Polymer Chemistry

Advancing Antimicrobial Polymer Development: A Novel Database and Accelerated Design via Machine Learning

Journal:	*Polymer Chemistry*
Manuscript ID	PY-ART-07-2024-000736.R2
Article Type:	Paper
Date Submitted by the Author:	28-Aug-2024
Complete List of Authors:	Zhao, Yuankai; RMIT University College of Science Engineering and Health, Enginerring Mulder, Roger; CSIRO, Manufacturing Eyckens, Daniel; CSIRO, Manufacturing Houshyar, Shadi; RMIT University, Engineering Le, Tu; RMIT University

RUNNING HEAD TITLE

Advancing Antimicrobial Polymer Development: A Novel Database and Accelerated Design via Machine Learning

Yuankai Zhao[a], Roger J Mulder[b] Daniel J. Eyckens[b], Shadi Houshyar[a] and Tu C. Le[a,*]

[a] School of Engineering, STEM College, RMIT University, GPO Box 2476, Melbourne, VIC 3001, Australia

[b] CSIRO Manufacturing, Research Way, Clayton, VIC 3168, Australia

[*] Tu.Le@rmit.edu.au

Abstract

The rapid growth of resistant microorganisms has caused serious public health issue and poses great pressure on the current healthcare system. In this environment, the necessity of new antibiotic materials is even more prominent. Antimicrobial polymers are a type of polymers that has the ability to eradicate or impede the proliferation of microbes on their surfaces or within their surrounding environment. The mechanism of action of antibacterial polymers also makes them a perfect fit for medical devices. Despite great growing needs, the design of new antibacterial polymer with desired antimicrobial properties is still challenging. In this work, we present the first open-source database for antimicrobial polymers which consists of 489 entries, with 177 unique polymers possessing diverse structures and properties. Multiple predictive models were also designed and trained to classify the antimicrobial property of these polymers. The best-performing random forest model showed an averaged accuracy of 86.7% in a 10-fold cross-validation test. We also developed multiple guiding pipelines for the design of novel antimicrobial polymers.

1 Introduction

Infections and diseases caused by different harmful microorganisms including bacterial, fungi and virus have increased significantly over the past 20 years, especially in the fields of medical devices, hospital surfaces, medicine, food packaging and dental equipment[1–5]. The emergence of drug-resistant microorganisms caused by the overuse of antibiotics complicates the situation and increases the pressure on the public health system[6]. Currently existing antibiotics are

unable to effectively kill drug-resistant microorganisms, which makes the treatment of many diseases difficult. Meanwhile, traditional disinfectants are also not effective in eliminating the proliferation of resistant microorganisms spread in the environment[7]. As a result, new effective disinfectants are urgently needed.

In this context, antimicrobial polymers (AMPs) have gained attention from both academia and industry[8–15]. AMPs are a class of polymers that have the ability to kill or inhibit the growth of microbes on their surface or within their surrounding environment[16]. AMPs have antibacterial properties themselves or can be used as a matrix filled with biocides. According to different mechanisms of action, AMPs can be divided into three categories: a) polymers with antibacterial properties themselves, b) polymers with antibacterial properties obtained through chemical modification, and c) polymers that do not have antibacterial properties themselves but can carry biocidal polymers[17]. Due to their antimicrobial properties, AMPs are considered one of the primary candidates for new antibiotics. Unlike traditional low-molecular antibacterial agents, AMPs has no toxicity to the surrounding environment and has a longer service life[18]. More importantly, AMPs destroy the membrane of microorganisms on the surface through electrostatic interactions, the hydrophobic effect and the chelate effect, thus are less likely to cause resistance of microorganisms[19].

Although there is a tremendous need for AMPs from industry and the medical system, the design of new AMPs with desired properties is challenging. This is mainly because the design of new polymers is often guided by intuitive and experimental experience. Such trail-and-error method is time and labour intensive, while unable to stably produce products with certain target properties[20–22]. Moreover, polymers exhibit a higher degree of unpredictability in their structure-property relationships compared to small molecule compounds, primarily attributable to their larger molecular mass and intricate structures[23].

One possible solution is the employment of Machine learning (ML) algorithms. In the last ten years, a rapidly increasing number of ML studies have been reported in material science such as antimicrobial peptides[24,25], ceramic materials[26–28], and nanomaterials[29–31]. These applications have proved the ability of ML to provide precise prediction of material property and generate structures with desired properties, thus can accelerate the material innovation. ML has also had

successful achievements in the field of polymer design. Significant achievements have been made for diverse polymer structures and properties such as polymer electrolytes[32], polymers with high thermal conductivity[33] or charge transfer properties[34], polymer–protein hybrids[35], copolymers[36] and polymer-blend materials[37]. Polymers with desired band gap, glass transition temperature, dielectric constant and other physicochemical properties have also been identified & synthesised with the aid of ML models[38–43]. However, the number of ML applications in the field of AMP is very limited despite the market demand and innovation capabilities[44].

ML is a data-driven method, and the lack of data is the main cause of the slow development of AMP using ML. Existing open-source polymer databases focus on the physical & chemical properties of polymers and no such AMP database has been built for this purpose. To address this problem, we have collected experimental AMP data from multiple resources which can be a critical foundation for the acceleration of AMP design with ML.

In this article, we will introduce this new AMP database and present our ML study focusing on the quantitative structure – activity relationship of AMPs.

2 Database

To ensure the quality of the database, we only collected experimental AMP data reported in peer-reviewed papers. A total of 489 data entries were collected, including 177 unique polymer structures. The properties of these polymers were summarized in a table format to include the following information: polymer structure, molecular weight (Mw), bacteria inhibitory function measurement, and type of bacteria. The data are presented in the Supplementary Information and some polymer structures in the compiled database are demonstrated in **Figure 1**.

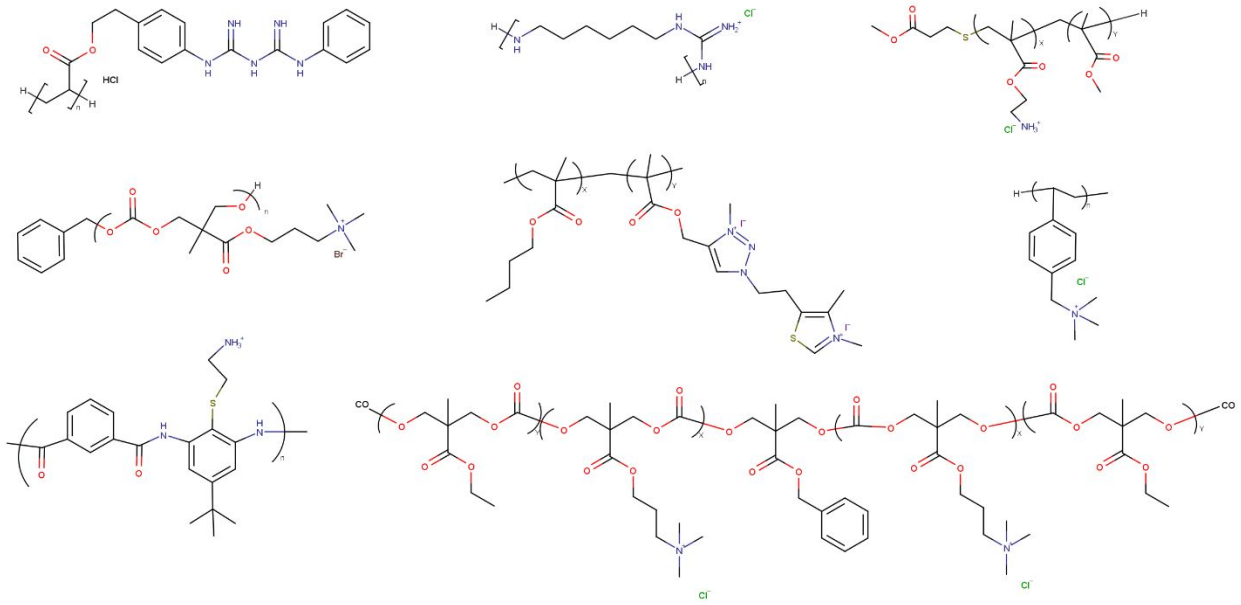


Figure 1. The chemical structure of eight polymer representatives of the database.

Molecular weight **(Mw)** refers to the average mass of the polymer molecules in a sample. It is a crucial parameter in polymer science and engineering because it directly affects the properties and behavior of polymers. In our database, the Mw value ranges from 0.19 to 4461.90 kg/mol, with the majority of the values being smaller than 50kg/mol as illustrated in **Figure 2** which shows the distribution of our data in terms of polymer Mw. In a few cases, the molecules have

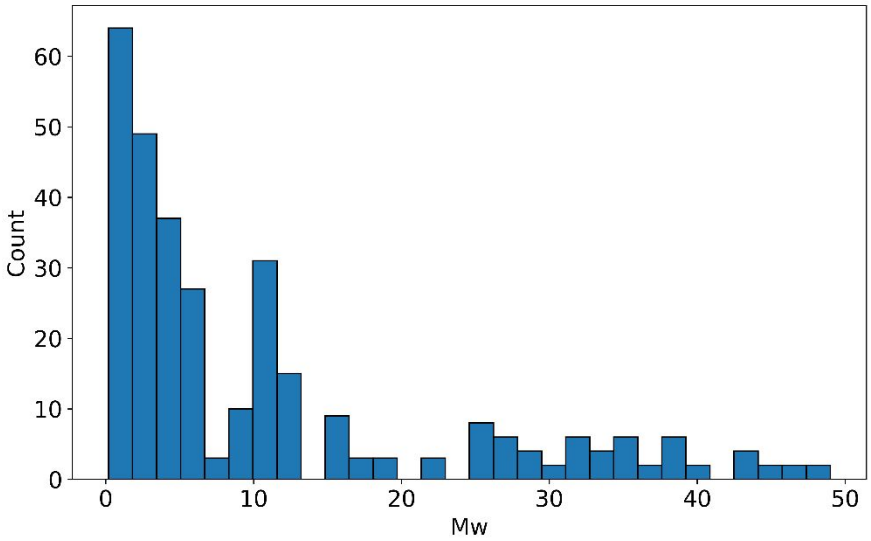


Figure 2. The molecular weight count of polymers in the compiled database.

very small number of repeating units, resulting in low Mw. These small molecules are part of a series where the number of the repeating units gradually increases to a large value.

