JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

ARTICLE

pubs.acs.org/jcim

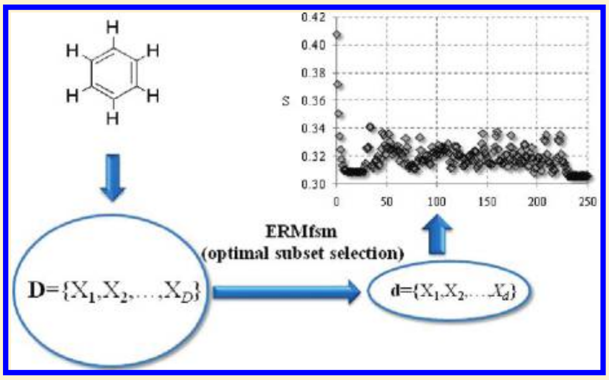
Advances in the Replacement and Enhanced Replacement Method in QSAR and QSPR Theories

Andrew G. Mercader,*,[†],[‡] Pablo R. Duchowicz,[†] Francisco M. Fernández,[†] and Eduardo A. Castro[†]

[†]Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas (INIFTA, UNLP, CCT La Plata-CONICET), Diag. 113 y 64, Sucursal 4, C.C. 16, 1900 La Plata, Argentina

[‡]PRALIB (UBA-CONICET), Facultad de Farmacia y Bioquímica, Universidad de Buenos Aires, Junín 956, C1113AAD Buenos Aires, Argentina

S *Supporting Information*



ABSTRACT: The selection of an optimal set of molecular descriptors from a much greater pool of such regression variables is a crucial step in the development of QSAR and QSPR models. The aim of this work is to further improve this important selection process. For this reason three different alternatives for the initial steps of our recently developed enhanced replacement method (ERM) and replacement method (RM) are proposed. These approaches had previously proven to yield near optimal results with a much smaller number of linear regressions than the full search. The algorithms were tested on four different experimental data sets, formed by collections of 116, 200, 78, and 100 experimental records from different compounds and 1268, 1338, 1187, and 1306 molecular descriptors, respectively. The comparisons showed that one of the new alternatives further improves the ERM, which has shown to be superior to genetic algorithms for the selection of an optimal set of molecular descriptors from a much greater pool. The new proposed alternative also improves the simpler and the lower computational demand algorithm RM.

INTRODUCTION

A generally accepted remedy for overcoming the lack of experimental data in complex chemical phenomena is the analysis based on quantitative structure−property/activity relationships (QSPR/QSAR).[1] Hence, there exists a permanently renewed interest focused on the development of such kinds of predictive techniques.[2−6] The essential role of QSPR/QSAR is to suggest mathematical models capable of estimating and predicting relevant properties or activities of interest, especially when those cannot be experimentally determined for some reason. These studies rely on the basic assumption that the structure of a compound determines entirely its properties, which can therefore be translated into so-called molecular descriptors.[7] These parameters are calculated through mathematical formulas obtained from several theories, such as chemical graph theory, information theory, quantum mechanics, etc.[8,9]

There are thousands of descriptors available in the literature,[7] and one has to decide how to select those that characterize the property/activity under consideration in the most efficient way. Thus the mathematical problem of selecting a subset $\mathbf{d}$ of d descriptors from a much larger set $\mathbf{D}$ of $D \gg d$ ones arises.

The search for the optimal set of descriptors is normally monitored by the minimization or maximization of a chosen statistical parameter; for instance, searching for the model that makes the standard deviation (S) as small as possible. In other words, the global minimum of $S(\mathbf{d})$ is sought, where $\mathbf{d}$ is a point in a space of $D!/[d!(D-d)!]$. As mentioned since D is very large, a full search (FS) of the optimal variables is impractical because it requires $D!/[d!(D-d)!]$ linear regressions.

Some time ago we proposed the replacement method (RM)[10,11] and later the enhanced replacement method (ERM)[12] that produce linear regression QSPR/QSAR models, presenting no relevant difference with FS (for small-sized descriptor data sets where FS can be calculated), using much less computational work.[12] These alternative techniques approach the minimum of S by judiciously taking into account the relative errors of the coefficients of the least-squares model given by a set of d descriptors $\mathbf{d} = \{X_1, X_2, ..., X_d\}$. The ERM gives models with better estimative and predictive ability than the forward stepwise regression procedure[13] and the more elaborated genetic algorithms[14] (GA).[15]

The RM is a rapidly convergent iterative algorithm that produces linear regression models with small S in remarkably little computer time.[16−18] However, in some cases, the RM can get trapped in a local minimum of S. Although such local minima provide acceptable models, as shown in all earlier applications of the RM,[16−18] there was still room for improvement, and the ERM was developed.

