A Wrapper-Based Feature Selection Method for ADMET Prediction Using Evolutionary Computing

Axel J. Soto[1,2], Rocío L. Cecchini[1], Gustavo E. Vazquez[1], and Ignacio Ponzoni[1,2]

[1] Laboratorio de Investigación y Desarrollo en Computación Científica (LIDeCC),
Departamento de Ciencias e Ingeniería de la Computación (DCIC)
Universidad Nacional del Sur – Av. Alem 1253 – 8000 – Bahía Blanca
Argentina
[2] Planta Piloto de Ingeniería Química (PLAPIQUI)
Universidad Nacional del Sur – CONICET
Complejo CRIBABB – Camino La Carrindanga km.7 – CC 717 – Bahía Blanca
Argentina
{saj,rlc,gev,ip}@cs.uns.edu.ar

Abstract. Wrapper methods look for the selection of a subset of features or variables in a data set, in such a way that these features are the most relevant for predicting a target value. In chemoinformatics context, the determination of the most significant set of descriptors is of great importance due to their contribution for improving ADMET prediction models. In this paper, a comprehensive analysis of descriptor selection aimed to physicochemical property prediction is presented. In addition, we propose an evolutionary approach where different fitness functions are compared. The comparison consists in establishing which method selects the subset of descriptors that best predicts a given property, as well as maintaining the cardinality of the subset to a minimum. The performance of the proposal was assessed for predicting hydrophobicity, using an ensemble of neural networks for the prediction task. The results showed that the evolutionary approach using a non linear fitness function constitutes a novel and a promising technique for this bioinformatic application.

Keywords: Feature Selection, Genetic Algorithms, QSAR, hydrophobicity.

1 Motivation

In the pharmaceutical industry, when a new medicine has to be developed, a 'serial' process starts where drug potency (activity) and selectivity are examined first [1]. Many of the candidate compounds fail at later stages due to ADMET (absorption, distribution, metabolism, excretion and toxicity) behavior in the body. ADMET properties are related to the way that a drug interacts with a large number of macromolecules and they correspond to the principal cause of failure in drug development [1]. In this way, a compound can be promising at first based on its molecular structure, but other factors such as aggregation, limited solubility or limited uptake in the human organism turn it useless as a drug.

Nowadays, the failure rate of a potential drug before reaching the market is still high. The main problem is that most of the rules that govern ADMET behavior in the

E. Marchiori and J.H. Moore (Eds.): EvoBIO 2008, LNCS 4973, pp. 188–199, 2008.

human body are unknown. For these reasons, interest in Quantitative Structure-Activity Relationships (QSAR) and Quantitative Structure-Property Relationships (QSPR) given by the scientific and industrial community has grown considerably in the last decades. Both of these approaches comprise the methods by which chemical structure parameters (known as descriptors) are quantitatively correlated with a well defined process, such as biological activity or any other experiment. QSAR has evolved over a period of 30 years from simple regression models to different computational intelligence models that are now applied to a wide range of problems [2], [3]. Nevertheless, the accuracy of the ADMET property estimations remains as a challenging problem [4].

In this context, hydrophobicity is one of the most extensively modeled physicochemical properties since the difficulty of experimentally determine its value, and also because it is directly related to ADMET properties [2], [5]. This property is traditionally expressed in terms of the logarithm of the octanol-water partition coefficient (logP).

QSAR methods developed by computer means are commonly named as *in silico* methods. These *in silico* methods, clearly cheaper than *in vitro* experiments, allow to examine thousands of molecules in shorter time and without the necessity of intensive laboratory work. Although *in silico* methods are not pretended to replace high-quality experiments at least in the short term, some computer methods have demonstrated to obtain as good accuracy as well-established experimental methods [6]. Moreover, one of the most important features of this approach is that a candidate drug (or a whole library) can be tested before being synthesized. Due to the gains in saved labour time, *in silico* predictions considerably help to reduce the large percentage of leads that fail in later stages of their development, and to avoid the amount of time and money invested in compounds that will not be successful.

In this context, machine learning methods are most preferred given the great amount of existing data and the little understanding of the pharmacokinetic rules of xenobiotics in the human body. Jónsdottir *et al.* [3] detail an extensive review of the many machine learning methods applied to bio- and chemoinformatics.

The major dilemma when logP is intended to be modeled by QSAR is that, thousands of descriptors could be measured for a single compound and also there is no general agreement on which descriptors are relevant or influence the hydrophobic behavior of a compound. This is an important fact, because overfitting and chance correlation could occur as a result of using more descriptors than necessary [7], [8]. On the other hand, poor models come as a result, when less descriptors than necessary are used. From an Artificial Intelligence (AI) perspective, this topic constitutes a particular case of the feature selection (FS) problem.

