QSAR Methods in the Discovery and Development of Antibacterials

B. Suay\textsuperscript{a}, J. I. Bueso\textsuperscript{a}, A. Falcó\textsuperscript{b}, M. T. Pérez\textsuperscript{a}, G. Antón\textsuperscript{a}, P. Alemán\textsuperscript{a}

\textsuperscript{a}Departamento de Farmacia
Universidad Cardenal Herrera-CEU, CEU Universities
C/ Ramón y Cajal s/n, 46115 Alfara del Patriarca (Valencia), Spain

\textsuperscript{b}ESI International Chair@CEU-UCH,
Departamento de Matemáticas, Física y Ciencias Tecnológicas,
Universidad Cardenal Herrera-CEU, CEU Universities
San Bartolomé 55, 46115 Alfara del Patriarca (Valencia), Spain

Abstract

With the pressing issue of antibiotic resistance, there is a constant need for new antibiotics. However, the fact that traditional methods of drug discovery are expensive and time-consuming has discouraged the pharmaceutical industry, leaving the burden of discovery to research institutions. This is where QSAR methods become a key tool in fighting multi-drug resistant bacteria, seeing as they provide useful information for the rational design of new active molecules at a minimal cost. A variety of linear and non-linear statistical methods are used to develop these models based on the 2D or 3D representations of the molecules. QSAR models have proven to be effective in rapidly providing lead compound candidates against resistant bacteria such as methicillin-resistant \textit{Staphylococcus aureus}, \textit{Escherichia coli}, \textit{Pseudomonas spp.}, \textit{Bacillus subtilis} or \textit{Mycobacterium tuberculosis}. Moreover, QSAR methods allow for a deeper analysis of a library of molecules, selecting those with not only the optimal activity, but also the most favorable pharmacokinetic and toxicological profiles. The information obtained from QSAR studies makes optimizing an existing drug simpler, which is a cost-effective approach to obtain new treatments against increasingly resistant bacteria.

Keywords: QSAR, antibiotic development, resistant bacteria, machine learning.
1. Introduction

The development of synthetic drugs and the discovery and improvement of antibiotics was a pharmacological revolution for the treatment of infectious diseases. However, the prevalence of infectious diseases continues to increase due to the extreme versatility and adaptability of microorganisms, which allows them to develop resistance mechanisms thus, protecting them against many of these antibacterial compounds [1]. The pressing issue of bacterial resistance is a public health threat worldwide, with antibiotics being antibiotics less effective against increasingly resistant bacteria [2]. This increase in bacterial resistances is resulting in higher mortality rates and health care expenses [3].

The issue with antibiotic resistance is such that the World Health Organization (WHO) has recently published the *Global Priority List of Antibiotic-Resistant Bacteria to Guide Research, Discovery and Development of New Antibiotics*, aiming to help researchers and the pharmaceutical industry in focusing their work towards combating these microorganisms [4]. Moreover, WHO has also pointed out that investment for the development of new antimicrobial compounds is insufficient [5]. It is here where QSAR (Quantitative Structure-Activity Relationships) methods play a key role, seeing as they provide useful information for the rational design of new molecules at a minimal cost [6].

Using QSAR methods for drug design has a series of advantages for the scientific community such as the decrease in spending on animal testing, sustainability, and economic and time saving. These advantages are key in the development of new antibacterial compounds, seeing as the constant increase in resistances requires efficient, effective and safe drugs to combat these constantly evolving bacteria [6].

Initially, drug discovery was a long and expensive process in which large libraries of compounds were synthesized in order to test them for a particular activity [7]. Virtual screening appears as a solution to this problem by allowing researchers to identify molecules with a high probability of being active from a virtual library of thousands of compounds. This way, only molecules with optimal theoretical pharmacological properties go on to synthesis and assays [8]. Since virtual libraries are usually big, filters are usually applied to discard molecules with low probability of becoming drugs. Lipinski was the first to apply these filters by describing the "rule of 5" in 1997 [9, 10]. This rule establishes that, in order for a synthetic molecule to have an appropriate oral absorption it must have:
a molecular weight \( \leq 500 \text{g/mol} \), seeing as those with higher molecular weights penetrate the cell membrane with more difficulty; lipophilicity expressed as the algorithm of the calculated partition coefficient, c-logP \( \leq 5 \); \( \leq 5 \) hydrogen bond donor groups (H atoms bonded to N or O); and 10 hydrogen bond acceptor groups (N or O atoms), seeing as molecules with many of these groups bond to water, interfering transport across the lipidic bilayer. With this rule, Lipinski created the first pre-filter previous to pharmacological screening.

The paper is organised as follows. In the next section we explain some of the QSAR modelling tools being used in the drug development. Section 3 is devoted to the use of the QSAR methodology in the antibacterial compound development. Finally, in Section 4, some conclusions are given.

