A General ANN-Based Multitasking Model for the Discovery of Potent and Safer Antibacterial Agents

ARTICLE *in* METHODS IN MOLECULAR BIOLOGY (CLIFTON, N.J.) · JANUARY 2015

Impact Factor: 1.29 · DOI: 10.1007/978-1-4939-2239-0_4 · Source: PubMed

Alejandro Speck Planche
Faculty of Sciences, University of P…

Natália D. S. Cordeiro
University of Porto

Chapter 4

A General ANN-Based Multitasking Model for the Discovery of Potent and Safer Antibacterial Agents

A. Speck-Planche and M.N.D.S. Cordeiro

Abstract

Bacteria have been one of the world's most dangerous and deadliest pathogens for mankind, nowadays giving rise to significant public health concerns. Given the prevalence of these microbial pathogens and their increasing resistance to existing antibiotics, there is a pressing need for new antibacterial drugs. However, development of a successful drug is a complex, costly, and time-consuming process. Quantitative Structure-Activity Relationships (QSAR)-based approaches are valuable tools for shortening the time of lead compound identification but also for focusing and limiting time-costly synthetic activities and in vitro/vivo evaluations. QSAR-based approaches, supported by powerful statistical techniques such as artificial neural networks (ANNs), have evolved to the point of integrating dissimilar types of chemical and biological data. This chapter reports an overview of the current research and potential applications of QSAR modeling tools toward the rational design of more efficient antibacterial agents. Particular emphasis is given to the setup of multitasking models along with ANNs aimed at jointly predicting different antibacterial activities and safety profiles of drugs/chemicals under diverse experimental conditions.

Key words Antibacterial activity, Drug resistance, QSAR, Topological indices, Ontology, Moving average approach, Artificial neural networks, mtk-QSBER models

1 Introduction

For centuries, mankind has been particularly affected by microbial diseases, those caused by bacteria being among the most dangerous and lethal. Even though antibiotics have saved millions of lives and alleviated the suffering of people for many years [1], drug-resistant, disease-causing bacteria have emerged and spread so much in recent decades [2–5] that they have created a global public health problem. This situation is so serious that nowadays bacteria cause high mortality rates not only in communities but also in healthcare environments and facilities [6]. Given the prevalence of the major gram-positive pathogens, especially those belonging to the genera *Staphylococcus*, *Streptococcus*, and *Enterococcus*, transferability of resistance genes is of special concern. This is exemplified

Hugh Cartwright (ed.), *Artificial Neural Networks*, Methods in Molecular Biology, vol. 1260,
DOI 10.1007/978-1-4939-2239-0_4, © Springer Science+Business Media New York 2015

by the ability of enterococci to interact with methicillin-resistant *S. aureus* (MRSA), transferring to the latter vancomycin-resistant genes [7]. Thus, MRSA can become resistant to vancomycin, which is the unique antibacterial chemotherapeutic alternative. On the other hand, gram-negative pathogens such as *Escherichia coli* and *Pseudomonas aeruginosa* exhibit intrinsic resistance to almost any antibiotic [8]. The most deadly bacterial disease has been tuberculosis (TB), which is prone also to multidrug resistance. Indeed, more than 9 million new TB cases and about 1.7 million deaths were reported in 2009 [9], while a recent report confirmed that in 2010, 5.7 million official cases of TB were notified, with 8.8 million incidents worldwide (equivalent to 128 cases *per* 100,000 population) [10].

Diseases and infections caused by bacteria are very complex and depend on multiple pathogenic and epidemiological factors [6, 8]. For this reason, the search for new antibacterial drugs is very urgent. This battle against bacterial diseases/infections will depend on the efficiency of chemotherapies and have a profound influence on human health.

High-throughput screening technologies [11] along with combinatorial chemistry [12] were expected to solve the drug discovery problem by a massive parallelization of the process. In practice, however, while the number of identified hits can substantially be increased using these methods, no corresponding growth in the number of drugs entering the market has been observed [13]. This fact has progressively led to a reconsideration and rationalization of the drug discovery process, in which chemoinformatics methods, led by Quantitative Structure-Activity Relationships (QSAR) tools [14], have gained a role of tremendous importance [15–17]. To be clearly effective, these methods should aim at targeting both pharmacological profiles and desirable ADMET (absorption, distribution, metabolism, elimination, toxicity) properties, as the latter are the major causes of non-approval of drugs [18–20]. Particularly, in recent years, promising multi-target (mt-QSAR) models have been reported, revolutionizing concepts regarding QSAR paradigm prediction [21–26]. These mt-QSAR models have been applied to simultaneously predict several features of biological activity by considering different biological targets such as biomolecules, cell lines, tissues, and organisms. The aforementioned mt-QSAR models have evolved to the point that nowadays it is possible to predict multiple biological profiles by considering different measures of the effects, many biological targets, and diverse levels of curation of the experimental information [27–32]. These new models, known as multitasking Quantitative Structure-Biological Effect Relationships (mtk-QSBER), have often employed classification techniques such as linear discriminant analysis (LDA) [33]. But sometimes, due to the complexity of the data and a lack of accuracy in the experiments, linear classification tools do not lead to models with enough statistical quality and predictivity. In such cases, artificial neural networks (ANNs) can instead be applied to deal with nonlinear behavior and usually do enhance the performance of the predictions [22, 29]. This chapter is devoted to an overview of the setup and use of mtk-QSBER models based on ANNs. Overall, specific insights from our perspective and personal experience will be provided.

2 General Procedure for the Setup of QSAR Models

Generally speaking, QSAR approaches seek to uncover possible relationships between chemical structures (*molecular descriptors*) and one or more endpoints of interest (e.g., biological activity, toxicity, or pharmacokinetic profiles) [14]. Several steps should be followed for establishing the QSAR models (Fig. 1), and these, if well devised, may help yielding more accurate results. Firstly, the data must be retrieved typically from a public source and subsequently curated in, for example, a spreadsheet application like Excel. Secondly, molecular descriptors are calculated from the molecular structures; finally, by applying a statistical modeling approach (either linear or nonlinear), the best QSAR model can be obtained.

2.1 Databases and Handling of the Retrieved Data

Databases are typically organized as public sources and contain large collections of data describing the most relevant aspects of phenomena and/or experiments [34, 35]. A well-known database is ChEMBL [34]—an online free and dynamic source available at

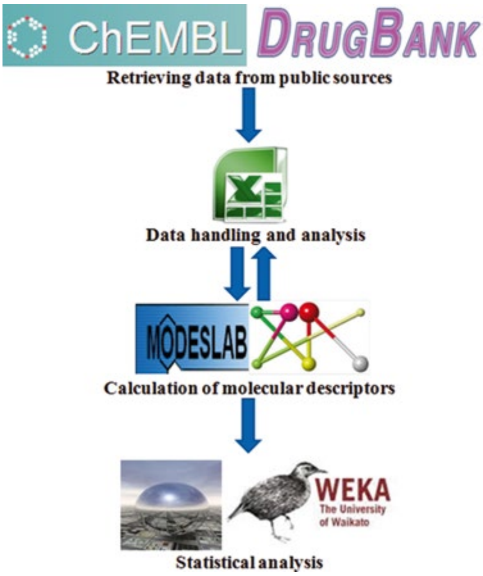
ChEMBL
DRUGBANK
Retrieving data from public sources
Data handling and analysis
MODESLAB
Calculation of molecular descriptors
WEKA
The University of Waikato
Statistical analysis

