

Multi-Objective Feature Selection in QSAR Using a Machine Learning Approach

Axel J. Soto^{a,b*}, Rocío L. Cecchini^a, Gustavo E. Vazquez^a, Ignacio Ponzoni^{a,b}

^a Laboratorio de Investigación y Desarrollo en Computación Científica (LIDeCC), Departamento de Ciencias e Ingeniería de la Computación, Universidad Nacional del Sur, Av. Alem 1253, 8000 Bahía Blanca, Argentina

^b Planta Piloto de Ingeniería Química (PLAPIQUI), Universidad Nacional del Sur, CONICET, Camino La Carrindanga km.7, CC 717, Bahía Blanca, Argentina

*e-mail: asoto@plapiqui.edu.ar

Keywords: Descriptor selection, Multi-objective evolutionary algorithms, Bayesian regularized neural networks, Computational chemistry, Medicinal chemistry

Received: May 11, 2009; Accepted: November 25, 2009

DOI: 10.1002/qsar.200960053

Abstract

The selection of descriptor subsets for QSAR/QSPR is a hard combinatorial problem that requires the evaluation of complex relationships in order to assess the relevance of the selected subsets. In this paper, we describe the main issues in applying descriptor selection for QSAR methods and propose a novel two-phase methodology for this task. The first phase makes use of a multi-objective evolutionary technique which yields interesting advantages compared to mono-objective methods. The second phase complements the first one and it enables to refine and improve the confidence in the chosen subsets of descriptors. This methodology allows the selection of subsets when a large number of descriptors are involved and it is also suitable for linear and nonlinear QSAR/QSPR models. The proposed method was tested using three data sets with experimental values for blood-brain barrier penetration, human intestinal absorption and hydrophobicity. Results reveal the capability of the method for achieving subsets of descriptors with a high predictive capacity and a low cardinality. Therefore, our proposal constitutes a new promising technique helpful for the development of QSAR/QSPR models.

1 Introduction

The advantages of using QSAR methods in the drug discovery process have been highly recognized. However, although over the last years the number of scientific publications in this subject remained high, prediction capacity of QSAR models still remains to be improved [1–3].


One of the key and first steps in the development of successful QSAR models is the selection of relevant descriptors that relate molecular and chemical information with the desired activity or property [4, 5]. Commonly, the descriptor selection task could not be completely achieved manually by experts in biology or chemistry, given that structure–activity relationships are usually complex and nonlinear [6]. Furthermore, the number of molecular descriptors that may be calculated for a single compound is very large. Thereby, it is im-

portant to have a computational procedure for the selection of the subset of descriptors to be used in a QSAR model. This kind of procedure is referred to as Feature Selection (FS) in the literature [7–9] and is a current research area given that, besides QSAR, applications with many variables have become frequent. Such is the case with gene selection from microarray data [10, 11].

1.1. Background: Feature Selection

In order to clarify the many concepts associated with FS methods, we shall highlight the common taxonomy and main related issues. A first distinction could be made in relation to whether FS is applied in a supervised or in an unsupervised way. Applying FS in a supervised way is related to selecting and assessing variables in terms of their capacity for predicting a target variable. Applying FS in an

Abbreviations: Abbreviations and Symbols: FS feature selection; MO multi-objective; DT decision trees; kNN *k*-nearest neighbors; NLR nonlinear regression; MLR multiple linear regression; ANN artificial neural network; ANNE artificial neural network ensemble

 Supporting information for this article is available on the WWW under www.qcs.wiley-vch.de

unsupervised way is related to applying clustering methods using selected variables in the absence of a target variable. We shall focus on the supervised scenario and for further reading on unsupervised FS, the reader may refer to reference [12]. Additionally, a FS method could be categorized as a filter, wrapper or embedded method [8]. The main difference between wrappers and filters is in how the usefulness of the selected subset of features is assessed. Wrappers use a statistical learning method that is trained using only the selected subset of variables, and consequently the prediction capacity of the subset is evaluated (e.g. by cross-validation using a machine learning method). In contrast, filter methods use 'proxy' measures in order to assess the relevance of the selected variables (e.g. information gain, χ^2 -test) [4, 13]. Finally, embedded methods are prediction methods that apply variable selection as part of their own training (e.g. decision trees with pruning).

Wrapper methods are the most common choice given that they are flexible, very effective at decreasing dimensionality of the feature space and at the same time at increasing predictive accuracy [14, 15]. However, they could lead to overfitting if they are not carefully applied [16].

1.2 Main Issues in Feature Selection

In QSAR, there are two important and coexisting issues that make the problem particularly hard to solve. Firstly, we have a huge number of available descriptors (n) and also little knowledge on which and how many descriptors are necessary (p). Using an exhaustive sequential forward selection, it is required to try $\binom{n}{1} + \binom{n}{2} + \dots + \binom{n}{p}$ times in order to find the suitable subset of descriptors, and thus its time complexity is $O(2^n)$. It has been shown that the FS problem is NP-complete [17]; hence a computational approach for FS must follow an heuristic method in order to be able to find an appropriate subset of variables in a reasonable time period. Moreover, chance correlations may occur when a number of descriptors are selected from a large pool of descriptors [18].

Secondly, given that structure–activity or structure–property relationships are often nonlinear, the methods for assessing the predictive capacity of the descriptor subsets are computationally expensive. In this sense, it is important to carry out the evaluation of the quality of a subset of variables using a method that is not only able to model any kind of function, but it is also computationally cheap in order to make a fast assessment of each descriptor subset. The latter is a key feature to allow a good coverage in the method's search over the space of feasible descriptor subsets. Clearly, there is a tradeoff between the accuracy in modeling any function and the time needed to train/generate the model.

1.3 Proposed Work

In this paper, we propose a methodology aimed at selecting relevant descriptors for QSAR models. The main objective of our approach is that it be able to be used when many descriptors have to be considered regardless of the linear or nonlinear complexity of the QSAR model. Our FS technique proposes a two-phase methodology. The first phase is a multi-objective (MO) wrapper that aims both, to maximize predictive capacity and to reduce the number of selected descriptors. The output of the first phase is used by the second phase, also called validation phase, in order to determine which subsets of descriptors are the most relevant for prediction.

The MO approach allows two simultaneous advantages: first, it is prone to favor subsets with minimal cardinality and second it selects the appropriate number of descriptors in an automatic way, i.e. without the necessity to sequentially iterate trying with different subset sizes. This is important because it can be shown that the linear increase in the number of selected features leads to an exponential increase in the number of feasible learning hypothesis [19].

The second phase was added in order to assess the predictive capacity of the subsets selected by the wrapper. The motivation for this two-phase procedure resides in the fact that the first phase is responsible for a coarse and fast selection of the subset of descriptors, and thus it allows evaluating the immensity of the feasible chemical search space. Thereby, the output of the wrapper is taken by the validation phase in order to have a more rigorous assessment of the obtained selected subsets and to apply a stronger and more accurate method of prediction than the ones used in the wrapper.

1.4 Related work

A large number of papers in the literature have investigated approaches towards the selection of descriptors in QSAR [4–6, 20–25]. Early works in the area were aimed at eliminating redundant or correlated variables, and even though this elimination is important and necessary it is not sufficient [19].

Many FS strategies use evolutionary algorithms [6, 23, 24], given that they allow a stochastic and parallel search of the possible solutions of a problem, and hence they are able to escape from local minima. There are also other approaches based on stepwise strategies which perform a greedy search of the best subsets of variables [22]. A recent work [26] addresses the benefits and tradeoffs in using deterministic and stochastic FS strategies in QSAR.