2. QSAR modelling tools: A learning based approach

QSAR modelling is one such technique that allows the interdisciplinary exploration of knowledge on compounds covering the aspects of chemistry, physics, biology, and toxicology. It provides a formalism for developing mathematical models involving the chemical features and the behavioural manifestations of (structurally) similar compounds. This framework is defined on the basis of a mathematical algorithm, and it provides a reasonable basis for establishing a predictive model.

QSAR models are constructed by a functional approximation in which activity data (outputs), namely \( O_1, O_2, \ldots, O_n \) are related with descriptors codifying chemical information (inputs), namely \( I_1, I_2, \ldots, I_m \). More precisely, the model is characterised by a set of mathematical functions \( f_1, \ldots, f_n \) satisfying that

\[
O_i = f_i(I_1, \ldots, I_m) \quad \text{for } 1 \leq i \leq m.
\]

The descriptors used to generate QSAR models can be classified according to their dimension. Within 2D descriptors one can find topological, structural and physiochemical indices [11]. On the other hand, 3D descriptors include electronic and spatial parameters, molecular shape analysis, molecular field analysis and receptor surface analysis [12].

Recently these quantitative techniques are denominated multi-way or multi-class classification, where the goal is to correctly predict one out of \( K \) classes for any data sample. In particular, a whole host of techniques such as artificial neural networks, decision trees, naïve Bayes, nearest neighbours, and support vector machines (SVMs) have been successfully applied for binary classification problems.
We also need to distinguish two cases. The first approach refers to cases in which the property \( O_i \) under study is quantitative and belongs to a continuous scale, while the second approach refers to cases in which the response \( O_i \) is qualitative.

### 2.1. 2D Strategies

Originally, QSAR studies were based solely on 2D descriptors. These descriptors, or indices, were calculated based on the discrete representation of the molecular shape. More precisely, the topological shape of the molecule is represented by using a graph [12]. Todeschini and Consonni then defined molecular indices as the result of a series of mathematical processes in which the chemical information of a molecule is assumed to be codified by the so-called graph invariants [13] (used in the mathematical Graph Theory) also called topological indices in the QSAR framework modelling.

The main idea is to associate the chemical properties of the molecule with its shape. For example, assume that we want to known when a particular molecule has a chemical property \( O \), by using a sample \((D_1,\ldots,D_m)\) from a particular set of \( m \)-chemical compounds of a topological descriptor \( D \). The idea is to construct a decision function \( d = d(D_1,\ldots,D_m) \) from the sample to a binary set, namely \( \{-1,1\} \), in such a way that when \( d = 1 \) we will say that the molecule has the chemical property \( O \), otherwise when \( d = -1 \) we say that the molecule does not have the chemical property, that is, it is not in \( O \), namely it is in \( O^c \). To test the performance of the decision function \( d \) we can use the following table of conditional probabilities:

| \( \Pr(O|d = 1) \) | \( \Pr(O^c|d = 1) \) |
| \( \Pr(O|d = -1) \) | \( \Pr(O^c|d = -1) \) |

and where the benchmark table for a decision function \( d \) is the following

| \( \Pr(O|d = 1) = 1 \) | \( \Pr(O^c|d = 1) = 0 \) |
| \( \Pr(O|d = -1) = 0 \) | \( \Pr(O^c|d = -1) = 1 \) |
These decision functions along with the nature of the chemical property are used to obtain QSAR models. In this review we discuss this relationship adopting a more recently approach based in the concept of learning process.

The goal of learning theory is to approximate a function, namely \( d \) (or some function features) from data samples, perhaps perturbed by noise. To attain this goal, learning theory draws on a variety of diverse subjects. Statistics is used to infer information from random samples. It also relies on the mathematical Approximation Theory, since our estimate of the function \( d \) must belong to a pre-specified class, and therefore the ability of this class to approximate the function accurately is of the essence. Finally, some algorithmic considerations are critical because our estimate of \( d \) is the outcome of algorithmic procedures, and the efficiency of these procedures is crucial in practice.

We explain the above discussion with a classical situation in science. We want to learn a physical law where a quantitative chemical feature \( Y \) is related with a descriptor \( D \). The learning process is driven by a set of \( k \)-values of \( D \), namely \( D_1, \ldots, D_m \) for which the corresponding \( Y \)-values \( Y_1, \ldots, Y_m \) are known. Assume that the law under consideration represented by a real function, namely \( f : \mathbb{R} \times \Theta \rightarrow \mathbb{R} \), has a specific form, namely

\[
f(D; \omega) = \sum_{j=1}^{N} \omega_j \phi_j(D),
\]