Fig. 1 Necessary steps required for building a QSAR model

https://www.ebi.ac.uk/chembl/, which contains more than 12 million assay outcomes. These biological assays derive from applying more than 1.3 million drugs/chemicals against at least one out of more than 9,300 biological targets (proteins, cell lines, microorganisms, mammals), and it is possible to obtain specific information about the diverse strains or breeds in the case of non-molecular biological entities. There is also information related to the type of the assay, i.e., if a test was carried out to measure binding (B), functional (F), or ADMET (A) properties. All data can be easily downloaded into Excel-like files. Most importantly, the ChEMBL database provides good transparency concerning the accuracy and/or reliability of the experimental information. For that, two columns of ChEMBL are of particular interest, namely, (1) a "CONFIDENCE SCORE" that contains values ranging from 4 to 9 for biomacromolecular targets, the highest number being given to a very accurately determined experiment, and (2) a "CURATED BY" that shows how much the experimental information has been verified, depending on the availability of data in the literature and certain details associated to the assay protocols, with three categories: autocuration (the lowest reliability), intermediate, and expert (highest reliability). In addition, in this database, each compound has one ChEMBL identifier and SMILES (Simplified Molecular Input Line Entry Specification, i.e., the 2D chemical structure) code. Therefore, ChEMBL is an interesting and potentially valuable resource for tackling drug discovery. In fact, it has been stated that mining ChEMBL is an excellent alternative to generate models for virtual screening of large libraries of drugs and compounds [36], allowing a greater coverage of the chemical space. With these considerations in mind, ChEMBL should be the database of first choice.

A further very important database is DrugBank, which is available at http://www.drugbank.ca/. This is a unique free source chemo-bioinformatics database that combines specific drug information (chemical structure, pharmacological, and therapeutic data) with details from the respective target(s) (sequences, structures, and pathways) [35]. DrugBank comprises 6,825 drug entries including 1,541 small-molecule drugs approved by the FDA (Food and Drug Administration) and other molecular entities such as 86 nutraceuticals, 5,082 experimental chemicals, and 150 FDA-approved biotech (protein/peptide) drugs. DrugBank is a suitable source of raw data for deep studies of drug-target and drug-drug interactions. QSAR modeling using this database can facilitate the discovery of new targets for antibacterial drugs. At the same time, studies focused on drug-drug interactions can be performed in advanced stages of the drug discovery process to analyze, e.g., possible chemicals with which an antibacterial drug should not be administrated. In addition, ChEMBL and DrugBank offer the possibility of downloading substantial volumes of data in the form of tabular-based files which can be easily opened, modified, and filtered/curated using Excel.

2.2 Molecular Descriptors and Their Applicability in QSAR Approaches

Molecular descriptors are essential hot spots of any QSAR study. As pointed up nicely some years ago by Todeschini and Consonni [37]: "The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardised experiment." The use of certain types of molecular descriptors depends on the kind of QSAR approach that is going to be employed. Molecular descriptors used in classical QSAR approaches may be experimentally determined (e.g., physicochemical properties like the Hammett σ constant, refractivity, etc.) via Hansch analysis or through the use of simple indicator descriptors (R-group substitutions of the parent core), Free-Wilson analysis [38]. 3D-QSAR approaches have emerged as a natural extension to the classical QSAR approaches pioneered by Hansch and Free-Wilson and are based on correlating the activity with non-covalent molecular interaction fields [38]. These approaches require mandatorily 3D structures, e.g., based on protein crystallography or molecule superimposition, and assume that all molecules interact with the bio-target in the same manner, i.e., with an identical binding mode. In 3D-QSAR approaches, pair potentials (e.g., Lennard-Jones) are calculated for the whole molecule to model the effects of shape and size (steric fields), the presence of hydrophobic/hydrophilic regions, and the behavior of the electronic density in different parts of the molecule (electrostatic fields). *CoMFA* (*Co*mparative *M*olecular *F*ield *A*nalysis) and *CoMSIA* (*Co*mparative *M*olecular *S*imilarity *I*ndex *A*nalysis) [39] are the most applied 3D-QSAR approaches, in particular, to support docking calculations. Over time, 3D-QSAR approaches have evolved to improved variants such as 4D-QSAR, additionally including ensemble of ligand conformations and protonation states in 3D-QSAR [40]; 5D-QSAR, adding induced-fit effects of the adaptation of the receptor-binding pocket to the ligand in 4D-QSAR [41]; and 6D-QSAR approaches, further incorporating different solvation models in 5D-QSAR [42].

As an alternative to 3D-QSAR approaches, the structures of the molecules can be optimized without any superimposition of alignment. Toward that end, calculations based on, e.g., density functional theory (DFT) and molecular dynamics (MD) are carried out [43], and both local (charge, polarizability HOMO-/LUMO-based parameters) and global descriptors (different energies, volumes, and areas) can be determined and used to obtain correlations with biological activities. Further, highly accurate quantum-mechanics calculations allow the study of more specific interactions such as the strength of hydrogen interactions, stability of coordination bonds, etc.

Several works using 3D-QSAR approaches or QM calculations have been reported in the field of antimicrobial research [44–48].

But from our point of view, QSAR research should be focused on developing models best suited for the analysis of a large number of diverse compounds and virtual screening of databases. For that purpose, 3D-QSARs and related approaches have several important drawbacks, namely:

- Although improved approaches such as 4D-, 5D-, and 6D-QSARs have emerged, uncertainties regarding the selection of the active conformations and the binding mode of ligands remain.
- 3D-QSARs are usually used to study the inhibitory activity of compounds related to in vitro data. For the successful application of these approaches to in vivo data, there must be a thorough understanding of the binding mechanisms of all underlying ligands.
- For 3D-QSAR approaches, optimization of the 3D structures of ligands can be particularly time consuming. Moreover, the lowest energy conformation of any ligand is usually considered to be the bioactive conformation and responsible for exerting the binding effects.
- In the case of QM-based models, the computational cost and required time to perform calculations can be very high, depending on the complexity of the molecules to be modeled, the QM method to be applied (ranging from molecular mechanics to semiempirical and DFT methods, including higher-level QM methods), and the availability of computational resources.
- If QM calculations are performed without alignment, there will be more uncertainty than in 3D-QSAR approaches for the selection of the active ligand conformation.

These handicaps can somehow be solved if 2D topological descriptors are applied [37]. More than 100 families of topological descriptors have been described in the literature [49]. They are among the most useful sets of molecular descriptors, as corroborated by the large number of QSAR applications reported to date [50–60]. However, these descriptors have been criticized for a lack of clarity regarding their physicochemical meaning and structural interpretation [38]. Definitely, these descriptors are unable to provide a complete understanding of the 3D structural aspects of molecules, which can be considered as a possible disadvantage if the analysis of the binding mode of a ligand with its bio-target is needed. But as proven by Estrada, although topological descriptors are derived from 2D representations of the molecular structures, connectivity indices, for example, encode information related to the molecular accessibility for the surrounding medium. For instance, first- and second-order connectivity indices represent molecular accessibility areas and volumes, respectively, while higher-order connectivity indices represent hyper-volumes [61]. The same author has also demonstrated that 2D topological descriptors can account for "pure" 3D structural parameters, such as the dihedral angle between phenyl rings in alkylbiphenyls [60].