It is noted that there are many different ways of reporting the antimicrobial performance. In general, the experiments evaluate the antimicrobial performance of the materials by comparing the differences in colonies before and after the application of the antimicrobial agent. Table 1 summarizes the different types of measurement reported in the database.

Table 1. Explanation of four different measurements and their proportion in the database.

Measurement	**Explanation**	**Proportion**
MIC(μg/mL)	Minimum inhibitory concentration (MIC) is defined as the lowest concentration of the antibiotic agent to inhibit the growth of 100% of the targeted microorganism.	76.2%
% of bacterial killed	The percentage of targeting microorganism that is killed.	14.5%
MIC90(μg/mL)	Minimum inhibitory concentration (MIC) is defined as the lowest concentration of the antibiotic agent to inhibit the growth of 90% of the targeted microorganism.	5.3%
Log reduction	$Log\ reduction = \log_{10}(\frac{c_i}{c_j})$ where c_i/c_j is the initial/final concentration. Log reduction of 1 equal to 90%, 2 equals to 99%, 3 equals to 99.9% and so on.	4%

The bacterial target of antimicrobial polymers is crucial. AMPs have different antimicrobial capabilities against different microorganisms. Based on different purposes, AMP can be designed to resist most microorganisms or have a high killing rate for certain types of microorganisms. In our database, the most commonly tested microorganisms are included,

such as E. coli, S. aureus and B. subtilis. The total number of data entries in the database against different targeting microorganisms are shown in **Figure 3**.

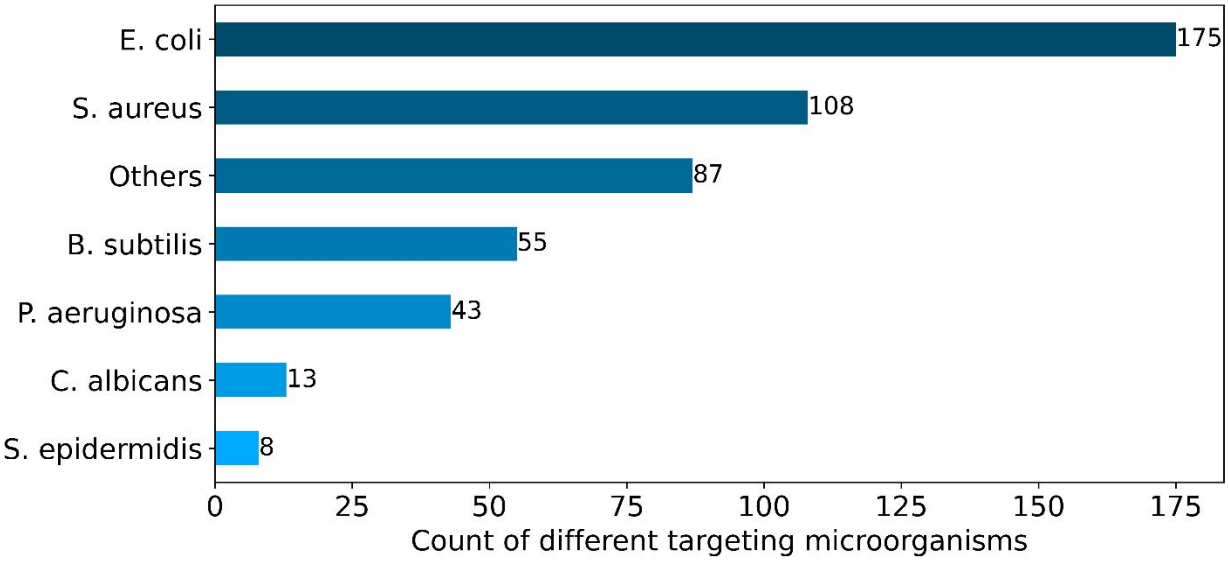


Figure 3. The total number of data entries against different targets in the compiled database.

3 Method

Our three-phase workflow for developing predictive ML models is summarized in **Figure 4**.

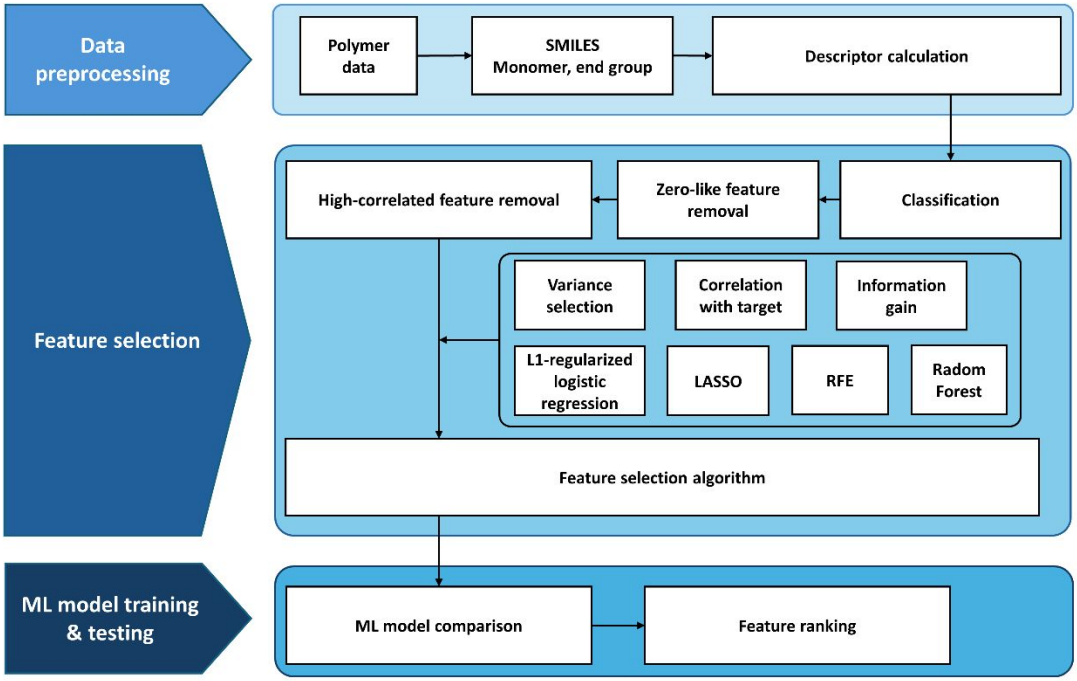


Figure 4. The three-phase machine learning workflow utilized in the study.

Preprocessing of the dataset. Polymers with MIC value of 200 μg/mL or less are categorized as active. Based on this threshold, the dataset was divided into a subset of 365 active polymers and a subset of 124 non-active polymers.

Polymer structure representation. As the molecular weight of AMPs is high and their structures contain many repeating units, it is unnecessary to calculate the descriptors for the whole molecule. It is a common approach to represent the whole polymer molecule using its repeating unit (monomer), and the end groups. In this study, the structure of each polymer is represented by its monomer and the two end groups and saved in the .mol file, one file format that stores information about the atoms, bonds and other information of a molecule. The corresponding files can be accessed through the supporting documents.

Descriptor generation. For simplicity, we used simplified molecular-input line-entry system (SMILES) to represent the repeating units and end groups[45]. It is worth noting that the structures are C-capped to allow for the calculation of descriptors.

In total, over 5666 polymer descriptors including constitutional, topological, geometrical and other descriptors have been computed based on the SMILES strings of the AMPs using the AlvaDesc software[46]. For polymers with multiple repeating units, the descriptors for each repeating unit were calculated independently and summed up based on their weight in the polymer. Additionally, for both end groups, 50 constitutional descriptors were calculated separately and within tandem with the descriptors for the repeating units. The molecular weight Mw (Kg/mol) reported in the original papers was also added as one descriptor. As a result, a total of 5767 (5666 + 50*2 +1) descriptors were calculated for each polymer.

Feature selection. To reduce the dimensionality of the dataset, firstly, descriptors with over 90% of the entries as zeroes were removed. These features were regarded as carrying unsignificant information. Then descriptors with very high correlations were then excluded. The correlation analysis was done by calculating the correlation matrix of features and, for each features pairs with a correlation of 0.95 or higher, the second feature is removed. Highly correlated features tend to present similar information and provide non-additional insights. To further reduce the number of features, we innovatively applied 7 feature selection algorithms that cover the main methods used in past studies, including variance selection, correlation with target, information gain, L1-regularized logistic regression, least absolute shrinkage and selection operator (LASSO), Random forest and recursive feature elimination (RFE). These algorithms covers the three main categories of feature selection methods: filter, embedded and wrapper method. Each algorithm was set to select features to a specific number. Features selected by most of the algorithms were finally selected.

Model Implementation. Several classification ML models were implemented using the scikit-learn[47] libraries and Python 3.10.6: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-nearest Neighbor, Naïve Bayes and Gradient Boosting[47].

- Logistic Regression[48] is a statistical model that estimates the probability of a binary outcome based on one or more predictor variables, utilizing a logistic function to make predictions.