where \( \mathcal{B} = \{\phi_1, \ldots, \phi_N\} \) is a set of basis (linearly independent) functions and the degrees of freedom are related with unknown parameter vector \( \omega = (\omega_0, \omega_1, \ldots, \omega_k) \). For example, by considering \( \mathcal{B} = \{1, D, \ldots, D^k\} \), we have that \( f \) is a degree \( k \) polynomial map, where \( N = k + 1 \) and \( Y = f(D) = \sum_{i=0}^{k} \omega_i D^i \). Thus, by taking into account the learning data set \( (Y_1, D_1), \ldots, (Y_m, D_m) \), we can compute the coefficients vector \( \omega \) by the least square method (the objective of the learning stage):

\[
\tilde{\omega} = \arg \min_{\omega} \sum_{\ell=1}^{m} \text{distance}(Y_\ell, f(D_\ell; \omega)),
\]

where, for algorithmic considerations, \( m > N = k + 1 \). If the measurements generating this set were exact, then \( Y = f(D, \tilde{\omega}) \). Otherwise, one expects that the measurements of \( Y \) are affected by noise. Thus it is assumed that \( Y = f(D, \tilde{\omega}) + \varepsilon \), where \( \varepsilon \) is a random variable (which may depend on \( D_i \)) with mean zero. Now the values \( Y_i \) are affected by noise. Then one might take as starting point, instead of the unknown function \( f \), a family of random variables \( \varepsilon_D \) on \( \mathbb{R} \) varying with
$D \in \mathbb{R}$. The only requirement on these variables is that the mean of $\varepsilon_D$ is equal to $f(D)$ for all $D \in \mathbb{R}$. Clearly, $Y_i$ is randomly drawn from $\varepsilon_{D_i} = Y_i - f(D_i, \bar{\omega})$. In this case we can use the following decision map:

$$
d(D_1, \ldots, D_m) = \begin{cases} 
+1 & \text{if } \mathbb{E}(\sum_{i=1}^m \varepsilon_{D_i}^2) \text{ is small enough,} \\
-1 & \text{otherwise,}
\end{cases}
$$

In practice, the training dataset may be split into two subsets, the training set and the cross-validation set. The learning algorithm may then be applied to the training set, and the regularisation parameter varied in order to minimize the error in the cross-validation set.

Another approach is to fix a target function $f_{\text{target}} : \mathbb{R} \rightarrow \mathbb{R}$ and consider that a network is given by a set of randomly chosen pairs $(Y_1, D_1), \ldots, (Y_m, D_m)$. Then by means a training algorithm a set of weights $\bar{\omega}$ is obtaining attempting to minimize some distance function from $f(D, \omega)$ to $f_{\text{target}}(D)$. Finally we can write our decision map in a general form as

$$
d(D_1, \ldots, D_m) = \begin{cases} 
+1 & \text{if distance}\left(f_{\text{target}}(D), f(D, \bar{\omega})\right) \text{ is small enough,} \\
-1 & \text{otherwise.}
\end{cases}
$$

### 2.2. Techniques based on Linear Regression

This techniques correspond to models with quantitative responses and it is based on a non-zero statistical correlation between the chemical feature, namely $D$, and the quantitative response, namely $Y$. The underlying idea is to see the statistical correlation as a classical euclidean inner product between unitary vectors and hence it coincides with the cosinus of the angle between of two given unitary vectors (see Figure 1.) Assuming non-zero correlation, that is $\theta \neq \pi/2$, we use the orthogonal projection of $Y$ over the linear span of $D$-measures to construct explicitly a linear relationship between the chemical $Y$-feature and the $D$-measures taking over similar compounds. In the framework introducing in the above section, we consider that $Y = f(D; \omega_1, \omega_2) = \omega_1 + \omega_2 D$, that is, the basis function under consideration is the set $\mathcal{B} = \{1, D\}$.

Multiple Linear Regression (MLR) is frequently used to obtain QSAR models due to its simplicity, transparency, reproducibility and easy interpretation [16]. The contribution of each descriptor to the model is easily interpreted seeing as it is directly dependent on the value of its coefficient and on its algebraic sign. In order for a model to be statistically reliable, the number of molecules and the number of indices must have a ratio of, at least, 5:1 [17]. A reliable MLR model
Figure 1: The correlation between $D$ and $Y$ coincides with $\cos \theta$, assuming that $Y$ and $D$ are unitary vectors.

will result in a scatter plot (Observed vs. Calculated), showing minimal deviation between points and the line of fit.

The Partial Least Squares (PLS) method is most adequate for cases in which a large number of correlated indices is used, and a limited number of molecules are available. PLS, being a generalization of MLR [18], works with latent variables, which are linear functions of the original variables. The keys to obtaining a quality QSAR model are a strict relevance test for each PLS component and finding the final point at which adding a new component does not improve the model [19].