In this way, this work presents a sound approach for inferring the subset of the most influential descriptors for physicochemical properties. The righteousness of the selection is assessed by the construction of a prediction model. Our technique is based in the application of a genetic algorithm (GA) where: different fitness functions, a different number of descriptors selected by GA and a different number of descriptors considered by the prediction method are compared. This work is organized as follows: next section discusses related issues of feature selection in AI and in chemoinformatics in particular. Section 3 expands the aforementioned idea by introducing the genetic algorithm proposed for descriptor selection. In Section 4, applied data and

methods are presented, followed by the obtained results. Finally, in Section 5, main conclusions and future work are discussed.

2 Introduction to Feature Selection

Feature selection is the common name used to comprise all the methods that select from or reduce the set of variables or features used to describe any situation or activity in a dataset. Some authors differentiate variables from features, assuming that variables are the raw entry data, whereas features correspond to processed variables. However, variables, features or descriptors will be used here without distinction.

Nowadays, FS is a current research area, given that applications with datasets of many (even hundreds or thousands) variables have become frequent. Most usual cases where this technique is applied are gene selection from microarray data [9], [10], [11] and text categorization [12], [13], [14]. Confronting dimensionality carries some recognized advantages like: reducing the measurement and storage requirements, facilitating visualization and understanding of data, diminishing training and predicting times and also improving prediction performance.

Special care has to be taken with the distinction between relevant or useful and redundant. As it can be elucidated, selecting most relevant variables may be suboptimal for a predictor, especially when relevant variables are redundant. On the other hand, a subset of useful variables for a predictor may exclude redundant, but relevant, variables [15], [16], [17]. Therefore, in FS it is important to know whether developing a predictor is a final objective or not.

FS methods may be applied in two main ways, in terms of whether variables are individually or globally evaluated. That is, the first of them, works ranking each variable in an isolated way, i.e. these methods rank variables according to their individual predictive power. However, a variable that is useless by itself could be useful in consideration with others variables [17]. In this way, more powerful learning models are obtained, when the FS model selects subsets of variables that jointly have good predictive capacity.

A refined division of FS methods, especially applied to the latter defined group, is commonly used. They are often divided into filters, wrappers and embedded methods. When variables are selected according to data characteristics (*e.g.* low variance or correlated variables) they correspond to filter-type FS methods. Wrappers utilize a learning machine technique of interest as a black box, as a pre-processing step, to score subsets of variables in terms of their predictive ability. Finally, embedded methods carry out FS in the process of the training of a learning method and are usually tailored to the applied learning method [17], [18].

A wrapper-based FS method generally consists of two parts: the objective function, which may be a learning (regression or classification) method and a searching function that selects variables to be evaluated by the objective function. The results of the learning method are used to guide the searching procedure in the selection of descriptors. Consequently, the selection procedure is closely tied to the learning algorithm used, whether in quality of selection or execution time. For instance, we may get very different behaviors whether we are using linear models or nonlinear techniques [18].

2.1 Feature Selection Applied to QSAR

Many several papers successfully applied the FS strategy in bioinformatics related areas, like: drug discovery, QSAR and gene expression patterns analysis. We decided to apply descriptor selection in our work in order to detect which and how many descriptors are the most useful ones for the prediction of logP. We agreed on the use of GAs as the searching function, given that they offer a parallel search of solutions, potentially avoiding local minima. Moreover, with a correct design of a fitness function, GA inherently guides the different generations of individuals to a good if not optimal solution. In this context, the objective function corresponds to the function used for the fitness of GA.

In this way, and as a result of the review about the related work in the area, we found some inspiring papers. In ref. [9], [18], [19], [20], [21], [22] different fitness functions are tested within a GA to determine a subset reduction. In [18], [23], [24] FS is applied using a neural network (NN) for the fitness function. However, we find that this proposal has the drawback of the great amount of time required by the NN for training and thus the execution time becomes prohibitive when the number of combination of feasible selections is large.

3 Wrapper Method

We implemented a GA for searching the space of the multiple feasible selections. We propose three appropriate fitness functions for guiding the search of GA, namely: decision trees, *k*-nearest neighbors (KNN) and a polynomic non linear function. According to the previous classification, our proposed FS method belongs to a wrapper method because statistical or machine learning methods are used in the fitness function for assessing the prediction capability of the selected subset.