- Decision Tree[49] is a hierarchical structure where nodes represent features, branches represent decisions, and leaves represent outcomes, serving as a versatile tool for both classification and regression tasks.
- Random Forest[50] is an ensemble learning method that constructs multiple decision trees during training and aggregates their predictions, reducing overfitting and improving accuracy by combining the results of multiple models.
- Support Vector Machine[51] (SVM) is a supervised learning algorithm that finds the optimal hyperplane to separate data into different classes, relying on a subset of training data points called support vectors to define the decision boundary.
- k-nearest Neighbor (KNN)[52] is a non-parametric algorithm that classifies new data points based on the majority class of their K closest neighbors, making predictions by identifying the most similar instances in the feature space.
- Naïve Bayes[53] is a probabilistic classifier that applies Bayes' theorem with the assumption of feature independence, making it computationally efficient and particularly well-suited for text classification tasks.
- Gradient Boosting[54] is an ensemble technique where weak learners, typically decision trees, are added sequentially to correct errors made by the previous models, resulting in a powerful predictive model with high accuracy.
- eXtreme Gradient Boosting[55] (XGBoost) is a powerful, scalable machine learning library for gradient boosting, optimized for speed and performance. It supports various objective functions, efficient handling of missing data, and parallel processing. XGBoost is widely used for classification and regression tasks, especially for small datasets.
- Artificial neural network (ANN) consists of interconnected layers of neurons whose weights are adjusted during training to learn patterns from data. Through multiple training runs and evaluations, ANNs can achieve high accuracy and adaptability across diverse applications[53].

Models obtained using these methods were then optimized through a Grid Search algorithm, together with a 10-fold cross-validation (GSCV) algorithm for evaluation.

4 Results and discussion

4.1 Identification of significant features

From the large of descriptors, zero-like and highly correlated features were excluded, resulting in a set of 494 features remaining. The number of descriptors selected by each of the seven feature selection algorithms, as shown in Table 2, was set to 100 ± 10. The main reason for this approach is to avoid the bias of each algorithm.

Table 2. Seven algorithms applied for feature selection.

Algorithm	**Number of features selected**	**Category**
Variance selection	100	Filter method
Correlation with target	100	Filter method
Information gain	95	Filter method
L1-regularized logistic regression	103	Embedded method
LASSO	100	Filter method
Radom Forest	94	Filter method
RFE	100	Wrapper method

Subsequently, features selected by the seven algorithms were divided into 7 groups based on the numbers of algorithm identifying them as relevant descriptors, as shown in **Figure 5**. Three descriptors chosen by all seven algorithms were 'P_VSA_m_2', 'P_VSA_ppp_L' and

'P_VSA_charge_7'. Seven descriptors chosen by six different algorithms were 'P_VSA_LogP_2', 'P_VSA_LogP_6', 'P_VSA_MR_6', 'P_VSA_s_4', 'P_VSA_charge_9', 'CATS 2D_05_LL' and 'AMR'.

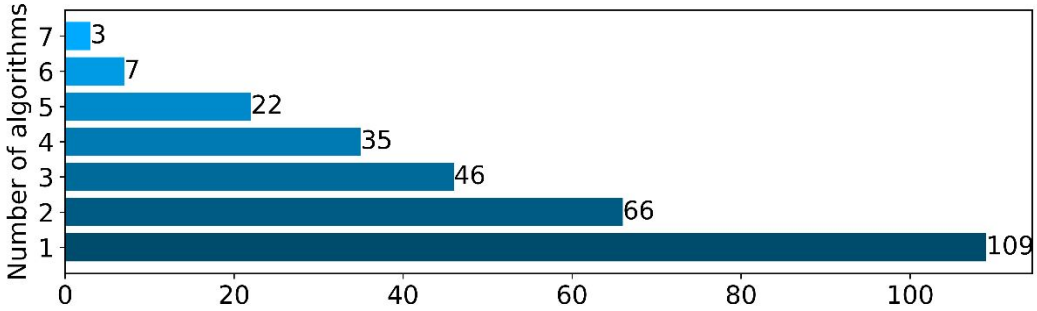


Figure 5. Count of descriptors selected by different features selection algorithms.

In this study, 32 descriptors selected by five or more algorithms were used to form the final set of input descriptors for the ML models. Most of descriptors are highly independent to each

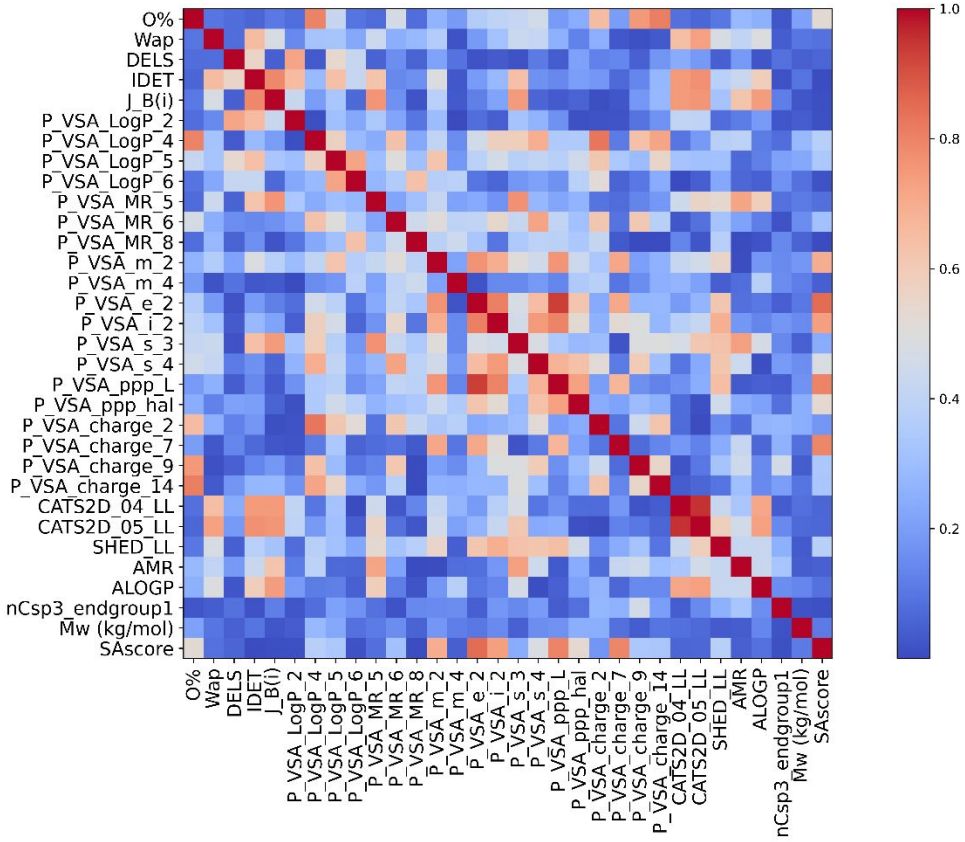


Figure 6. Heatmap of correlation coefficient between descriptors selected. Red color indicates the high correlation coefficient values, blue otherwise.

other and their relationship are visualized in a heat map showing the correlation coefficients (see **Figure 6)**.

Herein, we provide a detailed explanation of the 32 selected descriptors:

Mw (kg/mol)

Polymers are materials with large molecular weights and not surprisingly, this molecular weight descriptor was found to play an important role in affecting the antimicrobial property of such materials.

P_VSA-like

P_VSA-like descriptors are based on the sum of atomic contributions to van der Waals surface area, based on the atoms having a property in a defined range of values[56]:

$$VSA_i = 4 \cdot \pi \cdot R_i^2 - \pi \cdot R_i \cdot \sum_{j=1}^{nAT} a_{ij} \cdot \left(\frac{R_i^2 - (R_i - g_{ij})^2}{g_{ij}}\right) \quad (1)$$

$$g_{ij} = min\{max\{|R_i - R_j|, b_{ij}\}, (R_i + R_j)\} \quad (2)$$

$$b_{ij} = r_{ij} - c_{ij} \quad (3)$$

$$P_{VSA\_w\_k} = \sum_{i=1}^{nAT} VSA_i \cdot \delta(w_i \epsilon [a_{k-1}, a_k))\ k = 1,2,\ldots n \quad (4)$$

In equation (1), VSA_i is the van der Waals surface area of the *i-th* atom, R_i is the atomic van der Waals radius of the atom i, nAT is the number of atoms, and a_{ij} are the elements of the adjacency matrix.

g_{ij} is the max value in the max value of $|R_i - R_j|$ and b_{ij} and $(R_i + R_j)$.

b_{ij} denotes the ideal length of the bon formed by atom i and j.

r_{ij} is the reference bond length and c_{ij} is a correlation term related to the bond multiplicity: 0 for single bond, 0.1 for aromatic, 0.2 for double, and 0.3 for triple bonds. R_i and c_{ij} are pre-defined value, which can be accessed online at https://www.talete.mi.it/help/dproperties_help/index.html?p_vsa_like_descriptors.htm.

Based on equations 1, 2, and 3, P_VSA-like descriptors can be computed using equation (4) in which δ is a Kronecker delta function which equals to one for atoms with property value in the specified range, and zero otherwise. w_i denotes one of the weighting schemes including: 'logP' for log P (octanol/water), 'MR' for molar refractivity, 'm' for atomic mass, 'e' for Sanderson electronegativity, 'ppp' for potential pharmacophore points and 'charge' for partial charge. k is the bin number, indicating a pre-defined range.

O%

This descriptor calculates the relative occurrence frequency of O atom.

ALOGP

ALOGP is the Ghose-Crippen-Viswanadhan octanol-water partition coefficient, defined as:

$$LOGP = \sum_i n_i \cdot h_i$$

where n_i is the atom of type i and h_i is the corresponding hydrophobicity contribution[57].

CATS 2D

CATS represents Chemically Advanced Template Search descriptors. CATS 2D descriptors are a particular case of autocorrelation descriptors. They are defined as:

$$CATS2D_k(u,v) = \frac{1}{2} \cdot \sum_{i=1}^{|V|} \sum_{j=1}^{|V|} \delta(i;u) \cdot \delta(j;v) \cdot \delta(d_{ij};k)$$

where u,v represent two atom types, $\delta(i;u)$, $\delta(j;v)$ and $\delta(d_{ij};k)$ are three Kronecker delta functions equal to one if atom i is of type u, atom j is of type v, and the topological distance d_{ij} is equal to k, respectively, zero otherwise. In CATS 2D descriptors, the atom-type definition is related to the concept of potential pharmacophore points (PPP). PPP is a generalized atom type defined considering some physicochemical aspects. CATS 2D descriptors are calculated based on a topological distance varying from 0 to 9 and any atom of the molecule can be assigned to none, one, or two atom types (DD, DA, DP, DN, DL, AA, AP, AN, AL, PP, PN, PL, NN, NL, LL) resulting in a vector of 150 frequencies.