2.3. Techniques Based on Linear Classification

This technique can be used for models with either qualitative or quantitative responses. Linear discriminant analysis (LDA) is a pattern recognition method aiming to establish a linear combination of variables to separate two or more classes of objects and, therefore, be used in classification problems. The LDA method explicitly focuses on modelling the differences among data classes. In a two-group situation, predicted classification is obtained by calculating the value of a discriminant function (DF) for each case. Consequently, those cases in which the value of DF is lower than the cut-off value would be classified in one group, while those with values higher than the cut-off value would be classified in the other group. Once the model is created, any new case can be classified using the same DF [20]. The case of stepwise LDA is a special case in which the model is constructed step by step. In each step, all the variables are reviewed and evaluated to determine which contribute to a larger extent to the discrimination between
groups. The determination is made based on the F-Snedecor parameter, which relates the variance of the equation with the residual variance. Variables are selected one by one and the process is repeated until, when adding another index, the prediction ability of the model does not improve [21].

2.4. Techniques Based on Machine Learning

The techniques explained in 2.2 and 2.3 are based on linear combinations of fixed basis functions. These methods have useful computational properties. However, their applicability is limited by the so-called *the curse of dimensionality*. Thus in order to apply such models to large-scale problems it is necessary to adapt the choice of the basis function to the shape of data sets.

The machine learning approach is based in running an algorithm that can be expressed by a function \( \hat{O} = f(I_1, \ldots, I_n) \) which takes the chemical features to generate an output encoded in the same way that the target response \( O \). The precise form of the function \( f \) is determined during the training phase, also known as the *learning phase*, on the basis of the training data. Once the model is trained it can then determine the identity of new chemical composes, which are said to comprise a test set. The ability to categorize correctly new examples that differ from those used for training is known as *generalization*.

With the increasing number of public domain molecular data bases available and large number of descriptors, QSAR modelling has entered the big data world. This has led the scientific community to adopt machine learning methods in order to develop methods working with large amounts of data [20]. The most frequently machine learning based methods used in QSAR are: Artificial Neural Networks (ANN) and Random Forests (RF).

2.4.1. Artificial Neural Networks

Similarly to LDA, artificial neural networks (ANN) are algorithms used for pattern recognition. The network structure is based in an entry level with connected neurons, one or more hidden levels and an exit level (see Figure 2). In a standard design, each connection between two neurons has a specific weight. This weight varies during the training phase until the network learns how to connect entry and exit data [21]. More precisely, in this framework and for a two layer network (see Figure 3), equation (1) is given by

\[
f(D; \omega) = h^{(2)} \left( \omega^{(2)} \right)_2 h^{(1)} \left( \omega^{(1)} D + \omega^{(1)} \right) + \omega^{(2)}
\]  

(2)
where $h^{(1)}$ and $h^{(2)}$ are non-linear output activation functions and the degrees of freedom are related with unknown parameter vector $\omega = (\omega_0^{(1)}, \omega_1^{(1)}, \omega_0^{(2)}, \omega_1^{(2)})$. Here the superscript $(i)$, $i = 1, 2$, indicates that the corresponding parameters are in the $i$-th layer of the network.

### 2.4.2. Random Forests

Random forests (RF) are a class of learning algorithms used to solve pattern recognition problems [22]. Growing many different trees from a single data set and where usually each tree is built using training data of multiple indices for a series of molecules. In consequence, a variety of random forests exist, depending on how trees are built and how the randomness is introduced in the tree building process.
We first proceed dividing the set of all possible values of the descriptors \( \{D_1, \ldots, D_m\} \) into \( J \) distinct and non-overlapping regions, namely \( R_1, \ldots, R_J \). To perform a recursive binary splitting, such as active/inactive, we decompose the predictor space for each descriptor \( D_j \) into two regions

\[
R_1(j, s) = \{(D_1, \ldots, D_m) : D_j < s\} \quad \text{and} \quad R_2(j, s) = \{(D_1, \ldots, D_m) : D_j \geq s\}
\]

depending on a cut point \( s \) that leads to smaller possible residual reduction. Assume that for every observation that falls in the region \( R_k(j, s) \) we make the same prediction, denoted by \( \hat{Y}_k^{(j)} \), for the training observations in \( R_k(j, s) \) for \( k = 1, 2 \). Then we seek the value of \( j \), over a random sample from \( p \leq m \) of the set of \( m \)-predictors, and \( s \) that minimize the equation

\[
\sum_{j:R_1(j,s)} (Y_j - \hat{Y}_1^{(j)})^2 + \sum_{j:R_2(j,s)} (Y_j - \hat{Y}_2^{(j)})^2
\]