Other recent articles present a similar two-phase methodology [27, 28]. In one of these works [27], the subsets of descriptors selected by a genetic algorithm are then used by a neural network model. Unlike our approach, this sec-

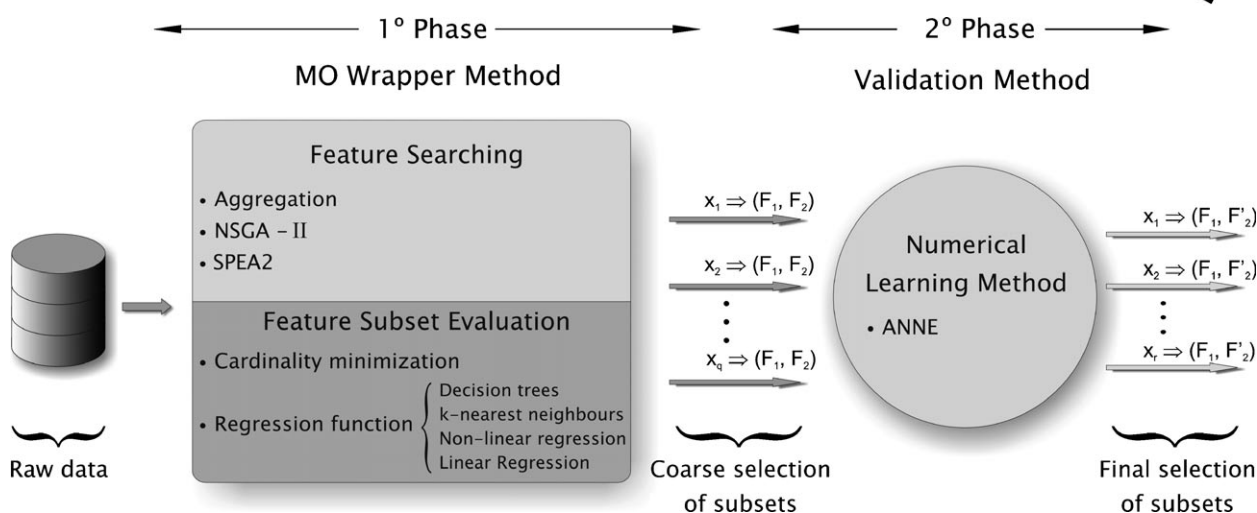


Figure 1. Scheme of the two-phase Feature Selection procedure.

ond phase is not part of the FS process. In the other two-phase work [28], a genetic algorithm is used in the second phase, after using a minimum redundancy maximum relevance criterion in the first phase.

MO FS methods were successfully applied to different scenarios, and are essentially motivated by the fact that larger feature sets will be prone to overfitting and hence to have a poor generalization performance [11, 15, 29]. In addition, MO approaches allow to reduce the search space, since they aim at giving preference to smaller subsets before bigger ones. Descriptor selection in QSAR is a new and suitable scenario for applying MO FS, given that the correct number of descriptors is not known in advance and also because models with few descriptors are more interpretable and less prone to produce overfitted models [30].

2 Methods

In this section, we shall describe the method for selecting subsets of descriptors for QSAR models by applying a methodology divided in two phases. The first phase applies MO optimization and acts as a coarse selector of subsets of descriptors, whereas the second phase is in charge of accurately assessing the subsets from the latter coarse selection. Figure 1 shows an outline of the overall procedure. It is worth mentioning that, the proposed methodology is not restricted to the specific methods used here for the wrapper and the validation phases.

2.1 First Phase: Multi-Objective (MO) Wrapper Method

Wrapper methods may be internally divided in two parts: (1) Feature Searching and (2) Feature Subset Evaluation. The former is responsible for doing the combinatorial

search among different feasible subset selections, whereas the second assesses the usefulness of the selected subset and hence guides the Feature Searching to a selection of relevant descriptors.

In the first phase, our MO wrapper is applied so that the predictive capacity of a chosen subset of variables is intended to be maximized, while maintaining the cardinality of the selected subset to a minimum. In contrast with mono-objective wrappers, the Feature Subset Evaluation of our MO wrapper is composed by two independent objectives and the Feature Searching must be able to be guided by more than one objective.

Continuing with this theme, we shall describe each component of the MO wrapper. The Feature Subset Evaluation component of the MO wrapper is conformed by a function that calculates the cardinality of the selected subsets and also by a regression or statistical function that computes an estimate of the prediction error when using a selected subset of descriptors. Although the first one is straightforward to develop, the latter could be applied with different methods [31, 32]. It is worth highlighting that the regression algorithm must be continuously invoked in order to assess the predictive capacity of each combination of descriptor subsets chosen by the Feature Searching. In this regard, it is advisable that this regression method have a good numerical learning ability and, at the same time, a 'fast' training time, so that time complexity of the wrapper does not make the problem computationally prohibitive. For example, neural networks would be ideal from the point of view of learning capacity, but taking into account its time complexity it would not be appropriate for a wrapper unless it is limited to be used for models with a small number of descriptors.

The Feature Searching function of a mono-objective wrapper is usually implemented by simulated annealing

[33] or evolutionary algorithms [34]. However, in the MO optimization field, studies on evolutionary algorithms and their application outnumber those on simulated annealing [35]. According to the literature, MO evolutionary algorithms may be applied following a Pareto or a non-Pareto strategy, such as aggregation [36].

2.1.1 Feature Subset Evaluation

In our evolutionary scenario a selected subset of descriptors is represented by an individual. As it was previously mentioned, in order to assess the relevance of the subset selection associated to an individual two objective functions were defined. The first objective function F_1 , calculates the number of selected descriptors associated to each individual. The second objective function F_2 , estimates the performance (i.e. accuracy) of a prediction method when a given set of descriptors is used. In particular, the function F_2 for the j^{th} individual is:

$$F_2(\mathcal{P}_{Z_1}, Z_2^j) = \frac{1}{n_2} \sum_{(\vec{x}_i, y_i) \in Z_2^j} (y_i - \mathcal{P}_{Z_1}(\vec{x}_i))^2 \quad (1)$$

This formula computes the mean square error of prediction (MSEP) applied to a set of molecules not used for training, where:

- Z is a matrix that represents the entire compound data set, where each row and column corresponds to a compound and a descriptor respectively. The last column of Z stores the experimental target values for each compound. This column vector is denoted as y .
- Z_1 and Z_2 are compound data sets, that are used here as training and validation sets respectively, with corresponding sizes $m_1 \times n$ and $m_2 \times n$. Also $Z_1 \cap Z_2 = \emptyset$ and $Z_1 \cup Z_2 = Z$.
- \mathcal{P}_{Z_1} is a predictor method trained with data set Z_1 .
- Superscript j , as in Z_2^j , is a filtered data set according to the descriptor selection encoded in the j^{th} individual. In other words, Z_2^j only contains those variables of Z_1 whose corresponding *loci* of the j^{th} individual's chromosome are set to '1' (see Section 2.1.2).
- \vec{x}_i is a vector that represents the values of the descriptors for the i^{th} compound of a given data set. In this way, $\mathcal{P}_{Z_1}(\vec{x}_i)$ is the predicted value of \vec{x}_i using \mathcal{P} trained with data set Z_1 .
- y_i is the target value for the i^{th} compound of a given data set.

Four different predictors are used here as \mathcal{P} , namely: decision trees (DT), k -nearest neighbors regression (kNN), a nonlinear regression function (NLR), and a multiple linear regression function (MLR). DT are used in this work as regression trees [31] without using any kind of pruning. The second method is kNN regression as used in reference [10]. A nonlinear regression model was also applied as one

of the predictors of the fitness function. It is conformed by a linear combination of nonlinear basis functions where their coefficients $\beta_{i,j}$ are adjusted with a nonlinear least-square fitting by the Gauss–Newton method (Eq. 2):

$$\sum_i^p \left(\sum_{j=1}^4 \beta_{i,j} x_i^j \right) + \beta_0 \quad (2)$$

Finally, an MLR method was also used to evaluate the predictive capacity of the molecular descriptors. The choice for these four predictors was based on their use in previous works on FS [10, 27, 37, 38]. A discarding of linear-correlated variables is applied prior to the application of the wrapper algorithm. This allows discarding linear redundancies among the descriptors, and hence making the wrapper's task easier.

We may see that when F_2 is calculated (Eq. 1) a set Z_1 is used for training the regression function \mathcal{P} , but a set Z_2 is used for assessing the predictive capacity of a subset. This data separation is applied so that the wrapper does not overfit the data used for training [39]. This data separation is applied once for each run of the evolutionary algorithm.

2.1.2 Feature Searching

We applied two different approaches for the Feature Searching: aggregation and Pareto. One important difference between these searching strategies resides in how the multiple objectives are managed. The first one aggregates all the objectives in a single fitness function, whereas the latter methods have as many fitness functions as the number of multiple objectives. Pareto-based methods optimize each objective separately according to the dominance ranking concept (we further describe this concept below).

As we said, aggregation approaches combine multiple objectives that result in a single fitness function. Despite their simplicity, these methods usually have good performance when tackling certain combinatorial optimization problems [40, 41]. Thereby, we propose the following aggregating formula:

$$F_{AG} = \alpha F_2 + (1 - \alpha) F_2 F_1 / p_m \quad (3)$$

Here, α is a weighting parameter for each objective ($0 \leq \alpha \leq 1$), and p_m is a parameter that represents an upper bound to the cardinality of a subset. The first term of the fitness function F_{AG} (Eq. 3) returns the prediction error obtained with a given \mathcal{P} and the second term reflects the ratio of selected descriptors to p_m scaled by F_2 . It is worth noting, that the aggregation strategy may be viewed as a regularization procedure that balances the complexity of the model (assessed by the number of descriptors used for the model) and the accuracy of prediction. Therefore, α controls the sparseness of the method.

