting Information.

Acknowledgement

This work was supported by the Portuguese Foundation for Science and Technology (FCT) under the scope of project POLYmix-POLYmic (2022.06595.PTDC), of the programmatic funding to CQUM (UID/QUI/00686/2020), and of the strategic funding to CEB (UIDB/04469/2020), of the contract 2020.00194.CEECIND, and by LABBELS - Associate Laboratory in Biotechnology, Bioengineering and Microelectromechanical Systems (LA/P/0029/2020).

References

(1) Global Priority List of Antibiotic-Resistant Bacteria to Guide Research, Discovery, and Development of New Antibiotics. WHO Press, Geneva, 2017.

(2) Duan, N.; Du, J.; Huang, C.; Li, H. Microbial Distribution and Antibiotic Susceptibility of Lower Respiratory Tract Infections Patients From Pediatric Ward, Adult Respiratory Ward, and Respiratory Intensive Care Unit. *Front. Microbiol.* **2020**, *11*.

(3) Garg, S. K.; Singh, O.; Juneja, D.; Tyagi, N.; Khurana, A. S.; Qamra, A.; Motlekar, S.; Barkate, H. Resurgence of Polymyxin B for MDR/XDR Gram-Negative Infections: An Overview of Current Evidence. *Crit. Care Res. Pract.* **2017**, *2017*, 1–10.

(4) Manioglu, S.; Modaresi, S. M.; Ritzmann, N.; Thoma, J.; Overall, S. A.; Harms, A.; Upert, G.; Luther, A.; Barnes, A. B.; Obrecht, D.; Müller, D. J.; Hiller, S. Antibiotic polymyxin arranges lipopolysaccharide into crystalline structures to solidify the bacterial membrane. *Nat. Commun.* **2022**, *13*.

(5) Jiang, X.; Yang, K.; Yuan, B.; Han, M.; Zhu, Y.; Roberts, K. D.; Patil, N. A.; Li, J.; Gong, B.; Hancock, R. E. W.; Velkov, T.; Schreiber, F.; Wang, L.; Li, J. Molecular dynamics simulations informed by membrane lipidomics reveal the structure–interaction relationship of polymyxins with the lipid A-based outer membrane of *Acinetobacter baumannii*. *J. Antimicrob. Chemother.* **2020**, *75*, 3534–3543.

(6) Nang, S. C.; Azad, M. A. K.; Velkov, T.; Zhou, Q. T.; Li, J. Rescuing the Last-Line Polymyxins: Achievements and Challenges. *Pharmacol Rev* **2021**, *73*, 679–728.

(7) Velkov, T.; Thompson, P. E.; Nation, R. L.; Li, J. Structure-Activity Relationships of Polymyxin Antibiotics. *J. Med. Chem.* **2009**, *53*, 1898–1916.

(8) Yu, Z.; Qin, W.; Lin, J.; Fang, S.; Qiu, J. Antibacterial Mechanisms of Polymyxin and Bacterial Resistance. *Biomed Res. Int.* **2015**, *2015*, 1–11.

(9) Deris, Z. Z.; Akter, J.; Sivanesan, S.; Roberts, K. D.; Thompson, P. E.; Nation, R. L.; Li, J.; Velkov, T. A secondary mode of action of polymyxins against Gram-negative bacteria involves the inhibition of NADH-quinone oxidoreductase activity. *J. Antibiot.* **2013**, *67*, 147–151.

(10) Yun, B.; Zhang, T.; Azad, M. A. K.; Wang, J.; Nowell, C. J.; Kalitsis, P.; Velkov, T.; Hudson, D. F.; Li, J. Polymyxin B causes DNA damage in HK-2 cells and mice. *Arch. Toxicol.* **2018**, *92*, 2259–2271.

(11) Vaara, M. Polymyxins and Their Potential Next Generation as Therapeutic Antibiotics. *Front. Microbiol.* **2019**, *10*.

(12) Vaara, M.; Siikanen, O.; Apajalahti, J.; Fox, J.; Frimodt-Møller, N.; He, H.; Poudyal, A.; Li, J.; Nation, R. L.; Vaara, T. A Novel Polymyxin Derivative That Lacks the Fatty Acid Tail and Carries Only Three Positive Charges Has Strong Synergism with Agents Excluded by the Intact Outer Membrane. *Antimicrob. Agents Chemother.* **2010**, *54*, 3341–3346.

(13) Rudilla, H.; Pérez-Guillén, I.; Rabanal, F.; Sierra, J. M.; Vinuesa, T.; Viñas, M. Novel synthetic polymyxins kill Gram-positive bacteria. *J. Antimicrob. Chemother.* **2018**, *73*, 3385–3390.

(14) Tsubery, H.; Ofek, I.; Cohen, S.; Eisenstein, M.; Fridkin, M. Modulation of the Hydrophobic Domain of Polymyxin B Nonapeptide: Effect on Outer-Membrane Permeabilization and Lipopolysaccharide Neutralization. *Mol. Pharmacol.* **2002**, *62*, 1036–1042.