DELS

The molecular electro topological variation (DELS) is calculated as:

$$DELS = \sum_{i}^{nSK} |\Delta_i|$$

where nSK is the number of non-H atoms. DELS index could be considered as a measure of the total charge transfer in the molecule[58].

J_B(i)

This descriptor is a Balaban-like index from the Burden matrix weighted by ionization potential. It is also highly correlated with multiple 2D matrix-based descriptors extracted from matrixes such as adjacency, topological distance, Laplace and Chi.

nCsp3_endgroup1

nCsp3 stands for the number of sp^3 hybridized carbons. This descriptor indicates the count of sp^3 hybridized carbons in one of the end groups (end group 1). This descriptor is highly correlated with the followings: Se (sum of atomic Sanderson electronegativities, scaled on Carbon atom), Si (sum of first ionization potentials, scaled on Carbon atom), nAT (total number of atoms), RBN (number of rotatable bonds) and nH (number of Hydrogen atoms).

4.2 Modelling results

To ensure the acquisition of generalized ML models, ten separate runs were performed for each ML algorithm and the average of accuracies was used to evaluate the performance of the obtained ML models. The variance of ten accuracies was also calculated as a metric to quantify the spread or dispersion of the accuracy values. Six out of seven models achieved an average accuracy of around 85% and only the Naïve Bayes model had a low accuracy of 67.2%, as shown in **Table 3**. The receiver – operating characteristic (ROC) curves for all algorithms are provided

in **Figure S2** of the Supplementary Information. For the dataset in this work, Random Forest method is among those that achieved the highest average accuracy while also has the lowest variance, indicating its excellent predictive capability and a steady performance across different data sets.

Table 3. Statistical results of different ML models for antimicrobial polymers.

ML model	Average accuracy	Variance
Logistic Regression	0.850	0.0015
Decision Tree	0.858	0.0012
Random Forest	0.877	0.0005
Support Vector Machine	0.875	0.0010
k-nearest Neighbor	0.836	0.0026
Naïve Bayes	0.666	0.0047
Gradient Boosting	0.863	0.0009
XGBoost	0.867	0.0008
ANN	0.854	0.0015

4.3 Antimicrobial design principles

Since the trained models can provide accurate predictions as well as reveal the relationships between descriptors and polymer properties, we analysed the modelling results and determine guidelines for the design of new AMPs.

4.3.1 Logistic Regression

Logistic regression is a statistical model used for binary classification tasks, predicting the probability of an event occurring based on input features. The weights in logistic regression represent the strength and direction of the relationship between each feature and the log-odds

of the event happening, allowing us to interpret the impact of each feature on the antimicrobial property.

As shown in **Figure 7**, a number of descriptors such as 'P_VSA_charge_2','P_VSA_charge_7', and

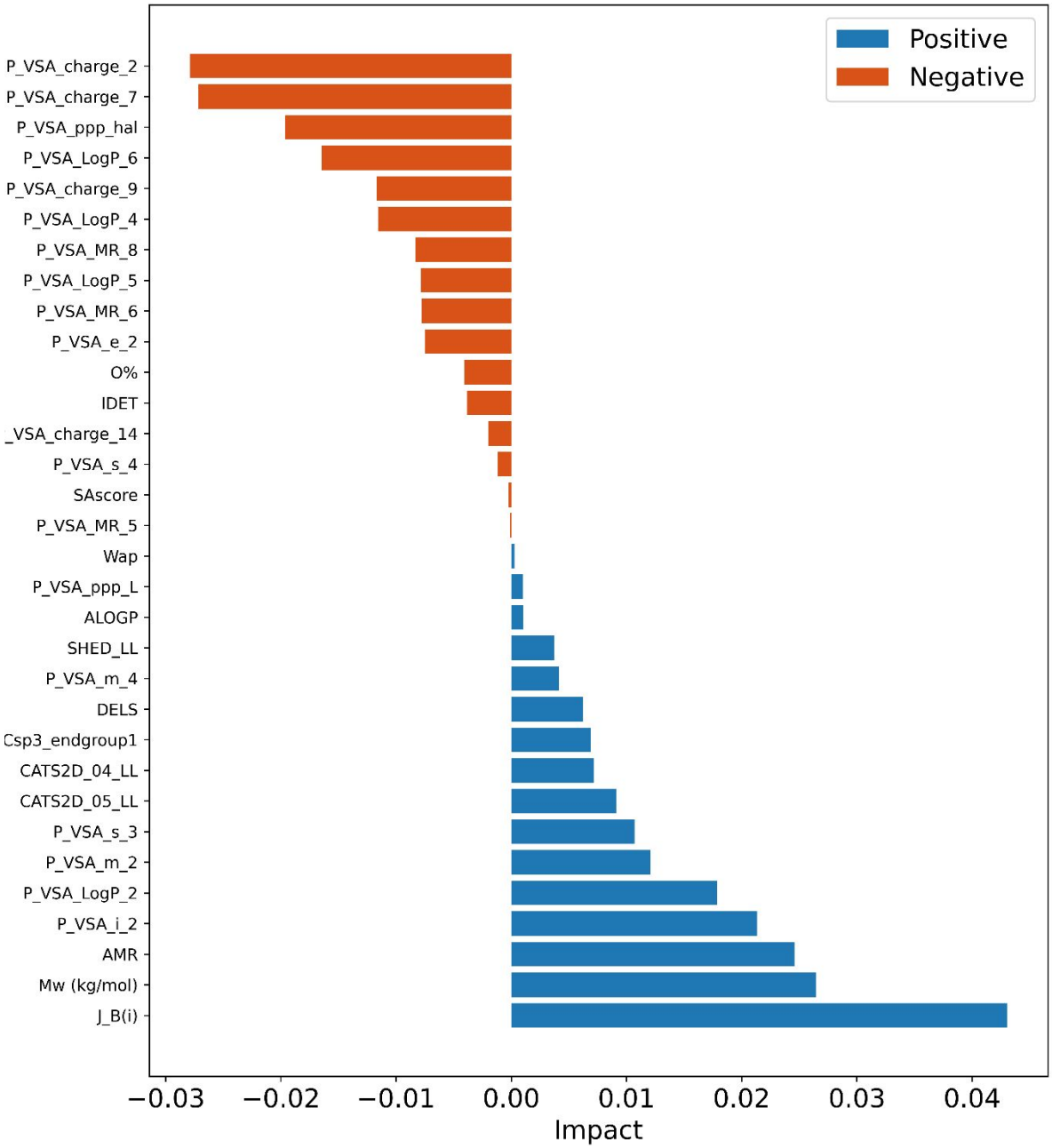


Figure 7. Impact of each descriptor to antimicrobial property.

'P_VSA_ppp_hal' showed a negative impact on the antimicrobial property. When the value of these descriptors decreases, the antimicrobial property tends to be stronger. On the other side,

the increase of the descriptors with positive impact such as 'J_B(i)', ' Mw (kg/mol)', 'AMR', and 'P_VSA_i_2' would contribute to higher antimicrobial activity.

4.3.2 Decision Tree

Decision tree model is a popular machine learning algorithm that mimics human decision-making by partitioning the data into subsets based on feature values. It iteratively selects the best features to split the data and creates a tree-like structure of decisions, making it interpretable and easy to understand. The decision tree obtained from this study is shown in **Figure S1**. Based on the decision boundary and classification accuracy, we can summarize the conditions that the descriptors need to meet to result in higher probability of active antimicrobial activities. It is worth noting that the guidelines consider multiple descriptors and will only work when the conditions for all of these descriptors are met.

Figures 8 to 10 illustrate different parts of the decision tree (subtrees 1 to 3) separately for clarification, and the classification is highlighted.

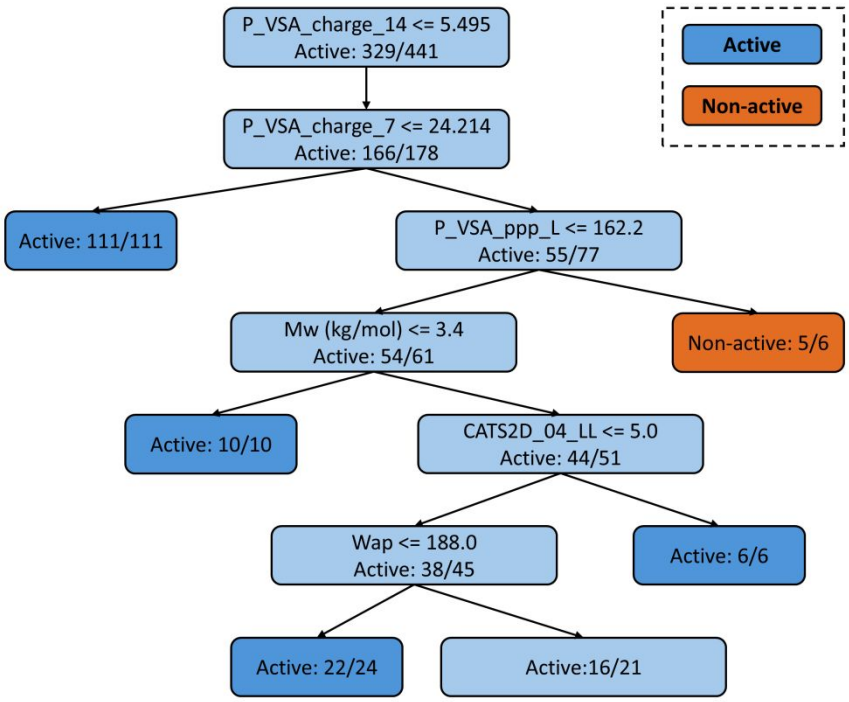


Figure 2. Decision boundary from decision tree (subtree 1).