Received: February 16, 2011
Published: June 06, 2011

Table 1. Number of Necessary Linear Regressions and Calculation Time To Carry Out a Full Search with $D = 1187$ and $N = 78$[a]

d	1	2	3	4	5	6	7
reg. number	1187	703891	2.8×10^{8}	$8.2 \times 10^{1}0$	1.9×10^{13}	3.8×10^{15}	6.5×10^{17}
minutes	**0.003**	**1.95**	7.7×10^{2}	2.3×10^{5}	5.4×10^{7}	$1.1 \times 10^{1}0$	1.8×10^{12}
hours	5.5×10^{-5}	3.2×10^{-2}	**12.8**	3791.6	9.0×10^{5}	1.8×10^{8}	$3.0 \times 10^{1}0$
days	2.3×10^{-6}	1.4×10^{-3}	5.3×10^{-1}	**158.0**	3.7×10^{4}	7.4×10^{6}	1.2×10^{9}
years	6.2×10^{-9}	3.7×10^{-6}	1.5×10^{-3}	4.3×10^{-1}	$\mathbf{1.0 \times 10^{2}}$	$\mathbf{2.0 \times 10^{4}}$	$\mathbf{3.4 \times 10^{6}}$

[a] Using an AMD Athlon 64 2800+ processor.

The ERM follows the same RM philosophy but exhibits less propensity for remaining in local minima and at the same time is less dependent on the initial set of descriptors. It has a resemblance with the simulated annealing (SA), which is an adaptation of the Metropolis−Hastings algorithm, a Monte Carlo method[19] that generates sample states of a thermodynamic system. The name and the inspiration come from annealing in metallurgy, a technique involving heating and controlled cooling of a material to increase the size of its crystals and reduce their defects. The heat causes the atoms to become unstuck from their initial positions (a local minimum of the internal energy) and wander randomly through states of higher energy; the slow cooling gives them more chances of finding configurations with lower internal energy than the initial one.[20]

ERM and RM have been compared to GA in several practical applications[21−23] and recently in a work that makes a more extensive and reliable comparison.[15] The results showed that ERM gives better results than GA and additionally is much simpler to implement. On the other hand the work showed that GA was slightly superior to RM, however simplicity and lower computational demand of RM still makes it an attractive methodology.

The QSAR/QSPR models obtained using ERM and RM can be analyzed, and the dependence of the activity or property on the descriptors can be visualized. This is because these methodologies do not require any descriptor transformation into different variables that lack physical meaning. In contrast partial least-squares (PLS) regression is a technique that combines features from and generalizes principal component analysis (PCA) and multiple linear regressions. In order to predict a set of dependent variables from a set of independent variables, this technique extracts from the descriptors a set of orthogonal factors called latent variables which have the best predictive power.[24]

The first step in RM and ERM does not use the same scheme as the rest of the algorithm; the current strategy was determined in the practical use of the algorithms as the best alternative. However there are other alternatives that we have recently developed. The main target of this work is to present these different alternatives and test them to determine if they provide a significant improvement to the algorithms.

METHODS

Algorithms. The following subsections briefly describe the theory of the present state of RM and ERM. All the algorithms were programmed in the computer system Matlab 5.0.[25] Tests were done using $d = 7$ for a high computational demanding search with a reasonable number of descriptors for a potential model in common QSPR/QSAR studies.

Comparisons of the algorithms were done performing 100 numerical tests for each of the 4 data sets (100 different random initial sets for RM/ERM and 600 additional random sets for RMfsm/ERMfsm/RMafs to get the same computational effort). The results were compared in terms of the number of times that the proposed alternative gave a smaller S than the previous algorithm and are presented in Tables 2, 4, and 6. Additionally the average S and regression coefficient (R) of the 100 models found with the algorithms were contrasted in Tables 3, 5, and 7, along with the cross-validation leave one out (loo)[26] S and R. In order to determine that the difference in the mean values of S for the different algorithms presented a statistical significance, a student's t-test[27] was performed. The results were offered in terms of the probability that the results have come about through mere random variability; lower than 0.05 probability values indicate statistical significance.

RM. An optimal subset $\mathbf{d}_m = \{X_{m1}, X_{m2}, ..., X_{md}\}$ of $d \ll D$ is chosen, from a large set $\mathbf{D} = \{X_1, X_2, ..., X_D\}$ of D descriptors provided by some available commercial program, with minimum standard deviation S:

$$S = \sqrt{\frac{1}{(N-d-1)} \sum_{i=1}^{N} res_i^{\,2}} \qquad (1)$$

where N is the number of molecules in the training set, and res_i the residual for molecule i (difference between the experimental and predicted property). The fact that $S(\mathbf{d}_n)$ is a distribution on a discrete space of $D!/d!(D-d)!$ disordered points ($\mathbf{d}_n$) should be noticed. The FS that consists of calculating $S(\mathbf{d}_n)$ on all those points always allows to arrive at the global minimum but as mentioned is computationally prohibitive if D is sufficiently large. For example, using for $d = 7$ and $D = 140$ or higher will take more than 1.8×10^{11} regressions, which translates to more than 1 year to complete only 1 calculation (using an AMD Athlon 64 2800+ processor) (refer to Table 1 for other examples). The RM briefly consists of the following steps:

- An initial set of descriptors $\mathbf{d}_k$ is selected from D at random, one of the descriptors is replaced, denoted as X_{ki}, with all the remaining $D - d$ descriptors, one by one, and the set with the smallest value of S is kept. What was done up to this point is defined as a 'step'.
- From this resulting set, the descriptor with the greatest standard deviation in its coefficient is chosen (the one changed previously is not considered) and substituted with all the remaining $D - d$ descriptors, one by one. This procedure is repeated until the set remains unmodified. In each of these cycles the descriptors replaced in previous steps are not taken into account. Thus, the candidate $\mathbf{d}_m^{(i)}$ that comes from the so-constructed path i is obtained. The 'paths' are consequently defined as all possible steps to start the algorithm from the initial set of descriptors.
- It should be noticed that if the replacement of the descriptor with the largest error by those in the pool does not decrease the value of S, then that descriptor is not changed.

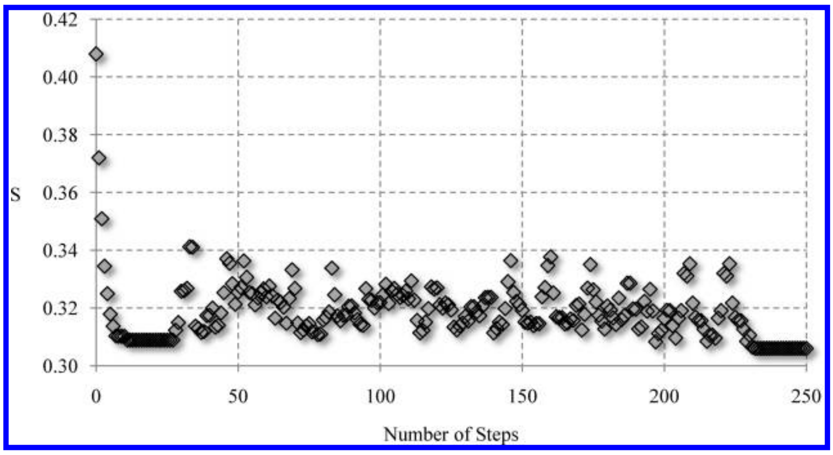


Figure 1. Standard deviation vs number of steps for the ERM.

- The above process is carried out for all the possible paths $i = 1, 2, ..., d$, and the point $\mathbf{d}_m$ with the smallest standard deviation: $\min_i S(\mathbf{d}_m^{(i)})$ is kept.

ERM. The ERM is a three step combination of two algorithms: first the RM already described above, then a modified RM (MRM), and finally a RM is used again. MRM follows the same strategy as RM except that in each step the descriptor with the largest error is substituted even if that substitution is not accompanied by a smaller value of S (the next smallest value of S is chosen). The main difference in MRM is that it adds some sort of noise that prevents the selected model to stay in a local minimum of S.[12]

NEW ALTERNATIVES

RM and ERM First Step Modification (RMfsm and ERMfsm). In both original algorithms the first step was taken without taking into account the relative standard deviation (rsd) of the coefficient of the descriptor in the model, instead all possible d paths were followed one at a time. This is because in the practical use of the algorithms, it was noticed that best results not always depended on the initial rsd of the path.[10−12,17,18,21−23,28−31] However there are different alternatives that may give better results.

In the new alternatives named RMfsm and ERMfsm, if only the path with higher rsd is chosen, then the computational cost is reduced by d times, but at the same time, the results will be poorer than selecting the best ones form all possible paths. In order to obtain algorithms that have the same computational cost and at the same time use only the initial descriptor, substitution with higher rsd is necessary to add $d - 1$ starting sets of descriptors.

RM Arbitrary First Step (RMafs). When using the first step modification, the improvement of the results may possibly come from the fact that several different starting sets are used instead of only one and not from the use of the path with higher rsd. For that reason a second alternative was proposed that is identical to the previous one with the only difference that the initial descriptor substitution is randomly chosen. This alternative was called RMafs.

RM and ERM with a Starting Set of High *S*. One may think that using a starting set of descriptors that present a very low S may be beneficial in further lowering it down. Nevertheless this may favor the algorithm to get trapped close to that starting point in a local minimum S. For that reason an option is to start the algorithms in a point with very high S far away from any such local minimum, hence having more chances to arrive to the global minimum. In order to do so, the point with high S was found using two options:

- Using an RM that maximizes S
- Using a forward stepwise regression (FSR) maximizing S

MATERIALS

Data Sets. Four different experimental data sets previously analyzed were used to test and contrast the performance of RM, ERM, and the new alternatives.

