Current Computer-Aided Drug Design. 2013; 9(2): 206-225

# Evolutionary computation and QSAR research

Vanessa Aguiar-Pulido[1], Marcos Gestal[1], Maykel Cruz-Monteagudo[2,3,4], Juan R. Rabuñal[1], Julián Dorado[1] and Cristian R. Munteanu[1]

[1] *Department of Information and Communication Technologies, Computer Science Faculty, University of A Coruña, 15071 A Coruña, Spain*
[2] *CIQ, Department of Chemistry and Biochemistry, Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal*
[3] *REQUIMTE, Department of Chemistry and Biochemistry, Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal*
[4] *Centro de Estudios de Química Aplicada (CEQA), Facultad de Química y Farmacia, Universidad Central "Marta Abreu" de Las Villas, Santa Clara, 54830, Cuba*
[5] *Molecular Simulation and Drug Design Group, Centro de Bioactivos Químicos (CBQ), Universidad Central "Marta Abreu" de Las Villas, Santa Clara, 54830, Cuba*

**Abstract:**
The successful high throughput screening of molecule libraries for a specific biological property is one of the main improvements in drug discovery. The virtual molecular filtering and screening relies greatly on quantitative structure-activity relationship (QSAR) analysis, a mathematical model that correlates the activity of a molecule with molecular descriptors. QSAR models have the potential to reduce the costly failure of drug candidates in advanced (clinical) stages by filtering combinatorial libraries, eliminating candidates with a predicted toxic effect and poor pharmacokinetic profiles, and reducing the number of experiments. To obtain a predictive and reliable QSAR model, scientists use methods from various fields such as molecular modeling, pattern recognition, machine learning or artificial intelligence. QSAR modeling relies on three main steps: molecular structure codification into molecular descriptors, selection of relevant variables in the context of the analyzed activity, and search of the optimal mathematical model that correlates the molecular descriptors with a specific activity. Since a variety of techniques from statistics and artificial intelligence can aid variable selection and model building steps, this review focuses on the evolutionary computation methods supporting these tasks. Thus, this review explains the basic of the genetic algorithms and genetic programming as evolutionary computation approaches, the selection methods for high-dimensional data in QSAR, the methods to build QSAR models, the current evolutionary feature selection methods and applications in QSAR and the future trend on the joint or multi-task feature selection methods.

**Keywords:** Evolutionary computation; Feature extraction; Genetic algorithms; Genetic programming; Molecular descriptors; Quantitative structure-activity relationships; QSAR; Variable selection.

## 1. INTRODUCTION

One of the main improvements of the drug discovery process is represented by the possibility of a high throughput screening (HTS) of molecule libraries for a specific biological property [1]. The virtual molecular filtering and screening relies greatly on quantitative structure-activity relationships (QSAR), proposed by Hansch in the early 1960s [2] and on quantitative structure-property relationships (QSPR). QSAR and QSPR generate computational models that predict a biological activity, chemical reactivity or physicochemical property from the molecular structure of a chemical compound. In chemistry, classification techniques were introduced as pattern recognition techniques, and were developed separate from the QSAR of Hansch by researchers from chemometrics such as Jurs [3]. The most general mathematical formula of a QSAR model is represented in Eq. 1.

Activity = *f*(physiochemical properties and/or structural properties) (1)

The use of QSAR models reduces the costly failure of drug candidates in clinical trials by filtering the combinatorial libraries, by eliminating the compounds with a predicted toxic effect and poor pharmacokinetic properties [4] and by reducing the number of experiments. The difficulty of obtaining QSAR models creates the necessity of using methods from various fields such as molecular modeling, pattern recognition, machine learning or artificial intelligence [5-8].

A QSAR model can be built by using the following steps: computation of molecular descriptors from the molecular structure, selection of the proper variables in the context of the analyzed activity, and search of the optimal mathematical model that correlates a specific biological activity with the molecular structure. The molecular descriptors are a series of numbers that characterize a molecule and contain the information derived from the atomic covalent structure [9]. The activity information is not explicitly included in the molecular structure and, therefore, the molecular descriptors [10] are based on different molecular properties that range from physicochemical and quantum-chemical to geometrical [11] and topological features [12].

The topological descriptors can be extended to macromolecules such as proteins [13] and nucleic acids [14, 15] by using the complex network or graph theory considering the nodes as amino acids [16] or nucleic acids [17] linked by chemical bonds. The same concept of QSAR can be extended to other complex system such as protein-protein interaction [18, 19], drug-protein [20], gene [21] or drug-parasite [22] networks by considering proteins, drugs, genes or parasites [23, 24] as the nodes of the network. In addition, the amplitudes of the proteome spectrum [25, 26] can be used as nodes of a spectrum graph in order to build quantitative proteome-disease relationship (QPDR) or quantitative structure-property relationship (QSPR) models [27, 30]. Several previous papers have presented applications in bioinformatics, complex networks, artificial intelligence, pharmacology, metabolism drug design or medicine [31, 34].

The selection of the molecular descriptors (MDs) and the search for finding the best mathematical model between a structure and activity use a large spectrum of methods, such as regression, linear discriminant models, artificial neural networks, genetic algorithms or machine learning. In this review we overview the main aspects of employing evolutionary computation [35-37] in QSAR research.

## 2. EVOLUTIONARY COMPUTATION

Scientific fields need to tackle with more complex problems as time goes by. In addition to this, the time and effort required to solve these problems when using conventional techniques also increase remarkably. This can happen either because, initially, the way to find a solution is unknown or due to the high complexity of the technique's implementation.

Many times, however, solutions may be found observing the environment thoroughly. In this sense, the survival of species and organisms living in it could be considered the greatest challenge that any system may raise; a challenge that nature has solved since the beginning of time, providing a great variety of valid solutions. At this point, the following question arises: how does nature find these solutions? This has already been answered by Darwin in his Theory of the Evolution of Species: using the mechanism of natural selection and with the survival of the fittest individuals [38]. In this context, the human race plays a major role due to its predominant position in nature. The reason for this is still unknown, but the most widely accepted assumption points toward the human cognitive activity: humans have a higher intellectual capacity.

### *2.1. Genetic Algorithms and Genetic Programming*

Evolutionary Computation (EC) is inspired in Darwin's theory. Thus, EC uses the concepts of evolution and genetics to provide solutions to problems, obtaining excellent results, especially for optimization tasks [35-37]. However, the work of Arthur Samuel and Alan Turing in the 50's regarding whether machines are capable of thinking and whether computers are capable of learning to solve problems without being explicitly programmed must also be taken into account.

Broadly speaking, EC methods can be defined as search and optimization techniques that apply heuristic rules based on natural evolution principles, that is, algorithms that look for solutions based on genetics and evolution properties. Among these properties, the survival of the fittest individuals (which implies that the best solutions to a problem will be maintained once they are found) and heterogeneity (basic heterogeneity so that algorithms have multiple types of information when generating solutions) become particularly relevant.

#### *2.1.1. Evolutionary Algorithm Performance*

Evolutionary algorithms work following a relatively simple outline, as shown in Fig. (1). The algorithm will iteratively refine solutions, progressively approaching a definitive solution to the problem to be solved.

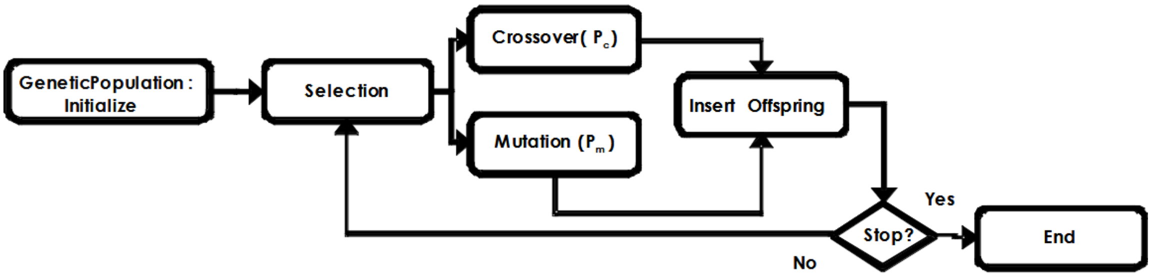

**Fig. (1)**. General outline of performance of an evolutionary algorithm.

Before implementing the evolutionary process in itself, some important decisions must be made: one regarding the encoding, that is, how solutions will be represented, and the other one regarding what is known as fitness function, which is what will determine how accurate a solution is.

Regarding the first one, there are mainly two options (see Fig.2): using a list of values (integer, real, bits, etc.) of fixed or variable length, or using a tree-shaped representation (in which the leaves usually represent the values and the intermediate nodes represent the operators). Depending on the encoding strategy chosen, a different technique will be used: Genetic Algorithm (GA) [39] or Genetic Programming (GP) [40, 41] respectively. Regardless of the technique used, the solutions obtained by the algorithms are known as genetic individuals. Each component (gene for GA or leaf node for PG) represents the variables or parameters involved in the problem.

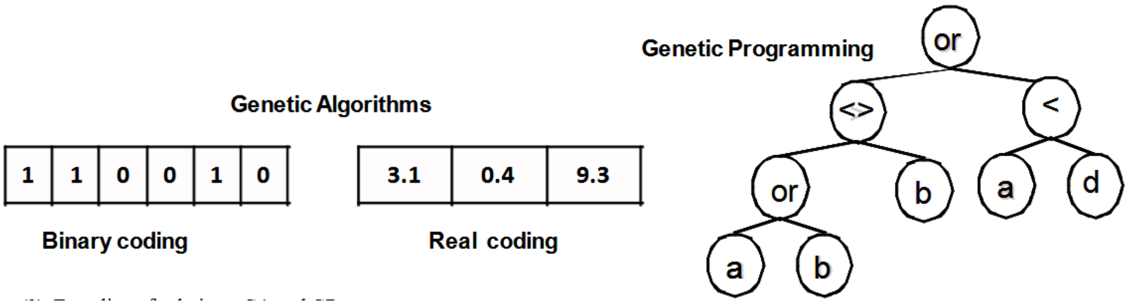

**Fig. (2)**. Encoding of solutions: GA and GP.

However, maybe the most critical part of implementing an evolutionary algorithm is the definition of the fitness function. This function will determine the accuracy of an individual, that is, how good it is. Therefore, each individual will be evaluated and rated with a real value. This function is responsible for guiding the search process towards one direction or another. Since this function is responsible for determining the goodness of each solution, this is feature is specific to the problem to be solved.

Through the application of the different genetic operators, crossover and mutation, the solutions will be evolved. These mechanisms simulate the analogue processes of sexual and asexual reproduction that take place in nature. Subsequently, each step involved in this kind of algorithm will be discussed.

### *2.1.2. Initialization*

The power of evolutionary algorithms lies in the large-scale parallel exploration of search space. Thus, multiple solutions representing the population will co-exist, each one exploring a region of the search space. The size of the population, in general, will remain stable along the whole process and the population will change as the generations go by. These changes will be directed by the fitness function in such a way that the different solutions meet constantly the global solution.

### *2.1.3. Selection*

Selection algorithms will choose which individuals will have the opportunity to reproduce or not, imitating nature's work. Accordingly, more opportunities will be given to the fittest individuals. It is clear then that selecting an individual will be related to its fitness value. However, less fit individuals should not be completely discarded: not including a part of them in this process would cause the population to become uniform in few generations.

Although there exist several selection algorithms, one of the best known ones may be the roulette-wheel selection algorithm (see Fig. 3), also called fitness proportional selection. In this algorithm, a proportion of the wheel is assigned to each of the possible selections based on their fitness value, being the sum of all percentages one unit. Hence, the best individuals will be assigned a bigger roulette-wheel proportion than the weakest ones, being less likely to be eliminated. In order to select an individual, a random number is generated within the interval [0,1], selecting the individual that corresponds to it. Thus, the selection is made in a similar way to how a roulette wheel is rotated.

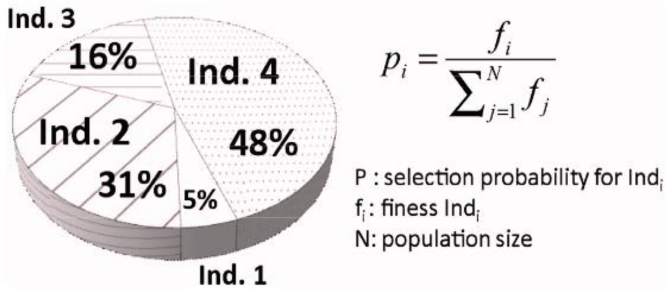

**Fig. (3)**. Roulette-wheel selection.

### *2.1.4. Crossover*

Once the individuals that will take part in the reproduction are selected, they are recombined to produce the offspring that will be inserted into the next generation. This mechanism is highly relevant for the transition between generations, being the usual crossover rates around 90%. The main idea of crossover is based on the fact that if two individuals, properly adapted to the environment, are selected and the offspring obtained share genetic information of both, it is likely that the inherited information is quite the cause of their parents' goodness. Since they share the good features of two individuals, the offspring, or at least part of them, should have better characters than each parent separately.

Different crossover operators will be used for GA and GP. In the case of GA, there exist numerous crossover algorithms, being maybe the following the most widespread ones: 1-point Crossover, 2-point Crossover and uniform Crossover (see Fig. 4). Regarding GP, the most extended crossover implementation is the exchange of sub-trees (see Fig. 5). In the 1-point crossover, once two individuals are chosen, their chromosomes are cut at an arbitrarily selected site in order to generate two segments that combine the head and tail of the parents as shown below. Hence, both descendants inherit genetic information from their parents. The 2-point crossover is similar to the 1-point crossover but performing two cuts. Thus, the offspring are generated by choosing the central segment of one of the parents and the lateral segments of the other parent.