3.1 Main Characteristics of GA

Binary strings are used to represent the individuals. Each string of length m stands for a feasible descriptor selection, where m is the number of considered descriptors. A nonzero value in the i^{th} bit position means that the i^{th} descriptor is selected. We have constrained to a model where p bits are active for each individual. In other words, each chromosome encodes its choice of the p selected descriptors.

The initial population is randomly generated by imposing the described restriction of exactly p active descriptors on each individual. A one-point crossover is used for the recombination [25]. Non feasible individuals could take place after crossover, because the number of nonzero bits may be different than p . This problem is solved by randomly setting or resetting bit locations as needed to be up to p active bits. Since the crossover scheme inherently incorporates bit-flip mutation, we abstained to use an additional scheme of mutation.

We did different experiments and we concluded that tournament method is appropriate for the selection of parents. Furthermore, this method is preferred than others because it is particularly easy to implement and its time complexity is $O(n)$ [25]. We

also included elitism, which protects the fittest individuals in any given generation, by moving them to the next generation.

3.2 Fitness Function

Taking into account that the GA objective is to determine the most relevant set of p descriptors for predicting a physicochemical property, the fitness function should estimate the accuracy of a prediction method when only the p descriptors are used. In particular, the general form of the fitness function employed is presented in the equation 1. This formula computes the mean square error of prediction (MSE):

$$\mathrm{F}(\mathcal{P}_{Z_{1,k}}, Z_{2,k}) = \frac{1}{n_2} \left[\sum_{(x_i, y_i) \in Z_{2,k}} (y_i - \mathcal{P}_{Z_{1,k}}(x_i))^2 \right]. \quad (1)$$

Where:

- **Z** is a matrix that represents a compound dataset, 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**.
- $\mathcal{P}_Z$ is a statistical method trained with the dataset **Z**. In the same way, $\mathcal{P}_Z(x)$ is the output for the $\mathcal{P}_Z$ method when the case x is presented.
- $\mathbf{Z_1}$ and $\mathbf{Z_2}$ are compound databases used as learning and validation sets respectively with corresponding sizes $\mathbf{n_1 \times m}$ and $\mathbf{n_2 \times m}$.
- $\mathbf{Z_{j,k}}$ is a filtered dataset in accordance with the descriptor selection encoded by the k^{th} individual. In other words, $\mathbf{Z_{j,k}}$ only contains those variables of $\mathbf{Z_j}$ whose values in the corresponding locations of the k^{th} individual's chromosome are 1.
- $\mathbf{x_i}$ is a vector that represents the values of the descriptors for the i^{th} compound of a given dataset.
- $\mathbf{y_i}$ is the target value for the i^{th} compound of a given dataset.

The first argument of the fitness function is the statistical method applied to a given learning set, while the second argument corresponds to a validation set, from where fitness value is calculated. In this work, three different predictor techniques were tested. The first one corresponds to decision trees (DT) (as regression trees) using Gini's diversity index for the splitting criteria and without using any kind of pruning [26]. The second is KNN regression as used in ref [9]. Both methods are local and usually applied for prediction or for FS purposes [27].

A non linear regression model was also applied in this paper as the first argument of the fitness function. A nonlinear expression is established where their coefficients ($\beta_{i,j}$) are adjusted with a nonlinear least-squares fitting by the Gauss-Newton method [28]. The corresponding and nonlinear regression model formula is presented in Equation 2, where x_i corresponds to the value of the $\boldsymbol{i}^{\text{th}}$ descriptor for any given compound. Non linear models are not generally applied given that they need the construction of a mathematical formula. Nevertheless, we propose it as an alternative for

NN, so that non linear regressions could be carried out. It is worth mentioning that this approach circumvent the necessity of a manual tuning of the architecture and training parameters as is the case with NN.

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

4 Methodology and Analysis of Results

Our proposal consists in the search of a selection of descriptors that minimizes the prediction error when they are used as input of a predictor method. This selection is fulfilled with the GA previously described. Moreover, a fair comparison is intended to be established in order to determine which fitness function works best with GA. It is worth mentioning that, as well as minimizing error, it is important to obtain relevant descriptors in a subset of minimal size.

4.1 Data Sets

Our FS method was applied to a data set of 440 organic compounds compiled from the literature [29] where their logP values at 25ºC conform the modeled target variable. The choice of the data set was supported by the possibility of comparison with the previous work and also for the heterogeneous compounds that it comprises (e.g. hydrocarbons, halogens, sulfides, anilines, alcohols, carboxylic acids amongst others).