A fluorophilicity data set (FLUOR), consisting of 116 organic compounds characterized by 1268 theoretical descriptors, was used. The fluorophilicity of a each compound was quantified through the associated partition coefficient (P) between fluorous ($CF_3C_6F_{11}$) and organic ($CH_3C_6H_5$) layers:

$$ln\ P = ln\left[\frac{c(CF_3C_6F_{11})}{c(CH_3C_6H_5)}\right] \qquad T = 298\text{ K} \qquad (1.2)$$

The tendency of an organic substance to dissolve in fluorous media has continuously gained importance after the disclosure of the fluorous biphase catalysis, as biphasic reactions take advantage of the fact that organic and fluorous phases are typically immiscible at room temperature but may homogenize at elevated temperatures.[17]

A growth inhibition (GI) data set, with growth inhibition values to the ciliated protozoan *Tetrahymena pyriformis* by 200 mechanistically diverse phenolic compounds and 1338 structural descriptors. The aqueous toxicities are expressed as $pIGC_{50} = \log(IGC_{50}^{-1})$, with IGC_{50} expressing the concentration $[\text{mmol}\cdot\text{L}^{-1}]$ producing a 50% growth inhibition on *T. pyriformis* under a static regime.[29]

A GABA receptor data set (GABA) contains 78 inhibition data for flavone derivatives and 1187 molecular descriptors. The data set consists of the logarithm of the experimental binding affinity constants ($\log_{10} K_i$ [μM]) of flavonoid ligands for the benzodiazepine site of the GABA(A) receptor complex in washed crude synaptosomal membranes from a rat cerebral cortex.[30]

And finally a data set that consists of 100 $\log_{10}$ ED_{50} mice antiepileptic experimental activity values for enaminones with 1306 descriptors. The activity ED_{50} represents the dose of the chemical compound for which 50% of the individuals reach the desired effect obtained by the 'maximal electroshock seizure' (MES) experimental method.[32]

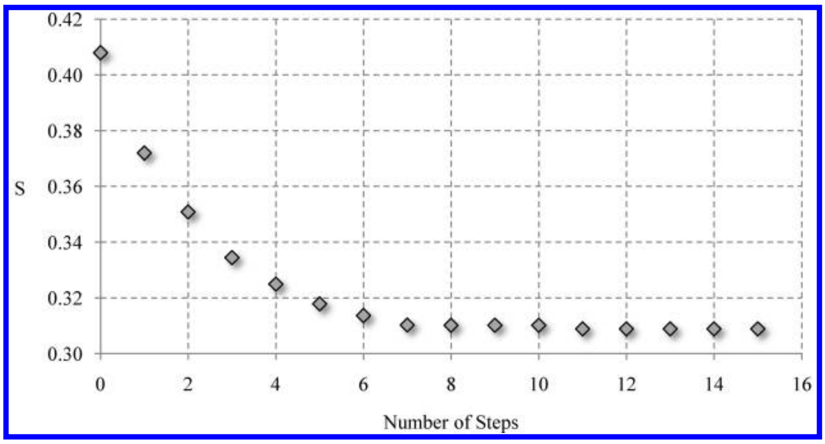


Figure 2. Standard deviation vs number of steps for the RM.

In all cases, the structures of the compounds were preoptimized with the molecular mechanics force field (MM+)[33] procedure included in Hyperchem version 6.03,[34] and the resulting geometries were further refined by means of the semi empirical method parametric method 3 (PM3)[35] using the Polak−Ribiere algorithm and a gradient norm limit of 0.01 kcal/Å. The molecular descriptors were calculated using the software Dragon 3.0,[36] including parameters of all types, such as constitutional, topological, geometrical, quantum mechanical, etc.

RESULTS AND DISCUSSION

With the purpose of providing a graphical visualization of the behavior of our two algorithms, Figures 1 and 2 show *S* as a function of the number of steps for both ERM and RM, respectively, and for the optimization of a seven-parameter model using the MES data set.[32] Figure 1 reveals that ERM has three sections: A first section where RM is performed, a second section that simulates a higher temperature or 'a higher noise' than the RM, although maintaining the overall decreasing tendency of the *S* function, and finally a third section where a second RM is used to further decrease *S*. This apparent thermal agitation makes the ERM less likely to get trapped in a local minimum.[12] The behavior of the new alternatives is similar to the one shown in Figures 1 and 2.

In order to compare the performance of RM against the alternative of using *d* initial sets following only the path with higher rsd (RMfsm), several numerical tests were carried out. The initial sets were chosen arbitrarily, with the precaution that the initial set used in RM is one of the *d* sets used in RMfsm. The comparisons were done for 100 different cases; 100 different initial sets for RM and an additional 600 sets for RMfsm were necessary in order to have the same computational cost.