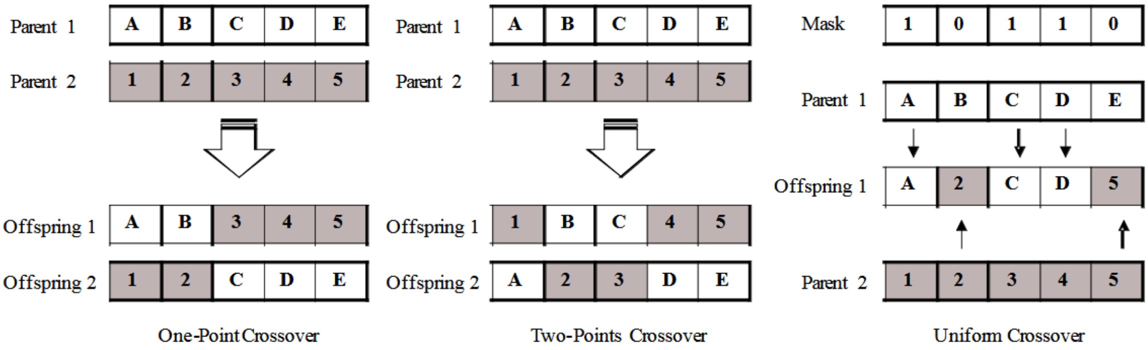

**Fig. (4)**. Crossover operator in genetic algorithms.

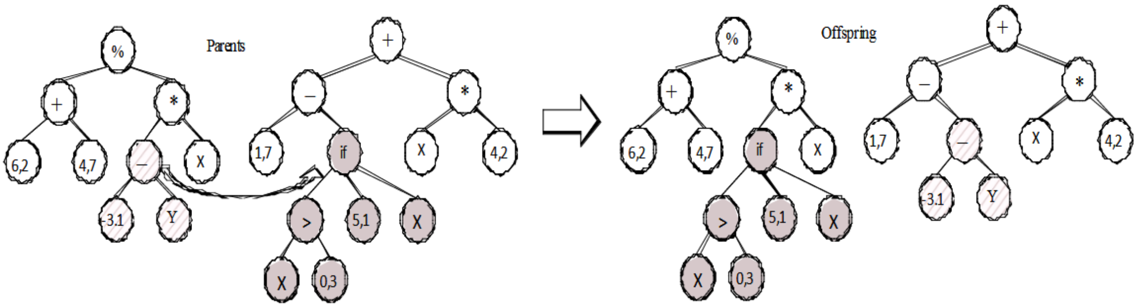

**Fig. (5).** Crossover operator in genetic programming.

Finally, the uniform crossover is totally different from the previous implementations. In this case, each gene of the offspring has the same likelihood of belonging to one or the other parent. Although there exist multiple implementations of this type of crossover, a crossover mask with binary values is always used. When there is a "1" in one of the mask positions, the gene located at that same position in one of the offspring is copied from the first parent. Otherwise, that is, when there is a "0", the gene is copied from the second parent. To generate the second offspring, there are two possibilities: either the parents' roles are exchanged or the interpretation of ones and zeros of the crossover mask is inverted. In the case of GP, the genetic recombination is generally performed by first choosing a non-terminal node from each parent and then exchanging the sub-trees hanging from these nodes.

### *2.1.5. Mutation*

The mutation operator, which is generally used in combination with the crossover operator, implies a modification of the value of one of the individual's genes or nodes. After the crossover has been performed, one or both of the offspring obtained will be mutated depending on a probability, *Pm*. This way, a behavior that occurs in Nature is reproduced: when the offspring are generated there is always some sort of error, normally without any effect on the transmission of the genetic information from parents to offspring. Occasionally, the mutation provokes a reduction in the fitness value of the individual (which might be rectified in successive generations). Nevertheless, the new information implies an important increase in the goodness of the solutions or should be taken into account for an improved solution in future generations. In general, the mutation is assigned a quite low probability, that is, less than 1%. This is mostly due to the fact that individuals tend to have a lower fitness after mutation.

Once more, the implementation of the mutation operator depends on the type of encoding used. In the case of GA, mutation usually involves randomly modifying the value of one or more genes or replacing a gene (see Fig. 6). In contrast, if using GP, the mutation operator can involve simple mutations (functional or terminal) or subtree mutations (see Fig. 7).

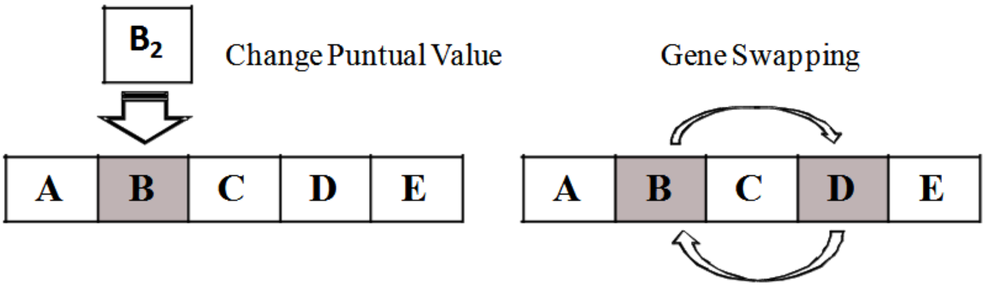

**Fig. (6).** Mutation operator in genetic algorithms.

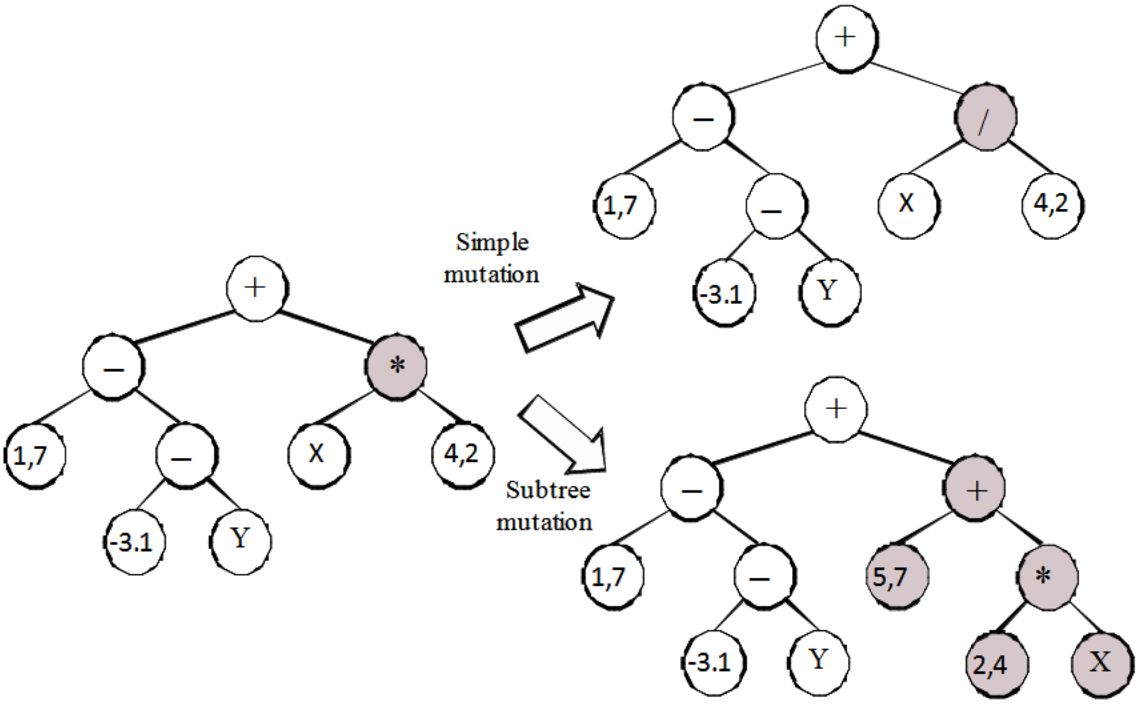

**Fig. (7).** Mutation operator in genetic programming.

*2.1.6. Stopping Criterion*

As mentioned before, solutions evolve through a highly iterative mechanism. Hence, there must exist a criterion that determines when the execution has concluded. Although there are different possibilities, below, the most frequently used are described:

- The fittest individuals that are part of the population represent solutions good enough to solve the problem.

- The population has converged, that is, the population has become very much alike. When this happens, the average goodness of the population is close to the goodness of the fittest individual.

- The difference between the best solutions from subsequent generations is very small. In the best possible scenario, this may mean that the population has already attained a global solution or, in contrast, that the algorithm has fallen into a local minimum.

- A fixed maximum number of generations has been reached.

After introducing main ideas behind GAs and GP, the next step is the setup of these two techniques by translating the QSAR problems into *evolutionary* terms. First of all, when GAs are used, two main approaches could be used to address the variable selection process. In the first one, the members of the population (chromosomes) will encode all the existing features as bit strings. Each bit will represent the either presence or absence of the feature, so if a bit is “active” (has a value 1) it will mean that the corresponding feature will take into account the next process (classification, ...). On the other hand, if the bit is “inactive” (has a value 0) the corresponding featuring will be discarded.

Another possible codification is obtained when each gene fully represents the feature and not only its presence/absence status, for example with an index, a character string with its name, while the chromosome length will inform about the total number of features selected. Here, fixed or variable-length genetic individuals could be used. In both codifications, the fitness function consists of the error of the models over the training data, based on the variables selected by the GA.

In the case that we use a GP-based codification that approach could be extended, allowing the genetic individuals to express a rule to perform the classification. In this case, the leaf-nodes in the best tree will represent the selected features. Furthermore, an expert would provide the different operators allowed and set the valid intervals to the constants used to perform comparisons between features values. It will allow the construction of more concrete and precise rules.

## 3. QSAR

QSARs are the final result of the process that starts with a suitable description of molecular structures and ends with some inference, hypothesis, and prediction on the properties of molecules in environmental, biological, and physicochemical systems in analysis [42].

The chemical structure is susceptible of many numerical representations, commonly known as molecular descriptors or MDs. These MDs map the structure of the molecules into a set of numerical or binary values that characterize specific molecular properties which can explain an activity. MDs are based on several different theories (quantum-chemistry, information theory, organic chemistry, graph theory, etc.) and are used to model several different properties of chemicals in many scientific fields (toxicology, analytical chemistry, physical chemistry, and medicinal, pharmaceutical, and environmental chemistry). Just to have an idea of the number of available MDs, the books *Molecular Descriptors for Chemoinformatics and Molecular Topology* collect definitions, formulas, and short comments of over 30,000 MDs known in chemical literature up to 2008 [43-50].

Depending on whether the included information is about the 3D orientation or conformation of molecules, the MDs are grouped into 2D and 3D QSAR descriptors. The 2D QSAR group contains the following:

- constitutional descriptors: molecular weight, total number of atoms in the molecule and number of atoms of different identity, total number of single, double, triple or aromatic type bonds, as well as number of aromatic rings;

- electrostatic and quantum-chemical descriptors that capture information on the electronic nature of the molecule, including descriptors containing information on atomic net and partial charges, molecular polarizability [51];

- topological descriptors: they treat the molecular structure as a graph where the nodes are the atoms and the connections are the chemical bonds; this group includes the Randiindices [52] (sum of geometric averages of edge degrees of atoms with paths of given lengths), Wiener index [53] (the total number of bonds in the shortest paths between all pairs of non-hydrogen atoms), Balaban’s index [54], Kier and Hall indices [55] and Gálvez indices [56], Diudea’s TI [57], new distance TI [58], Trinajstic’s TI [59], Pogliani’s valence TI [60], Bonchev’s information TI [61], Kier’s electrotopological state [62], Hosoya’s TI [63];

  - geometrical descriptors: information about molecular surface from the atomic van der Waals areas and their overlap, molecular volume, principal moments of inertia and gravitational indices, shadow areas, total solvent-accessible surface area [64];

  - fragment-based descriptors and molecular fingerprints: they are based on the substructural motifs such as BCI fingerprints [65], fast random elimination of descriptors/substructure keys (FRED/SKEYS) [66] or hologram QSAR (HQSAR) [67];

  - topological index-based pharmacophores: descriptors for sub-structural patterns (topological pharmacophores) that are linked with the biological property of the molecules [68].

The 3D QSAR descriptors can be alignment-dependent or alignment-independent. The first group is based on different computational methods due to superimposing structures in space such as the following:

- comparative molecular field analysis (CoMFA) is using the electrostatic (Coulombic) and steric (van der Waals) energy fields [69];

- comparative molecular similarity indices analysis (CoMSIA) [70] is similar with CoMFA but it uses a Gaussian-type function as potential function.

The alignment-independent 3D QSAR MDs are invariant to molecule rotation and translation in space, no superposition of compounds being required. They contain the following MDs: comparative molecular moment analysis (CoMMA) [71] descriptors that use the second-order moments of the mass distribution and charge distributions; weighted holistic invariant molecular (WHIM) descriptors [72] and molecular surface WHIM [73] descriptors that provide the invariant information by applying principal component analysis (PCA) on the centered coordinates of the atoms that constitute the molecule; VolSurf descriptors [74] are based on probing the grid around the molecule with specific probes (for example, hydrophobic interactions or hydrogen bond acceptor or donor groups); grid-independent descriptors (GRIND) [75] that overcome the problems with interpretability, which are common in alignment-independent descriptors.

There are many applications that can generate hundreds or thousands of different MDs such as MARCH-INSIDE [76], S2SNet [30], Dragon [77]or TOPS-MODE [78]. Only a couple of these descriptors can be linked with a specific molecular activity and many of them are intercorrelated. In addition, the use of a large MD set requires a large dataset in order to provide a good model. The inclusion of non-relevant variables, redundant information and highly correlated variables in the model usually leads to overfitting and lack of interpretability for the final QSAR model [79-81]. Consequently, most of the QSAR applications are based on high-dimensional multivariate data and require performing variable selection in order to find the best QSAR model.

### *3.1. Why we Need Variable Selection in QSAR*

At the inaugural lecture of the 18th *EuroQSAR* Symposium, Professor Hugo Kubinyi differentiated between “good” and “poor” QSARs based on four key elements (see Table1): i) nature of the variables or parameters; ii) number of variables to select; iii) number of variables in the model; and iv) the predictive ability of the model and the way it is evaluated [79]. Depending on the nature of the variables used, the final relationship can be not necessarily causative and consequently, the model lacks biophysical meaning, which may lead to misleading interpretations [82, 83]. Equally pathological is the presence of too many variables at the selection phase or in the final model. In both cases, there is a high probability of obtaining chance correlations, overfitted models, or in the best case, just lack of interpretability of the model. Equally determinant for the reliability of the QSAR model derived is its predictive ability and the way it is evaluated. Cross-validation (CV), using the original variables (leave-one-out CV, leave-many-out CV) is insufficient for model validation [84]. Yscrambling [85], either using the original variables or after performing variable selection, may be misleading. Leave-one-out CV with new variable selection in every CV run is misleading in larger data sets [79, 86]. Thus, the leave-many-out (up to 30%) cross-validation with new variable selection in every CV run is one reliable validation procedure [79].