The other approach for the Feature Searching uses the concept of domination. Dominance is a partial order that could be established among vectors defined over an R^k space, where k is the number of objectives to optimize. In our case, each individual is associated with a vector in R^2 , such that its first component is F_1 and the other one is F_2 . A nondominated set of an entire feasible search space is called Pareto-optimal set. The MO optimization algorithms that use the concept of domination for the selection mechanism to move a population toward the Pareto front are commonly named Pareto-based algorithms.

In this sense, Nondominated Sorting Genetic Algorithm-II (NSGA-II) [42] and Strength Pareto Evolutionary Algorithm 2 (SPEA2) [43] were proposed here as Pareto-based algorithms to be used within the wrapper.

The NSGA-II begins creating a random parent population P_0 of size S . The population is sorted out based on the nondomination concept. Each solution is assigned a rank equal to its nondomination level (1 if it belongs to the first front, 2 for the second front, and so on). After ranking the solutions, a population Q_0 of S offspring is created using binary tournament selection, recombination and mutation. The i -th generation follows three basic steps. First, a combined population $R_i = P_i \cup Q_i$ of size $2S$ is formed. Second, R_i is ordered according to its nondominance. Since all previous and current population members are included in R_i , elitism is ensured. Solutions belonging to the best front, i.e. $D_1(R_i)$, are the best solutions in the combined population R_i . Finally, if the size of $D_1(R_i)$ is smaller than S , all members of the set $D_1(R_i)$ are chosen for the new population P_{i+1} . The remaining members of the population P_{i+1} are chosen from subsequent nondominated fronts in the order of their ranking until no more sets can be accommodated. If $D_j(R_i)$ is the last front from which individuals can be accommodated in the population, but not all the members can enter in the population, then a decision needs to be made to choose a subset of individuals from D_j . In order to decide which members of this front will win a place in the new population, the NSGA-II uses a selection criterion based on a crowded-comparison operator that favors solutions located in less crowded regions. This crowded comparison is applied based on the objective space R^k .

The SPEA2 algorithm starts with an initial population P_0 , of size S and an empty external population \bar{P}_0 with a maximum capacity of \bar{S} . The i -th generation repeats four basic steps. First, the nondominated set of P_i , i.e. $D_1(P_i)$, is calculated and copied to \bar{P}_i (i.e., $\bar{P}_i = \bar{P}_i \cup D_1(P_i)$). Second, all dominated solutions of \bar{P}_i are removed. If the number of nondominated external solutions exceeds \bar{S} , \bar{P}_i is pruned selecting a *representative* by means of a clustering method. That is, the individuals are grouped in \bar{S} classes or clusters based on a crowded distance. Then, the individual with the lowest distance to the others within each cluster is selected as the *representative* of the cluster. In the third step, the fitness of each individual within $P_i \cup \bar{P}_i$ is calculated. Then, individuals from $P_i \cup \bar{P}_i$ are selected using binary tournament

selection until the mating pool is filled. Fourth, a population of S offspring, P_{i+1} , is created applying problem-specific recombination and mutation as usual.

The precedent steps of both algorithms are repeated until some termination criterion is reached. These Pareto-based evolutionary algorithms are two of the most prominent MO evolutionary algorithms used when comparing a newly designed MO algorithm [40].

We include here a summary of the implementation of the Feature Searching with a GA. Binary strings are used to represent the individuals. Each string of length n stands for a feasible descriptor selection, where n is the number of considered descriptors. A nonzero (zero) value in the i^{th} bit position means that the i^{th} descriptor is selected (not selected). The algorithm was developed using the MATLAB genetic algorithm library and the PISA (Platform and Programming Language Independent Interface for Search Algorithm) framework [44, 45].

The initial population is randomly generated with the number of nonzero bits initially set to a value between 0 and p_m . A one-point crossover is used for the recombination and a bit-flip is used for the mutation operation [46]. When as a result of an evolutionary operation the number of selected descriptors of an individual is greater than p_m , randomly selected *loci* are set to zero. In this way, the maximal cardinality of any individual is always bounded by p_m . This kind of domain constraint is commonly used in optimization problems, since it avoids wasting CPU cycles in solutions that will not be interesting for the problem [26, 47].

The selection scheme depends on the searching function applied for the MO wrapper. For the aggregation strategy, we performed different experiments with typical selection methods and we concluded that the tournament method is appropriate. Furthermore, this method is more preferred than others, because it is particularly easy to implement and its time complexity is $O(n)$ [46]. In the Pareto approach, selection operators correspond with the NSGA-II or the SPEA2 ones respectively. All Feature Searching functions include elitism, which protects the fittest individuals in any given generation, by moving them to the next generation. The replacement strategy works following the classical procedures used in evolutionary computing.

2.2 Second Phase: Validating Wrapper Results

After a combination of any Feature Searching and Subset Evaluation method is applied within the wrapper, a front of nondominated individuals is conformed from each independent run. It is worth noting that the conformation of this front of nondominated solutions is carried out regardless whether an aggregation or a Pareto-based strategy was used for the Feature Searching. In this way, all nondominated subsets are treated as the most 'interesting' set of selections obtained in that run by the wrapper. This is indicated in Figure 1 as the 'Coarse Selection of Subsets'.

Each subset of descriptors contained in the front is used for being assessed by a validation method. For this work, an Artificial Neural Network Ensemble (ANNE) was chosen as the validation method, since ANNEs are methods that were widely and successfully applied in the QSAR literature [48]. Thus, it is expected that the accuracy obtained using ANNEs is greater than the accuracy obtained using the regression methods applied within the wrapper. In addition, using a different method for validation in FS is important, given that selected descriptors may be optimal for a specific method but the same can not be guaranteed when using other methods [5]. Thereby, the ANNE performs the role of an independent assessment of the predictive capacity of the selected descriptors. Moreover, an MLR is also applied for this second stage in order to use it for comparison reasons.

In addition to the potential of ANNEs, this second stage is a more rigorous assessment of predictive capacity, because a same subset of descriptors is evaluated many times. This validation involves performing many trials of an f -fold cross validation, where in each trial each fold is randomly obtained. It is worth noting that the application by the wrapper of a large number of replications on a same subset would not be feasible, due to complexity reasons. Thereby, this replication is only possible and profitable for the second phase.

We used an ANNE consisting of 5 artificial neural networks (ANN), given that ensembles improve stability of the predictions [5, 49]. The ANN's architecture depends on the size of the data set and its complexity. Every ANN used for this work was trained using the Levenberg-Marquardt algorithm with a Bayesian regularization procedure [50] which is a learning procedure that was also applied in other QSAR proposals [21, 51]. Bayesian regularized neural networks tend to perform quite well and they do not need the ensemble approach to give good models. Nevertheless, yet at the expense of additional computing time, ensembles yield slight improvements in stability and prediction capacity of models.

3 Data

The three data sets used for our analysis were compiled from scientific publications which provided descriptors that were selected by FS methods. These data sets are attractive, since they have different ratios of the number of descriptors to the number of compounds. Small data sets were avoided, since their validation generates controversy [21, 39].

Data Set 1 (DS1): This data set was used by Konovalov et al. [25, 52], and named therein as 'KS289-log BB '. The target variable is log BB which is a common measure of the blood brain barrier (BBB) penetration. DS1 is composed of 289 compounds and 1501 descriptors, plus a de-

scriptor I_v that distinguishes whether log BB was calculated from an in vivo or an in vitro assay.

Data Set 2 (DS2): The compounds and descriptors of DS2 were also extracted from reference [25] and named therein as 'KS172-HIA', where the target variable in this data set is logHIA. This target variable is a nonlinear transformation of the intestinal absorption expressed as fraction absorbed (%HIA), i.e. percentage of dose appearing in the portal vein. This data set has 127 compounds and 1499 molecular descriptors. DS2 contains chemical entities with different degrees of %HIA, but in the work of Konovalov et al. [25], those with 0 and 100 %HIA were removed.

Data Set 3 (DS3): This set of compounds was extracted from reference [53] and the target variable is the logarithm of octanol/water partition coefficient (log P) at 25 °C. This data set has 442 organic compounds that belong to different chemical classes. In contrast to the other data sets, far fewer descriptors are here considered. The original article reports 12 descriptors that were selected by a FS method. We aimed at recreating their original condition, so we additionally incorporated 61 simple descriptors using Dragon [54].

4 Results

4.1 Design of the Experiments

We applied our proposed method to the three data sets mentioned in Section 3. The parameters of the evolutionary algorithm were fixed for all the data sets and Feature Searching functions. The population size was fixed to 145 individuals, the crossover probability was set to 0.75 and the mutation probability was established as $2/n$. As is recommended in the literature [55], a phenotypic criterion was selected for the stopping of the evolutionary algorithm: it stops when the improvement during 15 generations of the average fitness of the population is less than a given tolerance value ($\xi = 10^{-16}$). Additionally, the maximum number of generations was set to 200 generations. In particular, for the aggregation strategy the tournament size was set to 4 and the number of elite individuals per generation was set to 5.