Each compound was characterized by 73 molecular descriptors commonly used for logP [30], [31], [32]. Dragon 5.4 [33] was used for calculating descriptors of the: constitutional (41), functional groups (16), properties (2) and empiricals (3) families and we completed with 11 descriptors from [29] (Table 3). Previous to the use of the data, all descriptors were normalized, so each descriptor has a standard deviation of 1.

4.2 Genetic Algorithm Parameters

In order to assess the stability of the GA in the selection and to explore the sensitivity of the choice of p in the prediction, 45 independent runs were carried out for each choice of p, where p was set to 10, 20 and 30. This same procedure was made for the three considered fitness functions, making a total of 405 runs for the GA.

The chromosome size m is 73 according to the number of calculated descriptors. For the GA runs we used typical parameter values: population size=45; crossover probability=0.8; tournament size=3, elite members=2. A phenotypic stopping criterion is used; the GA stops when the highest fitness of the population does not improve during 15 generations or when the improvement of the average fitness of the population is less than a given tolerance value.

4.3 Prediction Method

NNs are probably one of the most widely used methods for QSAR modelling [2], [6], [34]. In order to evaluate the suitability of the selection, we used a neural network

ensemble (NNE) as an independent prediction method, *i.e.* it measures the accuracy of prediction for each proposed wrapper method. The number of descriptors (d) used as input for this independent prediction method is not necessarily the same as the p genes selected by the GA. With the intention of establishing a suitable (minimal cardinality and error) subset of descriptors this value was settled to 10 different values: 11, 12, 15, 20, 25, 30, 40, 50, 60 and 73. The d descriptors used for the predictor are selected from a ranking of the most selected descriptors obtained in the 45 re-runs of the GA. Each ensemble consists of three NNs, and all of them are of type feed-forward back-propagation. The specific architecture of each NN, was established according to the number picked for d . Principal Component Analysis (PCA) is applied prior to the training of the NNE, so the descriptors that contribute less than a 0.2% of the total variance are discarded and considered as redundant.

4.4 Results

With the purpose of evaluating the performance in the prediction achieved by the aforementioned fitness functions, we trained NNEs for each presented configuration of the GA and we obtained error prediction for different choices of d (Table 1, Fig. 1).

It is worth mentioning that it is not straightforward to obtain logP related works from the bibliography that allows a reproducibility or benchmarking of the results of the work, as it is the case of ref [29]. So, to enable a direct comparison with this work, the data set was identically divided into training, validation and test set, also using the same compounds in each set.

Our results were obtained after several different NN configurations and replicas, and the tendency was rather similar. Each reported error is an average over 5 replicas (15 NNs) applied to the test set.

In comparison with the backpropagation NN proposed in ref. [29], which obtains a 0.23 MAE and where similar conditions apply, our model of NNE with the assistance of the FS method has improved the accuracy of logP prediction, even when using one less descriptor (NL, p =10 fitness function).

Decision trees as fitness function have a better behavior in their variant with $p = 20$ descriptors, but with few descriptors for d , the performance is quite far from optimal. In the case of KNN, it looks like few descriptors for p is not appropriate at least when less than 20 descriptors are used for the NNE. For NL, the behavior is quite good when $p = 10$. As expected, in all models similar results are obtained when more than 25 descriptors are considered for d .

Considering the best alternative of p for each fitness function (Fig. 1 (d)) we highlight the performance of NL. It has a roughly equal behavior along all d values and takes a minimal prediction error when $d = 25$. KNN's behavior is similar to NL, except for the lowest values of d . In the case of the DT-based predictions, although they have a better performance than the previous two cases for large d , the bad performance with small d values, makes it not so valuable as an FS technique, at least for the present example.