Results were summarized in Tables 2 and 3, where it can be seen that RMfsm presented better results in all data sets for the same computational effort (number of initial sets on RMfsm = 7) with a statistically significant difference respect to RM, hence making it a preferable algorithm over RM. In addition, results using six, five and four initial sets in RMfsm were added to the tables. It can be appreciated that using six initial sets in RMfsm results are still better for all sets, nevertheless in some cases, the difference was not statistically significant. When using five and four initial sets in RMfsm, the results were in some cases better and some worse than RM, nevertheless in the cases that RM presented better results, the difference was not statistically significant. Hence the results of RMfsm with four and five initial sets have no statistical difference and less computational cost (4/7 and 5/7 times smaller, respectively) than RM, making it a more efficient algorithm.

Table 2. Number of Cases That *S* Is Lower Comparing RMfsm vs RM for 100 Cases Using the Four Data Sets[a]

	data set				
results	MES	GI	Fluor	GABA	total
			No. of Initial Sets on RMfs = 7		
RMfsm	**52**	**57**	**58**	**58**	**225**
RM	39	30	36	30	135
equal	9	13	6	12	40
			No. of Initial Sets on RMfs = 6		
RMfsm	**49**	**53**	**52**	**56**	**210**
RM	40	34	40	33	147
equal	11	13	8	11	43
			No. of Initial Sets on RMfs = 5		
RMfsm	41	**49**	44	**49**	**183**
RM	**48**	38	**46**	39	171
equal	11	13	10	12	46
			No. of Initial Sets on RMfs = 4		
RMfsm	33	**43**	38	41	155
RM	**55**	**43**	**52**	**48**	**198**
equal	12	14	10	11	47

[a] Bold face numbers indicate better results.

As mentioned, the enhancement in the performance of RMfsm could be a consequence of the increment of the number of initial sets that explore the solution space from different points and not only for using the path with higher rsd. Aiming to elucidate this, RMfsm was tested versus a similar algorithm that uses *d* initial sets, the first step is chosen randomly instead of the one with higher rsd (RMafs). The same 700 initial sets were use to get 100 different cases; the results are presented in Tables 4 and Table 5. It can be seen that even though RMfsm presented better results in most cases, indicating that it is preferable to chose the path with higher rsd over a random alternative, the difference was not very marked and had no statistical significance. This indicated that part of the improvement of RMfsm is because a higher number of initial sets are used, which is equivalent to

Table 3. Average Standard Deviation (S), Correlation Coefficient (R) of the Calibration, and Leave One Out (loo) Cross Validation Using RMfsm and RM for 100 cases of the Four Data Sets[a]

data set	algorithm	initial sets	S	R	S_{loo}	R_{loo}	S t-test pr.
MES	RMfsm	**7**	**0.3031**	**0.7464**	**0.3255**	**0.7014**	**0.0106**
		6	**0.3035**	**0.7456**	**0.3264**	**0.6995**	0.0518
		5	0.3045	0.7436	0.3275	0.6971	0.4963
		4	0.3053	0.7420	0.3280	0.6958	0.1247
	RM	7	0.3045	0.7436	0.3271	0.6978	
GI	RMfsm	**7**	**0.4493**	**0.8456**	**0.4710**	**0.8290**	**0.0004**
		6	**0.4500**	**0.8451**	**0.4721**	**0.0027**	**0.0027**
		5	**0.4509**	**0.8443**	**0.4734**	**0.8271**	**0.0342**
		4	**0.4522**	**0.8434**	**0.4748**	**0.8260**	0.2345
	RM	7	0.4530	0.8427	0.4755	0.8254	
Fluor	RMfsm	**7**	**0.4870**	**0.9812**	**0.5293**	**0.9777**	**0.0354**
		6	**0.4888**	**0.9811**	**0.5277**	**0.9779**	**0.1178**
		5	0.4927	0.9808	**0.5336**	**0.9774**	0.4289
		4	0.4951	0.9806	**0.5351**	**0.9772**	0.1622
	RM	7	0.4922	0.9808	0.5409	0.9766	
GABA	RMfsm	**7**	**0.4265**	**0.9188**	**0.4757**	**0.8982**	**0.0036**
		6	**0.4276**	**0.9184**	**0.4769**	**0.8977**	**0.0102**
		5	0.4324	0.9164	0.4823	0.8952	0.2985
		4	0.4360	0.9149	0.4861	0.8934	0.2282
	RM	7	0.4339	0.9158	0.4854	0.8938	

[a] T-test probability that the difference between the means of S values using RMfsm with respect to RM are not statistically significant. Bold face numbers indicate better results except for the case of S t-test pr., which indicates statistical significance.