We proposed 12 different MO wrapper methods, which come from the combination of the different Feature Searching (the aggregation strategy, NSGA-II, SPEA2) and the Feature Subset Evaluation functions (DT, kNN, NLR, MLR). In the first phase, we performed 10 runs of the MO wrapper for each one of these feasible combinations. In each run we only retained the solutions that belonged to the front of nondominated subsets. In the second phase of our method, we took each descriptor subset of the nondominated fronts, and we applied an ANNE using an f -fold cross validation that was performed 50 times and then the predictions were averaged. Tables with the com-

prehensive information and results obtained in every experiment may be accessed from the Supporting Information files.

In relation to the architecture of the ANNs, we used three-layer networks where the number of hidden nodes was first optimized according to the best descriptor subsets reported in the papers out of which we extracted the data sets [25, 53]. The number of hidden nodes was kept constant while testing all other subsets of descriptors of a same data set. Generally, a fixed internal network architecture might be considered as a design flaw, but rather than providing perfectly optimized regression models it increases comparability and saves computing time while providing insights into effects of descriptor selection. As a result, for DS1 and DS2 we used two hidden nodes, whereas for DS3 we used five hidden nodes. Also network inputs and targets were normalized so that the Bayesian regularization is correctly applied.

We compared our models to the models obtained from the descriptors reported in the papers out of which we extracted the data sets [25, 53]. Besides, we also compared our subsets of descriptors to the descriptors obtained from a Bayesian feature selection method which applies a regression method using a sparse prior (Jeffreys' prior in this case) [56]. This approach was also applied for selection of descriptor subsets in a recent descriptor selection article [21].

Since that in the work of Konovalov et al. [25] they reported predictive capacity in terms of an MLR method, it could be argued that a comparison between a neural network and an MLR method is not fair. However, it is worth highlighting that the ANNE models can be used here, since a preselection of subsets using 'fast' methods is first applied, and only a reduced number of potential relevant subsets are left to the ANNE. Anyway, we also provide the results obtained using an MLR for the second phase, and we trained ANNEs for the subsets of Konovalov et al. [25].

Regarding to the comparison with Figueredo's method [56], we applied an analogous procedure as the one applied for our evolutionary work. We used a validation data set (equivalent to Z_2) in order to determine the parameter σ , which controls the sparseness of the method. We performed 20 runs of this algorithm, where a new training-validation splitting was applied before each run. Data were standardized before applying the selection procedure and no transformation was applied to the design matrix.

For DS1 a special consideration was taken, that guarantees that descriptor 'Iv' be always considered for any model. This decision is based on the claim of Konovalov et al. [25] that a systematic difference of about 0.5 log BB units difference exists between the in vivo and the in vitro experimental values. On the basis of the selected subsets reported in this work, we set $p_m = 20$. In accordance with the work of Konovalov et al., the size of Z_1 was equal to the size of Z_2 for the MO wrapper; and the number of folds

for the Monte Carlo cross-validation of the second phase was set to 2 (50% for training, 50% for testing). The same considerations and parameters were kept for DS2.

For the last data set, the estimated number of necessary descriptors is expected to be higher than in both previous cases, so we set $p_m = 50$. Yaffe et al. [53] used a fixed hold-out set with 17% of the data. It can be showed, that this hold-out data are completely contained within the convex hull of the training/validation data, which does not allow an unbiased evaluation of the generalization capacity of the model. So, we considered that, given the size of the data set, a 5-fold Monte Carlo cross validation is more appropriate for evaluating the subsets in the second phase. In the same way for the wrapper, Z_1 and Z_2 comprise 80% and 20% of the data respectively.

4.2 Analysis of the Best Selected Subsets of Descriptors

Tables 1–3 comprise the information of the best selected subsets, the associated methods and the errors obtained in their validations for each data set. In each table, we show the best subset reported in the referenced articles [25, 53] (first row of Tables 1–3), the best subset using Figueredo's method (row 2) and the best subsets obtained with our method, showing two aggregation (rows 3 and 4) and two Pareto-based (rows 5 and 6) selections. Table 4 enumerates all the descriptors selected for Tables 1–3.

For DS1 (Table 1), we found that Subset III has better prediction capacity than the one reported in Konovalov et al. (Subset I) using the same number of descriptors. The best subset that was obtained using Figueredo's method (Subset II) is slightly better than Subset I when MLR is used for validation, but the number of descriptors is greater than any considered subset. Using more descriptors than in Subset I we find that Subsets IV, V and VI allow a better prediction capacity regardless whether they are predicted with an ANNE or with an MLR validation method.

For the second data set (Table 2), even though no subset was found with a strictly better prediction capacity than the proposed in Konovalov (Subset VII), very interesting subsets were found. Subsets IX, X and XII have a slightly worse prediction capacity than Subset VII, but they have much fewer descriptors. Subset XI also has a comparable prediction capacity and uses one less descriptor than Subset VII. When MLR was used to calculate the predictive capacity of the preceding subsets, all predictions were slightly worse compared with the MSEP using the ANNEs. The subset obtained using Figueredo's method (Subset VIII) was outperformed by the other presented subsets for this data set.

Unlike Konovalov et al. [25], we did not preselect any descriptor, except for Iv in DS1, which necessarily must be always taken into account.

DS3 results are motivating since the complexity of the model for the hydrophobicity prediction in terms of the selected descriptors is higher than the complexity of the

Table 1. Comparison results for selected subsets of DS1. MSEP and q^2 columns refer to results obtained on the hold-out fold. N_w/N_{eff} is the number of weights in the neural network and the number of effective weights in the model. Subset I corresponds to the best subset reported in reference [25], whilst Subset II was obtained using Figueredo's method [56].

Subset	Feature searching	Regression function	Cardinality	Validation method	MSEP	q^2	N_w/N_{eff}
I	MCVS [a]	MLR	6	ANNE*	0.1265	0.645	17/13
				MLR	0.1225	0.6752	–
II	Jeffreys' prior	MLR	20	ANNE	0.1302	0.6528	45/34
				MLR	0.121	0.6757	–
III	MO-Aggreg, $\alpha=0.3$	MLR	6	ANNE	0.1205	0.6816	17/15
				MLR	0.1281	0.6525	–
IV	MO-Aggreg, $\alpha=0.7$	MLR	15	ANNE	0.1103	0.7198	35/31
				MLR	0.1113	0.703	–
V	NSGA-II	MLR	8	ANNE	0.1140	0.6993	21/18
				MLR	0.1178	0.6727	–
VI	NSGA-II	MLR	11	ANNE	0.1052	0.7352	27/23
				MLR	0.1124	0.6821	–

[a] Monte Carlo Variable Selection.

* These ANNE results were not reported in the original article [25] but they were calculated for this work to allow a fairer comparison.

Table 2. Comparison results for selected subsets of DS2. MSEP and q^2 columns refer to results obtained on the hold-out fold. N_w/N_{eff} is the number of weights in the neural network and the number of effective weights in the model. Subset VII corresponds to the best subset reported in reference [25], whilst Subset VIII was obtained using Figueredo's method [56].

Subset	Feature searching	Regression function	Cardinality	Validation method	MSEP	q^2	N_w/N_{eff}
VII	MCVS [a]	MLR	8	ANNE*	0.1191	0.6813	21/17
				MLR	0.09	0.7532	–
VIII	Jeffreys' prior	MLR	4	ANNE	0.1715	0.5733	13/7
				MLR	0.1380	0.6404	–
IX	MO-Aggreg, $\alpha=0.1$	MLR	3	ANNE	0.0984	0.7421	11/9
				MLR	0.1282	0.65	–
X	MO-Aggreg, $\alpha=0.3$	DT	3	ANNE	0.1055	0.7092	11/9
				MLR	0.1512	0.57	–
XI	NSGA-II	MLR	7	ANNE	0.0915	0.6459	19/16
				MLR	0.1112	0.6623	–
XII	NSGA-II	kNN	2	ANNE	0.1013	0.6174	9/7
				MLR	0.1374	0.6186	–

[a] Monte Carlo Variable Selection.

* These ANNE results were not reported in the original article [25] but they were calculated for this work to allow a fairer comparison.

Table 3. Comparison results for selected subsets of DS3. MSEP and q^2 columns refer to results obtained on the hold-out fold. N_w/N_{eff} is the number of weights in the neural network and the number of effective weights in the model. Subset XIII corresponds to the best subset reported in reference [53], whilst Subset XIV was obtained using Figueredo's method [56].