Table 1. Prediction errors in terms of MAE, MSE and variance on 5 runs

	Method	*d* = 11	*d* = 12	*d* = 15	*d* = 20	*d* = 25	*d* = 30	*d* = 40	*d* = 50	*d* = 60	*d* = 73
MAE	*NL, p = 10*	0,1972	0,1922	0,1896	0,1928	0,1708	0,1720	0,1737	0,1855	0,2021	0,1840
MAE	*DT, p = 20*	0,2331	0,2362	0,2428	0,2114	0,1928	0,1742	0,1656	0,1626	0,1733	0,1847
MAE	*KNN, p = 30*	0,2562	0,2771	0,2986	0,2325	0,2079	0,1866	0,1925	0,1782	0,1827	0,1858
MSE	*NL, p = 10*	0,0740	0,0658	0,0738	0,0773	0,0566	0,0551	0,0593	0,0705	0,0895	0,0650
MSE	*DT, p = 20*	0,1013	0,1160	0,1173	0,0903	0,0745	0,0539	0,0522	0,0510	0,0521	0,0676
MSE	*KNN, p = 30*	0,1166	0,1266	0,1510	0,1066	0,0781	0,0730	0,0706	0,0613	0,0664	0,0635
Variance	*NL, p = 10*	1,330E-04	1,425E-04	7,278E-05	3,581E-04	6,015E-05	5,289E-05	8,314E-05	6,531E-05	2,197E-05	1,154E-04
Variance	*DT, p = 20*	4,437E-05	1,384E-04	2,064E-04	2,821E-04	7,206E-05	4,601E-05	9,563E-05	6,871E-05	3,635E-04	1,440E-04
Variance	*KNN, p = 30*	2,222E-04	1,079E-04	4,144E-04	6,366E-05	6,142E-05	1,032E-04	7,277E-05	1,880E-04	9,916E-05	6,228E-05

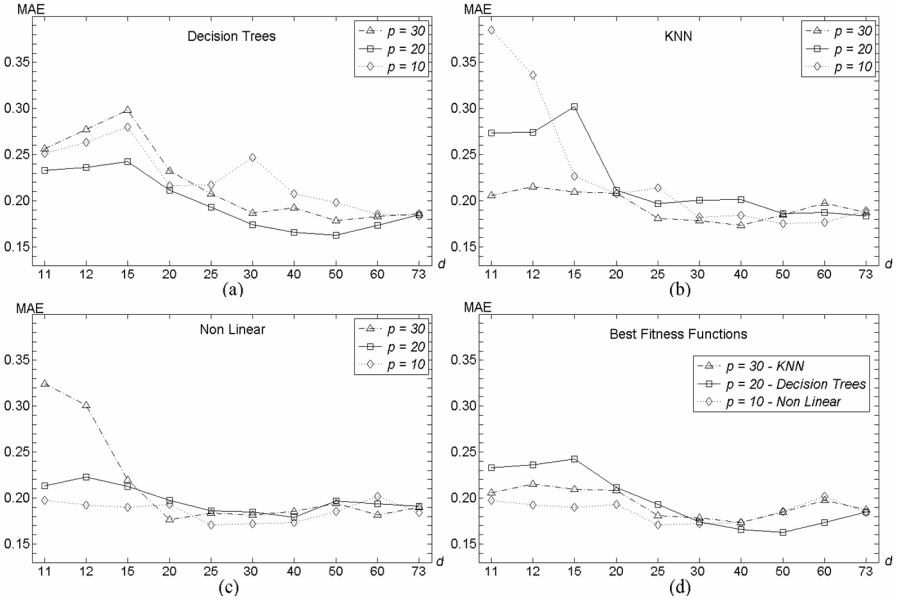

Fig. 1. NNE prediction error in terms of mean absolute error (MAE) considering different number of descriptors as input, and also for different GA-based selection methods: (a) decision trees, (b) k-nearest neighbors and (c) non linear (d) best fitness functions

Table 2. Two-way ANOVA for MAE of prediction of the three best methods and when few descriptors are used ($d = 11$, $d = 12$ and $d = 15$)

Source of Var.	Sum of Squares	D.F.	M.S.	F	p
BETWEEN	0,015629	8	0,0019536	14,7950419	2,622E-09
d factor	0,000056	2	0,0000279	0,21117041	0,8106
wrapper factor	0,015003	2	0,0075014	56,8094222	7,306E-12
Interaction	0,000570	4	0,0001426	1,0797875	0,3809
WITHIN	0,004754	36	0,0001320		
TOTAL	0,020383	44			

In order to formally support preceding facts, we analyze whether significant discrepancies exist among the different models by using a two-way ANOVA test (Table 2). The two involved factors are the FS method and the choice of d. Our comparison is focused on finding significant differences on the methods when using few descriptors

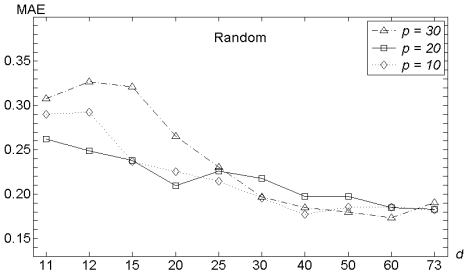

Fig. 2. NNE prediction error in terms of mean absolute error (MAE) considering different number of descriptors as input, and also for different random-based selections

Table 3. List of ranked descriptors according to wrapper method NL, $p = 10$. Descriptors with * are scaled on carbon atoms.