Table 4. Number of Cases That the S Is Lower Comparing RMfsm vs RMafs for 100 Cases Using the Four Data Sets[a]

	data set				
results	MES	GI	Fluor	GABA	total
RMfsm	**239**	**237**	246	**246**	**968**
RMafs	220	218	**250**	226	914
equal	240	244	203	227	914

[a] Bold face numbers indicate better results.

Table 5. Average Standard Deviation (S), Correlation Coefficient (R) of the Calibration, and Leave One Out (loo) Cross Validation Using RMfsm and RMafs for 100 Cases Using the Four Data Sets[a]

data set	algorithm	S	R	S_{loo}	R_{loo}	S t-test pr.
MES	RMfsm	0.3064	0.7397	**0.3290**	**0.6938**	0.4284
	RMafs	**0.3062**	**0.7403**	0.3298	0.6917	
GI	RMfsm	**0.4537**	**0.8422**	**0.4758**	**0.8251**	0.1030
	RMafs	0.4554	0.8409	0.4775	0.8237	
Fluor	RMfsm	**0.4992**	**0.9802**	**0.5429**	**0.9765**	0.2214
	RMafs	0.5015	0.9801	0.5504	0.9759	
GABA	RMfsm	**0.4399**	**0.9133**	**0.43990**	**0.8911**	0.3102
	RMafs	0.4413	0.9128	0.8907	0.8907	

[a] T-test probability that the difference between the means of S values using RMfsm with respect to RMafs are not statistically significant. Bold face numbers indicate better results.

Table 6. Number of Cases That the S Is Lower Comparing ERMfs vs ERM for 100 Cases Using the Four Data Sets[a]

	data set				
results	MES	GI	Fluor	GABA	total
	No. of Initial Sets on ERMfs = 7				
ERMfsm	**53**	**56**	**51**	**47**	**207**
ERM	33	25	27	30	115
equal	14	19	22	23	78
	No. of Initial Sets on ERMfs = 6				
ERMfsm	**50**	**50**	**46**	**45**	**191**
ERM	35	32	34	34	135
equal	15	18	20	21	74
	No. of Initial Sets on ERMfs = 5				
ERMfsm	**44**	**46**	**42**	**40**	**172**
ERM	40	37	37	39	153
equal	16	17	21	21	75
	No. of Initial Sets on ERMfs = 4				
ERMfsm	40	38	34	37	149
ERM	**43**	**43**	**42**	**47**	**175**
equal	17	19	24	16	76

[a] Bold face numbers indicate better results.

exploring the solution space in different sections. Since, even without a statistical significance, RMfsm presented better results than RMafs, it is recommended to use RMfsm over RMafs.

In the case of the ERM, the new alternative that uses d initial sets (ERMfsm) in principle could result in a lower improvement, since ERM has a lower dependence on the initial set of descriptors. Numerical tests were carried in a similar approach, 100 cases for ERM and 600 additional cases for ERMfsm using all four the data sets and $d = 7$, the results were presented in Tables 6 and 7. In the tables, the results using six, five, and four initial sets for ERMfsm are also shown. It can be appreciated that ERMfsm presented better results than ERM for the case of equal computational effort (seven initial sets on ERMfsm), with a statistically significant difference with respect to ERM. This implicates that ERMfsm is an algorithm that is even a more efficient than ERM presetting significant additional improvements. Results were also better for the lower computational cost cases using six and five initial sets on ERMfsm; nevertheless in some cases, the difference was not statistically significant. Only when the number of initial sets was lowered to four ERM offered better results in some of the data sets, but the difference was not significant in any of the statistical tests. Hence ERMfsm with four initial presented results with no statistical difference and less computational cost (4/7 times smaller) than ERM, making it a more efficient algorithm.

The last proposed alternative is a different approach that aims to find an optimal starting set with S as high as possible in order to try to avoid local S minimum. In this case, three initial alternatives on ERM were compared: The first alternative uses three random initial sets of descriptors, the second uses an initial set of maximum S found by RM, and the third one uses an initial set of maximum S found by FSR. The results for the four previously

Table 7. Average Standard Deviation (S), Correlation Coefficient (R) of the Calibration, and Leave One Out (loo) Cross Validation Using ERMfsm and ERM for 100 Cases Using the Four Data Sets[a]