Subset	Feature searching	Regression function	Cardinality	Validation method	MSEP	q^2	N_w/N_{eff}
XIII	GA [a]	[b]	12	ANNE*	0.247	0.884	71/61
				ANNE [c]	0.29	–	–
XIV	Jeffreys' prior	MLR	16	ANNE	0.2052	0.9097	91/80
				MLR	0.2724	0.8804	–
XV	MO-Aggreg, $\alpha=0.9$	MLR	24	ANNE	0.154	0.9297	131/95
				MLR	0.286	0.8795	–
XVI	MO-Aggreg, $\alpha=0.1$	MLR	13	ANNE	0.164	0.9317	76/64
				MLR	0.2617	0.8698	–
XVII	SPEA2	MLR	15	ANNE	0.1778	0.9135	86/74
				MLR	0.299	0.8649	–
XVIII	NSGA-II	MLR	20	ANNE	0.1696	0.9240	111/94
				MLR	0.3426	0.8496	–

[a] Genetic Algorithm.

[b] Regression function was not reported in the original work.

[c] Result reported in the original work using a fixed hold-out test set.

* These ANNE results were not reported in the original article [52] but they were calculated for this work to allow a fairer comparison.

models inferred from the previous data sets. However, comparisons with the original work are cumbersome to be established, given that different prediction and validation methods are involved for each work. Also, we incorporate descriptors not considered in the original work, so the subsets obtained for DS3 may not be directly compared with the results of the work of Yaffe et al. [53]. However, to quantify the improvement obtained with our new subset of descriptors, we took the descriptors chosen in the work of Yaffe et al. [53] (Subset XIII) and we applied the same validation and the same prediction method that was applied to our subsets.

Therefore, we may observe that any of the proposed subsets of descriptors (Subsets XV, XVI, XVII and XVIII) enhances the predictive capacity of the Subset XIII (Table 3). Even though our subset increased the number of descriptors the obtained difference in MSEP is important. The ANNE prediction capacity of the subset obtained using Figueiredo's method (Subset XIV) is better than the one of Yaffe's method, but worse than the subsets using our evolutionary approach.

For this data set, we may observe that the MSEP using MLR for the second phase are rather higher than the MSEP when using ANNEs. Thus, it may be inferred that the relationship among the selected descriptors and $\log P$ is highly nonlinear.

4.3 Analysis and Comparison of the Different MO Wrappers

In this subsection we will compare the performance of the subsets obtained after the second phase, in terms of the different combinations of the MO wrapper. Table 5 gives an idea about which regression function and searching strategy of the wrapper performs better in finding relevant subsets for each data set. In order to eliminate the effect of subsets with low predictive capacity, the 50th percentile of the data is considered for this table. We applied multiple comparison tests in order to statistically compare the average values of each combination of the wrapper. We used the Tukey-Kramer test with an experiment-wise error rate of 5%. When analyzing DS1 and DS2 we found that the MLR method was better than the remaining methods, and that in most cases Pareto-based strategies were better than the aggregation strategy. Also, no differences were found between NSGA-II and SPEA2 for DS1.

For DS3, when we analyzed the searching strategies for each regression method, we found that the aggregation strategy is the best searching strategy regardless of the applied regression method. Moreover, when we examined the regression methods for each searching strategy, we found that MLR is the best regression method for NSGA-II; whereas MLR and kNN were significantly better than the remaining methods – though no significant differences were found between them – for the aggregation strategy and SPEA2.

Taking this analysis a bit further, we applied two-way ANOVA tests so that focusing on variance contributions we may determine whether the regression methods or the searching strategy have a stronger impact on its prediction capacity. One factor was represented by the different searching strategies and the other one was represented by the different Feature Subset Evaluators. For the first two data sets we found that the selection of the regression method has the strongest impact (variance contribution to predictive capacity is 95.8% and 56.1% for DS1 and DS2, respectively). In the case of DS3, the selection of the searching strategy is the most important for the prediction capacity of the MO wrapper (variance contribution is 55.3%).

In view of the fact that Pareto-based methods performs better only for the first two data sets, it is quite likely that the cause of this outcome is due to the fact that the first two data sets require fewer descriptors than the $\log P$ data set. This is so because Pareto-based strategies look for minimums in any objective, irrespective of the value of the other objective. In other words, a subset that has fewer descriptors than the remaining individuals of the population will be in the nondominated front, even when its prediction capacity is very low. This feature turns these methods more prone to find subsets with lower cardinality. To exemplify this claim, the average cardinality of all the subsets obtained using NSGA-II for DS3 is 12.71, whereas for the same set using an aggregation approach and $\alpha=0.1$ is 19.98 (data not shown). In this sense, when the 'theoretical' optimal number of descriptors is high, Pareto-based methods have to face up with a broader feasible space of optimal selections in relation to the space in an aggregation strategy [40].

Considering the results about the good performance of MLR within the wrapper, this result was especially expected for DS1 and DS2, where the relationship among descriptors and the target properties are quite linear. For the $\log P$ data set, the good performance of the MLR is interesting, since it shows that MLR could be a very good method for identifying predictive capacity of the descriptors even when the relationship with the target variable is nonlinear [27].

Finally, we want to illustrate how α influences cardinality and prediction capacity when using the aggregation strategy. In Figure 2, we may appreciate that there is a tradeoff between these two objectives, but since we are more interested in prediction capacity, the choice of α should be based on this latter objective (only shown for the BBB data set). This figure allows to identify the range of values of α where most interesting subsets of descriptors may be found. Also, it may be appreciated that subsets obtained with $\alpha=1.0$ (mono-objective wrapper) have always lower predictive capacity than when using another α , such that $0 < \alpha < 1.0$. This leads us to state that models conformed by subsets with lower cardinality are more prone

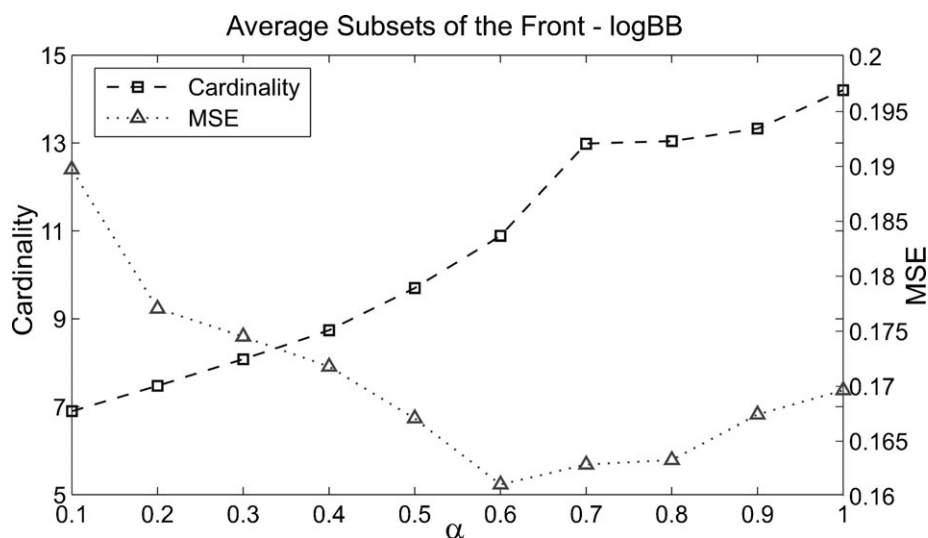


Figure 2. Average cardinality and MSE of prediction in terms of α for DS1. Results for the first iteration of the validation phase are considered.

Table 4. Selected molecular descriptors for each subset of Tables 1–3.

Subset	Descriptors [a] [b]
I	[Iv], [TPSA(NO)], [Ic], SRW09, BELv4, HATS7v / HATS8e
II	[Iv], TPSA(NO), MLOGP2, SRW09, nROH, EEig05d, C-034, nRCOOH, O-057, nArNR2, nArCOOH, H-051, Psychotic-80, nRCOOR, HATS8u, nN(CO)2, Infective-50, HATS7u, nArCOOR, Deppresant-50
III	[Iv], TPSA(NO), ALOGP, Mor21v, EEig12r, Ic
IV	[Iv], TPSA(NO), SRW07, O-057, MATS2p, nOHp, R3u, RDF020p, T(N..I), nArOH, Psychotic-80, RDF050m, HATS8u, Cl-087, G3u
V	[Iv], TPSA(NO), ALOGP, Mor16v, ISIZ, nN(CO)2, BELm4, Ic
VI	[Iv], TPSA(NO), ALOGP, R7u+, ARR, H4p, Mor20v, TE2, BELe3
VII	[ALOGP], LAI, Neoplastic-80, RDF045m, R5v+, DDI, N-074, IDE
VIII	Hy, GVWAI-80, Infective-80, nP(=O)O2R
IX	MLOGP, TPSA(NO), RDF130m
X	ALOGP, C-011, nPyridines
XI	ALOGP, Mor13e, RDF045p, Vs, nArCONHR, R5u+, PW5
XII	ALOGP, R1v+
XIII	MW, D_P, D_H, D_S, E2, EX, ELC, IP, PO, VMC1, VMC2, VMC4
XIV	D_H, IP, PO, Mv, nH, nN, nCL, nR06, nCp, nCaR, nOHp, nOHs, nROR, nRSR, nHDon, PSA
XV	D_S, E2, IP, Sv, Se, nAT, nBM, nDB, nAB, nH, nC, nO, nF, nCL, nI, nR03, nR06, nR11, nCp, nCs, nOHp, nOHs, nOHt, ARR
XVI	D_H, IP, PO, nBT, nBM, nAB, nO, nF, nX, nR06, nCaR, nHAcc, Ui
XVII	MW, E2, AMW, Mv, Mp, nBT, SCBO, nAB, nC, nBR, nCp, nCaR, nOHs, nHDon, ARR
XVIII	MW, E2, PO, VMC2, AMW, Mv, nBM, SCBO, nDB, nH, nC, nBR, nR03, nR06, nCp, nCaR, nOHp, nOHs, nHDon, ARR

[a] Complete names of the descriptors may be found in the Supporting Information files and in the E-Dragon web site [54].