At this point, it is easy to conclude that topological descriptors do not depend on the conformation and thus do not need the superimposition rules and alignments used in 3D-QSAR approaches. Further, topological descriptors can be used in QSAR studies involving both in vitro and in vivo assay data. When these descriptors are used to generate QSAR models based on statistical classification techniques like LDA, the databases can be formed by compounds belonging to dissimilar chemical families, because the mechanisms of action are not so important. These descriptors will afford the structural patterns to successfully separate the dataset of compounds into groups having the observed biological activity or not [21–32, 54, 59]. Topological descriptors have exhibited a great applicability for virtual screening of antibacterial agents [62–66]. Finally, extremely interesting kinds of topological descriptors are the graph-based spectral moments, designed according to the *TOPS-MODE* (*TOP*ological *S*ubstructural *MO*lecular *DE*sign) approach [67–69]. The greatest advantage of the *TOPS-MODE* approach, over other traditional QSAR methods, stems from its substructural nature. This means that one can transform the resulting QSAR model into a bond additive scheme and thus describe the endpoint activity as a sum of bond contributions related to different structural fragments of the molecules. Moreover, one can detect the fragments on a given molecule that contribute positively or negatively (by summing up bond contributions) to the underlying activity [51].

Many programs have been built for the calculation of molecular descriptors. One of the most widely applied is the MODESLAB software [70] developed by Estrada and Gutierrez for Windows. This software allows the calculation of physicochemical properties (e.g., polar surface area, van der Waals radii, etc.) and classical topological descriptors like atom and bond connectivity indices, as well as other indices [49]. The program also calculates topographical descriptors resulting from mixing topological indices with QM calculation features. But the most powerful set of descriptors implemented in MODESLAB is the spectral moments. The MODESLAB software is freely available through the webpage http://www.estradalab.org/links/index.html.

Another computer program which has attracted the attention of QSAR practitioners is PaDEL-Descriptor, devised by Wei [71]. The current version of this Java-based software calculates 905 descriptors (770 1D and 2D descriptors, as well as 135 3D descriptors) and ten types of fingerprints. These are calculated principally using the well-known Chemistry Development Kit. PaDEL-Descriptor is a free open-source software, which can work on all major platforms (Windows, Linux, MacOS), supporting more than 90 dissimilar file formats to encode chemical information of the molecules. This software can be downloaded from http://padel.nus.edu.sg/software/padeldescriptor/.

But perhaps the best known program for the calculation of molecular descriptors is DRAGON [72]. In the latest version of this program, 4,885 molecular descriptors can be calculated for small, medium, and even larger molecules. Furthermore, after computing the molecular descriptors for a given dataset, correlations between variables can be analyzed by applying the same software in order to exclude redundancies. At the same time, almost any version of DRAGON allows one to carry out principal component analysis (PCA) for feature selection, in order to identify the variables which best explain the variability of the data. The DRAGON software is available at http://www.talete.mi.it/products/dragon_description.htm, covering different academic and commercial licenses.

2.3 Modeling Techniques

The last stage pertains to the creation of a statistically significant QSAR model, where the molecular descriptors serve as the independent variables and the desired endpoint response(s) as the dependent variable(s). For that purpose, there are many different modeling techniques available in a variety of software packages. This section will just focus on two of the most commonly applied statistical techniques to generate QSAR models. The first group embraces regression approaches like partial least squares (PLS) [33], fundamentally used in 3D-QSAR studies, and the traditional multiple linear regression (MLR) [38]. The second group comprises classification approaches such as LDA and ANNs.

A very important aspect should be highlighted here, whenever one is retrieving a large volume of data to develop the QSAR model. Usually databases are compilations of biological and chemical data reported in the literature, which have been determined within different laboratories by diverse workers and using several types of assay protocols, consequently making it difficult to estimate the associated experimental errors. As regression techniques try to predict the real response values of the compounds, it is then evident that if one is working with a large and heterogeneous dataset of chemicals, it would be very difficult to find a good predictive QSAR model using such techniques because it is almost impossible to control the uncertainty of the data. The aim of classification techniques is to generate lines, planes, hyperplanes, and/or surfaces able to separate compounds into different groups (e.g., active/inactive). So, one has only to preestablish the cutoff value(s) of the endpoint response(s) under study to then categorize the compounds accordingly. Our personal experience, and looking at diverse published studies in the past, led us to conclude that when classification techniques are employed even for the prediction of diverse endpoints, there is a greater possibility of obtaining high-quality QSAR models [21–32, 54, 59]. This reflects the fact that when classification techniques are used, there is no need to predict the exact value(s) of the endpoint(s) under study, so problems related to the uncertainty of the data are significantly reduced.

However, as previously commented, sometimes simple modeling techniques like LDA are unable to cope with the complexity of the data, and thus ANNs are required for gathering a deeper knowledge about a particular target phenomenon [22, 29, 31]. Any ANN is an effort to emulate biological intelligent systems (*human brain*) by simulating their structure and/or functional aspects [73]. ANNs entail simple processing units (*neurons*) that are linked by weighted connections (*synapses*), with the output (*axon*) signal transmitted through the neuron's outgoing connection. To the best of our knowledge, three types of ANNs have been widely used (Fig. 2) in drug discovery [22, 29, 31], namely, linear neural networks (LNN), multilayer perceptron (MLP), and radial basis function (RBF) networks. LNN consists of a first entry layer, comprising the input nodes directly associated to the molecular descriptors, and a final output layer—the predicted response. MLP and RBF include additionally a second layer, known as the hidden