data set	algorithm	initial sets	S	R	S_{loo}	R_{loo}	S t-test pr.
MES	ERMfsm	**7**	**0.2942**	**0.7633**	**0.3182**	**0.7174**	**0.0470**
		6	**0.2946**	**0.7626**	**0.3187**	**0.7163**	0.0825
		5	**0.2950**	**0.7618**	**0.3189**	**0.7160**	0.1358
		4	**0.2955**	**0.7610**	**0.3192**	**0.7153**	0.2240
	ERM	7	0.2965	0.7576	0.3202	0.7107	
GI	ERMfsm	**7**	**0.4408**	**0.8520**	**0.4620**	**0.8361**	**0.0001**
		6	**0.4410**	**0.8518**	**0.4623**	**0.8359**	**0.0007**
		5	**0.4413**	**0.8516**	**0.4626**	**0.8356**	**0.0061**
		4	**0.4420**	**0.8511**	**0.4633**	**0.8351**	0.1991
	ERM	7	0.4424	0.8508	0.4635	0.8349	
Fluor	ERMfsm	**7**	**0.4411**	**0.9846**	**0.4714**	**0.9824**	**0.0001**
		6	**0.4424**	**0.9845**	**0.4735**	**0.9823**	**0.0013**
		5	**0.4439**	**0.9844**	**0.4752**	**0.9821**	**0.0178**
		4	0.4481	0.9841	0.4865	0.9811	0.4253
	ERM	7	0.4477	0.9841	0.4799	0.9818	
GABA	ERMfsm	**7**	**0.3979**	**0.9299**	**0.4451**	**0.9116**	**0.0030**
		6	**0.3989**	**0.9295**	**0.4461**	**0.0414**	**0.0414**
		5	**0.4006**	**0.9288**	**0.4486**	**0.9101**	0.3534
		4	0.4020	0.9283	0.4502	0.9094	0.2616
	ERM	7	0.4011	0.9286	0.4487	0.9100	

[a] T-test probability that the difference between the means of S values using ERMfsm with respect to ERM are not statistically significant. Bold face numbers indicate better results except for the case of S t-test pr., which indicates statistical significance.

Table 8. Standard Deviation (S) Found Using ERM with Different Starting Alternatives[a]

	initial set of descriptors		
database	random	S max. (RM)	S max. (FSR)
MES	**0.2896**	0.2933	0.2950
GI	**0.4367**	0.4421	0.4421
GABA	0.3961	**0.3929**	**0.3929**
FLUOR	**0.4328**	**0.4328**	0.4831

[a] Bold face numbers indicate better results.

mentioned databases are presented in Table 8. The results indicate that the best results were found using random initial sets of descriptors. In this case, there were three different initial sets used, indicating that the use of several sets gives better results than the more sophisticated alternatives that maximize S as a starting point. This is in line with the previously mentioned results. Comparing the results obtained using RM and FSR to maximize S indicates that the use of RM is preferable, probably because RM finds a higher starting point and hence further away from any possible local minimum of S.

CONCLUSIONS

In this paper we presented three different improving alternatives in the first step of the previously developed RM and ERM. The best alternative for both cases turned out to be the using of a set of d different initial sets of descriptors and using as a first step replacement of the descriptor with a higher relative standard deviation. This new alternative (ERMfsm) makes ERM an even superior algorithm; ERM already had shown to give better results than the more elaborated genetic algorithms (GA). The RM alternative (RMfsm) improves the RM, which has proven to be slightly inferior to ERM, nevertheless its simplicity and lower computational demand of RM still make it an attractive methodology.

In both cases the new alternatives (ERMfsm and RMfsm) are more efficient, since they either give models better statistical values for the same computational work or show statistically similar results using a lower computational demand with respect to the older algorithms.

APENDIX A

In order to illustrate the presented alternatives we will apply them to the fluorophilicity data set (FLUOR), which consists of 116 organic compounds characterized by 1268 theoretical descriptors. We will show the first steps in obtaining an optimal model with $d = 7$ topological descriptors out of the pool of $D = 1268$, using ERM or RM.

We arbitrarily choose the initial set $\mathbf{d} = \{X_1, X_2, X_3, X_4, X_5, X_6, X_7\}$, which yields $S_{(0)} = 0.771$ and has relative errors for the regression coefficients {90.12, 38.95, 59.59, 20.36, **194.94**, 84.91, 50.2}, respectively, to the previous list. Normally we would start RM or ERM by taking the first path and hence replacing the first descriptor X_1 by the rest of the available descriptors and by keeping the one which gives a lowest S. Of all of the 1261 $(D - d)$ variables, the substitution that minimizes S is (X_1, X_{1068}), yielding $S(1) = 0.689$ and relative errors for the regression coefficients {18.58, 41.89, 67.67, 15.74, 66.44, **796.66**, 35.89}. We now replace the variable with the greatest relative error X_6 with all the 1261descriptors (X_{1068} is now out of the descriptor pool, and X_1 is in it) and find that the substitution (X_6, X_{40}) yields the smallest standard deviation $S(2) = 0.634$. This process continues until the optimal set is found for path 1 and repeated in this case 7 times starting with all the rest of available paths.

In the presented alternatives the first path will be 5 since X_5 is the descriptor with higher relative error. Once the optimal set is found, no other path is used, instead 6 more initial arbitrary sets are used in the same mode.