[b] Brackets denote preselected descriptors.

to turn into more general models, and hence with higher predictive capacity.

4.4 Analysis of the Obtained Subsets of Descriptors

One should note here that according to the parallel and stochastic nature of the proposal, the final subsets of descriptors obtained from the second phase are different and not necessarily the same among different trials. Horvath

et al. [26] emphasized that this aspect of stochastic methods is not a limitation, but it represents the possibility to offer more than one relevant subset for prediction using a QSAR model.

Nonetheless, it is noticeable that some descriptors are repeatedly chosen and some of them are either theoretically known as relevant or they have also been selected in other FS works. As a brief analysis of the selected descriptors (Table 4), we may find that the *TPSA(NO)* (topologi-

Table 5. Performance of the regression functions of the wrapper: average MSEP of the 50th percentile.

Data Set	Feature Searching	DT	kNN	NLR	MLR
DS1	Aggregation [a]	0.1504	0.1486	0.1462	0.1261
	NSGA-II	0.1437	0.1382	0.1454	0.1277
	SPEA2	0.1385	0.1361	0.1368	0.1269
DS2	Aggregation [b]	0.1211	0.1285	0.1212	0.1161
	NSGA-II	0.1049	0.1052	0.105	0.101
	SPEA2	0.1018	0.1104	0.1064	0.0982
DS3	Aggregation [c]	0.1881	0.1855	0.1877	0.1787
	NSGA-II	0.2645	0.2592	0.3080	0.2222
	SPEA2	0.2120	0.2067	0.2073	0.1963

[a] Using $\alpha = 0.7$.

[b] Using $\alpha = 0.3$.

[c] Using $\alpha = 0.9$.

cal polar surface area using nitrogen and oxygen polar contributions) descriptor is highly present in most subsets for DS1 (see Supporting files for the entire list of selected descriptors). This fact was to be hoped, since the well-known importance of polar surface area in the prediction of BBB penetration [57]. Generally, lipophilic compounds are likely to cross the BBB, so it is also expectable that hydrophobicity descriptors (like *ALOGP*) and carboxylic acid groups descriptors (*nRCOOH* or *Ic*) are frequently present.

Similarly, it is not surprising to find that again descriptors related with water solubility such as *ALOGP* (Ghose-Crippen octanol-water partition coefficient) or *MLOGP* (Moriguchi octanol-water partition coefficient) are often selected in the subsets when predicting HIA. Also in the case of the DS3 subsets, we may detect that descriptors known to be related with the $\log P$ property are repeatedly chosen, e.g. *MW* (molecular weight), carbon-related descriptors (*nC* – number of carbon atoms; *nCar* – sum of all the carbons belonging to any aromatic and heteroaromatic structure; *nCs* – number of total secondary carbons), descriptors related with dipole moments (*D_H* – total dipole, hybridization; or *D_S* – total dipole, hybridization + point charge).

4.5 Assessing Probabilities of Chance Correlations

Equally relevant to the descriptor selection issue is the question of whether chance correlations are likely to occur when applying our methodology to the preceding data sets. There are articles [58, 59] that emphasize that chance correlations are inversely correlated with the object-variable ratio (number of compounds to the number of selected descriptors). According to this statement, we may affirm that chance correlations are not likely to occur in this work, since firstly, we used data sets with more than a hundred of compounds, and secondly, our MO FS method aims at minimizing the number of selected features.

Nonetheless, we have applied y -randomization experiments [58] in order to analyze the risk of chance correlations for our FS methodology. We ran the experiments with DS1 and DS2, since they have a larger pool of descriptors compared to the pool of descriptors available for DS3. Also, they have a smaller number of compounds, thus they are more prone to obtain chance correlations. We executed 10 runs with a different y -randomization, and in each run we applied the descriptor selection task in the same way as we did for our MO two-phase methodology. Then, we averaged the best results of each run using ANNE and MLR for the second phase. Using ANNE, MSEP and q^2 for DS1 were 0.3652, and 0.00685 respectively, whereas for DS2 they were 0.3713 and 0.018261. In terms of prediction capacity, the results using ANNE are particularly bad, since Bayesian neural networks tend to predict roughly a constant value when there is no clear relationship among descriptors and target property. On the other hand, using MLR for the second phase, MSEP and q^2 were 0.2352, and 0.3893 respectively, whereas for DS2 they were 0.2424 and 0.4579. All y -randomization results were considerably worse than the results obtained when the original target was used. These experiments highlight the fact that the results obtained in Sections 4.2 and 4.3 are unlikely of being obtained purely by chance.

4.6 Time-Complexity Analysis

As a coarse analysis of the time-complexity of our methodology, the first aspect to point out is that it is bounded by the time-complexity of the MO wrapper (first phase), since it corresponds to the part with the highest computational burden. Particularly, Pareto-based algorithms are more compute-intensive than the aggregation ones, since the first ones also need to do some additional tasks: manage an additional population of individuals, sort the individuals according to the nondomination criterion and calculate crowding distances among the individuals.

From reference [36], we may obtain that the worst case time-complexity for NSGA-II, is $O(kS^2)$ where k is the number of objectives to be optimized (2 for our methodology) and S is the population size (145 for our methodology). The worst case time-complexity for SPEA2 is $O((S + \bar{S})^3)$, where \bar{S} is the population size of the external population (also 145). Deb [36] suggests that this bound is rather pessimist and the average case time-complexity $O((S + \bar{S})^2 \log(S + \bar{S}))$ is much more realistic. These execution orders are defined for a single generation assuming no complexity cost for the calculation of the fitness function.

The time-complexity of the fitness functions are reduced to the time-complexity of the regression methods of the Feature Subset Evaluation. From empirical observations, the NLR is the most compute-intensive, and it has an order $O(c^3)$ where c is the number of coefficients to be adjusted ($4p + 1$ in our case, where p is the number of descriptors in the evaluated subset).

In order to state the worst case complexity of any specific combination of a searching strategy and a regression method for the MO wrapper, we define $O(\textit{Searching_Gen})$ as the order of execution of a generation in a searching strategy. Similarly, we define $O(F_2)$ as the order of execution of an objective function. Since the fitness function is computed at the beginning of a generation, the time-complexity of a single generation is $O(\max(S \cdot O(F_2), O(\textit{Searching_Gen})))$. The overall time of a MO wrapper run, will be the order of a single run multiplied by the number of generations, which is cumbersome to define in advance, since it will depend on the convergence/stopping criteria.

4.7 Analysis of our FS Methodology Using External Validation

All the results presented so far have been obtained from the prediction errors of the ANNE in the second phase. The main goal of this phase is to provide a stronger regression method in order to assess the subsets of descriptors selected by the first phase, but it is not intended to be a validation methodology that estimates the real predictive capacity of the QSAR models obtained from these descriptor subsets. We may observe that this second phase internally validates using data that was previously used for the selection of features in the first phase. This fact might be considered as a design flaw that introduces an overoptimistic estimation of the true prediction capacity of the subsets. However, we shall show that, even when this second phase is not a strict validation procedure, it is reliable enough to assess the predictive capacity of the selected subsets of descriptors.

In this sense, we applied an external validation procedure, in order to quantify how different the prediction errors in comparison with our internal validation are. We randomly segregated a data set Z_3 (with 20% of the data) previous to the first phase. Then, we applied the conventional procedure of the first and second phase to the remaining data, where we obtained the 20 most promising subsets of descriptors. The prediction capacity of these subsets was evaluated as it is mentioned in Section 2.2 and we also assessed the predictive performance of these 20 subsets of descriptors on the Z_3 data set.

We want to analyze for the top 20 subsets of descriptors whether the average errors obtained on Z_3 are not significantly worse than those obtained by our internal validation procedure. The standard practice to statistically compare the means of both validation procedures is carried out by using the *t*-student comparison test for correlated observations. In this way, we may establish confidence intervals for the averages of the differences between the average errors of both procedures.