For each finite sequence of random samples we can construct a binary tree. In Figure 4 we give an example for two descriptors \((D_1, D_2)\), \( p = m \) (that is there are not chance in the choice of sample) and where, for this particular case, the three regions are

\[
R_1 = \{(D_1, D_2) : D_2 < s_2\}, \quad R_2 = \{(D_1, D_2) : D_2 \geq s_2 \text{ and } D_1 < s_1\}, \quad \text{and} \quad R_2 = \{(D_1, D_2) : D_2 \geq s_2 \text{ and } D_1 \geq s_1\}.
\]

Once the regions and the tree are created, we can repeat the above procedure obtaining a finite sequence of trees called the random forest. Then we can predict the response for a given test observation according to the majority of votes obtained the region to which that test observations belongs among all the trees in the forest [23]. Observe that in constructing a random forest, at each split along the tree, the algorithm is not even allowed to consider a majority of the available predictors.

2.5. 3D Strategies

Recently, QSAR methods have experienced a great evolution from traditional 2D-QSAR to 3D-QSAR, which includes parameters such as molecular and spatial variety or protein flexibility [25]. Two of the most frequently used QSAR
methods are Comparative Molecular Field Analysis (CoMFA), Molecular Similarity Index Analysis (CoMSIA) and molecular docking. CoMFA was the first 3D QSAR method to be developed [26]. The construction of a CoMFA model consists of 5 steps:

1. Molecule selection and determination of bioactive conformations interacting with a certain receptor.

2. Superposition of the training group using traditional pharmacophore models.

3. Molecule aligning with a defined network distance and calculation of the field values around the molecules.

4. Correlation of field values with the bonding affinity using algorithms such as PLS.

5. Application of cross-validation to verify and optimize the final model.

Being a useful technique, CoMFA models have a series of limitations [26] that lead to the creation of Molecular Similarity Index Analysis (CoMSIA) [27]. The main difference between CoMFA and CoMSIA is that, while CoMFA uses fields based on networks to calculate indices, CoMSIA calculates similarity indices using algorithm encryption software (SEAL) [27]. This allows for the model to take into account steric, electrostatic, hydrophobic and hydrogen bond properties. Finally, molecular docking is a widely used 3D technique. It aims to predict the main binding mode of a ligand to a protein with a known 3D structure. A successful model seeks high-dimensional spaces effectively and uses a qualification function that correctly classifies docking of candidates. This molecular docking technique is frequently used to carry out virtual screening of large libraries of compounds.
and, by classifying the results, proposes structural hypotheses regarding molecular inhibition of the therapeutic target studied, which is key in the optimization of lead compounds [28]. The Table 1 given below shows some examples of drugs design using these techniques.

<table>
<thead>
<tr>
<th>Drug</th>
<th>Main activity</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boceprevir</td>
<td>Antiviral</td>
<td>28</td>
</tr>
<tr>
<td>Captopril</td>
<td>Antihypertensive</td>
<td>29</td>
</tr>
<tr>
<td>Crizotinib</td>
<td>Antineoplastic</td>
<td>30</td>
</tr>
<tr>
<td>Dorzolamide</td>
<td>Glaucoma</td>
<td>31</td>
</tr>
<tr>
<td>Flobufen</td>
<td>Anti-inflammatory</td>
<td>32</td>
</tr>
<tr>
<td>Ipconazole</td>
<td>Antifungal</td>
<td>33</td>
</tr>
<tr>
<td>Losartan</td>
<td>Antihypertensive</td>
<td>34</td>
</tr>
<tr>
<td>Metconazole</td>
<td>Antifungal</td>
<td>35</td>
</tr>
<tr>
<td>Norfloxacin</td>
<td>Antibiotic</td>
<td>36</td>
</tr>
<tr>
<td>Rilpivirin</td>
<td>Antiviral</td>
<td>37</td>
</tr>
<tr>
<td>Vemurafenib</td>
<td>Antineoplastic</td>
<td>38</td>
</tr>
<tr>
<td>Zolmitriptan</td>
<td>Anti-migraine</td>
<td>39</td>
</tr>
</tbody>
</table>

Table 1: Examples of commercial drugs design using QSAR methods.