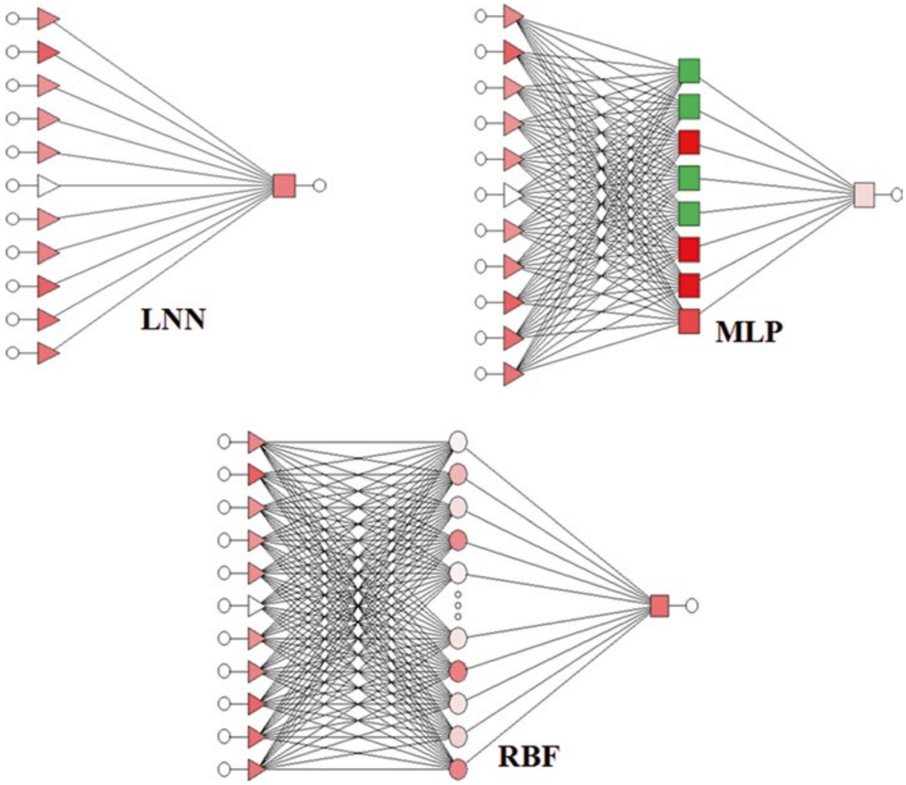

Fig. 2 Different architectures of artificial neural networks

layer, consisting of an array of neurons that receive, transform, and transmit the incoming signals from the first layer. The functions applied on the hidden layer are very important—the so-called activation functions [74], because they model the signals coming from the input layer, as well as the signal from the hidden layer to the output layer. LNN models are very similar to those which can be obtained through LDA, but the algorithms to optimize them are different. LNNs are conceived to establish direct relationships between the inputs (molecular descriptors) and the output, i.e., the endpoint activity usually expressed as a binary variable. MPLs use linear combinations of weights from the input layer to the hidden layer, and the signal from the hidden layer to the output is modeled by a nonlinear activation function (usually a sigmoid function). In RBF networks, nonlinear activation functions (usually Gaussian functions) are applied in the first step, i.e., from the input layer to the hidden layer. Following on, a linear combination of such functions is applied to generate the output layer. Regarding the nomenclature, as shown in Fig. 2, for instance, the second ANN has the form MLP 11:11-8-1:1, meaning that this is an MLP (multilayer perceptron) containing 11 variables (descriptors) which were used as entries to generate 11 input nodes in the first layer, 8 nodes in the second layer, and one node in the third (output) layer from which one response variable (endpoint activity) is to be predicted. Similarly, the first and the last ANNs shown have profiles LNN 11:11-1:1 and RBF 11:11-305-1:1, respectively. Despite the increasing use of ANNs in different areas [22, 29, 31], in the last 15 years, few papers have described the use of ANN-based models to predict antibacterial profiles of compounds [48, 75–77]. In these works, some models were derived with the aim of predicting a reduced series of analogue compounds [48, 75, 77] or heterogeneous but using relatively medium-sized datasets of chemicals [76]. This further illustrates the need to use ANNs with the aim of improving the discovery of desirable antibacterial drugs.

Of the software available for carrying out these modeling techniques, we will mention only two programs because of their applicability, user-friendly nature, and availability. One of the programs widely used in different fields of modern science is STATISTICA [73]. This is a software package developed by StatSoft, which affords data analysis and statistics, as well as data visualization and data mining procedures. The latest version of STATISTICA can handle large amounts of data and is able to tackle various file formats. This software can also generate codes for posterior programming after the QSAR model is built. There are different classification techniques implemented in STATISTICA such as LDA, ANNs, support vector machines (SVM), classification trees (CT), random forest (RF), and many others [73]. Several types of commercial and academic licenses are available for different versions of STATISTICA at http://www.statsoft.com/.

In the past, STATISTICA has often been applied to construct QSAR models aimed at predicting antibacterial activity of heterogeneous series of compounds [62–64, 66].

Another freely available program to perform statistical analysis is Weka [78]. This Java program can be viewed as a collection of machine learning algorithms for data mining purposes. The algorithms can either be applied directly to a dataset or accessed from the user's own Java code. Weka contains tools for data preprocessing, classification, regression, clustering, association rules, and visualization and a powerful tool for feature selection [78]. It is also well suited for developing new machine learning schemes. Weka can be downloaded from http://www.cs.waikato.ac.nz/ml/weka/downloading.html.

3 An mtk-QSBER Model Combined with ANN for Virtual Screening of Antibacterial Agents

Despite the wide applicability of alignment-free QSAR models based on topological descriptors, an unfavorable aspect of such models is that the antibacterial activity is predicted nonspecifically, meaning that any compound can be predicted to have antibacterial activity but without information regarding the bacteria against which it may be active [76]. What is more, antibacterial activity data should be integrated with other relevant properties such as those related to ADMET profiles to effectively discover reliable antibacterial agents. Of particular interest is that the developed models can potentially guide the screening and discovery of effective antibacterial agents. The next subsections will describe the series of steps involved in building up mtk-QSBER models based on ANNs, which in principle overcome these limitations. Details of the diverse steps for the development of mtk-QSBER-ANN models are depicted in the flowchart of Fig. 3. As a typical example application, we will describe recent work that reported an mtk-QSBER-ANN model whose role was to simultaneously predict the anti-enterococci activity and toxicity in laboratory animals of a set of compounds [27].

3.1 Curation of the ChEMBL Database

In the work referred to above [27], 13,073 endpoints belonging to more than 9,000 chemicals/drugs were retrieved from ChEMBL in an Excel-compatible file format [34]. Chemicals for which data were incomplete were eliminated, as were duplicates (i.e., the same chemicals assayed under the same experimental conditions), leading to a final dataset comprising 8,560 chemicals/drugs and to 10,918 statistical cases, considering the various experimental conditions (c_j). The experimental conditions c_j cover an ontology of the form $c_j \Rightarrow (m_e, b_t, a_i, l_c)$, where m_e describes the different measures of biological effects (anti-enterococci activity or toxicity),