Three different situations may occur:

1. the external validation error is significantly lower than the internal validation error,

2. there is no statistical evidence that the external validation errors are higher than the internal validation errors,
3. anything else different from Situation 1 or 2.

It is clear that Situation 1 may be easily identified (with $\alpha \leq 0.05$ or $\alpha \leq 0.01$). However, it is not possible to probabilistically quantify the second situation (type-II error). The usual procedure would be to accept the hypothesis that the external validation is not higher, when this hypothesis may not be rejected with a probability $\alpha \leq 0.025$, or when the average internal validation error is higher than average error on Z_3 and Situation 1 does not hold. Anything else will be assumed that the external validation error is higher than the internal validation error.

We applied this procedure to the three data sets. In order to take into account the effect of different Z_3 data separations, we replicated this same experimentation 10 times, where we used, without loss of generality, MLR and the aggregation strategy for the first phase.

We have obtained that in only 3 out of the 30 runs the Situation 3 holds (1 for DS1 and 2 times for DS3), whereas in 11 out of the 30 runs Situation 1 holds and in the remaining 16 runs Situation 2 holds. These results lead us to believe that the validation procedure of the second phase is a good estimator of the prediction capacity of the subsets. The main reason of these results arises from the fact that, when our internal validation procedure is applied, only one part of the remaining data is used for training (Z_1) in each iteration of the cross-validation. On the other hand, when external validation is applied, all the remaining data (80%) is used for training the model. There is nothing wrong in using all the remaining data for training with Bayesian neural networks, since it has been shown that these models do not tend to overfit the data used for training [51], and hence they do not need a validation set.

Even though external validation is the gold standard for assessing prediction capacity for a QSAR model, an external validation procedure was not used in this work for the presentation of the results, since it is dependent on the training-testing data splitting. To illustrate this point we applied a random effect ANOVA, where each one of the ten data separations is considered as a random factor that has the predictive capacity of the top 20 subsets of descriptors as their observations. Our goal is to determine how important the variance due to the data separation is. These ANOVA experiments (table not shown) reveal that with a probability less than 0.5×10^{-5} there is a source of variance due to the data separation in the three cases, where this factor represents 34.5%, 17.85% and 60.08% of the total variance, for DS1, DS2 and DS3 respectively. This variance effect due to random training-testing separation, require to perform a large number of trials in order to get rid of the variance in results. However, a large number of trials for each combination of the MO wrapper would make this work computationally unfeasible. On the other

hand, our internal validation is computationally cheaper and, at the same time seems to be comparable with the results of the external validation.

5 Discussion

A novel approach for addressing the selection of descriptors in QSAR/QSPR methods is presented. The main contributions of this paper reside in two main aspects. First, our method uses a multi-objective strategy within a wrapper method, and according to the authors' knowledge, a multi-objective descriptor selection was not previously applied in the QSAR/QSPR literature. Second, our method proposes a two-phase FS methodology that attempts to combine a wide searching of descriptors in the first phase with accurate methods, like ANNs, for assessing descriptor relevance. This rigorous assessment is only applied to the preselected subsets obtained by the wrapper. The second phase of the FS method, apart from improving accuracy, allows independency among the regression methods of the wrapper and thus the final subset selection is not biased in terms of the regression algorithm applied in the wrapper.

It would be unfair not to mention that our evolutionary methodology is compute-intensive compared with other FS methods [8, 22]. However, we argue that for feature selection methods CPU-time is not a crucial issue provided that the methodology may be executed in a reasonable polynomial time. One should not forget that feature selection is not aimed to be applied in real time nor in numerous opportunities.

Taking into account that our FS method is a product of different statistical or machine learning methods, our proposal agrees with the machine learning literature, which states that better results are obtained when combining different machine learning approaches than when using a single model [49, 60]. In addition, other works [5, 59] also emphasize the benefits of using ensembles and combination of methods in a FS procedure.

A subset of descriptors is considered as relevant here, depending on its predictive capacity determined by the MSE_P obtained by performing cross-validation several times. Utilization of MSE_P for quantifying error prediction with a cross-validation procedure is in accordance with the literature for assessing prediction capacity when using moderate-size data sets [25, 61, 62].

Authors are aware of the criticism and limitations associated to assessing relevance of descriptors by only relying on statistical validations obtained from a machine learning method [63]. We emphasize that an FS method does not pretend to give a definite solution to the problem of inferring which the best subset of molecular features that controls the variation of a biological activity or property is. However, in the absence of theoretical procedures, we argue that machine learning FS methods allow to circumvent the problems that emerge from not knowing the rules that

govern a given activity or property. Moreover, these methods should represent a tool for scientists, who in turn may contribute with their knowledge to make a final decision.

The multi-objective approach presented in this work is prone to obtain subsets with minimal cardinality. Thus, this favors human interpretability of the results and also it diminishes the number of learning hypothesis associated to a subset of descriptors [19].

We also emphasize that this proposal is not aimed at creating a ranking of descriptors according to their relevance. A descriptor is not relevant by itself, but its importance should be quantified by considering all the descriptors of a given subset as a whole. In this work, more than a single subset of descriptors is proposed for prediction; so, the final decision would be taken by those who will develop the QSAR model, who may also consider other objectives such as interpretability.

It is worth pointing out, that the selection of features for a prediction method is not a classical multi-objective optimization problem. The best subsets of descriptors (or the subsets contained in the Pareto front) obtained by the wrapper are not necessarily optimal selections but, as it was previously showed by the ANNE results, they provide subsets of variables that are expected to be relevant and nonredundant. We may expect them to be highly relevant, given that they have a good prediction capacity as verified by the validation with the ANNEs. In the same way, and having in mind that one of the objectives of the wrapper is to minimize cardinality, we may expect that they have low redundancy, given that if a subset had one or more redundant features, that subset would have a low probability of survival during all generations.

Even though statistical comparisons among results of the methods are not completely established, the results obtained for DS1 and DS2 show that better or comparable results than the Monte Carlo Variable Selection (MCVS) of Konovalov et al. were obtained with our proposal. Focusing on the hydrophobicity data set (DS3), this one has an additional difficulty, since the relationship among $\log P$ and the considered molecular descriptors is known to be nonlinear. Results with the $\log P$ data set show that the method presented here is appropriate regardless of the linear or nonlinear relationship among descriptors and target variable. A comparison with a competent deterministic method [56] was also provided and results show that our descriptor selection method performs better than the deterministic one, at least for the data sets under study.

Another contribution of this work lies in the comparison among different Feature Searching and Feature Subset Evaluation methods. Considering the results with the Pareto-based and aggregation strategies, we think that both Feature Searching strategies are an advisable way of doing FS, at least when using the proposed algorithms.

The parameter p_m , and also α when using an aggregation approach, have to be established manually. The first one is based on the theoretical knowledge of an upper bound to

the number of descriptors expected to be necessary. The second should be tuned according to the desired balance between the cardinality and the predictive capacity obtained for a given α value. Although setting manual parameters may result awkward, *No Free Lunch Theorems* indicate that there is no “best” FS method for any data set [64]. So, incorporating knowledge and restrictions is a common practice to make a method tailored and efficient to a specific task. In contrast, Pareto-based strategies have the advantage of not having to set a parameter that weighs both objectives. However, these Pareto-based methods are not the best choice when the number of necessary descriptors is expected to be large (Sec. 4.3).

In conclusion, we think that a two-phase strategy is advisable for feature selection in QSAR/QSPR, in order to apply light weight methods as a preselection of subsets and afterwards a stronger method as an ANNE to assess the predictive capacity of the subsets.

6 Acknowledgements

Authors thank Dr. Eladio Pardillo-Fontdevila for his valuable comments for this work. Anonymous reviewers are gratefully acknowledged for their suggestions and critical comments that led to improving the paper. Also we thank the SeCyT (UNS) for Grants PGI 24/ZN15 and PGI 24/ZN16.