(15) Velkov, T.; Roberts, K. D.; Nation, R. L.; Wang, J.; Thompson, P. E.; Li, J. Teaching 'Old' Polymyxins New Tricks: New-Generation Lipopeptides Targeting Gram-Negative 'Superbugs'. *ACS Chem. Biol.* **2014**, *9*, 1172–1177.

(16) Tsubery, H.; Ofek, I.; Cohen, S.; Fridkin, M. The Functional Association of Polymyxin

B with Bacterial Lipopolysaccharide Is Stereospecific: Studies on Polymyxin B Nonapeptide. *Biochem.* **2000**, *39*, 11837–11844.

(17) Vaara, M.; Fox, J.; Loidl, G.; Siikanen, O.; Apajalahti, J.; Hansen, F.; Frimodt-Møller, N.; Nagai, J.; Takano, M.; Vaara, T. Novel Polymyxin Derivatives Carrying Only Three Positive Charges Are Effective Antibacterial Agents. *Antimicrob. Agents Chemother.* **2008**, *52*, 3229–3236.

(18) Abdalla, M.; McGaw, L. Natural Cyclic Peptides as an Attractive Modality for Therapeutics: A Mini Review. *Molecules* **2018**, *23*, 2080.

(19) Brown, P.; Dawson, M. J. Development of new polymyxin derivatives for multi-drug resistant Gram-negative infections. *J. Antibiot.* **2017**, *70*, 386–394.

(20) Frecer, V.; Ho, B.; Ding, J. L. De Novo Design of Potent Antimicrobial Peptides. *Antimicrob. Agents Chemother.* **2004**, *48*, 3349–3357.

(21) Talat, A.; Khan, A. U. Artificial intelligence as a smart approach to develop antimicrobial drug molecules: A paradigm to combat drug-resistant infections. *Drug Discov Today* **2023**, *28*, 103491.

(22) Helguera, A.; Combes, R.; Gonzalez, M.; Cordeiro, M. N. Applications of 2D Descriptors in Drug Design: A DRAGON Tale. *Curr Top Med Chem* **2008**, *8*, 1628–1655.

(23) Lepore, C.; Silver, L.; Theuretzbacher, U.; Thomas, J.; Visi, D. The small-molecule antibiotics pipeline: 2014–2018. *Nat. Rev. Drug Discovery* **2019**, *18*, 739–739.

(24) Chen, C.; Zuo, Y.; Ye, W.; Li, X.; Deng, Z.; Ong, S. P. A Critical Review of Machine Learning of Energy Materials. *Adv. Energy Mater.* **2020**, *10*, 1903242.

(25) Scott-Fordsmand, J. J.; Amorim, M. J. Using Machine Learning to make nanomaterials sustainable. *Sci Total Environ* **2023**, *859*, 160303.

(26) Xiang, M.; Cao, Y.; Fan, W.; Chen, L.; Mo, Y. Computer-Aided Drug Design: Lead Discovery and Optimization. *Comb. Chem. High Throughput Screening* **2012**, *15*, 328–337.

(27) McComb, M.; Bies, R.; Ramanathan, M. Machine learning in pharmacometrics: Opportunities and challenges. *Brit J Clin Pharmaco* **2021**, *88*, 1482–1499.

(28) Helguera, A. M.; Pérez-Garrido, A.; Gaspar, A.; Reis, J.; Cagide, F.; Vina, D.; Cordeiro, M. D.; Borges, F. Combining QSAR classification models for predictive modeling of human monoamine oxidase inhibitors. *Eur. J. Med. Chem.* **2013**, *59*, 75–90.

(29) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2023 update. *Nucleic Acids Res.* **2022**, *51*, D1373–D1380.

(30) RDKit: Open-source cheminformatics. http://www.rdkit.org.

(31) Mauri, A.; Consonni, V.; Todeschini, R. *Handbook of Computational Chemistry*; Springer International Publishing, 2017; pp 2065–2093.

(32) Hall, L. H.; Kier, L. B. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; John Wiley & Sons, Inc., 2007; Vol. 2; pp 367–422.

(33) Stanton, D. T. Evaluation and Use of BCUT Descriptors in QSAR and QSPR Studies. *J Chem Inf Comp Sci* **1998**, *39*, 11–20.

(34) Walker, S. H.; Duncan, D. B. Estimation of the Probability of an Event as a Function of Several Independent Variables. *Biometrika* **1967**, *54*, 167.

(35) Breiman, L. *Classification and regression trees*; Chapman & Hall, 1993; p 358.

(36) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32.

(37) Hastie, T.; Rosset, S.; Zhu, J.; Zou, H. Multi-class AdaBoost. *Stat Interface* **2009**, *2*, 349–360.

(38) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(39) Chicco, D.; Tötsch, N.; Jurman, G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *Biodata Min.* **2021**, *14*, 13.

(40) Kier, L. B. An Index of Molecular Flexibility from Kappa Shape Attributes. *Quant. Struct.-Act. Relat.* **1989**, *8*, 221–224.

(41) Molnar, C. *Interpretable Machine Learning A Guide for Making Black Box Models Explainable*; Independently Published, 2022.