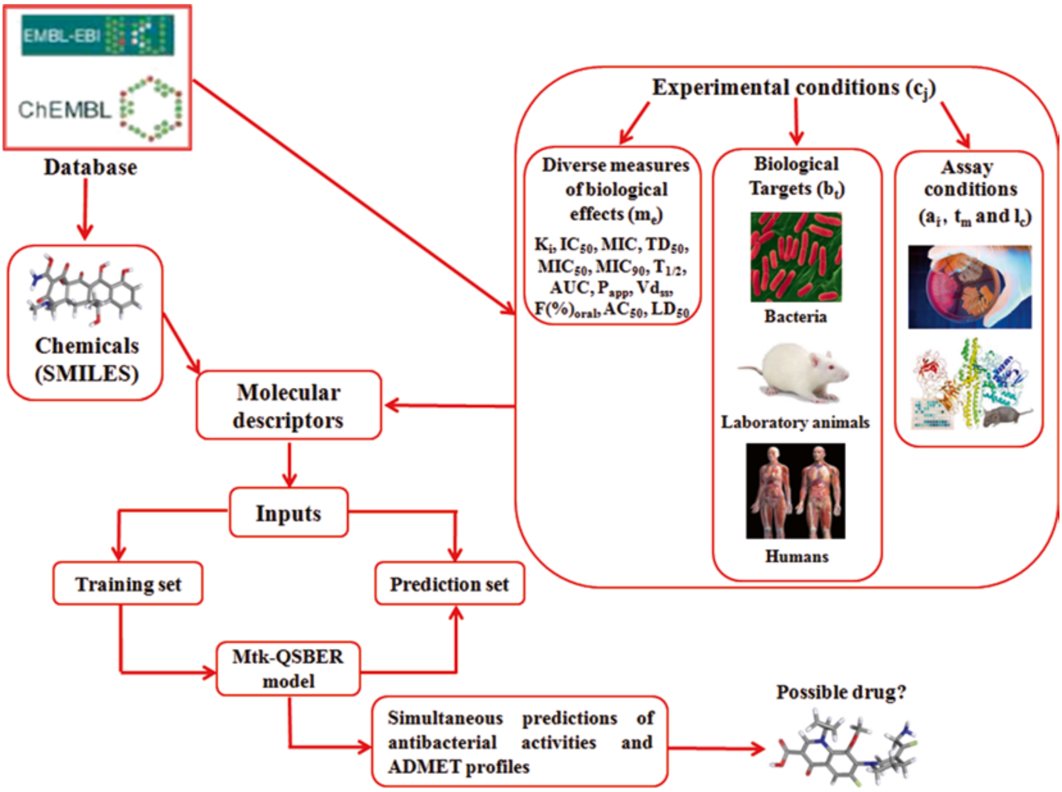
EMBL-EBI
ChEMBL
Database
Chemicals (SMILES)
Molecular descriptors
Inputs
Training set
Prediction set
Mtk-QSBER model
Simultaneous predictions of antibacterial activities and ADMET profiles
Possible drug?
Experimental conditions (c_j)
Diverse measures of biological effects (m_e)
K_i, IC_{50}, MIC, TD_{50}, MIC_{50}, MIC_{90}, $T_{1/2}$, AUC, P_{app}, Vd_{ss}, $F(\%)_{oral}$, AC_{50}, LD_{50}
Biological Targets (b_t)
Bacteria
Laboratory animals
Humans
Assay conditions
(a_i, t_m and l_c)

Fig. 3 General steps involved in the development of an mtk-QSBER model

and b_t denotes the dissimilar biological targets such as enterococci (including their respective strains), human immune system cells (lymphocytes), and laboratory animals like mice (*Mus musculus*) and rats (*Rattus norvegicus*). Additionally, a_i refers to the specific assay information, meaning that it provides details whether the assay focused on the assessment of functional (F) or ADMET profiles (A), and l_c describes the degree of curation or verification of the experimental information of a particular assay. Therefore, the elements m_e, b_t, a_i, and l_c embody factors that may be modified to obtain a specific experimental condition.

One should point out also here that the 10,918 cases were the results of using 8,560 chemicals/drugs to assess one out of 18 measures of biological effects, against at least one out of 131 biological targets, by considering at least one out of two labels of assay information, with at least one out of three categories of curation of the experimental information. An important detail to take into account is that the measures of antibacterial activity (e.g., the minimum inhibitory concentration, MIC) were often expressed in different units, such as nM or μg/mL, and something similar applies also to the toxicity measures (e.g., the median lethal dose, LD_{50}). It is clearly essential that all data for a defined measure be in the same units; for instance, antibacterial activity should be converted to nM or a multiple such as μM. All these conversions can be easily made and are necessary if one is to correctly compare all of the compounds' biological effects (anti-enterococci potency and toxicological profiles). Further, each of the 10,918 cases was assigned to one out of two possible groups related with the biological effect [$\mathrm{BE}_i(c_j)$] of a defined compound i in a specific experimental condition c_j. That is, a chemical/drug was assigned to the group of positives [$\mathrm{BE}_i(c_j) = +1$] when its value of biological effect fulfilled certain requirements within a preestablished cutoff [27]; otherwise, the compound was considered as negative [$\mathrm{BE}_i(c_j) = -1$].

3.2 Moving Average Approach

A useful predictive mtk-QSBER model clearly depends on the molecular structure of the compound and on the experimental conditions/ontology c_j under which the compound was assayed. But if the original descriptors are calculated as previously described (*see* Subheading 2.2), one will not be able to differentiate the biological effects for a given molecule when different elements of the ontology c_j are modified. For this reason, a new set of molecular descriptors was computed by applying the moving average approach [73]:

$$D_i\left(c_j\right) = D_i - D_i\left(c_j\right)_{\mathrm{avg}} \tag{1}$$

In Eq. 1, D_i is the molecular descriptor of the compound i. The descriptor $D_i(c_j)_{\mathrm{avg}}$ characterizes a defined set of n_j compounds tested under the same experimental conditions c_j. This descriptor is calculated as the average of the D_i values for the compounds in subset n_j that were considered as positive (desirable) cases [$\mathrm{BE}_i(c_j) = +1$] for the same element of c_j. It should be clarified here that, though we have resorted to an arithmetic mean for computing the $D_i(c_j)_{\mathrm{avg}}$ descriptors, which in fact is one of the best known measures of the central tendency of a list of data, other measures like geometric, harmonic means or standard scores could have been used instead or in addition.

To sum up, $\Delta D_i(c_j)$ are clearly very important descriptors because they encode information based on the chemical structure and the characteristics of c_j. That is, each $\Delta D_i(c_j)$ represents, in structural terms, how much a compound deviates from the group of molecules which were classified as positive.

3.3 Setup of the mtk-QSBER Model

Firstly, the dataset under study (10,918 cases) was randomly divided into two series: training and prediction sets. The training set contained 8,298 cases, 4,217 assigned as positive and 4,081 as negative. The prediction or external set included 2,620 cases, 1,353 positive and 1,267 negative cases.

Then, the original molecular descriptors D_i were calculated—i.e., in this case, MODESLAB spectral moments [70] and those of the type $\Delta D_i(c_j)$ determined by applying Eq. 1. The total final number of $\Delta D_i(c_j)$ descriptors can be found by multiplying the number of original descriptors by the number of elements of the ontology c_j. In this work, we have thus started with 120 $\Delta D_i(c_j)$ descriptors. As usually happens, this is a large number of molecular descriptors, and so the next task encompasses selecting a proper subset of descriptors to then build the mtk-QSBER model. In this work, we have resorted to program Weka [79], which contains a series of powerful algorithms for variable selection. Notice that the final set of descriptors should yield the best, or at least a very good, discrimination between positive and negative groups.

Two algorithms are usually followed for selecting the variable subset using Weka. In the first, known as the filter algorithm (attribute evaluator), an independent assessment based on the general characteristics of the data is performed. The second algorithm is used in combination with the first one, being focused on the subset evaluation using a defined machine learning algorithm. The latter is called the wrapper (search) algorithm, because a machine learning algorithm is wrapped into the selection procedure. More details about the filter and wrapper algorithms implemented in Weka can be found in Witten et al. [80]. We suggest using at least a minimum of two combinations of filter and wrapper algorithms, "running" the data with such combinations at least five times. In so doing, one is able to easily select the variables common to all such runs, which embrace the most important attributes. In the case under study [27], combinations of the attribute evaluator CFsSubsetEval and the wrapper algorithms called BestFirst and GeneticSearch were employed. Another possibility might be to use the forward stepwise (FS) technique as the variable selection strategy, generally along with LDA. After that, the chosen descriptors can be applied as inputs to derive the best mtk-QSBER model based on ANNs. Our previous analyses and results have shown us that when LDA is used with FS, there is a high probability of then obtaining ANNs of the type RBF. In our opinion, several techniques of variable selection should be combined to provide a solid background on which molecular descriptors do have more influence in the final model.