**Table 1**. Positive and Negative Aspects in QSAR (a General Perspective)

| “Poor” QSAR | “Good” QSAR | Pathologies of “Poor” QSAR Models Related to High Dimensionality |
|---|---|---|
| Artificial parameters | Parameters with biophysical relevance | Non-qualitative (biophysical) model; non-causal relationship; lack of interpretability |
| Too many variables to select | Few variables to select | Chance correlations; overfitting; lack of interpretability |
| Many variables in the model | Few variables in the model | Chance correlations; overfitting; lack of interpretability |
| No test set predictivity (“Kubinyi paradox”) | Leave-many-out cross-validation | Statistical Unicorns |

If the predictive ability of the model is not properly validated and it is based on variables lacking some structural or biophysical sense when using a high dimensional QSAR data, the probability of generating a statistically significant but unreliable QSAR models [80, 81, 83, 87, 88] will be very high. As suggested by Unger and Hansch [89] we may generate what they call "statistical unicorns", that is, beasts that only exist on papers. As it can be noted, all four key elements to generate predictive, biophysically meaningful and interpretable QSAR models pass through dimensionality reduction and variable selection.

### *3.2. Selection Methods for High-Dimensional Data in QSAR*

Variable selection, also known as feature selection, attribute selection or variable subset selection, is the process of selecting, from the whole set of variables, the most relevant ones, so that as much information as possible is contained in a reduced amount of variables or features. Hence, this phase will be crucial. Due to the large number of existing variable selection techniques, identifying the best one becomes a difficult task. For this reason, it is desirable for these techniques to be fast, automated, and applicable to large data sets of structurally diverse compounds. Automatic selection methods of the best descriptors for the construction of a QSAR model can be grouped into filtering, wrapper and hybrid methods [90].

Filtering methods do not depend on the method used to build the QSAR model and consist of the following:

- Correlation-based methods – based on the Pearson's correlation coefficients in order to rule out the intercorrelated descriptors and consist in creating clusters of descriptors, for which correlation coefficients are higher than a certain threshold and retaining only one. This descriptor can be randomly chosen among the members of each cluster [91] or as a result of estimating the correlations between pairs of descriptors (if the estimation exceeds a threshold, it will be randomly done ruling out one of the descriptors). Nevertheless, it is important to note that this method could be applied only if the value of descriptors and experimental data are normally distributed [92].
- Methods based on the Information Theory – the information content of the MD is defined in terms of the entropy of descriptor and treated as a random variable. The method defines various ranking measures, obtained from the information shared between two descriptors or between a descriptor and the activity such as the mutual information [93].
- Statistical Criteria – can be used to rank the descriptors: the correlation between pairs of Fisher's ratio (ratio of the between class variance to the within-class variance) [94] or the quality of MD based on the Kolmogorov-Smirnov [95].

Wrapper methods work linked to the mapping algorithm for the QSAR model building [96]. Therefore, the error of the mapping algorithm for a given subset measured, for example, using cross-validation will be used to choose the best subset of descriptors. This group of variable selection methods contains the following:

- Principal component analysis (PCA) [97], which is the dominant approach included in industrial applications, tries to obtain a new set of variables (referred to as principal components) that will describe the data in order of decreasing variance. PCA can be seen as a method used to determine the natural dimensionality of the dataset, allowing subsequent embedding of the data into a space of lower dimensionality within a margin of prescribed original variance percentage. This mathematical procedure converts a set of observations of possibly correlated variables into a set of values of uncorrelated variables by means of an orthogonal transformation. These variables are called principal components. Hence, with this procedure, the number of variables will be reduced. This selection method is used with the building of linear models with projections to latent structures by means of partial least squares (PLS) [98]. This can be extended to non-linear systems using neural networks or kernel-based methods such as support vector machines (SVM) [99].
- Genetic algorithms (GA) [100] which, as it was previously showed, are efficient methods that mimic natural evolution by modeling a dynamic population of solutions. Chromosomes evolve by means of crossover and mutation, enhancing the survival and reproduction of the fittest ones. Choosing the initial population is an important aspect and it can be based on Shannon's entropy combined with graph analysis [101].
- Simulated annealing (SA) [102], which is another stochastic method designed for function optimization utilized in QSAR, and represents an evolutionary approach where the minimized

function is the error of the model built using the subset of descriptors. This algorithm iteratively finds a new subset of descriptors obtained as a result of modifying the current-best one. Subsequently, this method decides whether to adopt the new solution as the current optimal solution or not, based on the evaluation of the new subset's prediction error. Thus, the new solution will be selected if it leads to lower error rates than the current one. In addition, the worst solution can replace the current-best one by using a given probability, based on the Boltzmann distribution. This replacement allows this method to escape from local minima of the error function.

- Sequential feature forward selection, which is a deterministic method implementing a search throughout the feature subsets [103]. First, a single feature that leads to the best prediction is selected. Then, each feature is individually added to the current subset and the errors of the resulting models are quantified. The feature that reduces the most the error is added to the subset. The selection will stop when a previously provided number of variables is reached or by incorporating an artificial random feature [104].
- Sequential backward feature elimination [103], which is another sequential method that uses as a starting point the entire set of MDs. During each step eliminates one feature with the highest error until a specific number of MDs is obtained. Due to the time consuming disadvantage, a new method was proposed such as the recursive feature elimination for support vector machines (SVM) [105]. In this method, the learning method is executed once using all the features except one, choosing the feature to be removed based on these results.

The hybrid methods use the fusion of the filtering and wrapper methods. The feature selection methods can also be included in the mapping methods that generate the final QSAR model. A hybrid method particularly suitable for genomic and biological datasets is the multiple inclusion criterion (MIC) [106]. This method is a simplification of stepwise feature selection which selects features that are helpful across multiple tasks and permits adding each feature to none, some or all the tasks. This method is particularly appropriate for selecting a small set of predictive features from a large set of potential ones. The next section of the review will shortly present the main QSAR methods used to obtain prediction models for specific properties/activities.

### *3.3. Methods to Build QSAR Models*

After selecting the proper descriptors, the QSAR model can be constructed with a series of methods as a linear or non-linear function. Depending on the nature of the activity variable, the mapping methods are different. If the activity is a continuous value, the regression method will be used. In contrast, if the activity is categorical (i.e. inactive or active compounds) the resulting model will be defined by a decision boundary, separating the classes in the descriptor space.

Linear methods have been used from the beginning of QSAR and are straightforwardly interpretable and adequately accurate for small datasets of similar compounds. The model function is a linear combination of the MDs as follows:

- Multiple linear regression (MLR) [107, 108], are models in which the activity to be predicted is a linear function including all the descriptors and the coefficients of the function are estimated based on the training set. These free parameters are chosen to minimize the squares of the errors between the predicted and the current activity. However, MLR analysis presents some limitations: large descriptors-to-compounds ratio or multicollinear descriptors in general make the problem ill-conditioned and make the results unstable.
- Partial least squares (PLS) [109] linear regression is a technique that surmounts the problems existing in MLR related to multicollinear or over-abundant descriptors. This method assumes that despite the large number of descriptors, the modeled process is directed by a relatively small number of latent independent variables. PLS linear regression decomposes the input matrix of descriptors into scores and loadings with the aim of indirectly obtaining knowledge on the latent variables. The scores are orthogonal and they also allow predicting well the activity while they are able to capture the descriptor information [110, 111].
- Linear discriminant analysis (LDA) [112] linearly transforms the original feature space into a new space that maximizes the interclass separability and minimizes the within-class variance. This technique works solving a generalized eigenvalue problem based on the between-class and within-

class covariance matrices. Hence, in order to avoid ill-conditioning of the eigenvalue problem [113, 114], there has to be a significantly smaller number of features than of observations.

However, interpreting non-linear models is harder and there may be overfitting even though they become more accurate, in particular for large and diverse datasets. The main types are the following:

- Bayes classifier, which is derived from the Bayes rule concerning the posterior probability of a class in relation to its overall probability, the probability of the observations and the probability of a class in relation to the observed variables [115, 116]. In this method, the predicted class corresponds to that one which minimizes the posterior probability. However, in real-world problems, these probabilities remain unknown, so they must be estimated. In other to overcome this, the probabilities of the classes in relation to the different descriptors are assumed to be independent. This is the basis of the naïve Bayes classifier (NBC) [117]. The Bayesian approach can be used to learn an optimal non-linear classifier and the most relevant subset of predictor variables (or features) regarding the classification task at the same time [118]. The approach presented by these authors utilizes heavy-tailed priors in order to promote sparsity in the usage of basis functions as well as features; these priors regulate the likelihood function, enhancing accurate classification of the training data. As a result, an expectation-maximization (EM) algorithm is obtained. This algorithm is capable of efficiently computing a maximum a posteriori (MAP) point estimate of the various parameters. This algorithm extends already existing sparse Bayesian classifiers, which can be seen as the Bayesian equivalent to support vector machines. Experiments involving kernel classifiers applied to several artificial and benchmark data sets showed parsimonious feature selection and excellent classification accuracy.
- k-Nearest neighbor (k-NN) [119, 120] is a type of instance-based learning and can be conserided one of the simplest machine learning algorithms. It requires very little training, converging to the lowest prediction error when more training data is used. Each instance (in this case a compound in the descriptor space) will be classified based on the *k* closest training examples (in this case compounds) from the training set. Thus, the output of the algorithm, in this context, will be the activity class that is most highly represented among the *k* nearest neighbors [121, 122].
- Artificial neural networks (ANN) [123, 124] which are prediction models inspired from biology, based on the architecture of a network of neurons. The most frequent ANNs are feed-forward networks, such as those based on perceptron or radial-basis function. In these networks, information flows in one single direction: from the input neurons to the output passing through a set of hidden layers [125, 127]. Two examples of ANNs are the followings:
- The multi-layer perceptron (MLP) network is composed of interconnected perceptrons distributed in layers [128]. Each perceptron is able to linearly combine its input values, returning either a binary or continuous value as output, by means of a transfer function. It must be highlighted that each input of the perceptron has an associated weight that represents its relevance.
- The radial-basis function (RBF) neural networks [129] are composed of three layers: an input layer, a hidden layer and an output layer. Unlike MLPs, the output obtained by the neurons belonging to the hidden layer is not calculated as the product of the weights and the input values. Each neuron of the hidden layer is characterized by its center and its output is computed as the distance between the input compound in the descriptor space and the neuron's center.
- Decision trees (DT) [130, 131] which are derived from logic-based and expert systems, where each classification tree corresponds to a set of predictive rules in Boolean logic [132].
- Support vector machines (SVMs) [133] which are derived from the structural risk minimization principle, being the linear support vector classifier the most basic element. SVMs builds a hyperplane in a high- or infinite-dimensional space, that can be utilized for classification, regression, or other tasks [134]. The advantage of this model compared with ANNs is that the objective function is unimodal. Hence, this function can be efficiently optimized for the global optimum.
- Evolutionary algorithms (EA) [35-37, 135] which are inspired by mechanisms of biological evolution such as selection, reproduction, recombination and mutation. Possible solutions to an optimization problem correspond to the individuals of a population, being the fitness function what characterizes the environment within which these solutions evolves [136]. Examples of these methods are the following: genetic algorithm, which can be considered the most popular type of EA, genetic programming (GP), evolutionary programming, evolution strategy and neuroevolution [137, 138]. The genetic function approximation (GFA) algorithm [139, 141] is an earlier strictly QSAR application of GP and it was conceived to be applied to the function approximation problem. When it receives a large number of potential factors influencing a response, including

several powers and other functions of the raw inputs, it should obtain the subset of terms that correlates best with the response.

- Ensemble methods which represent an approach to QSAR analysis focused on building a single predictive model; consists of bagging [142], random subspace [143] methods or boosting [144].

The current review is focused on the description and applications of the evolutionary computation methods.

### *3.4. Current Evolutionary Feature Selection Methods and Aplications in QSAR*

Evolutionary techniques have usually provided excellent results dealing with complex problems. One of the areas where these results are especially good is drug design. Different approaches are tested, but the most usual are related to two key tasks on the drug design process: combinatorial library design and deriving QSAR models. Combinatorial libraries are the result of combinatorial synthesis, with the goal of synthesizing large numbers of compounds in parallel. The main problem is the huge amount of compounds that could be synthesized, so it is not possible to check them individually. Thus, methods are required for selecting appropriate subsets of compounds. The large amount of possibilities involved requires a method to optimize the tests (preprocessing and selecting the compounds) and makes this problem ideal for applying approaches based on evolutionary algorithms.

It is important to keep in mind that QSAR methods provide intelligible models by means of a relation between the structure and the activity of the molecule. One of the first approaches demonstrated that thermodynamic or electronic information could be used to explain the biological activity by means of a simple regression equation [2, 145]. Since this first approach has been stated by Hansch, several significant key-points happened: first, the availability of a large number of easily computable molecular compounds and, second, a great number of sophisticated techniques (i.e. evolutionary techniques) are available to improve the first linear regression approach. By using a large number of molecular descriptors, overfitting of the experimental data can result in QSAR models with poor predictive ability. Thus, the evolutionary algorithms have successfully been applied to select descriptors that give rise to good QSAR models [146].

Feature selection can be based on two main approaches. The first option or forward-step methods starts with a unique descriptor and the rest of them are included (step by step) according to any kind of fitness evaluation. The process will stop when the addition of new descriptors does not improve the model. The second option, called backward-elimination, used an opposite approach. Here, all the descriptors are included in the first model and one of them is removed in each step of the method [147].

Evolutionary techniques can be used to address this kind of optimization. First of all, as previously shown, a correct representation must be established. Using genetic algorithms, one of the most simple (and used) representations consists of a binary string where each bit represents a different feature (descriptor, compound). If the bit has a value of 1, the feature will be *active* (it is present), on the other hand if the bit has a value of 0, the feature will be *inactive* (it will not be used in the model). Using this representation and performing a least-squares regression to regenerate the coefficients, Rogers and Hopfinger [139] developed the genetic function approximation (GFA) method. It was successfully applied over the Selwood dataset, which has become a standard test for QSAR applications and contains 31 compounds and 53 descriptors. Another approaches use genetic algorithms in combination with other techniques, such as artificial neural networks [148] to improve the search and the model building [136, 149].