Rank	Symbol	Definition	Freq.	Rank	Symbol	Definition	Freq.
1	nHDon	number of donor atoms for H-bonds (N and O)	87%	38	nR06	number of 6-membered rings	42%
2	Hy	hydrophilic factor	84%	39	nR03	number of 3-membered rings	40%
3	D_S	total dipole (sum = point charge + hybridization)	78%	40	nAT	number of atoms	38%
4	nCp	number of total primary C(sp3)	78%	41	nR10	number of 10-membered rings	38%
5	D_H	total dipole (hybridization)	73%	42	nCq	number of total quaternary C(sp3)	38%
6	nH	number of Hydrogen atoms	73%	43	nCIR	number of circuits	36%
7	nSK	number of non-H atoms	71%	44	nRSR	number of sulfides	36%
8	RBN	number of rotatable bonds	71%	45	SCBO	sum of conventional bond orders (H-depleted)	33%
9	nC	number of Carbon atoms	71%	46	nCXr	number of X on ring C(sp3)	33%
10	nCs	number of total secondary C(sp3)	71%	47	nDB	number of double bonds	31%
11	nCaR	number of aromatic C(sp2)	69%	48	nBT	number of bonds	29%
12	D_P	total dipole (point charge)	67%	49	nS	number of Sulfur atoms	27%
13	IP	ionization potential	67%	50	nR09	number of 9-membered rings	24%
14	VMC1	first-order valence molecular connectivity index	67%	51	nOHs	number of secondary alcohols	24%
15	nF	number of Fluorine atoms	67%	52	nBnz	number of benzene-like rings	22%
16	Sp	sum of atomic polarizabilities [*]	64%	53	nSH	number of thiols	22%
17	Ms	mean electrotopological state	64%	54	nR=CX2	number of R=CX2	22%
18	nBO	number of non-H bonds	64%	55	nR07	number of 7-membered rings	20%
19	Ui	unsaturation index	64%	56	Mv	mean atomic van der Waals volume [*]	18%
20	nX	number of halogen atoms	62%	57	Mp	mean atomic polarizability [*]	18%
21	nHAcc	number of acceptor atoms for H-bonds (N, O, F)	62%	58	nTB	number of triple bonds	18%
22	Sv	sum of atomic van der Waals volumes [*]	60%	59	nR=CHX	number of R=CHX	18%
23	nO	number of Oxygen atoms	60%	60	nI	number of Iodine atoms	16%
24	VMC2	second-order valence molecular connectivity index	58%	61	ARR	aromatic ratio	16%
25	nCL	number of Chlorine atoms	58%	62	nR11	number of 11-membered rings	13%
26	VMC4	fourth-order valence molecular connectivity index	56%	63	PSA	fragment-based polar surface area	11%
27	nAB	number of aromatic bonds	56%	64	nOHt	number of tertiary alcohols	7%
28	nBM	number of multiple bonds	53%	65	Me	mean atomic Sanderson electroneg. [*]	2%
29	nROR	number of ethers (aliphatic)	53%	66	MW	molecular weight	0%
30	nCt	number of total tertiary C(sp3)	51%	67	E2	total two-center energy	0%
31	nOHP	number of primary alcohol	51%	68	EX	exchange energy (two-center term)	0%
32	Se	sum of atomic Sanderson electroneg. [*]	49%	69	ELC	total electrostatic interaction (two-center term)	0%
33	nCIC	number of rings	49%	70	PO	average polarizability	0%
34	nBR	number of Bromine atoms	49%	71	Ss	sum of Kier-Hall electrotopological states	0%
35	nN	number of Nitrogen atoms	47%	72	RBF	rotatable bond fraction	0%
36	AMW	average molecular weight	42%	73	MR	Ghose-Crippen molar refractivity	0%
37	nR05	number of 5-membered rings	42%				

for the NNE ($d = 11, d = 12$ and $d = 15$). Given that there is not strong evidence of an interaction factor, we can separately analyze both factors. The ANOVA test shows that there is no evidence of differences on using 11, 12 or 15 descriptors for one same wrapper method (d factor near 1), and also that significant differences are found for the choice of the method for feature selection (p-value of wrapper factor ≈ 0). Finally, we

also apply Bonferroni multiple comparison procedure to determine which method differs from which. With a global level of error $\alpha = 0.03$ we found that all methods differ from each other (data not shown).