To find the best mtk-QSBER model based on ANNs, LNN, MLP, and RBF profiles were considered. A fourth ANN called a probabilistic neural network (PNN) was also analyzed. The software STATISTICA was used to carry out this procedure [81]. This program has a package called "Intelligent Problem Solver," which contains a huge number of tools, and one of them allows the evaluation of the most important variables, by performing a task known as sensitivity analysis. That is, with the latter, only the variables with sensitivity values higher than one are chosen to enter in the final model. "Intelligent Problem Solver" should be run at least five times to check enough different architectures and quality of the resulting ANNs. Another important aspect is to ensure that at least one descriptor belonging to each element of the ontology c_j is in the final mtk-QSBER model. Moreover, possible correlation between descriptors should be analyzed because, if some of the selected descriptors are redundant, they should be discarded, since the stability of the model as well as its ability to make future predictions might be affected.

In this work, the best model for the simultaneous prediction of anti-enterococci activity and toxicity in laboratory animals was found to be an ANN with an RBF 5:5-767-1:1 profile [27]. As can be judged from the results in Table 1, the RBF ANN had a far better accuracy and predictive power than the LNN—overall accuracy of around 59 %, two MLPs—overall classification ranging in the interval 72–77 %, and PNN—accuracy of around 61 %.

The quality and predictive power of the final mtk-QSBER model were also assessed by checking overall and group-specific performance measures on the training and prediction sets, respectively (Table 2). These included the *sensitivity*, percentage

Table 1
Comparative analysis of the different ANNs exploited in this study

	Training set		Prediction set	
ANN[a]	Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)
LNN 5:5-1:1	58.88	58.61	58.17	61.01
MLP 5:5-8-1:1	72.11	72.38	72.28	73.72
MLP 5:5-7-10-1:1	77.35	77.82	77.01	78.06
RBF 5:5-767-1:1	92.98	93.58	90.76	92.11
PNN 5:5-8298-2-2:1	95.04	27.96	94.38	28.02

[a] The first MLP is a three-layer perceptron (TLP), while the second MLP is a four-layer perceptron (FLP)

Table 2
Quality and classification performance of the mtk-QSBER model

Classification[a]	Training set	Prediction set
Sensitivity/TP (%)	92.98	90.76
Specificity/TN (%)	93.58	92.11
Accuracy (%)	93.28	91.41
MCC	0.866	0.828
AUROC	0.981	0.965

[a] TP, TN, MCC, and AUROC stand for the true positive, true negative, the Matthews correlation coefficient, and for the area under the ROC curves, respectively

of actual positives that are correctly identified as such or true positive rate; *specificity*, true negative rate; *accuracy*, percentage of correct classification for all cases; the Matthews correlation coefficient (MCC) [82]; and the area under the receiver-operating characteristic (ROC) curve [83]. The latter allows one to confirm that the model is not a random classifier, that is, a model that only correctly predicts half of the cases (with an area under the ROC curve of 0.5).

As can be seen in Table 2, the areas under the ROC curves for the mtk-QSBER model in training and predictions sets were 0.981 and 0.965, respectively, indicating that the model is not a random classifier. Also, the attained MCC values further corroborate the very good quality and performance of the proposed mtk-QSBER model, since for both the training and prediction sets they are near to one. MCC will return values between −1 and +1, with +1 representing a perfect prediction, 0 a random prediction, and −1, a total disagreement between observed and predicted biological effects.

3.4 Virtual Screening of Antibacterial Agents

Many scientists argue that QSAR models are not entirely validated as long as no novel compounds are synthesized and biologically tested. Yet a QSAR model is feasibly validated if the external cases (not used to build the model) are correctly predicted. That is the major reason for splitting the datasets into training and prediction series. The critical point here, however, is the chemical space that a derived model can cover. The greater the number of different families of compounds used to build the model, the greater will be the chemical space in which the model can be used prospectively. Besides, a well-validated QSAR model should also show promising results when applied on the virtual screening of chemicals/drugs. The best way to reveal the promising applications of the present model was to predict the multiple biological effects of the antibacterial agent known as BC-3781 [84]. This investigational drug was reported to exhibit high inhibitory activity against different strains belonging to the genus *Enterococcus*. By applying our model, BC-3781 was predicted as a possible antibacterial agent against different enterococci using multiple experimental conditions, in good agreement with the experimental evidence. No toxicological data could be obtained from the literature for this investigational drug, but the mtk-QSBER model led us to infer that BC-3781 should not be potentially harmful for laboratory animals and, in principle, toxicologically safe for humans as well. Notice here that although there is still no experimental data about the toxicity of BC-3781, our toxicity predictions explain why this compound is already undergoing phase II clinical trials and that no significant toxic/adverse effects have been reported so far for humans.

4 Conclusions and Future Perspectives

Nowadays, modern societies are aware of the devastating power of bacterial diseases and infections. The process of creating innovative antibacterial chemotherapies can be accelerated by using chemoinformatics approaches such as QSAR tools supported by powerful statistical techniques like ANNs. In this chapter, we have presented a general overview of the evolution of QSAR models in antibacterial drug discovery. Particular attention was paid to test the ability of a recent ANN-based mtk-QSBER model in the discovery and virtual screening of antibacterial agents. It was shown as well how these types of models are able to participate in dissimilar stages of drug discovery. In our opinion, future perspectives regarding the use of mtk-QSBER models combined with ANNs may be focused on extending them by forward-integrating data of other biological assays against relevant bacteria such as staphylococci, gram-negative pathogens, bacteria causing diseases like pneumonia, or those involved in the appearance and development of nosocomial infections. As a final point, one should remark here that ANNs can be viewed as graphs of interconnected nodes. For this reason, the use of complex network theory may well be useful to analyze more deeply the topology of these machine learning techniques. Perhaps, new horizons could be opened and applied to the field of antibacterial research.

Acknowledgments

A. Speck-Planche acknowledges the Portuguese Fundação para a Ciência e a Tecnologia (FCT) and the European Social Fund for financial support (grant SFRH/BD/77690/2011). This work has been further supported by FCT through grant N° Pest-C/EQB/LA0006/2011.