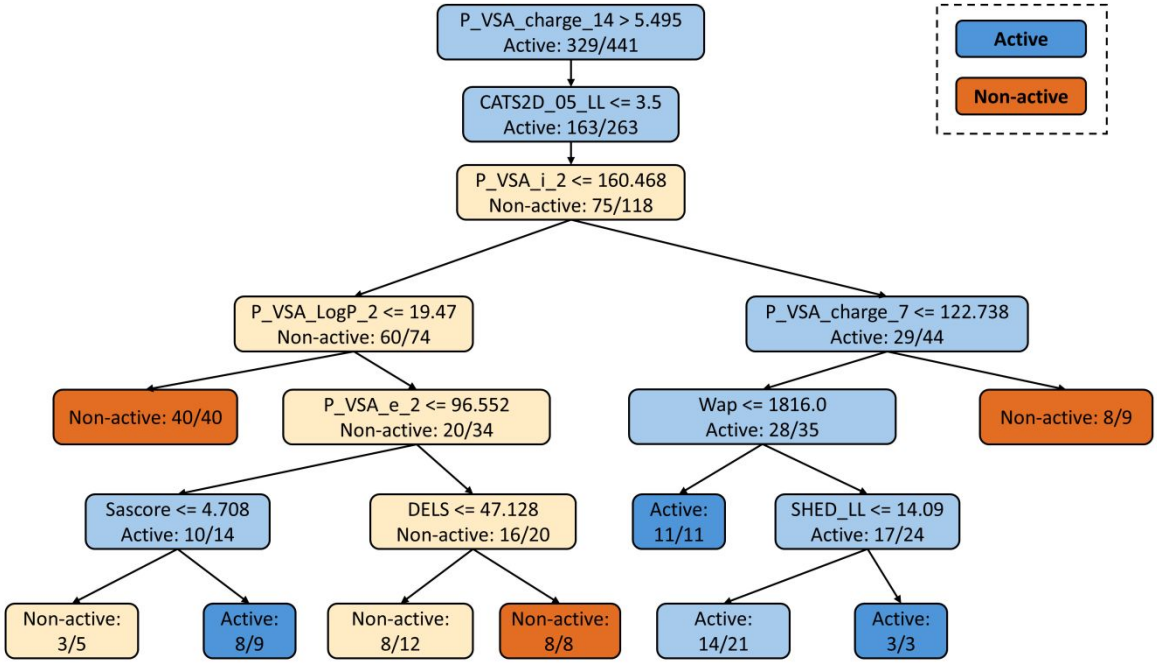


Figure 3. Decision boundary from decision tree (subtree 2).

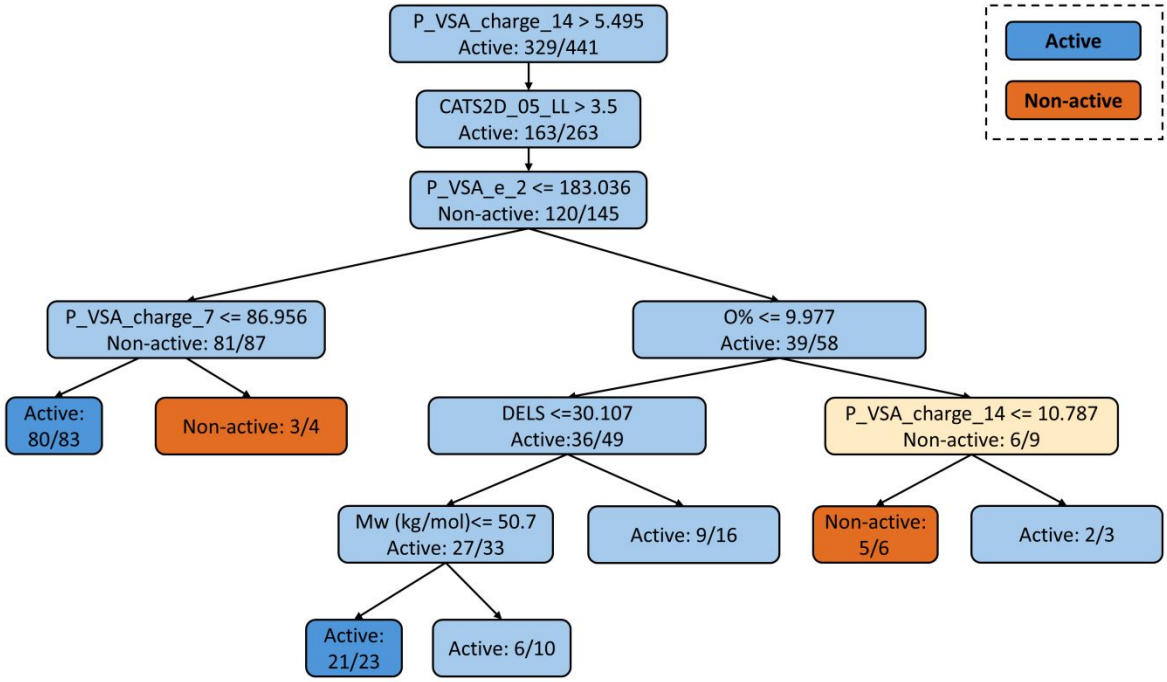


Figure 4. Decision boundary from decision tree (subtree 3).

In the decision tree, each node corresponds to a set of polymers. When the majority of polymers in each node (more than 50%) are active, the node has a blue color, otherwise yellow. The number at the bottom of the node shows the exact number of active/non-active polymers in the node. For example, in the first node of subtree 1 (**Figure 8**), there are 329 active polymers out of a total number of 441 polymers. The number at the top of the node shows the condition for the descriptor values. Polymers satisfying the condition will be transferred to the lower left node of the subtree and otherwise to lower right one. When most of the polymer in a node are classified correctly, there will be no subsequent conditions. Following the different branches of conditions, we can find different sets of conditions for active and non-active polymers.

As shown in subtree 1 (**Figure 8**), there are 111 active polymers which have P_VSA_charge_14 of no more than 5.495 and P_VSA_charge_7 of no more than 24.214. In contrast, polymers with P_VSA_charge_7 of greater than 24.214 and P_VSA_ppp_L of greater than 162.2 are suggested to be non-active and should be avoided in the design process. The subtree 1 also shows that polymers with P_VSA_charge_14 of no more than 5.495, but P_VSA_charge_7 and P_VSA_pp_L of greater than 24.214 and 162.2 respectively are more likely to be non-active.

Going down the subtree, it can also be seen that most of polymers (54/61 or 88.5%) with P_VSA_pp_L of no greater than 162.2 are classified as active. And all of the remaining 10 polymers with Mw (kg/mol) smaller than 3.4 are also active. For polymers with larger Mw, they could still be active if they have CATS2D_04_LL of smaller than 5.0 and Wap of smaller than 188.

Following the same pattern, we can also find conditions for designing active polymers using subtrees 2 and 3. We summerise all the conditions for active polymers or non-active polymers in **Table 4**.

Table 4. Conditions for designing antimicrobial active and non-active polymers using decision tree algorithm.

	Condition 1: P_VSA_charge_14 <= 5.495 & P_VSA_charge_7 <= 24.214

Conditions for active polymers	Condition 2: P_VSA_charge_14 <= 5.495 & P_VSA_charge_7 > 24.214 & P_VSA_pp_L <= 162.2 & Mw (kg/mol) <= 3.4
	Condition 3: P_VSA_charge_14 <= 5.495 & P_VSA_charge_7 > 24.214 & P_VSA_pp_L <= 162.2 & Mw (kg/mol) > 3.4 & CATS2D_04_LL <= 5.0 & Wap <= 188
	Condition 4: P_VSA_charge_14 > 5.495 & CATS2D_05_LL <= 3.5 & P_VSA_i_2 > 160.468 & P_VSA_charge_7 <= 112.738 & Wap <= 1816
	Condition 5: P_VSA_charge_14 > 5.495 & CATS2D_05_LL <= 3.5 & P_VSA_e_2 <= 183.036 & P_VSA_charge_7 <= 86.956
	Condition 6: P_VSA_charge_14 > 5.495 & CATS2D_05_LL <= 3.5 & P_VSA_e_2 > 183.036 & O% <= 9.997 & DELS <= 30.107 & Mw (kg/mol) <= 50.7
Conditions for non-active polymers	Condition 1: P_VSA_charge_14 <= 5.495 & P_VSA_charge_7 > 24.214 & P_VSA_pp_L > 162.2
	Condition 2: P_VSA_charge_14 > 5.495 & CATS2D_05_LL <= 3.5 & P_VSA_i_2 <= 160.468 & P_VSA_LogP_2 <= 19.47
	Condition *3: P_VSA_charge_14 > 5.495 & CATS2D_05_LL <= 3.5 & P_VSA_i_2 <= 160.468 & P_VSA_LogP_2 > 19.47 & P_VSA_e_2 > 96.552 & DELS > 47.128*
	Condition 4: P_VSA_charge_14 > 5.495 & CATS2D_05_LL <= 3.5 & P_VSA_i_2 > 160.468 & P_VSA_charge_7 > 112.738
	Condition 5: P_VSA_charge_14 > 5.495 & CATS2D_05_LL <= 3.5 & P_VSA_e_2 <= 183.036 & P_VSA_charge_7 > 86.956
	Condition 6: P_VSA_charge_14 > 5.495 & CATS2D_05_LL <= 3.5 & P_VSA_e_2 > 183.036 & O% > 9.997 & P_VSA_charge_14 <= 10.787

4.3.3. Random Forest

Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and outputs the mode of the classes for classification or the average prediction for regression[59]. Different from a single decision tree, Random Forest mitigates overfitting and

enhances accuracy by combining predictions from multiple decision trees trained on bootstrapped samples and using random subsets of features for each tree. Based on the best-performing Random Forest model, the influence of each descriptor on the polymers' antimicrobial property can be quantified and ranked. These top-ranking descriptors could be critical factors to the antimicrobial properties and can assist with future design of AMPs. In **Table 5**, we presented seven most relevant descriptors that could be used as a guideline for AMP design. The full list can be found in Table S1 of the Supplementary Information.

Table 5. Top ranking descriptors by Random Forest algorithm and their feature importance values.