The work published by Jalali-Heravi et al. [150] presents a new nonlinear feature selection technique called genetic algorithm-kernel partial least square (GA-KPLS). This technique takes advantage of the power of an optimization method (a genetic algorithm, GA) and the robustness of a nonlinear statistical method for variable selection (KPLS). Then, with the aim of obtaining a nonlinear QSAR model for substituted aromatic sulfonamide (such as carbonic anhydrase II inhibitors) activity prediction, the feature selection technique is combined with ANNs. It can be concluded from the results that this approach works well for performing variable selection in nonlinear systems.

Hasegawa et al. [151] developed a GA-based PLS method, called GAPLS, to apply variable selection in QSAR studies. The modified genetic algorithm (GA) proposed by Leardi is capable of obtaining PLS models with high internal predictivity employing a reduced number of variables. The method was tested on data involving inhibitory activity of calcium channel antagonists with the objective of observing its performance for variable selection. This way, structural requirements for the inhibitory activity could be efficiently estimated by selecting variables that contribute significantly to it. Results show that the structural requirements nicely correspond to those derived by the MLR analysis, and the utility of GAPLS was demonstrated.

A novel genetic algorithm guided selection technique (GAS) is described by Cho et al. [152]. In order to build a QSAR/QSPR model, a simple encoding scheme which can represent compound subsets and descriptors as variables is utilized. As part of this technique, a genetic algorithm is used to concurrently optimize these variables. Thus, this technique builds several models each applying to a subset of the compounds. The method is able to correctly identify artificial data points belonging to various subsets. The analysis of the XLOGP data set shows that the subset selection method can be useful in improving a QSAR/QSPR model when the variable selection method fails.

Hasegawa et al. [153] designed a new variable selection method for comparative molecular field analysis (CoMFA) modeling (GARGS, genetic algorithm-based region selection). In order to assess whether the method was capable of identifying known molecular interactions in 3D space, GARGS was applied to a data set containing acetylcholinesterase (AChE) inhibitors. GARGS successfully gave the best model whose coefficient contour maps were consistent with the steric and electrostatic properties of the active site in AChE. GARGS may facilitate the prediction of binding affinities if one already has a series of compounds and assumes the same binding mode. Moreover, multivariate 3D-QSAR analysis can also be done using the GRID software. The amount of interaction field variables from GRID can be reduced and simplified by GARGS, and then the predictive 3D-QSAR model simulating the physical or chemical environment in receptor may be obtained.

Pavan et al. [154] developed GA-VSS: genetic algorithm-variable subset selection. This method was designed to look for the best ranking models within an extensive set of variables. Order-ranking strategies appear to be a valuable instrument, both to explore the data and to obtain order-ranking models, a promising option to usual QSAR methods.

A genetic programming approach is also available. In this case, a genetic programming individual will be a tree structure where the internal nodes will usually represent mathematical operators and the terminal nodes will represent variable and constant values. In QSAR problems these internal nodes will represent the sum, quadratic and cubic power operators whereas the terminal nodes will represent the molecular descriptors available for the dataset. The fitness function in both cases will be the same (or at least very similar). It will use the model encoded in the internal representation (bit string or tree structure) to get a perform value using linear regression or any other method to allow the evolutionary method assess the quality of the solution.

Venkatraman et al. [155] proposed a novel approach for analyzing QSAR data based on genetic evolutionary algorithms coupled with information theoretic approaches such as mutual information. These approaches have been used to find near-optimal solutions to such multicriteria optimization problems. The applicability of the method is demonstrated using a highly skewed dataset. In particular, it has been shown that the method can be used to realize a logic circuit capable of predicting the structure-function relationship.

Shen et al. [156] perform variable selection in QSAR studies with MLR and PLS modeling utilizing the evolution algorithm (EA). In this study, the Cp statistic has been adapted and used as the objective function in the EA search for diverse combinations of molecular descriptors. The process described was employed for predicting aromatic amine carcinogenicity. The QSAR analysis shows satisfactory prediction performance for the proposed methodology.

Genetic algorithms and genetic programming are the most used evolutionary techniques, but there are other useful natural computing algorithms. For example, ant colony optimization tries to imitate how ant colonies search food to look for a valid solution for the problem. In the QSAR field there are successful solutions based on this approach [157-159].

Shamsipur et al. [160] presented an innovative approach for the use of external memory in ant colony optimization strategy. The approach was applied to the descriptor selection problem in quantitative structure-activity/property relationship studies. More specifically, the QSAR/QSPR studies involved rate constants of o-methylation of 36 phenol derivatives and activities of 31 antifilarial antimycin compounds. Results showed that the speed and the quality of the solution are enhanced in contrast to traditional ant colony system algorithms. The authors also established that numerous models with high statistical quality are constructed from a single run of the method, which can infer the structure-activity/property relationship more easily and more accurately.

Patil et al. [161] present a filter/wrapper search method based on ant colony optimization (ACO)/random forest. This method goes across the search space and chooses a feature subset which can classify with high accuracy. The authors applied the proposed algorithm to four widely studied CoEPrA (Comparative Evaluation of Prediction Algorithms, http://coepra.org) datasets in order to observe its performance. They found that their method was capable of effectively obtaining small feature subsets with excellent classification accuracy. Therefore, they concluded that their method can be used to solve feature subset selection problems with a high level of confidence.

There exist other approaches which are based on the social behavior of bird flocking or fish schooling named particle swarm optimization (PSO). Applications of this kind of algorithms within QSAR environments can be found in refs. [157, 162, 163].

A modification of particle swarm optimization (PSO) was developed by Shen et al. [164]. The PSO algorithm was modified for its application to the discrete combinatorial optimization problem and to diminish the likelihood of falling into local optima. The proposed algorithm was designed to select variables in multiple linear regression (MLR) and partial least-squares (PLS) modeling and to predict antagonism of angiotensin II antagonists. Experimental results proved that this algorithm quickly converges towards the optimal solution and that it is a valuable tool for variable selection.

Agrafiotis et al. [163] proposed a novel method for feature selection in QSAR and QSPR studies. The method is a modification of particle swarms that uses a binary encoding. It was applied to build parsimonious QSAR models based on feed-forward neural networks and was tested on three classical data sets from the QSAR literature. Results show that the method works as well as simulated annealing and that it is capable of obtaining better and more varied solutions given the same amount of simulation time.

The greatest advantage of all this type of techniques is related to the intrinsic operation of evolutionary techniques. The same technique offers a pattern to encode the problem and a set of operators to guide the search. The most important aspect when using this kind of techniques is related to the idea that the user must know only a way to check if a solution is valid or not (and nothing about how the solution should be constructed), but there are also limitations. The most important one is derived from their own operation process. Evolutionary techniques offer (usually) excellent solutions to a given problem, but they neither explain why that solution is selected nor how it was made up.

The application of evolutionary computation techniques to solve problems related to QSAR started in 1980's. The first application in this field used two methods based on genetic programming to find useful QSAR models [165]. One of them penalizes model complexity (that is, the depth of the trees that represents the solutions is limited) and it is designed to generate only one linear model. The second one employs a multiobjective optimization approach (in this case the fitness function tries to optimize more than one parameter at the same time: model fitting, total number of terms, etc.).

Compared with the global optimization techniques [137], evolutionary algorithms present several advantages [135]:

- They are conceptually simple.
- Unlike other numerical techniques, which may be only appropriate for continuous values or other constrained sets, the performance of these algorithms does not depend on the representation.
- Incorporating prior knowledge is easy, yielding a more efficient exploration of the state space of possible solutions.
- Evolutionary algorithms can also be combined with more conventional optimization techniques (i.e. the combination of gradient minimization after primary search with an evolutionary algorithm

to perform the fine tuning of weights of an evolutionary neural network) or it may involve applying other algorithms in parallel, such as hybridizing with simulated annealing.

- Solutions can be simultaneously evaluated and only selection (which requires at least pair-wise competition) needs to be sequentially processed. Many global optimization algorithms, such as simulated annealing, do not allow implicit parallelism.
- Traditional optimization methods do not present robustness against dynamic changes, usually requiring a complete restart in order to provide a solution (e.g. dynamic programming). In contrast, evolutionary algorithms can be utilized to adjust solutions to changing circumstances.
- One of the greatest advantages of evolutionary algorithms is related to their capacity to deal with problems involving non-human experts. Even though human expertise should be used if available, it often proves unsuitable for automating problem solving routines.

One of the first approaches using genetic algorithms can be found in Ref. [166] where the authors use the evolutionary approach to guide the analysis of high throughput screening data to conclude that the results indicate that genetic algorithms are a very effective variable selection approach for binary QSAR analysis. Therefore, genetic algorithms have been used for optimizing robust mathematical models such as Bayesian-regularized artificial neural networks and support vector machines applied to diverse drug design problems [167].

A more specific application (due the datasets employed) consists of using the Genetic Algorithms and a similar technique called simulated annealing to obtain anti-tubercular activity prediction [168]. They make the previous variable selection that later will be used by different kind of neural networks to perform the QSAR modeling of quinoxaline compounds. The same schema of previous variable selection by means of Genetic Algorithms and later modeling with Kohonen neural networks is used in Ref. [169]. In the same work, Partial Least Square and Support Vector Machine are also used to construct different models for the Human Intestinal Absorption process.

Taha et al. [170] employed genetic algorithms (GAs) and multiple linear regression to construct different QSAR models. Thus, the GA was used to improve pharmacophor moleding. Pan et al. [171] present a method for performing quantitative structure-based design using genetic algorithm optimization and backward elimination multidimensional regression to achieve/perform data reduction, QSAR model construction and identification of possible pharmacophore sites. Zaheer-ul-Haq et al. [172] use a genetic algorithm to perform CoMFA, which is used with CoMSIA to construct 3D-QSAR. The authors found this approach to be the best. Moreover, results prove that using a genetic algorithm for ligand-based and receptor-based modeling is a potent approach to build 3D-QSAR models. Thus, these data can be used to direct the optimization process for inhibition enhancement, which is important for drug discovery and Alzheimer's disease development.

A genetic algorithm was proposed by Asadollahi et al. [173] to improve the performance of partial least squares (PLS) modeling. In this work, several methods are compared, demonstrating the superiority of the GA-PLS approach in terms of prediction capacity. The in silico screening technique was applied to the proposed QSAR model and the structure and potency of new compounds were predicted. The generated models proved to be useful to estimate $pIC_{50}$ of CXCR2 receptors for which no experimental data is available. The electron conformational-genetic algorithm (EC-GA) was used by Yanmaz et al. [174] to perform 4D-QSAR studies on a series of 87 penicillin analogues. In this algorithm, a matrix composed of electron structural parameters and interatomic distances was built to describe each conformation of the molecular system. A GA was used to identify the most relevant descriptors and to predict the theoretical activity of the training (74 compounds) and test (13 compounds, commercial penicillins) sets.

Zhou et al. [175] proposed a new method that combines particle swarm optimization algorithm (PSO) and genetic algorithm (GA) to optimize the kernel parameters of support vector machine (SVM) and determine the optimized features subset in parallel. These authors applied their method to four peptide datasets for quantitative structure-activity relationship (QSAR) research. The structural and physicochemical characteristics of peptides from amino acid sequences were employed to represent peptides for QSAR. The results obtained in this work suggested that the method might be a valuable instrument in peptide QSAR and protein prediction research.

Reddy et al. [176] investigated QSAR prediction models developed on a dataset of 170 HIV protease enzyme inhibitors. A hybrid-GA optimization technique is used for descriptor space reduction. QSAR prediction models were built using the previously selected descriptors. The approaches presented in this paper yield the QSAR with good prediction performance. The models obtained can be helpful to predict biological activity of new untested HIV protease inhibitors and virtual screening in order to identify new lead compounds.

A novel inductive data mining method design to automatically generate decision trees from data (genetic programming tree, GPTree) is presented by Ma et al. [177]. In order to model the original continuous endpoint by adaptively finding appropriate ranges to describe the endpoints during the tree induction process, avoiding the need for discretization prior to tree induction and enabling the ordinal nature of the endpoint to be considered in the generated models, the YAdapt method is created as an extension of the GPTree approach. The methods developed in this work were applied to QSAR modeling to predict chemical eco-toxicity and to analyze a historical database for a wastewater treatment plant.

Archetti et al. [178] used GP for a QSAR investigation of docking energy. Thus, the authors present a GP-based framework on which specific strategies can be built, proving to be a helpful tool for the problem. Experimental results obtained by GP are compared to those obtained by other “non-evolutionary” machine learning methods (including support vector machines, artificial neural networks, linear and least square regression), confirming that the technique proposed shows potential in terms of accuracy of the proposed solutions, of generalization capacity and of the correlation between predicted and data and correct ones.

Finally, Zhou et al. [179] developed GA-GP, which is an algorithm for dealing with all kinds of nonlinear function relationships using GP to create new individuals and GA to optimize them at the same time making, thus, full advantage of both techniques. The authors applied this algorithm in QSAR in the medicinal field. The results obtained show that GA-GP is better than ANN. Moreover, it was found that the expressions obtained by the method clearly express and explain QSAR of the medicine. Thus, GA-GP can have very wide applications in the medicinal field of the molecule design.

## 4. FUTURE TRENDS: JOINT OR MULTI-TASK FEATURE SELECTION METHODS

Improvement of the profile of a drug candidate involves finding the middle ground between several, usually competing objectives. As a matter of fact, the perfect drug should to have the maximum therapeutic efficacy, the maximum bioavailability, and the minimum toxicity, which shows the multi-objective nature of the drug discovery and development process. The fine-tuning of the multiple criteria in hit-to-lead identification and lead optimization is considered to be an important achievement in the rational drug discovery process. The objective of this change of paradigm is the rapid identification and removal of candidate molecules that have a low probability of surviving later stages of discovery and development. This novel approach will decrease clinical attrition, and as a result, the global cost of the drug development process [80, 81, 180-188].

Unfavorable absorption, distribution, metabolism and elimination (ADME) properties have been recognized as one of the main causes of failure for candidate molecules in drug development. Computational chemistry and informatics have had provable influence on synthesis and screening in drug discovery. Until recently, drug discovery was mainly linear: ADME/Tox and drug-likeliness were assessed only in later stages. Computational chemistry, jointly with high-throughput methods, has helped in making the process multi-dimensional, in combining the process for lead generation and optimization taking into account ADME/Tox and drug-likeliness. Nowadays, the majority of drugs are discovered through programs that start by identifying a biomolecular target of potential therapeutic value during biological studies.