References

1. Grayson ML, Crowe SM et al (eds) (2010) Kucers' the use of antibiotics. A clinical review of antibacterial, antifungal, antiparasitic, and antiviral drugs, 6th edn. CRC Press, Taylor & Francis Group, LLC, Boca Raton, FL
2. Shatalin K, Shatalina E et al (2011) H2S: a universal defense against antibiotics in bacteria. Science 334:986–990
3. Cordero OX, Wildschutte H et al (2012) Ecological populations of bacteria act as socially cohesive units of antibiotic production and resistance. Science 337:1228–1231
4. Rossolini GM, Mantengoli E (2008) Antimicrobial resistance in Europe and its potential impact on empirical therapy. Clin Microbiol Infect 14(Suppl 6):2–8
5. Gonzales R, Corbett KK et al (2008) Drug resistant infections in poor countries: a shrinking window of opportunity. BMJ 336: 948–949
6. Lautenbach E, Abrutyn E (2009) Healthcare-acquired bacterial infections. In: Brachman PS, Abrutyn E (eds) Bacterial infections of humans: epidemiology and control, 4th edn. Springer Science + Business Media, LLC, New York, NY, pp 543–575
7. Rigottier-Gois L, Alberti A et al (2011) Large-scale screening of a targeted Enterococcus

faecalis mutant library identifies envelope fitness factors. PLoS One 6:e29023

8. Tenover FC, McGowan JE Jr (2009) The epidemiology of bacterial resistance to antimicrobial agents. In: Brachman PS, Abrutyn E (eds) Bacterial infections of humans: epidemiology and control, 4th edn. Springer Science + Business Media, LLC, New York, NY, pp 91–104
9. Feuerriegel S, Oberhauser B et al (2012) Sequence analysis for detection of first-line drug resistance in Mycobacterium tuberculosis strains from a high-incidence setting. BMC Microbiol 12:90
10. Lienhardt C, Glaziou P et al (2012) Global tuberculosis control: lessons learnt and future prospects. Nat Rev Microbiol 10:407–416
11. Hann MM, Oprea TI (2004) Pursuing the leadlikeness concept in pharmaceutical research. Curr Opin Chem Biol 8:255–263
12. Lazo JS, Wipf P (2000) Combinatorial chemistry and contemporary pharmacology. J Pharmacol Exp Ther 293:705–709
13. Bleicher KH, Bohm HJ et al (2003) Hit and lead generation: beyond high-throughput screening. Nat Rev Drug Discov 2:369–378
14. Hansch C, Leo A (1995) Exploring QSAR: fundamentals and applications in chemistry and biology. American Chemical Society, Washington, DC
15. Jorgensen WL (2004) The many roles of computation in drug discovery. Science 303: 1813–1818
16. Oprea T (2005) Chemoinformatics in drug discovery. WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
17. Brown N, Lewis RA (2006) Exploiting QSAR methods in lead optimization. Curr Opin Drug Discov Devel 9:419–424
18. Borchardt RT, Kerns EH et al (eds) (2006) Optimizing the "drug-like" properties of leads in drug discovery. Springer Science + Business Media, LLC, New York, NY
19. Croes S, Koop AH et al (2012) Efficacy, nephrotoxicity and ototoxicity of aminoglycosides, mathematically modelled for modelling-supported therapeutic drug monitoring. Eur J Pharm Sci 45:90–100
20. Hau J, Schapiro SJ (2011) Handbook of laboratory animal science: essential principles and practices. CRC Press, Taylor & Francis Group, LLC, Boca Raton, FL
21. Vina D, Uriarte E et al (2009) Alignment-free prediction of a drug-target complex network based on parameters of drug connectivity and protein sequence of receptors. Mol Pharm 6:825–835
22. Prado-Prado FJ, Garcia-Mera X et al (2010) Multi-target spectral moment QSAR versus ANN for antiparasitic drugs against different parasite species. Bioorg Med Chem 18: 2225–2231
23. Garcia I, Fall Y et al (2011) First computational chemistry multi-target model for anti-Alzheimer, anti-parasitic, anti-fungi, and anti-bacterial activity of GSK-3 inhibitors in vitro, in vivo, and in different cellular lines. Mol Divers 15:561–567
24. Speck-Planche A, Kleandrova VV et al (2012) Fragment-based approach for the in silico discovery of multi-target insecticides. Chemometr Intell Lab Syst 111:39–45
25. Speck-Planche A, Kleandrova VV et al (2012) In silico discovery and virtual screening of multi-target inhibitors for proteins in Mycobacterium tuberculosis. Comb Chem High Throughput Screen 15:666–673
26. Speck-Planche A, Kleandrova VV et al (2012) Chemoinformatics in anti-cancer chemotherapy: multi-target QSAR model for the in silico discovery of anti-breast cancer agents. Eur J Pharm Sci 47:273–279
27. Speck Planche A, Cordeiro MNDS (2013) In Chemoinformatics in drug design. Artificial neural networks for simultaneous prediction of anti-enterococci activities and toxicological profiles. Proceedings of the 5th International joint conference on computational intelligence, NCTA-International conference on neural computation theory and applications, Vilamoura, Algarve, Portugal, 20–22 Sept, pp 458–465
28. Luan F, Cordeiro MNDS et al (2013) TOPS-MODE model of multiplexing neuroprotective effects of drugs and experimental-theoretic study of new 1,3-rasagiline derivatives potentially useful in neurodegenerative diseases. Bioorg Med Chem 21:1870–1879
29. Tenorio-Borroto E, Penuelas Rivas CG et al (2012) ANN multiplexing model of drugs effect on macrophages; theoretical and flow cytometry study on the cytotoxicity of the anti-microbial drug G1 in spleen. Bioorg Med Chem 20:6181–6194
30. Speck-Planche A, Kleandrova VV et al (2013) New insights toward the discovery of antibacterial agents: multi-tasking QSBER model for the simultaneous prediction of anti-tuberculosis activity and toxicological profiles of drugs. Eur J Pharm Sci 48:812–818
31. Speck-Planche A, Kleandrova VV et al (2013) Chemoinformatics for rational discovery of safe antibacterial drugs: simultaneous predictions of biological activity against streptococci