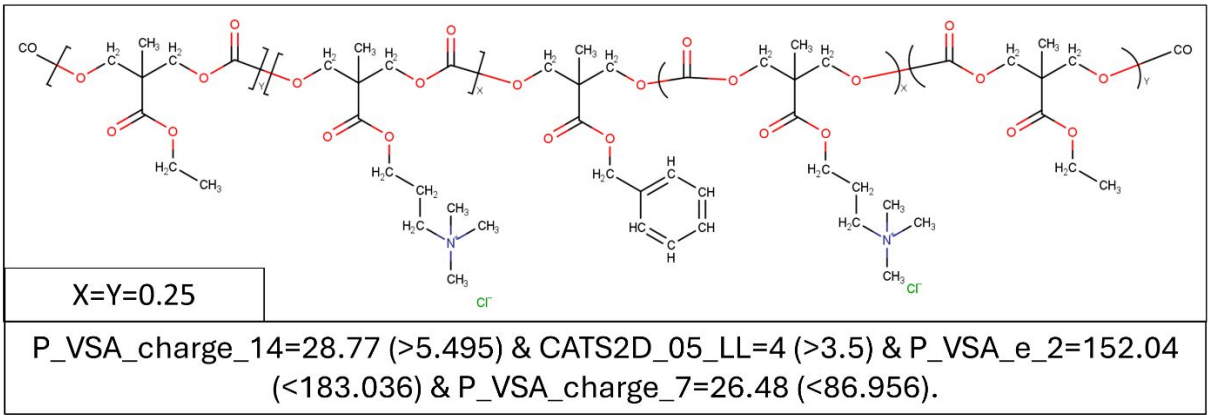
Descriptor	**Importance**	**Description**	**Type**
Mw (kg/mol)	0.121	Molecular weight	Constitutional indices
P_VSA_LogP_2	0.060	P_VSA-like on logP, bin 2	P_VSA-like descriptors
O%	0.043	percentage of O atoms	Constitutional indices
ALOGP	0.036	Ghose-Crippen octanol-water partition coefficient (logP)	Molecular properties
CATS 2D_05_LL	0.034	CATS 2D Lipophilic-Lipophilic at lag 05	Pharmacophore descriptors
DELS	0.033	Molecular electro topological variation	Topological indices
J_B(i)	0.013	Balaban-like index from Burden matrix weighted by ionization potential	2D matrix-based descriptors

4.4. Evaluation of the Design Principles

As presented above, the relationship between the polymer structure and the antimicrobial activity is complex and multidimensional. Inverse engineering requires careful consideration of various factors. In this section, we will illustrate how the design principles obtained using different ML techniques can be reflected in the compiled database using four polymers, two of which have very low MIC (high activity) (polymers 1 and 2, **Figure 11**) and 2 with very high MIC (low activity) (polymers 3 and 4, **Figure 13**).

As shown in **Figure 12**, polymer 1 satisfies condition 3 of the decision tree guidelines while polymer 2 satisfies condition 4. These are classified by the model to be antimicrobial-active. It should also be noted that although the structure of polymer 2 is similar to polymer 1, polymer 2 has higher Mw and O% but lower CATS2D_05_LL, SHED_LL, P_VSA_LogP_2 and ALOGP. Logistic regression and random forest algorithms ranking suggest that polymer 2 has a higher MIC value than polymer 1.

Similarly, as shown in Fig. 14, polymer 3 satisfies the non-active condition 1, while polymer 4 satisfies the non-active condition 2. They are classified by the decision tree model to be non-active polymers.



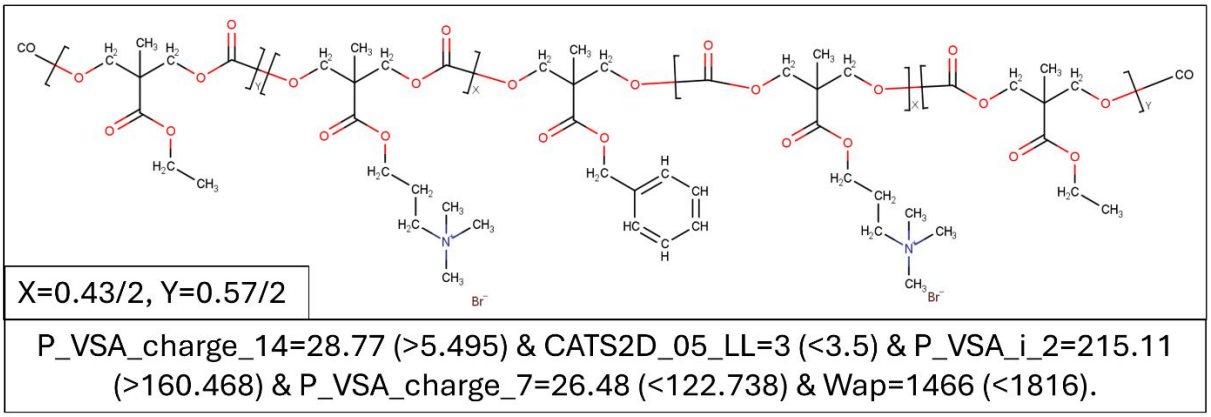


Figure 5. Chemical structures and descriptor values of two most active polymers in the database (polymer 1 (top) and polymer 2 (bottom)).

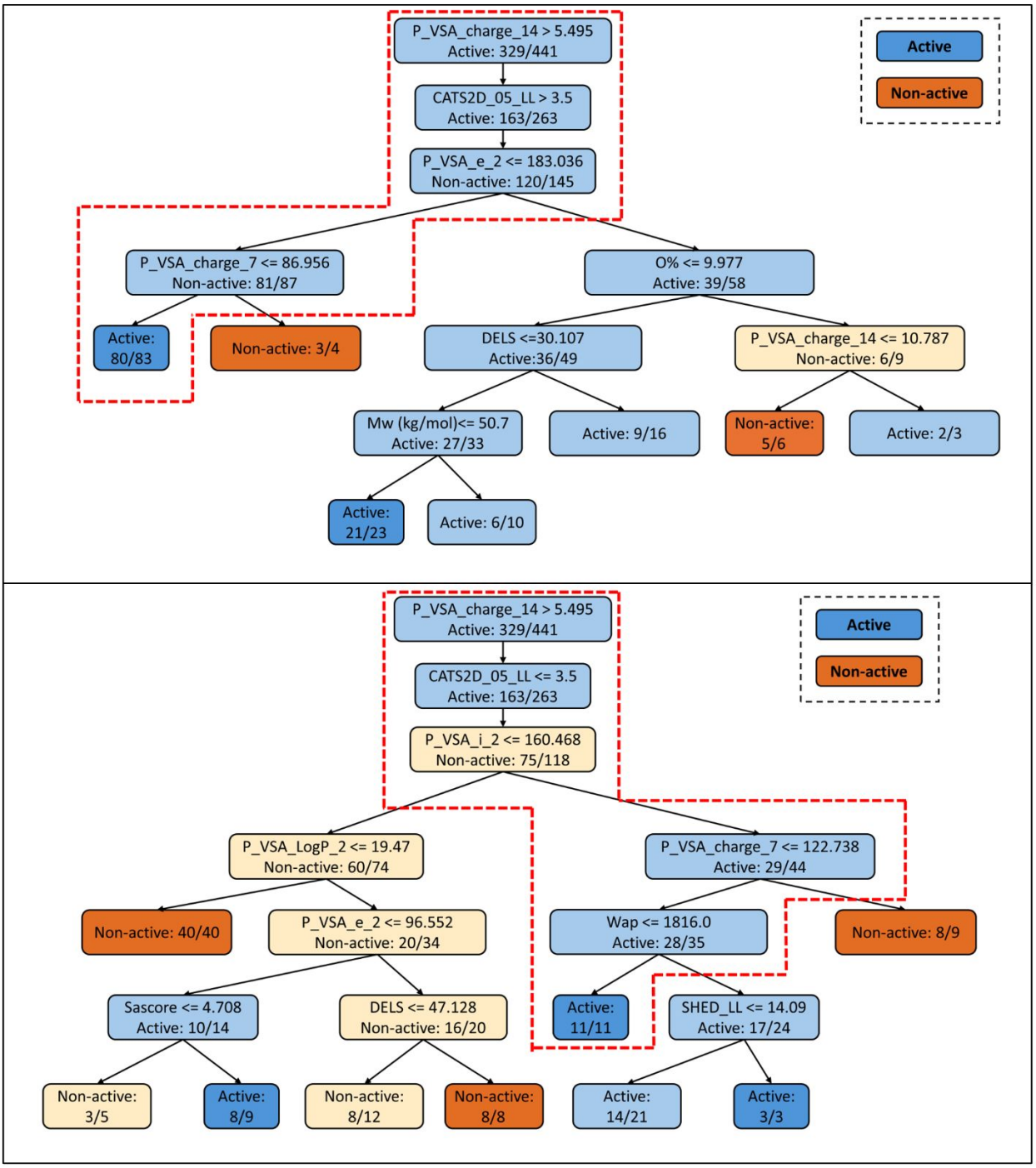
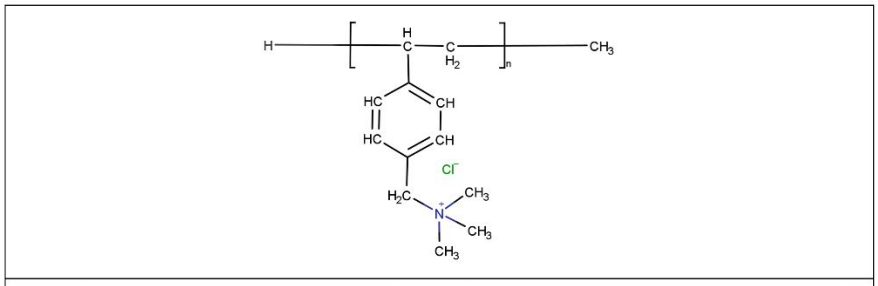
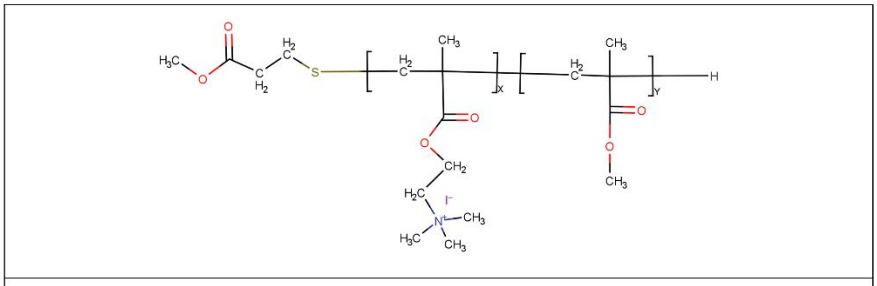


Figure 6. Decision tree active conditions that are met by polymers 1 (top) and 2 (bottom).