3. QSAR in antibacterial compound development

The development of synthetic antibiotics goes back to the early 21st century, when Paul Ehrlich and Alfred Bertheim discovered Salvarsan [41]. This discovery marked the beginning of selective therapies, being the first chemotherapy agent obtained as a result of rational screening. QSAR methods have been widely used in the discovery and development of new antibacterial agents. Zanni et al. [42] elaborated a QSAR model capable of selecting antibacterial molecules that were not susceptible to certain resistance mechanisms. The model selected 19 candidates, of which 5 showed good antibacterial activity against *Staphylococcus spp.* with MIC values between 16 and 31mg/L. Similarly, Bueso-Bordils et al. [43] developed a topological model using LDA to predict general antibacterial activity. This model was applied to a virtual library of 6375 compounds, having selected 263 as theoretically active. After a bibliographical search, 40% of these compounds were found to have proven antibacterial activity, proving the validity of the model. The remaining 158 compounds were proposed as candidates to be developed as new antibacterial agents. The fact that QSAR models provide a fast
route for drug discovery has led many research groups to use these techniques in order to find new active drugs against multi-resistant bacteria causing nosocomial infections in the hospital setting [44]–[46]. Speck-Planche et al. have focused their work in developing multi-task models to establish relationships between the chemical structure of a compounds and its biological effect (mkt-QSBER) [44, 45]. They developed two models, one targeting active compounds against *Escherichia coli* [44] and another one targeting active compounds against *Pseudomonas spp.* [45]. In both cases, the models simultaneously predicted antibacterial activity and ADMET profile (absorption, distribution, metabolism, elimination and toxicity). Both models were developed using structurally heterogeneous databases with 20000-30000 molecules. These models are particularly useful seeing as knowing the ADMET profile of theoretically active compounds allows for a better selection of drug candidates, because those with undesirable pharmacokinetic characteristics would be immediately discarded. The authors demonstrated the validity of these models by virtually predicting the properties of known active compounds against *E. coli* (avarofloxacin) and *Pseudomonas spp.* (delafloxacin). On another note, methicillin-resistant *Staphylococcus aureus* (MRSA) is another important pathogen involved in nosocomial infections. For this reason, Bueso-Bordils et al. [46] developed a topological model using LDA capable of classifying a compound according to its anti-MRSA activity. The model established structure-activity relationships, highlighting the role of chlorine atoms, cyclopropyl, substituted cycles and tertiary amines. This model was also used to screen a virtual combinatorial library of 1001 6-fluoroquinolones (Figure 1), of which 177 were selected as theoretically active. Six of these compounds were assayed, of which three showed activity similar to that of ciprofloxacin (0.5-1mg/L).

In this same work, the model was also applied to 263 theoretical antibacterial agents [43], of which 34 were selected as theoretically active against MRSA. Nine compounds were bibliographically verified to be correctly selected, of the 25 remaining, 3 were randomly selected for assays. The results showed that mitomycin C is more active than ciprofloxacin against MRSA. The study also proves that, when QSAR models are developed using global connectivity models instead of applying traditional fragmentary analysis, these models can be applied to libraries with great structural diversity, regardless of the fact that they are developed using structurally related compounds. Thomann et al. [47] developed a QSAR model focused on the inhibition of the quinolone signal quorum sensing (PQS-QS) of *Pseudomonas spp.* Using the model, they established structure-activity relationships of 2-sulfonylpyrimidines that had previously been identified as double targets of the PqsR receptor of PQS and the PQS-synthetase PqsD. The activity of
Figure 5: General structure of the 1001 fluoroquinolones screened using the anti-MRSA model developed by Bueso-Bordils et al. Compounds 1-3, showed similar activity to that of ciprofloxacin. A compound was determined by the efficiency of the ligand and its lipophilicity. Those compounds selected by the model were rationally modified by using the Hansch analysis. These modified inhibitors proved to have the theoretical ability to decrease biofilm mass and extracellular DNA, both important factors in bacterial resistance appearance. Similarly, ANN and RF were applied for the creation of a prediction model of Pseudomonas aeruginosa and Bacillus subtilis inhibitors using bibliographical data of imidazolium-based ionic liquids (47). Twenty theoretically active synthesis compounds of 1,3-dialkylimidazolium ionic liquids were assayed. It was observed that, for asymmetric compounds, only those with at least one radical containing a 12-carbon alkyl chain showed high antibacterial activity. On the other hand, the contrary was observed for symmetric compounds, being those with shorter chains (8 carbon atoms), most active (Figure 2).

Figure 6: General structure of 2-sulfonylpirymidines inhibitors of Pseudomonas aeruginosa and Bacillus subtilis, studied by Hodyna et al.

Likewise, He et al. [49] used an integrative protocol using 2D and 3D QSAR to design antimicrobial peptides containing unnatural amino acids (AMP-UAAs).
Once the model was developed, they designed and synthesized 12 AMP-UAAs, of which two proved to be especially active, with MICs < 10mg/ml against multi-resistant *P. aeruginosa* and MRSA. The fact that QSAR models can be used to predict the activity of structurally different molecules has led to the identification of lead compounds outside of the conventional antibacterial families. This is the case of a study in which a 2D QSAR model based on molecular fingerprints was developed using DNA-gyrase inhibitors with new structures and new mechanisms of action [50]. In the initial phase, bioactive molecular fingerprints were extracted from a DNA-gyrase inhibitor database. These fingerprints were converted into molecular fragments that, when recombined, generated a compound library. After a virtual screening of the combinatorial library using *LigandFit Gold docking*, a possibly active candidate was selected (Figure 3), which was later studied by pharmacophore validation and binding mode analysis.