Lately, decisions taken during the identification of small-molecule modulators of protein function and the transformation these into high-content lead series, as they have far-reaching consequences later due to increasing downstream costs due to high clinical failures, have gained more attention [180]. Thus, a more flexible model is needed: the sequential process of identification, evaluation and refinement activities must change towards a more integrated parallel one.

Although chemoinformatics has accomplished numerous achievements in diversity analysis, such as structure-activity relationship (SAR) and virtual screening during the last ten years, the authors of Ref. [181] also consider parallel optimization of potency, selectivity and ADMET properties using computational models is challenging. However, in silico ADMET models are not largely accepted as they are not robust enough and do not cover all the medicinal chemists' concerns, but more than acceptable performance has been achieved in drug-likeliness, solubility and lipophilicity. According to this paper, potential users of ADMET models frequently cite them as non-significant because they are based on small sets of chemical compounds, as, paradoxically, the numerous pharmaceutical companies in which these same potential users work do not share their in-house experimental data. Another problem cited in this paper is that when the compounds under investigation are more diverse chemically, SAR models are less likely to exist and to be uncovered. In addition, the increase of the chemical space boundaries and the diversity of the investigated compounds will cause the information content of a SAR model (if it exists) to increase. Thus, the authors propose solving this problem by combining the mechanism-based approaches with other data mining approaches. Technologies such as HTS or combinatorial chemistry (CC) produce great amounts of data. Therefore, it could be said that current drug design is data-driven and the most important objective is to discover knowledge from raw data. This process takes the raw, experimental results from data mining and then transforms them into useful and understandable information, which is not typically retrieved by standard techniques.

There is not any approach capable of predicting all the desired ADME properties. Nevertheless, identifying the most suitable one for modeling a specific property is challenging. Thus, physiologically-based pharmacokinetic (PBPK) simulations that predict pharmacokinetic properties have been developed [182]. Existing simulations are often based on compound data obtained in vivo or in vitro. However, according to this paper, the results of predictive ADME models may be utilized, in the near future, as inputs to PBPK simulations, providing precise in silico prediction of PK properties from compound structure alone. Thus, compounds with unsuitable PK will be removed at the beginning of the drug discovery process and those designed with optimal properties will progress to development.

Conventional chemistry will be integrated with data mining [183]. The integrated and iterative use of these techniques, together with the significant data from HTS (including ADME/Tox and drug-like property data), to make up models that can be used to prioritize further library screening and synthesis through virtual screening activities, already has noteworthy impact on lead discovery and exploration.

In silico screening methods are incredibly beneficial to drug discovery, being fast as well as practical for a seamless integration into daily research routines. The researchers concluded that the hit rates of virtual screening are greater than those from random screening [184]. Thus, the development of fast feedback-driven methods, using either purely in silico vHTS or integrated in silico/in vitro engines, are expected to further enhance the impact of virtual screening in the drug discovery process. That virtual screening can be considered as a valuable instrument to enrich libraries and compound collections [80, 81]. Preprocessing properly the compound database is of major importance and, for this purpose, stepwise procedures (filters, pharmacophore searches, docking and scoring, visual inspection) are the most effective. Fragment-based approaches show a great potential in lead structure search and optimization. HTS strategies are directed towards increasing the number of compounds per plate in order to augment the number of screens per day, but this exponential race is heading nowhere. For this reason, more effort has been put into finding a more rational design for chemical libraries in relation to diversity, drug-like behavior and reagent selection. Thus, due to the size of the problem, efficient drug discovery nowadays relies primarily on computational methods [185].

Physiologically-based pharmacokinetic modeling is a step further to total modeling, describing the whole body by a number of anatomical compartments. If we want to understand the real limiting factors in drug action in the whole body, a stepwise multiple QSAR technique should be considered [186]. That is, each step in drug action should be analyzed using a quantitative method, thus permitting one to fully conceive an overall QSAR. There is little doubt that the introduction of pharmacokinetic parameters in QSAR is a step towards a more rational drug design, that is, in the development of drugs and chemicals with an optimal wanted effect and a certain acceptable toxicity-profile, both predicted and understood. However, pharmacokinetic parameters are only useful when used properly. The weakest point in de novo design is the reduced transferability of the scoring functions [187]. Another general issue is the artificial accessibility of the constructs. However, it is probable that the most developed techniques (part of hybrid methods) will continue to advance and become more broadly used in drug discovery. In 2000, the global estimated R&D cost per new drug was US$ 802 million [188]. Results were validated in many different manners by analyzing independently derived published data on the pharmaceutical industry. Including an estimate of the cost per approved new drug for R&D carried out after approval increases total R&D cost to approximately US$ 900 million. All these arguments put forward the need for approaches capable of integrating drug- or lead-likeness, toxicity and bioavailability criteria at early stages of the drug discovery process as an emergent issue [180-183]. That is, methods capable of taking into account additional criteria for the early simultaneous treatment of the most relevant properties, potency, safety, and bioavailability, determining the pharmaceutical profile of a drug candidate [165, 189-196].

An interesting work presents two new methods based on GP [165] whose aim is to identify useful QSAR models that represent a compromise between model accuracy and complexity,. The first one, genetic QSAR (GPQSAR), penalizes model complexity. This method obtains only one linear model that represents an appropriate balance between the variance and the number of descriptors selected for the model. The second one, multi-objective genetic QSAR (MoQSAR), is based on multi-objective GP and shows a new point of view for QSAR. Another goal, called chemical desirability was added. This goal rewards those models that consist of descriptors which are easily interpretable by chemists. Thus, MoQSAR is capable of identifying models that are at least as good as models obtained by standard statistical approaches and, in addition, these models yield more chemical interpretability.

De novo drug design implies searching an enormous space of possible, drug-like molecules to choose those that are more likely to become drugs by means of computational technology. Multi-objective evolutionary graph algorithm (MEGA) [193] is a novel multi-objective optimization de novo design algorithmic framework that can be used to design structurally diverse molecules fulfilling one or more objectives. This algorithm integrates evolutionary techniques with graph-theory to directly manipulate graphs and performs an efficient global search for potential solutions. MEGA was applied for designing molecules that selectively bind to a known pharmaceutical target using the ChillScore interaction score family. This algorithm is capable of obtaining structurally diverse candidate molecules that represent an extensive variety of compromises of the supplied constraints and thus can be used to support expert chemists "giving ideas".

The study of a new data mining technique for multi-objective optimization of chemical properties was presented and combines hierarchical classification and visualization of multidimensional data [194]. Recursive partitioning is modified to use averaged information gains for multiple objective variables as a quality-of-split criterion. As a result of this, a hierarchical classification tree model is obtained.

The desirability theory, which is a recognized multi-criteria decision-making approach, was used to interpret prediction models with the aim of extracting useful information [189]. Applying the belief theory enabled quantifying the reliability of the predicted desirability of a compound according to two inverse and independent but complementary prediction approaches. This information has demonstrated of great utility as a ranking criterion in a ligand-based virtual screening study.

Some authors proposed a multi-objective optimization algorithm developed for automated integration of structure- and ligand-based molecular design [190]. Guided by a genetic algorithm, this algorithm allowed detecting a number of trade-off QSAR models accounting for two independent goals at the same time: the first one leaded towards the best regressions among docking scores and biological affinities and the second one minimizes the atom displacements from a properly established crystal-based binding topology. Taking into account the concept of dominance, 3D QSAR alike models profiled the Pareto frontier and were designated as non-dominated solutions of the search space. K-means clustering was then applied to choose a representative subset of the available trade-off models.

The multi-objective optimization (MOOP) method based on Derringer's desirability function permits carrying out global QSAR studies taking into account at the same time the pharmacological, pharmacokinetic and toxicological profile of a set of molecule candidates [191]. The value of the technique is shown with its application to simultaneous optimization of the analgesic, anti-inflammatory and ulcerogenic properties of a library of fifteen specific compounds. The levels of the predictor variables producing in parallel the best possible compromise between these properties are found and used to design a set of novel optimized drug candidates. MOOP methods initiate a new philosophy to attain optimality on the basis of compromises among the diverse objectives. These methods are aimed at achieving the global optimal solution by optimizing various dependent properties concurrently. Pareto MOOP can handle each property independently. However, optimization tasks need reliable fitness functions that can be derived from QSAR models. In this sense, simple descriptions (as few molecular descriptors as possible) of each property are desired. Thus, a common subset of features (as few as possible) able to accurately describe the k properties determining the final pharmaceutical profile of a drug candidate would be an ideal result. The application of joint or multi-task feature selection methods based on evolutionary computation methods like Pareto MOOP could be a substantial advance in this sense [106, 118, 197-204].

The ranking method permits drug candidates with unknown pharmaceutical properties from combinatorial libraries according to the degree of resemblance with the formerly obtained optimal candidate [192]. Thus, filtering the most promising drug candidates of a library (the best-ranked candidates) is possible, that is, those with the best pharmaceutical profile. The authors also propose both a method to validate the ranking process a quantitative and a measure of the quality of a ranking, the ranking quality index. Results show that the combined use of the desirability-based methods of MOOP and the proposed ranking appears to be a precious instrument for rational drug discovery and development.

The MOOP-DESIRE methodology and a variation of this was used to study the arylpiperazine derivates that could interact with 5-HT$_{1A}$ and 5-HT$_{2A}$, serotonin receptor subtypes with the aim of designing more selective molecules for the 5-HT$_{1A}$ receptor [195]. It was shown that the model results are in conformity with the existing pharmacophore descriptions, assuring an appropriate structural correlation and demonstrating that the methodology is helpful for solving the selective drug design problem. In addition, a combined strategy-based on MOOP and ranking for the prioritization of HIV-1 NNRTIs hits with appropriate trade-offs between inhibitory efficacy over the HIV-1 RT and toxic effects over MT4 blood cells was proposed [196]. A study comparing the sequential, parallel and multi-objective virtual screening showed that the multi-objective approach works better than the other approaches.

Modeling data from candidate cancer biomarkers taking into account newly developed ideas by the machine learning community is considered [200]. More specifically, combined objectives of feature selection and classification are considered. Estimation procedures are developed for analyzing immunohistochemical profiles by means of the least absolute selection and shrinkage operator. These lead to innovative and flexible models and algorithms for compositional data analysis.

A new method for simultaneous feature selection and classifier learning by means of a sparse Bayesian approach was proposed [201]. These two tasks are performed by optimizing a global loss function that includes a term related to the empirical loss and another function that represents a feature selection and regularization constraint on the parameters. To minimize this function, the Boosted Lasso algorithm is used, following the regularization path of the empirical risk associated with the loss function. An algorithm for the relevance vector machine (a well-known non-parametrical classification method) was also developed.

The QSAR models for molecules have been extended to non-molecular systems and descriptors for microarrays, imaging, spectra, and new species discovery. A multi-task feature selection filter that obtains strength from auxiliary microarray classification data sets utilizes Kruskal-Wallis test on auxiliary data sets and ranks genes based on their aggregated p-values [199]. Expressions of the top-ranked genes are used as features to build a classifier on the target data set.

The computerized system was preclinically evaluated in terms of robustness for breast lesion characterization on two breast magnetic resonance imaging (MRI) databases that were obtained by means of scanners from two different manufacturers [197]. Once a breast lesion has been identified by the radiologist, the system automatically performs segments the lesion and performs feature extraction, obtaining as a result an estimated probability of malignancy. A Bayesian neural network with automatic relevance determination for combined feature selection and classification was used.

A hot topic is the potential use of Bayesian neural network (BNN) with automatic relevance determination (ARD) priors for combined feature selection and classification in computer-aided diagnosis (CAD) of medical imaging [202]. Simulations show that the ARD-BNN approach has the ability to select the optimal subset of features on the designed non-linear feature spaces on which the linear approach fails. Moreover, ARD-BNN has the ability to recognize features that have high ideal observer performance.

A methodology for pre-processing spectra and extracting putatively meaningful features before applying feature selection and classification algorithms was developed and tested [203]. This methodology involves a HMM-based latent spectrum extraction algorithm for fusing the information from several replicate spectra obtained from a single tissue sample, a simple algorithm for baseline correction based on a segmented convex hull, a peak identification and quantification algorithm, and a peak registration algorithm to align peaks from multiple tissue samples into common peak registers.

Another method proposes an automated feature selection and classification system which includes logistic regression with controlled false discovery rate with the aim of addressing the taxonomic research need impediment in new species discovery [198]. Unlike conventional taxonomic practice, the system proposed by these authors is capable of automatically selecting body shape features from specimen samples with landmarks that unite populations within species, as well as distinguishing among species. It also associates probabilities to classification accuracy using the selected features in new species identification.

A recent work [204] has presented a genetic process for the selection of an appropriate set of features in a fuzzy rule-based classification system (FRBCS) and to automatically learn the whole database definition using a non-linear scaling function to adapt the fuzzy partition contexts and to determine an appropriate granularity for each of them. An ad-hoc data covering learning technique is used to obtain the rule base. The method utilizes a multi-objective genetic algorithm in order to obtain a good trade-off between accuracy and interpretability.

## 5. CONCLUSION

The virtual screening for new molecules with a specific pharmacokinetic activity is the first step in the drug design industry. This computational predictions help to avoid high cost and time-consuming experiments. The general process consists of building a mathematical model that correlate the molecular structure with a specific property/activity (QSAR models) by using molecular descriptors. Among the different and complex methods to select the proper variables and to build the best QSAR model, the evolutionary computation plays an essential role. The genetic algorithms provide excellent results by using the natural concept of the evolution. Evolutionary algorithm optimizers are global optimization methods and scale well to higher dimensional problems. They are robust with respect to noisy evaluation functions, and handling evaluation functions which do not yield a sensible result in a given period of time is straightforward. Its simplicity, flexibility, capacity to be mixed with other methods and accessibility to non-human experts make the evolutionary methods one of the best solutions in QSAR.

## CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

## ACKNOWLEDGEMENTS

Vanessa Aguiar-Pulido and Munteanu C.R. thank sponsorship from the “Plan I2C” and Isidro Parga Pondal research programs respectively, both funded by Xunta de Galicia (Spain) and the European Social Fund (ESF). This work is supported by the following projects: grant Ref. PIO52048 and RD07/0067/0005 funded by the Carlos III Health Institute and “PHR2.0: Registro Personal de Salud en Web 2.0” (Ref. TSI-020110-2009-53) funded by the Spanish Ministry of Industry, Tourism and Trade, and “Development of new image analysis techniques in 2D Gel for biomedical research” (Ref. 10SIN105004PR ) funded by Xunta de Galicia.