P_VSA_charge_14=0 (<5.495) & P_VSA_charge_7=72.37 (>24.214) & P_VSA_ppp_L=165.96 (>162.2)



P_VSA_charge_14=14.38 (>5.495) & CATS2D_05_LL=0 (<3.5) & P_VSA_i_2=115.17 (<160.468) & P_VSA_LogP_2=11.77 (<19.47)

Figure 7. Chemical structures and descriptor values of two least active polymers in the database (polymer 3 (top) and polymer 4 (bottom)).

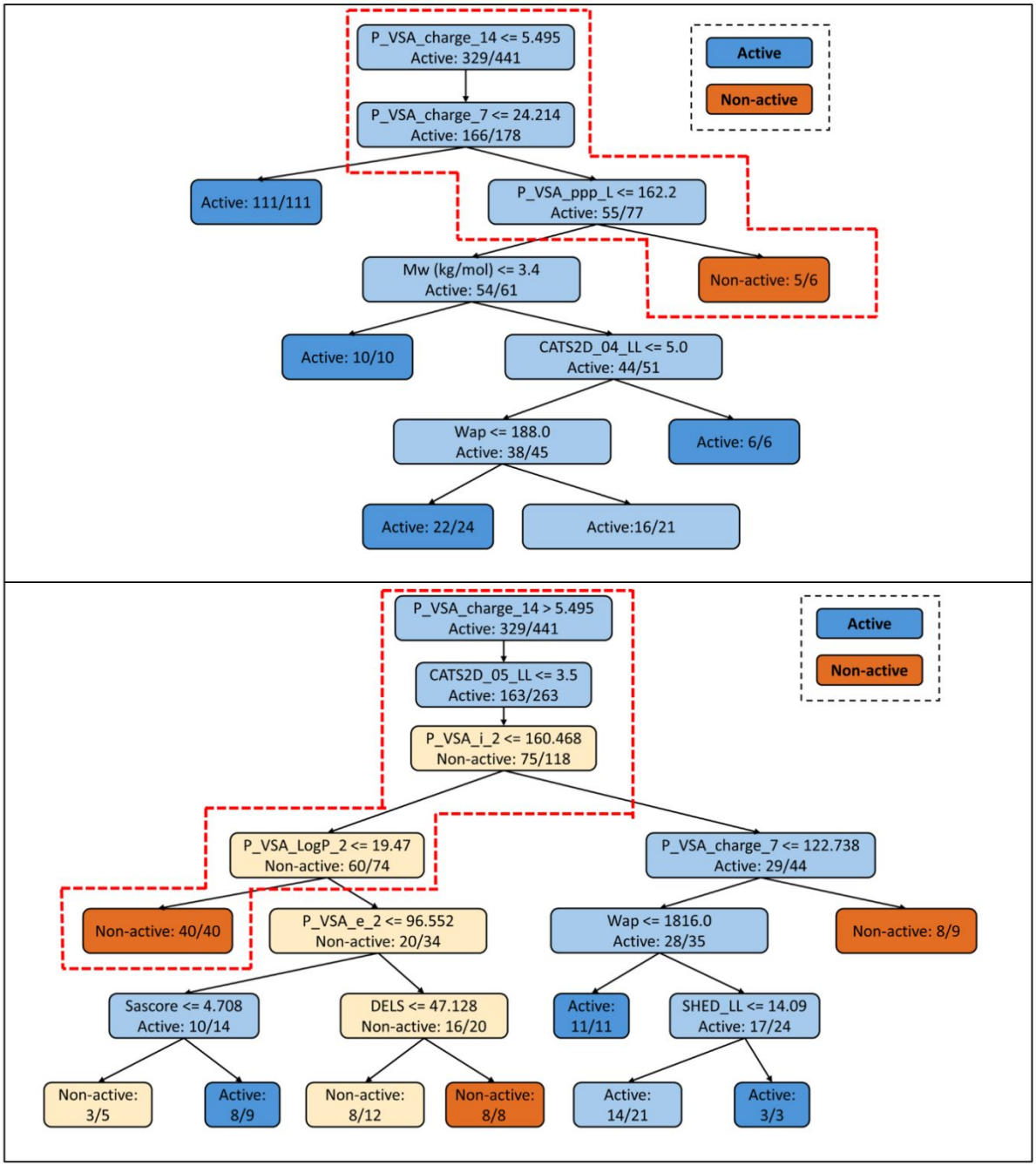


Figure 8. Decision tree non-active conditions that are met by polymers 3 (top) and 4 (bottom).

Conclusion

We have compiled and presented the first AMP database with experimental data from multiple peer-reviewed articles. This database is expected to contribute to the rapid, data-driven development of AMP for both research advancement and industry applications. We applied

innovative algorithms and identified 32 significant descriptors based on 7 different feature selection approaches. We also proposed multiple ML models with high predictive accuracy (around 85%) for antimicrobial properties. Furthermore, we determined the impact and importance of descriptors affecting the antimicrobial property of polymers. A guideline was proposed for the design of highly active AMPs. We hope this database and the discovery of influential descriptors can provide a solid and informative foundation for researchers in the field to explore new AMPs in the future.

Author Contributions

Yuankai Zhao: Conceptualization, investigation, data curation, visualization, coding, model development, writing the original draft.

Shadi Houshyara: reviewing & editing, supervision.

Roger J Mulder: data validation, reviewing & editing, supervision.

Eyckens Dan: data validation, reviewing & editing

Tu C. Le: Conceptualization, data validation, writing, review & editing, project administration, supervision.

Conflicts of interest

The authors declare no conflicts of interest.

Acknowledgements

Yuankai Zhao acknowledges the CSIRO-RMIT scholarship program.

References

1 T. U. Berendonk, C. M. Manaia, C. Merlin, D. Fatta-Kassinos, E. Cytryn, F. Walsh, H. Bürgmann, H. Sørum, M. Norström, M.-N. Pons, N. Kreuzinger, P. Huovinen, S. Stefani, T. Schwartz, V. Kisand, F. Baquero and J. L. Martinez, Tackling antibiotic resistance: the environmental framework, *Nat Rev Microbiol*, 2015, **13**, 310–317.

2 F. Siedenbiedel and J. C. Tiller, Antimicrobial polymers in solution and on surfaces: overview and functional principles, *Polymers*, 2012, **4**, 46–71.

3 H. M. Lode, Clinical impact of antibiotic-resistant Gram-positive pathogens, *Clinical Microbiology and Infection*, 2009, **15**, 212–217.

4 M. J. Garcia-Fernandez, L. Martinez-Calvo, J. Ruiz, M. R. Wertheimer, A. Concheiro and C. Alvarez-Lorenzo, Loading and Release of Drugs from Oxygen-rich Plasma Polymer Coatings, *Plasma Processes & Polymers*, 2012, **9**, 540–549.

5 S. I. Hay, P. C. Rao, C. Dolecek, N. P. J. Day, A. Stergachis, A. D. Lopez and C. J. L. Murray, Measuring and mapping the global burden of antimicrobial resistance, *BMC Med*, 2018, **16**, 78.

6 A. Muñoz-Bonilla and M. Fernández-García, Polymeric materials with antimicrobial activity, *Progress in Polymer Science*, 2012, **37**, 281–339.

7 H. Takahashi, I. Sovadinova, K. Yasuhara, S. Vemparala, G. A. Caputo and K. Kuroda, Biomimetic antimicrobial polymers—Design, characterization, antimicrobial, and novel applications, *WIREs Nanomed Nanobiotechnol*, 2023, **15**, e1866.

8 P. Pham, S. Oliver and C. Boyer, Design of antimicrobial polymers, *Macromolecular Chemistry and Physics*, **n/a**, 2200226.

9 M. Haktaniyan and M. Bradley, Polymers showing intrinsic antimicrobial activity, *Chem. Soc. Rev.*, 2022, **51**, 8584–8611.

10 K.-S. Huang, C.-H. Yang, S.-L. Huang, C.-Y. Chen, Y.-Y. Lu and Y.-S. Lin, Recent advances in antimicrobial polymers: a mini-review, *IJMS*, 2016, **17**, 1578.

11 M. R. E. Santos, A. C. Fonseca, P. V. Mendonça, R. Branco, A. C. Serra, P. V. Morais and J. F. J. Coelho, Recent developments in antimicrobial polymers: a review, *Materials*, 2016, **9**, 599.

12 M. M. Konai, B. Bhattacharjee, S. Ghosh and J. Haldar, Recent progress in polymer research to tackle infections and antimicrobial resistance, *Biomacromolecules*, 2018, **19**, 1888–1917.

13 J. R. Smith and D. A. Lamprou, Polymer coatings for biomedical applications: a review, *Transactions of the IMF*, 2014, **92**, 9–19.

14 T. Potta, Z. Zhen, T. S. P. Grandhi, M. D. Christensen, J. Ramos, C. M. Breneman and K. Rege, Discovery of antibiotics-derived polymers for gene delivery using combinatorial synthesis and cheminformatics modeling, *Biomaterials*, 2014, **35**, 1977–1988.

15 G. N. Tew, R. W. Scott, M. L. Klein and W. F. DeGrado, De novo design of antimicrobial polymers, foldamers, and small molecules: from discovery to practical applications, *Acc. Chem. Res.*, 2010, **43**, 30–39.