Besides, in order to evidence the advantages and the differences of the application of a FS technique, we analyzed the performance when a random selection is carried out (Fig. 2). As expected, the prediction error decreases when more descriptors are considered for the NNE. On the other hand, with large d values, error is not so bad given that all descriptors are related with the target property.

Our last analysis of results is about which descriptors were selected by GA, and their frequency of selection. Table 3 shows the list of descriptors, ranked according to NL – $p = 10$ criteria, with the percentage of the times selected in the 45 runs of the GA. From a chemical perspective, it is interesting to note that the first three top-ranked descriptors are considered as reasonably influential for logP [32].

5 Conclusions

The present work proposes a methodology to detect which descriptors are the most influential to the prediction of the molecule hydrophobicity. This detection of relevant features allows a decrease in the prediction error and also a better understanding of the structure-property relationships. The key contributions of our work are the proposal of a non linear function adjusted with least squares in the fitness function and the rigorous comparison carried out by the different combinations of the wrapper variants.

Despite the unknown of the general form of the function that governs the structure-property relationship, the fourth-order polynomial function works well for the wrapper, since it captures the nonlinearity of the model, as well as it maintains an acceptable execution time performance. Besides, the GA's behavior is quite stable given the low variance of the prediction errors and the high frequency associated with the top-ranked descriptors.

According to the authors' knowledge, we did not find previous works with a ranked list of relevant features for predicting hydrophobicity. It is worth noting that in the FS step, relevant but redundant variables can be selected. However, since PCA is applied before the training of the NNE, any redundant feature is thus discarded.

Our proposal is not restricted to logP, because this method could also be applied to any physicochemical property. It would be interesting to experiment this proposal with the aggregation of other descriptor families. In this context, we are evaluating other descriptors that express interactions between functional groups in molecules. Moreover, the GA could also be developed to directly detect the most adequate number of descriptors in a multi-objective way, instead of fixing to a specific number. At this moment, we are also planning to extend the comparison with other combinations of AI methods.

References

1. Selick, H.E., Beresford, A.P., Tarbit, M.H.: The Emerging Importance of Predictive ADME Simulation in Drug Discovery. Drug Discov 7(2), 109–116 (2002)
2. Taskinen, J., Yliruusi, J.: Prediction of Physicochemical Properties Based on Neural Network Modeling. Adv. Drug Deliver. Rev. 55(9), 1163–1183 (2003)

3. Jónsdottir, S.Ó., Jørgensen, F.S., Brunak, S.: Prediction Methods and Databases Within Chemoinformatics: Emphasis on Drugs and Drug Candidates. Bioinformatics 21, 2145–2160 (2005)
4. Tetko, I.V., Bruneau, P., Mewes, H.-W., Rohrer, D.C., Poda, G.I.: Can we estimate the accuracy of ADME-Tox predictions? Drug Discov. Today 11, 700–707 (2006)
5. Huuskonnen, J.J., Livingstone, D.J., Tetko, I.V.: Neural Network Modeling for Estimation of Partition Coefficient Based on Atom-Type Electrotopological State Indices. J. Chem. Inf. Comput. Sci. 40, 947–995 (2000)
6. Agatonovic-Kustrin, S., Beresford, R.J.: Basic Concepts of Artificial Neural Network (ANN) Modeling and its Application in Pharmaceutical Research. J. Pharmaceut. Biomed. 22(5), 717–727 (2000)
7. Tetko, I.V., Livingstone, D.J., Luik, A.I.: Neural Networks Studies. 1. Comparison of Over-fitting and Overtraining. J. Chem. Inf. Comput. Sci. 35, 826–833 (1995)
8. Topliss, J.G., Edwards, R.P.: Chance Factors in Studies of Quantitative Structure-Activity Relationships. J. Med. Chem. 22(10), 1238–1244 (1979)
9. Li, L., Weinberg, C.R., Darden, T.A., Pedersen, L.G.: Gene selection for sample classifica-tion based on gene expression data: Study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics 17(12), 1131–1142 (2002)
10. Tan, T., Fu, X., Zhang, Y., Bourgeois, A.G.: A genetic algorithm-based method for feature subset selection. Soft Comput 12(2), 111–120 (2008)
11. Zhu, Z., Ong, Y., Dash, M.: Markov blanket-embedded genetic algorithm for gene selection. Pattern Recognition 40(11), 3236–3248 (2007)
12. Forman, G.: An extensive empirical study of feature selection metrics for text classification. JMLR 3, 1289–1306 (2003)
13. Lin, K., Kang, K., Huang, Y., Zhou, C., Wang, B.: Naive bayes text categorization using improved feature selection. Journal of Computational Information Systems 3(3), 1159–1164 (2007)
14. Montañés, E., Quevedo, J.R., Combarro, E.F., Díaz, I., Ranilla, J.: A hybrid feature selection method for text categorization. International Journal of Uncertainty, Fuzziness and Knowlege-Based Systems 15(2), 133–151 (2007)
15. Kohavi, R., John, G.: Wrappers for feature selection. Artificial Intelligence 97, 273–324 (1997)
16. Blum, A., Langley, P.: Selection of relevant features and examples in machine learning. Artificial Intelligence 97, 245–271 (1997)
17. Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection. JMLR 3, 1157–1182 (2003)
18. Dutta, D., Guha, R., Wild, D., Chen, T.: Ensemble Feature Selection: Consistent Descriptor Subsets for Multiple QSAR Models. J. Chem. Inf. Model. 47, 989–997 (2007)
19. Liu, S., Liu, H., Yin, C., Wang, L.: VSMP: A novel variable selection and modeling method based on the prediction. J. Chem. Inf. Comp. Sci. 43(3), 964–969 (2003)
20. Wegner, J.K., Zell, A.: Prediction of aqueous solubility and partition coefficient optimized by a genetic algorithm based descriptor selection method. J. Chem. Inf. Comp. Sci. 43(3), 1077–1084 (2003)
21. Kah, M., Brown, C.D.: Prediction of the adsorption of lonizable pesticides in soils. J. Agr. Food Chem. 55(6), 2312–2322 (2007)
22. Bayram, E., Santago, P., Harrisb, R., Xiaob, Y., Clausetc, A.J., Schmittb, J.D.: Genetic algorithms and self-organizing maps: A powerful combination for modeling complex QSAR and QSPR problems. J. of Comput.-Aided Mol. Des. 18, 483–493 (2004)