## REFERENCES

[1] Goodnow, R.A.; Guba, W.; Haap, W. Library design practices for success in lead generation with small molecule libraries. Comb. Chem. High Throughput Screen, 2003, 6, 649-660.

[2] Hansch, C.; Fujita, T. ρ-σ-π analysis. A method for the correlation of biological activity and chemical structure. J. Am. Chem. Soc., 1964, 86, 1616-1626.

[3] Stuper, A.J.; Jurs, P.C. ADAPT: A computer system for automated data analysis using pattern recognition techniques. J. Chem. Inf. Comput. Sci., 1976, 16, 99-105.

[4] Hodgson, J. ADMET - turning chemicals into drugs. Nat. Biotechnol., 2001, 19, 722-726.

[5] Ivanciuc, O.; Ivanciuc, T.; Cabrol-Bass, D.; Balaban, A.T. Evaluation in quantitative structure-property relationship models of structural descriptors derived from information-theory operators. J. Chem. Inf. Comput. Sci., 2000, 40, 631-643.

[6] Balaban, A.T.; Beteringhe, A.; Constantinescu, T.; Filip, P.A.; Ivanciuc, O. Four new topological indices based on the molecular path code. J. Chem. Inf. Model., 2007, 47, 716-731.

[7] Ivanciuc, O. New neural networks for structure-property models. In: QSPR/QSAR Studies by Molecular Descriptors, Diudea, M.V., Ed. Nova Science Publishers: Huntington, NY, 2001; pp 213-231.

[8] Ivanciuc, O. Quantitative structure-activity relationships (QSAR) with the MolNet molecular graph machine. Curr. Bioinform., 2011, 6, 261-268.

[9] Randi, M. On characterization of molecular attributes. Acta Chim. Slov., 1998, 45, 239-252.

[10] Randi, M. Orthogonal molecular descriptors. New J. Chem., 1991, 15, 517-525.

[11] Randi, M. Molecular profiles. Novel geometry-dependent molecular descriptors. New J. Chem., 1995, 19, 781-791.

[12] Ivanciuc, O.; Ivanciuc, T.; Klein, D.J. Quantitative structure-property relationships generated with optimizable even/odd Wiener polynomial descriptors. SAR QSAR Environ. Res., 2001, 12, 1-16.

[13] Ivanciuc, O.; Oezguen, N.; Mathura, V.S.; Schein, C.H.; Xu, Y.; Braun, W. Using property based sequence motifs and 3D modeling to determine structure and functional regions of proteins. Curr. Med. Chem., 2004, 11, 583-593.

[14] Randi, M.; Vrako, M.; Nandy, A.; Basak, S.C. On 3-D graphical representation of DNA primary sequences and their numerical characterization. J. Chem. Inf. Comput. Sci., 2000, 40, 1235-1244.

[15] Randi, M.; Zupan, J. Highly compact 2D graphical representation of DNA sequences. SAR QSAR Environ. Res., 2004, 15, 191-205.

[16] Randi, M.; Butina, D.; Zupan, J. Novel 2-D graphical representation of proteins. Chem. Phys. Lett., 2006, 419, 528-532.

[17] Randi, M.; Vrako, M.; Zupan, J.; Novi, M. Compact 2-D graphical representation of DNA. Chem. Phys. Lett., 2003, 373, 558-562.

[18] Rodriguez-Soca, Y.; Munteanu, C.R.; Dorado, J.; Pazos, A.; Prado-Prado, F.J.; González-Díaz, H. Trypano-PPI: A web server for prediction of unique targets in trypanosome proteome by using electrostatic parameters of protein-protein interactions. J. Proteome Res., 2010, 9, 1182-1190.

[19] Rodriguez-Soca, Y.; Munteanu, C.R.; Dorado, J.; Rabuñal, J.; Pazos, A.; González-Díaz, H. Plasmod-PPI: A web-server predicting complex biopolymer targets in Plasmodium with entropy measures of protein-protein interactions. Polymer, 2010, 51, 264-273.

[20] Viña, D.; Uriarte, E.; Orallo, F.; González-Díaz, H. Alignment-free prediction of a drug-target complex network based on parameters of drug connectivity and protein sequence of receptors. Mol. Pharm., 2009, 6, 825-835.

[21] Fujita, A.; Sato, J.R.; Garay-Malpartida, H.M.; Sogayar, M.C.; Ferreira, C.E.; Miyano, S. Modeling nonlinear gene regulatory networks from time series gene expression data. J. Bioinform. Comput. Biol., 2008, 6, 961-979.

[22] Prado-Prado, F.J.; Garcia-Mera, X.; González-Díaz, H. Multi-target spectral moment QSAR versus ANN for antiparasitic drugs against different parasite species. Bioorg. Med. Chem., 2010, 18, 2225-2231.

[23] Concu, R.; Dea-Ayuela, M.A.; Perez-Montoto, L.G.; Prado-Prado, F.J.; Uriarte, E.; Bolas-Fernandez, F.; Podda, G.; Pazos, A.; Munteanu, C.R.; Ubeira, F.M.; González-Díaz, H. 3D entropy and moments prediction of enzyme classes and experimental-theoretic study of peptide fingerprints in Leishmania parasites. Biochim. Biophys. Acta, 2009, 1794, 1784-1794.

[24] Munteanu, C.R.; Vazquez, J.M.; Dorado, J.; Sierra, A.P.; Sanchez-Gonzalez, A.; Prado-Prado, F.J.; González-Díaz, H. Complex network spectral moments for ATCUN motif DNA cleavage: First predictive study on proteins of human pathogen parasites. J. Proteome Res., 2009, 8, 5219-5228.

[25] Randi, M. A graph theoretical characterization of proteomics maps. Int. J. Quantum Chem., 2002, 90, 848-858.

[26] Randi, M. Quantitative characterizations of proteome: Dependence on the number of proteins considered. J. Proteome Res., 2006, 5, 1575-1579.

[27] González-Díaz, H. Quantitative studies on structure-activity and structure-property relationships (QSAR/QSPR). Curr. Top. Med. Chem., 2008, 8, 1554-1554.

[28] Cruz-Monteagudo, M.; Munteanu, C.R.; Borges, F.; Cordeiro, M.N.D.S.; Uriarte, E.; González-Díaz, H. Quantitative proteome-property relationships (QPPRs). Part 1: Finding biomarkers of organic drugs with mean Markov connectivity indices of spiral networks of blood mass spectra. Bioorg. Med. Chem., 2008, 16, 9684-9693.

[29] Munteanu, C.R.; Magalhães, A.L.; Uriarte, E.; González-Díaz, H. Multi-target QPDR classification model for human breast and colon cancer-related proteins using star graph topological indices. J. Theor. Biol., 2009, 257, 303-311.

[30] Vázquez, J.M.; Aguiar, V.; Seoane, J.A.; Freire, A.; Serantes, J.A.; Dorado, J.; Pazos, A.; Munteanu, C.R. Star graphs of protein sequences and proteome mass spectra in cancer prediction. Curr. Proteomics, 2009, 6, 275-288.

[31] González-Díaz, H. Network topological indices, drug metabolism, and distribution. Curr. Drug Metab., 2010, 11, 283-284.

[32] González-Díaz, H. QSAR and complex networks in pharmaceutical design, microbiology, parasitology, toxicology, cancer and neurosciences. Curr. Pharm. Des., 2010, 16, 2598-2600.

[33] Munteanu, C.R.; Fernandez-Blanco, E.; Seoane, J.A.; Izquierdo-Novo, P.; Rodriguez-Fernandez, J.A.; Prieto-Gonzalez, J.M.; Rabunal, J.R.; Pazos, A. Drug discovery and design for complex diseases through QSAR computational methods. Curr. Pharm. Des., 2010, 16, 2640-2655.

[34] Ivanciuc, T.; Ivanciuc, O.; Klein, D.J. Network-QSAR with reaction poset quantitative superstructure-activity relationships (QSSAR) for PCB chromatographic properties. Curr. Bioinform., 2011, 6, 25-34.

[35] Michalewicz, Z. Genetic Algorithms + Data Structures = Evolution Programs. Springer: 1996.

[36] Fogel, D.B. What is evolutionary computation? IEEE Spectr., 2000, 37, 26-32.

[37] Goldberg, D.E. Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley: Reading, Massachusetts, 1989; p xiii, 412 p.

[38] Darwin, C. On the Origin of Species by Means of Natural Selection. John Murray: 1859.

[39] Holland, J. Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence. University of Michigan Press: 1975.

[40] Koza, J.R. Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press: 1992.

[41] Koza, J.R. Genetic Programming III: Darwinian Invention and Problem Solving. Morgan Kaufmann: 1999.

[42] Dudek, A.Z.; Arodz, T.; Gálvez, J. Computational methods in developing quantitative structure-activity relationships (QSAR): A review. Comb. Chem. High Throughput Screen, 2006, 9, 213-228.

[43] Todeschini, R.; Consonni, V. Handbook of Molecular Descriptors. Wiley-VCH: 2002.

[44] Xue, L.; Bajorath, J. Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. Comb. Chem. High Throughput Screen, 2000, 3, 363-372.

[45] Diudea, M.V.; Gutman, I.; Jäntschi, L. Molecular Topology. Nova Science Pub Inc: Huntington, NY, 2001; p 332.

[46] Skvortsova, M.I.; Fedyaev, K.S.; Palyulin, V.A.; Zefirov, N.S. Molecular design of chemical compounds with prescribed properties from QSAR models containing the Hosoya index. Internet Electron. J. Mol. Des., 2003, 2, 70-85.

[47] Roy, K.; Saha, A. Comparative QSPR studies with molecular connectivity, molecular negentropy and TAU indices. Part 2. Lipid-water partition coefficient of diverse functional acyclic compounds. Internet Electron. J. Mol. Des., 2003, 2, 288-305.

[48] Roy, K.; Ghosh, G. Introduction of extended topochemical atom (ETA) indices in the valence electron mobile (VEM) environment as tools for QSAR/QSPR studies. Internet Electron. J. Mol. Des., 2003, 2, 599-620.

[49] Milievi, A.; Nikoli, S.; Plavi, D.; Trinajsti, N. On the Hosoya Zindex of general graphs. Internet Electron. J. Mol. Des., 2003, 2, 160-178.

[50] Gute, B.D.; Basak, S.C.; Mills, D.; Hawkins, D.M. Tailored similarity spaces for the prediction of physicochemical properties. Internet Electron. J. Mol. Des., 2002, 1, 374-387.

[51] Mulliken, R.S. Electronic population analysis on LCAO-MO molecular wave functions. I. J. Chem. Phys., 1955, 23, 1833-1840.

[52] Randi, M. Characterization of molecular branching. J. Am. Chem. Soc., 1975, 97, 6609-6615.

[53] Wiener, H. Structural determination of paraffin boiling points. J. Am. Chem. Soc., 1947, 69, 17-20.

[54] Balaban, A.T. Highly discriminating distance-based topological index. Chem. Phys. Lett., 1982, 89, 399-404.

[55] Kier, L.B.; Hall, L.H. Derivation and significance of valence molecular connectivity. J. Pharm. Sci., 1981, 70, 583-589.

[56] Gálvez, J.; Garcia, R.; Salabert, M.T.; Soler, R. Charge indexes. New topological descriptors. J. Chem. Inf. Comput. Sci., 1994, 34, 520-525.

[57] Costescu, A.; Diudea, M.V. QSTR study on aquatic toxicity against Poecilia reticulataand Tetrahymena pyriformisusing topological indices. Internet Electron. J. Mol. Des., 2006, 5, 116-134.

[58] Lui, B.; Nikoli, S.; Trinajsti, N. Distance-related molecular descriptors. Internet Electron. J. Mol. Des., 2008, 7, 195-206.

[59] Trinajsti, N. A life in science. Internet Electron. J. Mol. Des., 2003, 2, 413-434.

[60] Pogliani, L. The evolution of the valence delta in molecular connectivity theory. Internet Electron. J. Mol. Des., 2006, 5, 364-375.

[61] Bonchev, D. My life-long journey in mathematical chemistry. Internet Electron. J. Mol. Des., 2005, 4, 434-490.

[62] Kier, L.B. My journey through structure: The structure of my journey. Internet Electron. J. Mol. Des., 2006, 5, 181-191.

[63] Hosoya, H. The topological index Z before and after 1971. Internet Electron. J. Mol. Des., 2002, 1, 428-442.

[64] Ciubotariu, D.; Medeleanu, M.; Vlaia, V.; Olariu, T.; Ciubotariu, C.; Dragos, D.; Corina, S. Molecular van der Waals space and topological indices from the distance matrix. Molecules, 2004, 9, 1053-1078.

[65] Barnard, J.M.; Downs, G.M. Chemical fragment generation and clustering software. J. Chem. Inf. Comput. Sci., 1997, 37, 141-142.

[66] Waller, C.L.; Bradley, M.P. Development and validation of a novel variable selection technique with application to multidimensional quantitative structure-activity relationship studies. J. Chem. Inf. Comput. Sci., 1999, 39, 345-355.

[67] Winkler, D.A.; Burden, F.R. Holographic QSAR of benzodiazepines. Quant. Struct.-Act. Relat., 1998, 17, 224-231.

[68] Van Drie, J.H. Monty Kier and the origin of the pharmacophore concept. Internet Electron. J. Mol. Des., 2007, 6, 271-279.

[69] Cramer, R.D.; Patterson, D.E.; Bunce, J.D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. J. Am. Chem. Soc., 1988, 110, 5959-5967.

[70] Klebe, G.; Abraham, U.; Mietzner, T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. J. Med. Chem., 1994, 37, 4130-4146.

[71] Silverman, B.D.; Platt, D.E. Comparative molecular moment analysis (CoMMA): 3D-QSAR without molecular superposition. J. Med. Chem., 1996, 39, 2129-2140.

[72] Todeschini, R.; Lasagni, M. New molecular descriptors for 2D and 3D structures. Theory. J. Chemometr., 1994, 8, 263-272.