16 H. B. Kocer, I. Cerkez, S. D. Worley, R. M. Broughton and T. S. Huang, Polymeric Antimicrobial *N* -Halamine Epoxides, *ACS Appl. Mater. Interfaces*, 2011, **3**, 2845–2850.

17 A. Jain, L. S. Duvvuri, S. Farah, N. Beyth, A. J. Domb and W. Khan, Antimicrobial Polymers, *Adv. Healthcare Mater.*, 2014, **3**, 1969–1985.

18 E.-R. Kenawy, S. D. Worley and R. Broughton, The Chemistry and Applications of Antimicrobial Polymers: A State-of-the-Art Review, *Biomacromolecules*, 2007, **8**, 1359–1384.
19 L. Timofeeva and N. Kleshcheva, Antimicrobial polymers: mechanism of action, factors of activity, and applications, *Appl Microbiol Biotechnol*, 2011, **89**, 475–492.
20 D. J. Audus and J. J. de Pablo, Polymer Informatics: Opportunities and challenges, *ACS Macro Lett.*, 2017, **6**, 1078–1082.
21 L. Chen, G. Pilania, R. Batra, T. D. Huan, C. Kim, C. Kuenneth and R. Ramprasad, Polymer informatics: current status and critical next steps, *Materials Science and Engineering: R: Reports*, 2021, **144**, 100595.
22 Danishuddin and A. U. Khan, Descriptors and their selection methods in QSAR analysis: paradigm for drug design, *Drug Discovery Today*, 2016, **21**, 1291–1302.
23 G. E. Wnek, Structure–Property Relationships of Small Organic Molecules as a Prelude to the Teaching of Polymer Science, *J. Chem. Educ.*, 2017, **94**, 1647–1654.
24 G. J. Gabriel, A. E. Madkour, J. M. Dabkowski, C. F. Nelson, K. Nüsslein and G. N. Tew, Synthetic mimic of antimicrobial peptide with nonmembrane-disrupting antibacterial properties, *Biomacromolecules*, 2008, **9**, 2980–2983.
25 J. Xu, F. Li, A. Leier, D. Xiang, H.-H. Shen, T. T. Marquez Lago, J. Li, D.-J. Yu and J. Song, Comprehensive assessment of machine learning-based methods for predicting antimicrobial peptides, *Briefings in Bioinformatics*, 2021, **22**, bbab083.
26 Y.-C. Hsu, C.-H. Yu and M. J. Buehler, Using Deep Learning to Predict Fracture Patterns in Crystalline Solids, *Matter*, 2020, **3**, 197–211.
27 R. W. K. Li, T. W. Chow and J. P. Matinlinna, Ceramic dental biomaterials and CAD/CAM technology: State of the art, *Journal of Prosthodontic Research*, 2014, **58**, 208–216.
28 J. R. Jones, Review of bioactive glass: From Hench to hybrids, *Acta Biomaterialia*, 2013, **9**, 4457–4486.
29 M. J. Kratochvil, A. J. Seymour, T. L. Li, S. P. Paşca, C. J. Kuo and S. C. Heilshorn, Engineered materials for organoid systems, *Nat Rev Mater*, 2019, **4**, 606–622.
30 Q. Liu, S. Zheng, K. Ye, J. He, Y. Shen, S. Cui, J. Huang, Y. Gu and J. Ding, Cell migration regulated by RGD nanospacing and enhanced under moderate cell adhesion on biomaterials, *Biomaterials*, 2020, **263**, 120327.
31 S. Mitragotri and J. Lahann, Physical approaches to biomaterial design, *Nature Mater*, 2009, **8**, 15–23.
32 Y. Wang, T. Xie, A. France-Lanord, A. Berkley, J. A. Johnson, Y. Shao-Horn and J. C. Grossman, Toward Designing Highly Conductive Polymer Electrolytes by Machine Learning Assisted Coarse-Grained Molecular Dynamics, *Chem. Mater.*, 2020, **32**, 4144–4151.
33 S. Wu, Y. Kondo, M. Kakimoto, B. Yang, H. Yamada, I. Kuwajima, G. Lambard, K. Hongo, Y. Xu, J. Shiomi, C. Schick, J. Morikawa and R. Yoshida, Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm, *npj Comput Mater*, 2019, **5**, 66.
34 S. Ye, N. Meftahi, I. Lyskov, T. Tian, R. Whitfield, S. Kumar, A. J. Christofferson, D. A. Winkler, C.-J. Shih, S. Russo, J.-C. Leroux and Y. Bao, Machine learning-assisted exploration of a versatile polymer platform with charge transfer-dependent full-color emission, *Chem*, 2023, **9**, 924–947.

35 M. J. Tamasi, R. A. Patel, C. H. Borca, S. Kosuri, H. Mugnier, R. Upadhya, N. S. Murthy, M. A. Webb and A. J. Gormley, Machine Learning on a Robotic Platform for the Design of Polymer–Protein Hybrids, *Advanced Materials*, 2022, **34**, 2201809.
36 C. Kuenneth, W. Schertzer and R. Ramprasad, Copolymer Informatics with Multitask Deep Neural Networks, *Macromolecules*, 2021, **54**, 5957–5961.
37 Z. Liang, Z. Li, S. Zhou, Y. Sun, J. Yuan and C. Zhang, Machine-learning exploration of polymer compatibility, *Cell Reports Physical Science*, 2022, **3**, 100931.
38 Y. Zhao, R. J. Mulder, S. Houshyar and T. C. Le, A review on the application of molecular descriptors and machine learning in polymer design, *Polym. Chem.*, 2023, **14**, 3325–3346.
39 H. Doan Tran, C. Kim, L. Chen, A. Chandrasekaran, R. Batra, S. Venkatram, D. Kamal, J. P. Lightstone, R. Gurnani, P. Shetty, M. Ramprasad, J. Laws, M. Shelton and R. Ramprasad, Machine-learning predictions of polymer properties with Polymer Genome, *Journal of Applied Physics*, 2020, **128**, 171104.
40 A. Mannodi-Kanakkithodi, G. Pilania, T. D. Huan, T. Lookman and R. Ramprasad, Machine learning strategy for accelerated design of polymer dielectrics, *Sci Rep*, 2016, **6**, 20952.
41 J. Liang, S. Xu, L. Hu, Y. Zhao and X. Zhu, Machine-learning-assisted low dielectric constant polymer discovery, *Mater. Chem. Front.*, 2021, **5**, 3823–3829.
42 L. A. Miccio and G. A. Schwartz, From chemical structure to quantitative polymer properties prediction through convolutional neural networks, *Polymer*, 2020, **193**, 122341.
43 P. R. Duchowicz, S. E. Fioressi, D. E. Bacelo, L. M. Saavedra, A. P. Toropova and A. A. Toropov, QSPR studies on refractive indices of structurally heterogeneous polymers, *Chemometrics and Intelligent Laboratory Systems*, 2015, **140**, 86–91.
44 S. M. McDonald, E. K. Augustine, Q. Lanners, C. Rudin, L. Catherine Brinson and M. L. Becker, Applied machine learning as a driver for polymeric biomaterials design, *Nat Commun*, 2023, **14**, 4838.
45 D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, *J. Chem. Inf. Comput. Sci.*, 1988, **28**, 31–36.
46 A. Mauri, in *Ecotoxicological QSARs*, ed. K. Roy, Springer US, New York, NY, 2020, pp. 801–820.
47 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss and V. Dubourg, Scikit-learn: Machine learning in Python, *the Journal of machine Learning research*, 2011, **12**, 2825–2830.
48 Z. Bursac, C. H. Gauss, D. K. Williams and D. W. Hosmer, Purposeful selection of variables in logistic regression, *Source Code Biol Med*, 2008, **3**, 17.
49 S. R. Safavian and D. Landgrebe, A survey of decision tree classifier methodology, *IEEE Transactions on Systems, Man, and Cybernetics*, 1991, **21**, 660–674.
50 G. Biau and E. Scornet, A random forest guided tour, *TEST*, 2016, **25**, 197–227.
51 C. Cortes and V. Vapnik, Support-vector networks, *Mach Learn*, 1995, **20**, 273–297.
52 T. Cover and P. Hart, Nearest neighbor pattern classification, *IEEE Transactions on Information Theory*, 1967, **13**, 21–27.
53 S. J. Russell and P. Norvig, *Artificial intelligence: a modern approach*, Prentice Hall/Pearson Education, Upper Saddle River, N.J, 2nd ed., 2003.

54 J. H. Friedman, Greedy Function Approximation: A Gradient Boosting Machine, *The Annals of Statistics*, 2001, **29**, 1189–1232.

55 J. H. Friedman, Stochastic gradient boosting, *Computational Statistics & Data Analysis*, 2002, **38**, 367–378.

56 P. Labute, A widely applicable set of descriptors, *Journal of Molecular Graphics and Modelling*, 2000, **18**, 464–477.

57 A. K. Ghose, V. N. Viswanadhan and J. J. Wendoloski, Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods, *J. Phys. Chem. A*, 1998, **102**, 3762–3772.

58 P. Gramatica, M. Corradi and V. Consonni, Modelling and prediction of soil sorption coefficients of non-ionic organic pesticides by molecular descriptors, *Chemosphere*, 2000, **41**, 763–777.

59 L. Breiman, Random Forests, *Machine Learning*, 2001, **45**, 5–32.

Advancing Antimicrobial Polymer Development: A Novel Database and Accelerated Design via Machine Learning

Yuankai Zhao[a], Roger J Mulder[b] Daniel J. Eyckens[b], Shadi Houshyar[a] and Tu C. Le[a,*]

[a] School of Engineering, STEM College, RMIT University, GPO Box 2476, Melbourne, VIC 3001, Australia

[b] CSIRO Manufacturing, Research Way, Clayton, VIC 3168, Australia

[*] Tu.Le@rmit.edu.au

The authors declare that the data supporting the findings of this study are available within the paper and its Supplementary Information files. Should any raw data files be needed in another format they are available from the corresponding author upon reasonable request.