Similarly, Mor et al. synthesized 20 tetracyclic 1,4-benzotiacines and studied their activity against 2 Gram-positive bacteria (*B. subtilis* and *Staphylococcus epidermidis*), 2 Gram-negative bacteria (*E. coli* and *P. aeruginosa*) and 2 fungi (*C. albicans* and *Aspergillus niger*) [51]. Using these data, MLR was applied to establish antimicrobial activity QSAR models against these microorganisms [52]. After determining that WHIM parameters were dominant in determining the compound’s activity, some of the molecules used in developing the model were modified in order to improve their activity. Another important contribution of QSAR models to antibacterial development is the improvement of existing drugs. This is an interesting approach to drug design, seeing as the improvement of an already existing molecule is less expensive than developing a new one [53]. By analyzing key components of active molecules using QSAR models, we can achieve a better understanding of the molecular basis of antibacterial activity. Petnepapun et al. [54] used CoMFA and CoMSIA to study dicoumarol derivatives with substituted benzene rings in the bis-methylene position. Analysis results showed that steric
repulsion in the *para* position could decrease antibacterial activity. Zhang et al. [55] used genetic algorithms and CoMFA to identify the indices that influenced the biological activity of nitazoxanide derivatives (Figure 4). The results obtained indicated that the antibacterial activity of these compounds against *Clostridium difficile* could be improved by increasing the molecular connectivity, local charge and sharp indices while decreasing the molecular flexibility index.

Similarly, Karhikeyan et al. [56] developed QSAR models to predict the antibacterial activity of novel dispirpyrrolidines against *B. subtilis, S. aureus, Salmonella typhi, P. aeruginosa* and *Proteus vulgaris*. QSAR studies showed that topology, shape, charge distribution and hydrophobicity were determinant for the antimicrobial activity of the molecule. Another study focused on synthesis and QSAR analysis of Ni(II) sylfonyl hydrazone complexes active against *E. coli* and *S. aureus* [57] concluded that the nucleophilic reaction index for Ni and O atoms as well as the HOMO-LUMO energy gap played key roles in the antimicrobial activity. Abdelrahman et al. also studied the SAR of a group of pyridine and quinolone hydrazone derivatives using genetic algorithms and MLR in order to understand the determinants of their activity against bacteria, fungi and tuberculosis [58]. Fluoroquinolones are an antibacterial family with considerable scientific and clinical interest due to their broad spectrum. A new series of pefloxacin hydrazones has been studied using 3D QSAR to analyze their activity against *S. aureus* by binding to the DNA-gyrase [59]. Results showed that those compounds with aromatic rings substituted with electron- donor groups had good molecular docking scores and showed better affinity than ciprofloxacin which suggests that this modification would optimize the activity of pefloxacin derivatives (Figure 5).

Furthermore, another study focused on the antibacterial activity of fluoroquinolones containing bulky arenosulfonyl fragments determined that the presence of methyl groups in the arenosulfonly group decreased the antibacterial activity against Gram-positive bacteria, while the presence of methoxy groups in-
creased it (Figure 6) [60]. As for Gram-negative bacteria, the QSAR study determined that the presence of methyl groups and chlorine atoms decreased the antibacterial activity of these compounds.

The threat of multi-resistant *Mycobacterium tuberculosis* strains is such that there is a constant need for the discovery of new active molecules. Virtual screening has proven to be a very useful tool for the rapid discovery of molecules against this pathogen. Rajkhowa et al. [61] combined several computational techniques to screen 62 triclosan derivatives searching for compounds with better bioavailability. Similarly, QSAR methods were used to screen a virtual library of more than 50000 chemical compounds aiming to identify proteasome inhibitors for *M. tuberculosis* [62]. Of the 50 compounds that were selected by the model as theoretically active, *in vitro* assays proved that 15 had a high activity range, meaning they could be studied as possible drug candidates. Magnati et al. [63] used CoMFA and CoMSIA to develop a model with good statistical parameters for the identification of *M. tuberculosis* aryl acid adenylating enzyme inhibitors. This model was used to
screen a virtual library of 230755 compounds, which resulted in 13 possible drug candidates.

As it has been previously mentioned, these QSAR methods are commonly used in the development of treatments for tuberculosis. Abdel-Aziz et al. [64] constructed three QSAR models using genetic algorithms from halophenyl bys-hyrazones synthesized for the treatment of tuberculosis. These three models against *B. subtilis*, *Klebsiella pneumoniae* and *M. tuberculosis* determined that the bioactivity of these compounds was affected by the \( jurs \) descriptor and the \( shadow \) indices. Moreover, Punkvang et al. [65] studied the requirements for aminopyrimidine derivates to be active against *M. tuberculosis*. Their CoMSIA-based model determined that the \(-\text{NH} \) bond and the substituent group in \( R1 \) are crucial for the efficacy of these molecules against tuberculosis (Figure 7).