23. So, S.-S., Karplus, M.: Evolutionary Optimization in Quantitative Structure-Activity Relationship: An Application of Genetic Neural Networks. J. Med. Chem. 39, 1521–1530 (1996)
24. Fernández, M., Tundidor-Camba, A., Caballero, J.: Modeling of cyclin-dependent kinase inhibition by 1H-pyrazolo[3,4-d] pyrimidine derivatives using artificial neural network ensembles. J. Chem Inf. and Model. 45(6), 1884–1895 (2005)
25. Goldberg, D.E., Deb, K.: A comparative analysis of selection schemes used in genetic algorithms. In: Foundations of Genetic Algorithms, pp. 69–93. Morgan Kaufmann, San Mateo, CA (1991)
26. Breiman, L.: Classification and Regression Trees. Chapman & Hall, Boca Raton (1993)
27. Trevino, V., Falciani, F.: GALGO: An R package for multivariate variable selection using genetic algorithms. Bioinformatics 22(9), 1154–1156 (2006)
28. Madsen, K., Nielsen, H.B., Tingleff, O.: Methods for Non-Linear Least Squares Problems. Technical University of Denmark, 2nd edn. (April, 2004)
29. Yaffe, D., Cohen, Y., Espinosa, G., Arenas, A., Giralt, F.: Fuzzy ARTMAP and back-propagation neural networks based quantitative structure - property relationships (QSPRs) for octanol: Water partition coefficient of organic compounds. J. Chem. Inf. Comp. Sci. 42(2), 162–183 (2002)
30. Linpinski, C.A., Lombardo, F., Dominy, B.W., Freeny, P.: Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. Adv. Drug Deliv. Rev. 23, 3–25 (1997)
31. Duprat, A., Huynh, T., Dreyfus, G.: Towards a principled methodology for neural network design and performance evaluation in qsar; application to the prediction of logp. J. Chem. Inf. Comp. Sci. 38, 586–594 (1998)
32. Wang, R., Fu, Y., Lai, L.: A new atom-additive method for calculating partition coefficients. J. Chem. Inf. Comp. Sci. 37(3), 615–621 (1997)
33. Tetko, I.V., Gasteiger, J., Todeschini, R., Mauri, A., Livingstone, D., Ertl, P., Palyulin, V.A., Radchenko, E.V., Zefirov, N.S., Makarenko, A.S., Tanchuk, V.Y., Prokopenko, V.V.: Virtual computational chemistry laboratory - design and description. J. Comput. Aid. Mol. Des. 19, 453–463 (2005)
34. Winkler, D.A.: Neural networks in ADME and toxicity prediction. Drug. Future 29(10), 1043–1057 (2004)