and toxicological profiles in laboratory animals. Bioorg Med Chem 21:2727–2732

32. Speck-Planche A, Cordeiro MNDS (2013) Simultaneous modeling of antimycobacterial activities and ADMET profiles: a chemoinformatic approach to medicinal chemistry. Curr Top Med Chem 13:1656–1665
33. van de Waterbeemd H (1995) Chemometrics methods in molecular design. VCH Publishers, Weinheim
34. Gaulton A, Bellis LJ et al (2012) ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 40:D1100–D1107
35. Knox C, Law V et al (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic Acids Res 39:D1035–D1041
36. Mok NY, Brenk R (2011) Mining the ChEMBL database: an efficient chemoinformatics workflow for assembling an ion channel-focused screening library. J Chem Inf Model 51: 2449–2454
37. Todeschini R, Consonni V (2000) Handbook of molecular descriptors. WILEY-VCH Verlag GmbH, Weinheim
38. Kubinyi H (1993) QSAR: Hansch analysis and related approaches. VCH Publishers, Weinheim
39. Kubinyi H, Folkers G et al (eds) (2002) 3D QSAR in drug design: recent advances. Kluwer Academic Publishers, New York
40. Klein CD, Hopfinger AJ (1998) Pharmacological activity and membrane interactions of antiarrhythmics: 4D-QSAR/QSPR analysis. Pharm Res 15:303–311
41. Vedani A, Dobler M (2002) 5D-QSAR: the key for simulating induced fit? J Med Chem 45:2139–2149
42. Vedani A, Dobler M et al (2005) Combining protein modeling and 6D-QSAR. Simulating the binding of structurally diverse ligands to the estrogen receptor. J Med Chem 48:3700–3703
43. Carloni P, Alber F (eds) (2003) Quantum medicinal chemistry. WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
44. Li P, Yin J et al (2013) Synthesis, antibacterial activities, and 3D-QSAR of sulfone derivatives containing 1,3,4-oxadiazole moiety. Chem Biol Drug Des 82:546–556
45. Lu X, Lv M et al (2012) Pharmacophore and molecular docking guided 3D-QSAR study of bacterial enoyl-ACP reductase (FabI) inhibitors. Int J Mol Sci 13:6620–6638
46. Uddin R, Lodhi MU et al (2012) Combined pharmacophore and 3D-QSAR study on a series of Staphylococcus aureus Sortase A inhibitors. Chem Biol Drug Des 80:300–314
47. Bhonsle JB, Venugopal D et al (2007) Application of 3D-QSAR for identification of descriptors defining bioactivity of antimicrobial peptides. J Med Chem 50:6545–6553
48. Bucinski A et al (2004) Artificial neural networks for prediction of antibacterial activity in series of imidazole derivatives. Comb Chem High Throughput Screen 7:327–336
49. Todeschini R, Consonni V (2009) Molecular descriptors for chemoinformatics. WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
50. Estrada E, Matamala AR (2007) Generalized topological indices. Modeling gas-phase rate coefficients of atmospheric relevance. J Chem Inf Model 47:794–804
51. Estrada E, Uriarte E et al (2000) A novel approach for the virtual screening and rational design of anticancer compounds. J Med Chem 43:1975–1985
52. Roy K, Ghosh G (2004) QSTR with extended topochemical atom indices. 2. Fish toxicity of substituted benzenes. J Chem Inf Comput Sci 44:559–567
53. Roy K, Ghosh G (2008) QSTR with extended topochemical atom indices. 10. Modeling of toxicity of organic chemicals to humans using different chemometric tools. Chem Biol Drug Des 72:383–394
54. Castillo-Garit JA, Vega MC et al (2011) Ligand-based discovery of novel trypanosomicidal drug-like compounds: in silico identification and experimental support. Eur J Med Chem 46:3324–3330
55. Casañola-Martin GM, Marrero-Ponce Y et al (2010) Bond-based 2D quadratic fingerprints in QSAR studies: virtual and in vitro tyrosinase inhibitory activity elucidation. Chem Biol Drug Des 76:538–545
56. Barigye SJ, Marrero-Ponce Y et al (2013) Event-based criteria in GT-STAF information indices: theory, exploratory diversity analysis and QSPR applications. SAR QSAR Environ Res 24:3–34
57. Barigye SJ, Marrero-Ponce Y et al (2013) Relations frequency hypermatrices in mutual, conditional and joint entropy-based information indices. J Comput Chem 34:259–274
58. Vazquez-Prieto S, Gonzalez-Diaz H et al (2013) A QSPR-like model for multilocus genotype networks of Fasciola hepatica in Northwest Spain. J Theor Biol 343C:16–24
59. Alonso N, Caamano O et al (2013) Model for high-throughput screening of multi-target drugs in chemical neurosciences; synthesis, assay and theoretic study of rasagiline carbamates. ACS Chem Neurosci 4:1393–1403

60. Estrada E, Molina E et al (2001) Can 3D structural parameters be predicted from 2D (topological) molecular descriptors? J Chem Inf Comput Sci 41:1015–1021
61. Estrada E (2002) Physicochemical interpretation of molecular connectivity indices. J Phys Chem A 106:9085–9091
62. Molina E, Diaz HG et al (2004) Designing antibacterial compounds through a topological substructural approach. J Chem Inf Comput Sci 44:515–521
63. Gonzalez-Diaz H, Torres-Gomez LA et al (2005) Markovian chemicals "in silico" design (MARCH-INSIDE), a promising approach for computer-aided molecular design III: 2.5D indices for the discovery of antibacterials. J Mol Model 11:116–123
64. Marrero-Ponce Y, Marrero RM et al (2006) Non-stochastic and stochastic linear indices of the molecular pseudograph's atom-adjacency matrix: a novel approach for computational in silico screening and "rational" selection of new lead antibacterial agents. J Mol Model 12:255–271
65. Marrero-Ponce Y, Medina-Marrero R et al (2005) Atom, atom-type, and total nonstochastic and stochastic quadratic fingerprints: a promising approach for modeling of antibacterial activity. Bioorg Med Chem 13:2881–2899
66. Speck-Planche A, Scotti MT et al (2009) Design of novel antituberculosis compounds using graph-theoretical and substructural approaches. Mol Divers 13:445–458
67. Estrada E (1996) Spectral moments of the edge adjacency matrix in molecular graphs. 1. Definition and applications for the prediction of physical properties of alkanes. J Chem Inf Comput Sci 36:844–849
68. Estrada E (1997) Spectral moments of the edge adjacency matrix in molecular graphs. 2. Molecules containing heteroatoms and QSAR applications. J Chem Inf Comput Sci 37:320–328
69. Estrada E (1998) Spectral moments of the edge adjacency matrix in molecular graphs. 3. Molecules containing cycles. J Chem Inf Comput Sci 38:23–27
70. Estrada E, Gutiérrez Y (2002–2004) MODESLAB. v1.5, Santiago de Compostela
71. Yap CW (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem 32:1466–1474
72. Todeschini R, Lasagni M et al (1994) New molecular descriptors for 2D and 3D structures. Theory. J Chemometr 8:263–272
73. Hill T, Lewicki P (2006) STATISTICS methods and applications. A comprehensive reference for science, industry and data mining. StatSoft, Tulsa
74. Suzuki K (ed) (2011) Artificial neural networks: methodological advances and biomedical applications. InTech, Rijeka
75. Sabet R, Fassihi A et al (2012) Computer-aided design of novel antibacterial 3-hydroxypyridine-4-ones: application of QSAR methods based on the MOLMAP approach. J Comput Aided Mol Des 26:349–361
76. Garcia-Domenech R, de Julian-Ortiz JV (1998) Antimicrobial activity characterization in a heterogeneous group of compounds. J Chem Inf Comput Sci 38:445–449
77. Lata S, Sharma BK et al (2007) Analysis and prediction of antibacterial peptides. BMC Bioinformatics 8:263
78. Hall M, Frank E et al (2009) The WEKA data mining software: an update. SIGKDD Explor 11:10–18
79. Hall M, Frank E et al (1999–2013) WEKA. Waikato Environment for Knowledge Analysis. v3.6.9, Hamilton
80. Witten IH, Frank E et al (2011) Data mining: practical machine learning tools and techniques. Morgan Kaufmann Publishers, Elsevier, Amsterdam
81. StatSoft (2001) STATISTICA 6.0. Data analysis software system
82. González-Díaz H, Pérez-Bello A et al (2007) Chemometrics for QSAR with low sequence homology: Mycobacterial promoter sequences recognition with 2D-RNA entropies. Chemometr Intell Lab Syst 85:20–26
83. Hanczar B, Hua J et al (2010) Small-sample precision of ROC-related estimates. Bioinformatics 26:822–830
84. Sader HS, Biedenbach DJ et al (2012) Antimicrobial activity of the investigational pleuromutilin compound BC-3781 tested against Gram-positive organisms commonly associated with acute bacterial skin and skin structure infections. Antimicrob Agents Chemother 56: 1619–1623