![Figure 11: General structure of antituberculous aminopyrimidine derivatives studied by Punkvang et al.](image)

QSAR methods have proven to be key in the fight against resistant M. tuberculosis, seeing as they contribute to the design of new active molecules. Ventura et al. [66] combined MLR and ANN to elaborate a model capable of predicting the activity of hydrazides against tuberculosis. In a first stage, the MLR and ANN models were analyzed separately to evaluate their performance. The ANN model showed better learning ability and accuracy. However, the MLR model was capable of selecting with more precision key data for drug design. During the second stage of the study, both methods were combined to rationally design active compounds. Two of these compounds were synthesized and their *in vitro* assays validated the utility of the model (Figure 8).

Likewise, Gomes et al. [67] developed a binary QSAR model aiming to find new heteroarylchalones with antituberculous activity. After applying the model, 33 molecules were selected to be synthesized and assayed. Of the 33 molecules, 10 compounds (Figure 9) showed nanomolar antituberculous activity, as well as low activity against commensal bacteria, showing its selectivity for *M. tuberculosis*. These results suggest that the selected compounds are promising candidates
Figure 12: Active compounds against M. tuberculosis designed using the MLR and ANN models developed by Ventura et al.

for the treatment of tuberculosis.

Figure 13: Heteroarylchalcones with antituberculous activity designed by Gomes et al.

Many of the antibacterial compounds currently used have a natural origin. Consequently, some research groups focus their work on the discovery of antibiotics in nature [68]–[70]. Thai et al. have developed a linear QSAR model to analyse a collection of 182 flavonoids and the Chinese Medicine Database searching for new S. aureus NorA pump inhibitors [68]. Similarly, Pereira et al. developed a QSAR model for the search of antibacterial compounds in PubChem (1804 compounds) and the AntiMarin Database (418 compounds) [69, 70]. In both cases, the virtual screening resulted in a group of theoretically active compounds that were later confirmed in the literature.
4. Conclusion

QSAR models are presented as a key tool in the development of new antibacterial compounds in light of the increasing appearance of multi-resistant bacteria and the lack of investment from the pharmaceutical industry. These studies provide useful information for a better understanding of how the molecular structure may affect the antibacterial activity of a compound. Having this information, it is simpler to optimize existing drugs, which is a cost-effective approach to obtain new treatments against resistant bacteria. QSAR methods are an efficient tool to carry out fast virtual screenings of possible antibacterial compounds. Moreover, this chemoinformatic approach allows the prediction of other pharmacokinetic and toxicological properties directly related with the activity, in order to identify safer and more efficient drugs.

References


[34] https://www3.epa.gov/pesticides/chem_search/reg_actions/registration/fs_PC-125618_01-Sep-04.pdf [last access: November, 24 2017]


25
[60] Abdel-Aziz AA, Asiri YA, Al-Agamy MH (2011) Design, synthesis and anti-
bacterial activity of fluoroquinolones containing bulky arenesulfonyl frag-

computational strategies to identify potential compounds. J Mol Graph
Model. 62:56-68.

FG, Nargotra A (2015) Discovery of new proteasome inhibitors using a
knowledge-based computational screening approach. Mol Divers. 19:1003-
1019.

[63] Maganti L, OSDD Consortium, Ghoshal N (2015) 3D-QSAR studies and
shape based virtual screening for identification of novel hits to inhibit MbtA

[64] Abdel-Aziz HA, Eldehna WM, Fares M, Al-Rashood ST, Al-Rashood KA,
Abdel-Aziz MM, Soliman DH (2015) Synthesis, biological evaluation and
2D-QSAR study of halophenylbis-hydrazones as antimicrobial and antitu-

the structural requirements of aminopyrimidine derivatives for good potency
against both purified enzyme and whole cells of M. tuberculosis: combina-
tion of HQSAR, CoMSIA and MD simulation studies. J Biomol Struct Dyn.
34:1079-1091.

[66] Ventura C, Latino DA, Martins F (2013) Comparison of Multiple Linear
Regressions and Neural Networks based QSAR models for the design of

[67] Gomes MN, Braga RC, Grzelak EM, Neves BJ, Muratov E, Ma R, Klein LL,
Cho S, Oliveira GR, Franzblau SG, Andrade CH (2017) QSAR-driven de-
sign, synthesis and discovery of potent chalconederivatives withantitubercu-

[68] Thai KM, Ngo TD, Phan TV, Tran TD, Nguyen NV, Nguyen TH, Le MT
(2015) Virtual screening for novel Staphylococcus aureus NorA efflux pump