[73] Bravi, G.; Gancia, E.; Mascagni, P.; Pegna, M.; Todeschini, R.; Zaliani, A. MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: A comparative 3D QSAR study in a series of steroids. J. Comput.-Aided Mol. Des., 1997, 11, 79-92.

[74] Crivori, P.; Cruciani, G.; Carrupt, P.A.; Testa, B. Predicting blood-brain barrier permeation from three-dimensional molecular structure. J. Med. Chem., 2000, 43, 2204-2216.

[75] Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. GRid-INdependent descriptors (GRIND): A novel class of alignment-independent three-dimensional molecular descriptors. J. Med. Chem., 2000, 43, 3233-3243.

[76] González-Díaz, H.; Gia, O.; Uriarte, E.; Hernádez, I.; Ramos, R.; Chaviano, M.; Seijo, S.; Castillo, J.A.; Morales, L.; Santana, L.; Akpaloo, D.; Molina, E.; Cruz, M.; Torres, L.A.; Cabrera, M.A. Markovian chemicals 'in silico' design (MARCH-INSIDE), a promising approach for computer-aided molecular design I: discovery of anticancer compounds. J. Mol. Model., 2003, 9, 395-407.

[77] DRAGON for Windows (software for molecular descriptor calculations), http://www.talete.mi.it., 2005.

[78] Perez Gonzalez, M.; Morales Helguera, A. TOPS-MODE versus DRAGON descriptors to predict permeability coefficients through low-density polyethylene. J. Comput.-Aided Mol. Des., 2003, 17, 665-672.

[79] Kubinyi, H. Lecture at the 18th EuroQSAR, The Long Road from QSAR to Virtual Screening. Rhodes, Greece, 2010.

[80] Kubinyi, H. Virtual Screening (Lectures of the Drug Design Course) 2006. http://kubinyi.de/dd-19.pdf.

[81] Kubinyi, H. Virtual Screening - The Road to Success. In XIX International Symposium on Medicinal Chemistry, Available from: http://kubinyi.de/istanbul-09-06.pdf:Istambul, 2006.

[82] Sies, H. A new parameter for sex education. Nature, 1988, 332, 495.

[83] Doweyko, A.M. QSAR: dead or alive? J. Comput.-Aided Mol. Des., 2008, 22, 81-89.

[84] Golbraikh, A.; Tropsha, A. Beware of q2! J. Mol. Graph. Model., 2002, 20, 269-276.

[85] Lindgren, F.; Hansen, B.; Karcher, W.; Sjöström, M.; Eriksson, L. Model validation by permutation tests: Applications to variable selection. J. Chemom., 1996, 10, 521-532.

[86] Tropsha, A.; Golbraikh, A. Predictive QSAR modeling workflow, model applicability domains, and virtual screening. Curr. Pharm. Des., 2007, 13, 3494-3504.

[87] Tropsha, A. Best practices for QSAR model development, validation, and exploitation. Mol. Inf., 2010, 29, 476-488.

[88] Doweyko, A.M. 3D-QSAR illusions. J. Comput.-Aided Mol. Des., 2004, 18, 587-596.

[89] Unger, S.H.; Hansch, C. On model building in structure-activity relationships. A reexamination of adrenergic blocking activity of -halo--arylalkylamines. J. Med. Chem., 1973, 16, 745-749.

[90] González, M.P.; Terán, C.; Saíz-Urra, L.; Teijeira, M. Variable selection methods in QSAR: an overview. Curr. Top. Med. Chem., 2008, 8, 1606-1627.

[91] Merkwirth, C.; Mauser, H.; Schulz-Gasch, T.; Roche, O.; Stahl, M.; Lengauer, T. Ensemble methods for classification in cheminformatics. J. Chem. Inf. Comput. Sci., 2004, 44, 1971-1978.

[92] Guha, R.; Jurs, P.C. Development of QSAR models to predict and interpret the biological activity of artemisinin analogues. J. Chem. Inf. Comput. Sci., 2004, 44, 1440-1449.

[93] Liu, Y. A comparative study on feature selection methods for drug discovery. J. Chem. Inf. Comput. Sci., 2004, 44, 1823-1828.

[94] Lin, T.H.; Li, H.T.; Tsai, K.C. Implementing the Fisher's discriminant ratio in a k-means clustering algorithm for feature selection and data set trimming. J. Chem. Inf. Comput. Sci., 2004, 44, 76-87.

[95] Massey, F.J.J. The Kolmogorov-Smirnov test of goodness of fit. J. Am. Stat. Assoc., 1951, 46, 68-78.

[96] Kohavi, R.; John, G. Wrappers for feature subset selection. Artif. Intell., 1997, 97, 273- 324.

[97] Pearson, K. On lines and planes of closest fit to systems of points in space. London Edinburgh Dublin Philos. Mag. J. Sci., 1901, Sixth Series 2, 559-572.

[98] Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S. Multi- and Megavariate Data Analysis. Principles and Applications. Umetrics Academy: Umea, Sweden, 2001.

[99] Saltelli, A.; Chan, K.; Scott, E. Sensitivity Analysis. Wiley: Baffins Lane, Chichester, UK, 2001.

[100] Bayram, E.; Santago, P., 2nd; Harris, R.; Xiao, Y.D.; Clauset, A.J.; Schmitt, J.D. Genetic algorithms and self-organizing maps: a powerful combination for modeling complex QSAR and QSPR problems. J. Comput.-Aided Mol. Des., 2004, 18, 483-493.

[101] Wegner, J.K.; Frohlich, H.; Zell, A. Feature selection for descriptor based classification models. 1. Theory and GA-SEC algorithm. J. Chem. Inf. Comput. Sci., 2004, 44, 921-930.

[102] Kirkpatrick, S.; Gelatt, C.D.; Vecchi, M.P. Optimization by simulated annealing. Science, 1983, 220, 671-680.

[103] Sun, Y.; Todorovic, S.; Goodison, S. Local-learning-based feature selection for high-dimensional data analysis. IEEE Trans. Pattern Anal. Mach. Intell., 2010, 32, 1610-1626.

[104] Bi, J.; Bennett, K.; Embrechts, M.; Breneman, C.; Song, M. Dimensionality reduction viasparse support vector machines. J. Mach. Learn. Res., 2003, 3, 1229-1243.

[105] Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. Mach. Learn., 2002, 46, 389-422.

[106] Dhillon, P.S.; Tomasik, B.; Foster, D.; Ungar, L. Multi-task feature selection using the multiple inclusion criterion (MIC). In: Machine Learning and Knowledge Discovery in Databases. Lecture Notes in Computer Science, Buntine, W.; Grobelnik, M.; Mladenic, D.; ShaweTaylor, J., Eds. 2009; Vol. 5781, pp 276-289.

[107] Livingstone, D.J.; Salt, D.W. Judging the significance of multiple linear regression models. J. Med. Chem., 2005, 48, 661-663.

[108] Afantitis, A.; Melagraki, G.; Sarimveis, H.; Koutentis, P.A.; Markopoulos, J.; Igglessi-Markopoulou, O. A novel simple QSAR model for the prediction of anti-HIV activity using multiple linear regression analysis. Mol. Divers., 2006, 10, 405-414.

[109] Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. Chemometrics Intell. Lab. Syst., 2001, 58, 109-130.

[110] Luco, J.M.; Ferretti, F.H. QSAR based on multiple linear regression and PLS methods for the anti-HIV activity of a large group of HEPT derivatives. J. Chem. Inf. Comput. Sci., 1997, 37, 392-401.

[111] Olah, M.; Bologa, C.; Oprea, T.I. An automated PLS search for biologically relevant QSAR descriptors. J. Comput.-Aided Mol. Des., 2004, 18, 437-449.

[112] Fisher, R.A. The use of multiple measurements in taxonomic problems. Ann. Eugenics, 1936, 7, 179-188.

[113] Mattioni, B.E.; Jurs, P.C. Prediction of dihydrofolate reductase inhibition and selectivity using computational neural networks and linear discriminant analysis. J. Mol. Graph. Model., 2003, 21, 391-419.

[114] Reibnegger, G.; Weiss, G.; Werner-Felmayer, G.; Judmaier, G.; Wachter, H. Neural networks as a tool for utilizing laboratory information: comparison with linear discriminant analysis and with classification and regression trees. Proc. Natl. Acad. Sci. USA, 1991, 88, 11426-11430.

[115] Qu, K.; McCue, L.A.; Lawrence, C.E. In Bayesian protein family classifier, ISMB 1998. Proceedings of the 6th International Conference on Intelligent Systems for Molecular Biology, AAAI Press: 1998; pp 131-139.

[116] Polley, M.J.; Winkler, D.A.; Burden, F.R. Broad-based quantitative structure-activity relationship modeling of potency and selectivity of farnesyltransferase inhibitors using a Bayesian regularized neural network. J. Med. Chem., 2004, 47, 6230-6238.

[117] Sun, H. A naive Bayes classifier for prediction of multidrug resistance reversal activity on the basis of atom typing. J. Med. Chem., 2005, 48, 4031-4039.

[118] Krishnapuram, B.; Hartemink, A.J.; Carin, L.; Figueiredo, M.A. A bayesian approach to joint feature selection and classifier design. IEEE Trans. Pattern Anal. Mach. Intell., 2004, 26, 1105-1111.

[119] Cover, T.M.; Hart, P.E. Nearest neighbor pattern classification. IEEE Trans. Inf. Theory, 1967, 13, 21-27.

[120] Geva, S.; Sitte, J. Adaptive nearest neighbor pattern classification. IEEE Trans. Neural Netw., 1991, 2, 318-322.

[121] Chou, K.C.; Shen, H.B. Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-Nearest Neighbor classifiers. J. Proteome Res., 2006, 5, 1888-1897.

[122] Kauffman, G.W.; Jurs, P.C. QSAR and k-nearest neighbor classification analysis of selective cyclooxygenase-2 inhibitors using topologically-based numerical descriptors. J. Chem. Inf. Comput. Sci., 2001, 41, 1553-1560.

[123] Goodacre, R.; Neal, M.J.; Kell, D.B. Quantitative analysis of multivariate data using artificial neural networks: A tutorial review and applications to the deconvolution of pyrolysis mass spectra. Zentralbl. Bakteriol., 1996, 284, 516-539.

[124] Winkler, D.A. Neural networks as robust tools in drug lead discovery and development. Mol. Biotechnol., 2004, 27, 139-167.

[125] Niculescu, S.P.; Atkinson, A.; Hammond, G.; Lewis, M. Using fragment chemistry data mining and probabilistic neural networks in screening chemicals for acute toxicity to the fathead minnow. SAR QSAR Environ. Res., 2004, 15, 293-309.

[126] Devillers, J. A new strategy for using supervised artificial neural networks in QSAR. SAR QSAR Environ. Res., 2005, 16, 433-442.

[127] Hemmateenejad, B.; Safarpour, M.A.; Miri, R.; Nesari, N. Toward an optimal procedure for PC-ANN model building: Prediction of the carcinogenic activity of a large set of drugs. J. Chem Inf. Model., 2005, 45, 190-199.

[128] Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. Psychol. Rev., 1958, 65, 386-408.

[129] Mulgrew, B. Applying radial basis functions. IEEE Signal Process. Mag., 1996, 13, 50-65.

[130] Fineberg, H.V. Decision trees: construction, uses, and limits. Bull. Cancer, 1980, 67, 395-404.

[131] Quinlan, J.R. Induction of decision trees. Mach. Learn., 1986, 1, 81-106.

[132] Yang, Z.R. Mining SARS-CoV protease cleavage data using non-orthogonal decision trees: a novel method for decisive template selection. 2005, 21, 2644-2650.

[133] Cortes, C.; Vapnik, V. Support vector networks. Mach. Learn., 1995, 20, 273-297.

[134] Ivanciuc, O. Applications of support vector machines in chemistry. In: Reviews in Computational Chemistry, Lipkowitz, K.B.; Cundari, T.R., Eds. Wiley-VCH: Weinheim, 2007; Vol. 23, pp 291-400.

[135] Abraham, A. Evolutionary computation. In: Handbook for Measurement Systems Design, Sydenham, P.; Thorn, R., Eds. John Wiley and Sons Ltd.: London, 2005; pp 920-931.

[136] So, S.-S.; Karplus, M. Evolutionary optimization in quantitative structure-activity relationship: An application of genetic neural networks. J. Med. Chem., 1996, 39, 1521-1530.

[137] Fogel, D.B. Evolutionary Computation: Toward a New Philosophy of Machine Intelligence. 2nd ed.; IEEE Press: Piscataway, NJ, 1999.

[138] Eberbach, E. Toward a theory of evolutionary computation. Biosystems, 2005, 82, 1-19.

[139] Rogers, D.; Hopfinger, A.J. Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relations. J. Chem. Inf. Comput. Sci., 1994, 34, 854-866.

[140] Raichurkar, A.V.; Kulkarni, V.M. 3D-QSAR of cyclooxygenase-2 inhibitors by genetic function approximation. Internet Electron. J. Mol. Des., 2003, 2, 242-261.

[141] Vadlamudi, S.M.; Kulkarni, V.M. 3D-QSAR of protein tyrosine phosphatase 1B inhibitors by genetic function approximation. Internet Electron. J. Mol. Des., 2004, 3, 586-609.

[142] Breiman, L. Bagging predictors. Mach. Learn., 1996, 24, 123-140.

[143] Ho, T.K. The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell., 1998, 20, 832-844

[144] Freund, Y.; Schapire, R.E. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci., 1997, 55, 119-139.

[145] Hansch, C. Quantitative structure-activity relationships in drug design. In: Drug Design, Ariëns, E.J., Ed. Academic Press: New York, 1971; Vol. 1, pp 271-342.

[146] Gillet, V.J. Applications of evolutionary computation in drug design. In: Applications of Evolutionary Computation in Chemistry, Johnston, R.L., Ed. Springer: Berlin, 2004; Vol. 110, pp 133-152.

[147] Guyon, I.; Elisseefi, A.; Kaelbling, L.P. An introduction to variable and feature selection. J. Mach. Learn. Res., 2003, 3, 1157-1182.

[148] Haykin, S. Neural Networks: A Comprehensive Foundation. Prentice Hall: NJ, USA, 1998.

[149] Wu, J.; Mei, J.; Wen, S.; Liao, S.; Chen, J.; Shen, Y. A self adaptive genetic algorithm artificial neural network algorithm with leave one out cross validation for descriptor selection in QSAR study. J. Comput. Chem., 2010, 31, 1956-1968.