7 References

- [1] I. V. Tetko, P. Bruneau, H.-W. Mewes, D. Rohrer, G. Poda, *Drug Discov. Today* **2006**, *11*, 700–707.
- [2] T. Hou, J. Wang, *Expert. Opin. Drug Metab. Toxicol.* **2008**, *4*, 759–770.
- [3] J. Gola, O. Obrezanova, E. Champness, M. Segall, *QSAR Comb. Sci.* **2006**, *25*, 1172–1180.
- [4] Y. Liu, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1823–1828.
- [5] D. Dutta, R. Guha, D. Wild, T. Chen, *J. Chem. Inf. Model* **2007**, *47*, 989–997.
- [6] O. Nicolotti, A. Carotti, *J. Chem. Inf. Model* **2006**, *46*, 264–276.
- [7] A. L. Blum, P. Langley, *Artif. Intell.* **1997**, *97*, 245–271.
- [8] I. Guyon, A. Elisseeff, *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
- [9] I. Guyon, A. Elisseeff, in *Feature Extraction: Foundations and Applications* (Eds: I. Guyon, S. Gunn, M. Nikravesh, L. A. Zadeh), Springer, Heidelberg **2006**, pp. 1–25.
- [10] L. Li, C. R. Weinberg, T. A. Darden, L. G. Pedersen, *Bioinformatics* **2001**, *17*, 1131–1142.
- [11] K. Deb, A. Raji Reddy, *Biosystems* **2003**, *72*, 111–129.
- [12] H. Liu, L. Yu, *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 1–12.
- [13] W. Duch, in *Feature Extraction: Foundations and Applications* (Eds: I. Guyon, S. Gunn, M. Nikravesh, L. A. Zadeh), Springer, Heidelberg **2006**, pp. 89–117.
- [14] R. Kohavi, G. H. John, *Artif. Intell.* **1997**, *97*, 273–324.
- [15] J. Handl, J. Knowles, *IJCIR* **2006**, *2*, 217–238.
- [16] J. Loughrey, P. Cunningham, Trinity College Dublin, Dublin, **2005**, TCD-CS-2005-41.
- [17] S. Davies, S. Rusell, in *Proc. 1994 AAAI Fall Symp. on Relevance*, AAAI Press, New Orleans, **1994**, pp. 37–39.
- [18] J. G. Topliss, *J. Med. Chem.* **1979**, *22*, 1238–1244.
- [19] H. Liu, H. Motoda, in *Computational Methods of Feature Selection*, (Eds: H. Liu, H. Motoda), Chapman & Hall/CRC, Boca Raton **2008**, pp. 1–17.
- [20] J. W. Godden, J. Bajorath, *QSAR Comb. Sci.* **2003**, *22*, 487–497.
- [21] F. R. Burden, D. A. Winkler, *QSAR Comb. Sci.* **2009**, *28*, 645–653.
- [22] H. Fröhlich, J. K. Wegner, A. Zell, *QSAR Comb. Sci.* **2004**, *23*, 311–318.
- [23] J. Shen, Y. Du, Y. Zhao, G. Liu, Y. Tang, *QSAR Comb. Sci.* **2008**, *27*, 704–717.
- [24] T.-H. Lin, S.-H. Chiu, K.-C. Tsai, *J. Chem. Inf. Model* **2006**, *46*, 1604–1614.
- [25] D. A. Konovalov, L. E. Llewellyn, Y. Vander Heyden, D. Coomans, *J. Chem. Inf. Model* **2008**, *48*, 2081–2094.
- [26] D. Horvath, F. Bonachera, V. Solov'ev, C. Gaudin, A. Varnek, *J. Chem. Inf. Comput. Sci.* **2007**, *47*, 927–939.
- [27] F. Gharagheizi, *QSAR Comb. Sci.* **2008**, *27*, 165–170.
- [28] S.-S. Yang, W.-C. Lu, T.-H. Gu, L.-M. Yan, G.-Z. Li, *QSAR Comb. Sci.* **2009**, *28*, 175–182.
- [29] L. S. Oliveira, R. Sabourin, F. Bortolozzi, C. Y. Suen, *Int. J. Pattern. Recogn.* **2003**, *17*, 903–929.
- [30] D. M. Hawkins, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1–12.
- [31] R. Guha, P. C. Jurs, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2179–2189.
- [32] T. Fox, J. M. Kriegl, *Curr. Top. Med. Chem* **2006**, *6*, 1579–1591.
- [33] S. Kirkpatrick, C. D. Gelatt, M. P. Vecchi, *Science* **1983**, *220*, 671–680.
- [34] *Genetic Algorithms in Search, Optimization and Machine Learning* (Ed: D. E. Goldberg), Addison-Wesley, New York **1989**.
- [35] C. A. Coello Coello, *IEEE Comput. Intell. Mag.* **2006**, *1*, 28–36.
- [36] *Multi-Objective Optimization using Evolutionary Algorithms* (Ed: K. Deb), Wiley, Chichester **2004**.
- [37] V. Trevino, F. Falciani, *Bioinformatics* **2006**, *22*, 1154–1156.
- [38] A. J. Soto, R. L. Cecchini, G. E. Vazquez, I. Ponzoni, in *Machine Learning and Data Mining in Bioinformatics – EvoBIO 2008, 6th Eur. Conf. Evolutionary Comput.* (Eds: E. Marchiori, J. H. Moore), Springer, Heidelberg **2008**, pp. 188–199.
- [39] P. Gramatica, *QSAR Comb. Sci.* **2007**, *26*, 694–701.
- [40] *Evolutionary Algorithms for Solving Multi-Objective Problems* (Eds: C. A. Coello Coello, G. B. Lamont, D. A. Van Veldhuizen), Springer Science&Business Media, LLC, New York, **2007**.
- [41] A. Jaszkiwicz, in *Metaheuristics for Multiobjective Optimization* (Eds: X. Gandibleux, M. Sevaux, K. Sörensen, V. T'kindt), Springer, Berlin **2004**, pp. 65–89.
- [42] K. Deb, A. Pratap, S. Agrawal, T. Meyrivan, *IEEE Trans. Evol. Comp.* **2002**, *6*, 182–197.
- [43] E. Zitzler, M. Laumanns, L. Thiele, in *Evolutionary Methods for Design, Optimisation and Control with Application to Industrial Problems* (Eds: K. C. Giannakoglou, D. Tsahalis, J. Periaux, P. Papilou, T. Fogarty), International Center for Numerical Methods in Engineering (Cmine), Athens, Greece **2002**, pp. 95–100.

- [44] *Matlab*, Version 6.0, The Mathworks; <http://www.mathworks.com>
- [45] S. Bleuler, M. Laumanns, L. Thiele, E. Zitzler, in *Evolutionary Multi-Criterion Optimization* (Eds: C. M. Fonseca, P. J. Fleming, E. Zitzler, K. Deb) Springer, Heidelberg **2003**, pp. 494–508.
- [46] D. E. Goldberg, K. Deb, in *Foundations of Genetic Algorithms* (Ed: G. J. E. Rawlins), Morgan Kaufmann, San Mateo, CA **1991**, pp. 69–93.
- [47] Z. Michalewicz, M. Schoenauer, *Evol. Comput.* **1996**, *4*, 1–31.
- [48] D. A. Winkler, *Drug. Future* **2004**, *29*, 1043–1057.
- [49] T. Arodz, D. A. Yuen, A. Z. Dudek, *J. Chem. Inf. Model* **2006**, *46*, 416–423.
- [50] D. J. C. MacKay, *Neural Comput.* **1992**, *4*, 415.
- [51] F. R. Burden, D. A. Winkler, *J. Med. Chem.* **1999**, *42*, 3183–3187.
- [52] D. A. Konovalov, D. Coomans, E. Deconinck, Y. V. Heyden, *J. Chem. Inf. Model* **2007**, *47*, 1648–1656.
- [53] D. Yaffe, Y. Cohen, G. Espinosa, A. Arenas, F. Giralt, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 162–183.
- [54] R. Todeschini, V. Consonni, A. Mauri, M. Pavan, *E-Dragon for VCCLAB*, <http://micchem.disat.unimib.it/chm/Help/edragon/index.html>
- [55] M. D. Safe, J. A. Carballido, I. Ponzoni, N. B. Brignole, in *Advances in Artificial Intelligence – SBIA 2004, 17th Brazilian Symposium on Artificial Intelligence* (Eds: A. Bazzan, S. Labidi), Springer, Heidelberg **2004**, pp. 405–413.
- [56] M. A. T. Figueiredo, *IEEE Trans. Pattern Anal. Mach. Intell.* **2003**, *25*, 1150–1159.
- [57] P. Ertl, in *Molecular Drug Properties: Measurement and Prediction* (Ed: R. Mannhold), Wiley-VCH, Weinheim **2008**, pp. 111–126.
- [58] C. Rücker, G. Rücker, M. Meringer, *J. Chem. Inf. Model.* **2007**, *47*, 2345–2357.
- [59] K. Baumann, *QSAR Comb. Sci.* **2005**, *24*, 1033–1046.
- [60] J. Gama, P. Brazdil, *Mach. Learn.* **2000**, *41*, 315–343.
- [61] A. Tropsha, P. Gramatica, V. K. Gombar, *QSAR Comb. Sci.* **2003**, *22*, 69–77.
- [62] D. M. Hawkins, S. C. Basak, D. Mills, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 579–586.
- [63] S. R. Johnson, *J. Chem. Inf. Model* **2008**, *48*, 25–26.
- [64] D. H. Wolpert, W. G. Macready, *IEEE Trans. Evol. Comp.* **1997**, *1*, 67–82.