[150] Jalali-Heravi, M.; Kyani, A. Application of genetic algorithm-kernel partial least square as a novel nonlinear feature selection method: activity of carbonic anhydrase II inhibitors. Eur. J. Med. Chem., 2007, 42, 649-659.

[151] Hasegawa, K.; Miyashita, Y.; Funatsu, K. GA strategy for variable selection in QSAR studies: GA-based PLS analysis of calcium channel antagonists. J. Chem. Inf. Comput. Sci., 1997, 37, 306-310.

[152] Cho, S.J.; Hermsmeier, M.A. Genetic algorithm guided selection: Variable selection and subset selection. J. Chem. Inf. Comput. Sci., 2002, 42, 927-936.

[153] Hasegawa, K.; Kimura, T.; Funatsu, K. GA strategy for variable selection in QSAR studies: Application of GA-based region selection to a 3D-QSAR study of acetylcholinesterase inhibitors. J. Chem. Inf. Comput. Sci., 1999, 39, 112-120.

[154] Pavan, M.; Mauri, A.; Todeschini, R. Total ranking models by the genetic algorithm variable subset selection (GA-VSS) approach for environmental priority settings. Anal. Bioanal. Chem., 2004, 380, 430-444.

[155] Venkatraman, V.; Dalby, A.R.; Yang, Z.R. Evaluation of mutual information and genetic programming for feature selection in QSAR. J. Chem. Inf. Comput. Sci., 2004, 44, 1686-1692.

[156] Shen, Q.; Jiang, J.H.; Shen, G.L.; Yu, R.Q. Variable selection by an evolution algorithm using modified Cp based on MLR and PLS modeling: QSAR studies of carcinogenicity of aromatic amines. Anal. Bioanal. Chem., 2003, 375, 248-254.

[157] Shen, Q.; Jiang, J.-H.; Tao, J.-C.; Shen, G.-L.; Yu, R.-Q. Modified ant colony optimization algorithm for variable selection in QSAR modeling: QSAR studies of cyclooxygenase inhibitors. J. Chem. Inf. Model., 2005, 45, 1024-1029.

[158] Izrailev, S.; Agrafiotis, D.K. Variable selection for QSAR by artificial ant colony systems. SAR QSAR Environ. Res., 2002, 13, 417-423.

[159] Goodarzi, M.; Freitas, M.P.; Jensen, R. Ant colony optimization as a feature selection method in the QSAR modeling of anti-HIV-1 activities of 3-(3,5-dimethylbenzyl)uracil derivatives using MLR, PLS and SVM regressions. Chemometrics Intell. Lab. Syst., 2009, 98, 123-129.

[160] Shamsipur, M.; Zare-Shahabadi, V.; Hemmateenejad, B.; Akhond, M. An efficient variable selection method based on the use of external memory in ant colony optimization. Application to QSAR/QSPR studies. Anal. Chim. Acta, 2009, 646, 39-46.

[161] Patil, D.; Raj, R.; Shingade, P.; Kulkarni, B.; Jayaraman, V.K. Feature Selection and Classification Employing Hybrid Ant Colony Optimization/Random Forest Methodology. Comb. Chem. High Throughput Screen, 2009, 12, 507-513.

[162] Lin, W.-Q.; Jiang, J.-H.; Shen, Q.; Shen, G.-L.; Yu, R.-Q. Optimized block-wise variable combination by particle swarm optimization for partial least squares modeling in quantitative structure-activity relationship studies. J. Chem. Inf. Model., 2005, 45, 486-493.

[163] Agrafiotis, D.K.; Cedeño, W. Feature selection for structure-activity correlation using binary particle swarms. J. Med. Chem., 2002, 45, 1098-1107.

[164] Shen, Q.; Jiang, J.-H.; Jiao, C.-X.; Shen, G.-L.; Yu, R.-Q. Modified particle swarm optimization algorithm for variable selection in MLR and PLS modeling: QSAR studies of antagonism of angiotensin II antagonists. Eur. J. Pharm. Sci., 2004, 22, 145-152.

[165] Nicolotti, O.; Gillet, V.J.; Fleming, P.J.; Green, D.V.S. Multiobjective optimization in quantitative structure-activity relationships: Deriving accurate and interpretable QSARs. J. Med. Chem., 2002, 45, 5069-5080.

[166] Gao, H.; Lajiness, M.S.; Van Drie, J. Enhancement of binary QSAR analysis by a GA-based variable selection method. J. Mol. Graph. Model., 2002, 20, 259-268.

[167] Fernández-Blanco, E.; Dorado, J.; Rabuñal, J.; Gestal, M.; Pedreira, N. A computational morphogenesis approach to simple structure development. In: Advances in Artificial Life. 9th European Conference, ECAL 2007, Lisbon, Portugal, September 10-14, 2007. Proceedings. Lecture Notes in Computer Science, Almeida e Costa, F.; Rocha, L.M.; Costa, E.; Harvey, I.; Coutinho, A., Eds. 2007; Vol. 4648, pp 825-834.

[168] Ghosh, P.; Bagchi, M.C. QSAR modeling for quinoxaline derivatives using genetic algorithm and simulated annealing based feature selection. Curr. Med. Chem., 2009, 16, 4032-4048.

[169] Yan, A.X.; Wang, Z.; Cai, Z.Y. Prediction of human intestinal absorption by GA feature selection and support vector machine regression. Int. J. Mol. Sci., 2008, 9, 1961-1976.

[170] Taha, M.O.; Qandil, A.M.; Zaki, D.D.; AlDamen, M.A. Ligand-based assessment of factor Xa binding site flexibility viaelaborate pharmacophore exploration and genetic algorithm-based QSAR modeling. Eur. J. Med. Chem., 2005, 40, 701-727.

[171] Pan, D.H.; Tseng, Y.F.; Hopfinger, A.J. Quantitative structure-based design: Formalism and application of receptor-dependent RD-4D-QSAR analysis to a set of glucose analogue inhibitors of glycogen phosphorylase. J. Chem. Inf. Comput. Sci., 2003, 43, 1591-1607.

[172] Zaheer-ul-Haq; Uddin, R.; Yuan, H.B.; Petukhov, P.A.; Choudhary, M.I.; Madura, J.D. Receptor-based modeling and 3D-QSAR for a quantitative production of the

butyrylcholinesterase inhibitors based on genetic algorithm. J. Chem. Inf. Model., 2008, 48, 1092-1103.

[173] Asadollahi, T.; Dadfarnia, S.; Shabani, A.M.H.; Ghasemi, J.B.; Sarkhosh, M. QSAR models for CXCR2 receptor antagonists based on the genetic algorithm for data preprocessing prior to application of the PLS linear regression method and design of the new compounds using in silicovirtual screening. Molecules, 2011, 16, 1928-1955.

[174] Yanmaz, E.; Saripinar, E.; Sahin, K.; Gecen, N.; Copur, F. 4D-QSAR analysis and pharmacophore modeling: Electron conformational-genetic algorithm approach for penicillins. Bioorg. Med. Chem., 2011, 19, 2199-2210.

[175] Zhou, X.A.; Li, Z.C.; Dai, Z.; Zou, X.Y. QSAR modeling of peptide biological activity by coupling support vector machine with particle swarm optimization algorithm and genetic algorithm. J. Mol. Graph. Model., 2010, 29, 188-196.

[176] Reddy, A.S.; Kumar, S.; Garg, R. Hybrid-genetic algorithm based descriptor optimization and QSAR models for predicting the biological activity of Tipranavir analogs for HIV protease inhibition. J. Mol. Graph. Model., 2010, 28, 852-862.

[177] Ma, C.Y.; Buontempo, F.V.; Wang, X.Z. Inductive data mining: Automatic generation of decision trees from data for QSAR modelling and process historical data analysis, In: 18th European Symposium on Computer Aided Process Engineering. Computer Aided Chemical Engineering, Lyon, France, Braunschweig, B.; Joulia, X., Eds. Elservier Science: Lyon, France, 2008; pp 581-586.

[178] Archetti, F.; Giordani, I.; Vanneschi, L. Genetic programming for QSAR investigation of docking energy. Appl. Soft. Comput., 2010, 10, 170-182.

[179] Zhou, X.-D.; Li, T.-H.; Bian, F.; Quian, J.-L. Application of genetic programming coupling with genetic algorithm. Chem. J. Chin. Univ., 2000, 21, 216-218.

[180] Bleicher, K.H.; Bohm, H.J.; Muller, K.; Alanine, A.I. Hit and lead generation: Beyond high-throughput screening. Nat. Rev. Drug Discov., 2003, 2, 369-378.

[181] Xu, J.; Hagler, A. Chemoinformatics and drug discovery. Molecules, 2002, 7, 566-600.

[182] Butina, D.; Segall, M.D.; Frankcombe, K. Predicting ADME properties in silico: methods and models. Drug Discov. Today, 2002, 7, S83-S88.

[183] Manly, C.J.; Louise-May, S.; Hammer, J.D. The impact of informatics and computational chemistry on synthesis and screening. Drug Discov. Today, 2001, 6, 1101-1110.

[184] Seifert, M.H.J.; Wolf, K.; Vitt, D. Virtual high-throughput in silico screening. BIOSILICO, 2003, 1, 143-149.

[185] Lahana, R. How many leads from HTS? Drug Discov. Today, 1999, 4, 447-448.

[186] Mayer, J.M.; van de Waterbeemd, H. Development of quantitative structure-pharmacokinetic relationships. Environ. Health Perspect., 1985, 61, 295-306.

[187] Jorgensen, W.L. The many roles of computation in drug discovery. Science, 2004, 303, 1813-1818.

[188] DiMasi, J.A.; Hansen, R.W.; Grabowski, H.G. The price of innovation: new estimates of drug development costs. J. Health Econ., 2003, 22, 151-185.

[189] Cruz-Monteagudo, M.; Cordeiro, M.; Teijeira, M.; Gonzalez, M.P.; Borges, F. Multidimensional drug design: simultaneous analysis of binding and relative efficacy profiles of N6-substituted-4'-thioadenosines A3adenosine receptor agonists. Chem. Biol. Drug Des., 2010, 75, 607-618.

[190] Nicolotti, O.; Giangreco, I.; Miscioscia, T.F.; Carotti, A. Improving quantitative structure-activity relationships through multiobjective optimization. J. Chem. Inf. Model., 2009, 49, 2290-2302.

[191] Cruz-Monteagudo, M.; Borges, F.; Cordeiro, M. Desirability-based multiobjective optimization for global QSAR studies: Application to the design of novel NSAIDs with improved analgesic, antiinflammatory, and ulcerogenic profiles. J. Comput. Chem., 2008, 29, 2445-2459.

[192] Cruz-Monteagudo, M.; Borges, F.; Cordeiro, M.; Fajin, J.L.C.; Morell, C.; Ruiz, R.M.; Canizares-Carmenate, Y.; Dominguez, E.R. Desirability-based methods of multiobjective optimization and ranking for global QSAR studies. filtering safe and potent drug candidates from combinatorial libraries. J. Comb. Chem., 2008, 10, 897-913.

[193] Nicolaou, C.A.; Apostolakis, J.; Pattichis, C.S. De novo drug design using multiobjective evolutionary graphs. J. Chem. Inf. Model., 2009, 49, 295-307.

[194] Yamashita, F.; Hara, H.; Ito, T.; Hashida, M. Novel hierarchical classification and visualization method for multiobjective optimization of drug properties: Application to structure-activity relationship analysis of cytochrome P450 metabolism. J. Chem. Inf. Model., 2008, 48, 364-369.

[195] Machado, A.; Tejera, E.; Cruz-Monteagudo, M.; Rebelo, I. Application of desirability-based multi(bi)-objective optimization in the design of selective arylpiperazine derivates for the 5-HT1Aserotonin receptor. Eur. J. Med. Chem., 2009, 44, 5045-5054.

[196] Cruz-Monteagudo, M.; PhamThe, H.; Cordeiro, M.; Borges, F. Prioritizing hits with appropriate trade-offs between HIV-1 reverse transcriptase inhibitory efficacy and MT4 blood cells toxicity through desirability-based multiobjective optimization and ranking. Mol. Inf., 2010, 29, 303-321.

[197] Chen, W.J.; Giger, M.L.; Newstead, G.M.; Bick, U.; Jansen, S.A.; Li, H.; Lan, L. Computerized assessment of breast lesion malignancy using DCE-MRI: Robustness study on two independent clinical datasets from two manufacturers. Acad. Radiol., 2010, 17, 822-829.

[198] Chen, Y.X.; Huang, S.Q.; Chen, H.M.; Bart, H.L. Joint feature selection and classification for taxonomic problems within fish species complexes. Pattern Anal. Appl., 2010, 13, 23-34.

[199] Lan, L.; Vucetic, S. A multi-task feature selection filter for microarray classification, In: BIBMW 2009. Proceedings of the 2009 IEEE International Conference on Bioinformatics and Biomedicine Workshop, 2009; pp 160-165.

[200] Ghosh, D.; Chakrabarti, R. Joint variable selection and classification with immunohistochemical data. Biomark. Insights, 2009, 4, 103-110.

[201] Lapedriza, A.; Segui, S.; Masip, D.; Vitria, J. A sparse Bayesian approach for joint feature selection and classifier learning. Pattern Anal. Appl., 2008, 11, 299-308.

[202] Chen, W.; Zur, R.M.; Giger, M.L. Joint feature selection and classification using a Bayesian neural network with "automatic relevance determination" priors: Potential use in CAD of medical imaging. In: Medical Imaging 2007: Computer-Aided Diagnosis. Proc. SPIE, Giger, M.L.; Karssemeijer, N., Eds. 2007; Vol. 6514, p 65141G.

[203] Pratapa, P.N.; Patz, E.F., Jr.; Hartemink, A.J. Finding diagnostic biomarkers in proteomic spectra. Pac. Symp. Biocomput., 2006, 279-290.

[204] Cordón, O.; Del Jesus, M.J.; Herrera, F.; Magdalena, L.; Villar, P. A multiobjective genetic learning process for joint feature selection and granularity and contexts learning in fuzzy rule-based classification systems. In: Interpretability Issues in Fuzzy Modeling, Casillas, J.; Cordón, O.; Herrera Triguero, F.; Magdalena, L., Eds. Springer: Berlin, 2003; pp 79-99.