ASSOCIATED CONTENT

S Supporting Information. The descriptors matrixes of the used data sets along with their property vector and descriptor name string are available. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: amercader@inifta.unlp.edu.ar or andrewgmercader@gmail.com. Telephone: (+54)(221)425-7430.

ACKNOWLEDGMENT

The authors want to thank the National Research Council of Argentina (CONICET, project PIP11220100100151) and Universidad de Buenos Aires (UBA) for financial support. A.G.M.

thanks CONICET for a Post-Doctoral Research Fellowship in PRALIB (CONICET, UBA), on leave from INIFTA (CONICET, UNLP). P.R.D. thanks the National University of La Plata.

■ REFERENCES

(1) Hansch, C.; Leo, A. *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington, D.C., 1995.

(2) Shao, L.; Wu, L.; Fan, X.; Cheng, Y. Consensus Ranking Approach to Understanding the Underlying Mechanism With QSAR. *J. Chem. Inf. Model.* **2010**, *50*, 1941–1948.

(3) Wassermann, A. M.; Nisius, B.; Vogt, M.; Bajorath Identification of Descriptors Capturing Compound Class-Specific Features by Mutual Information Analysis. *J. Chem. Inf. Model.* **2010**, *50*, 1935–1940.

(4) Yu, H.; Kühne, R.; Ebert, R.-U.; Schüürmann, G. Comparative Analysis of QSAR Models for Predicting pKa of Organic Oxygen Acids and Nitrogen Bases from Molecular Structure. *J. Chem. Inf. Model.* **2010**, *50*, 1949–1960.

(5) Helgee, E. A.; Carlsson, L.; Boyer, S.; Norinder, U. Evaluation of Quantitative Structure-Activity Relationship Modeling Strategies: Local and Global Models. *J. Chem. Inf. Model.* **2010**, *50*, 677–689.

(6) Agarwal, S.; Dugar, D.; Sengupta, S. Ranking Chemical Structures for Drug Discovery: A New Machine Learning Approach. *J. Chem. Inf. Model.* **2010**, *50*, 716–731.

(7) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley VCH: Weinheim, Germany, 2000.

(8) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. QSPR: The Correlation and Quantitative Prediction of Chemical and Physical Properties from Structure. *Chem. Rev. Soc.* **1995**, *24*, 279–287.

(9) Trinajstic, N. *Chemical Graph Theory*; CRC Press: Boca Raton, FL, 1992.

(10) Duchowicz, P. R.; Castro, E. A.; Fernández, F. M.; González, M. P. A New Search Algorithm of QSPR/QSAR Theories: Normal Boiling Points of Some Organic Molecules. *Chem. Phys. Lett.* **2005**, *412*, 376–380.

(11) Duchowicz, P. R.; Castro, E. A.; Fernández, F. M. Alternative Algorithm for the Search of an Optimal Set of Descriptors in QSAR-QSPR Studies. *MATCH Commun. Math. Comput. Chem.* **2006**, *55*, 179–192.

(12) Mercader, A. G.; Duchowicz, P. R.; Fernandez, F. M.; Castro, E. A. Modified and enhanced replacement method for the selection of molecular descriptors in QSAR and QSPR theories. *Chemom. Intell. Lab. Syst.* **2008**, *92*, 138–144.

(13) Draper, N. R.; Smith, H. *Applied Regression Analysis*; John Wiley&Sons: New York, 1981.

(14) So, S.-S.; Karplus, M. Evolutionary Optimization in Quantitative Structure-Activity Relationship: An Application of Genetic Neural Networks. *J. Med. Chem.* **1996**, *39*, 1521–1530.

(15) Mercader, A. G.; Duchowicz, P. R.; Fernández, F. M.; Castro, E. A. Replacement Method and Enhanced Replacement Method Versus the Genetic Algorithm Approach for the Selection of Molecular Descriptors in QSPR/QSAR Theories. *J. Chem. Inf. Model* **2010**, *50*, 1542–1548.

(16) Helguera, A. M.; Duchowicz, P. R.; Pérez, M. A. C.; Castro, E. A.; Cordeiro, M. N. D. S.; González, M. P. Application of the Replacement Method as Novel Variable Selection Strategy in QSAR. 1. Carcinogenic Potential. *Chemometr. Intell. Lab.* **2006**, *81*, 180–187.

(17) Mercader, A. G.; Duchowicz, P. R.; Sanservino, M. A.; Fernandez, F. M.; Castro, E. A. QSPR analysis of fluorophilicity for organic compounds. *J. Fluorine Chem.* **2007**, *128*, 484–492.

(18) Duchowicz, P. R.; González, M. P.; Helguera, A. M.; Cordeiro, M. N. D. S.; Castro, E. A. Application of the Replacement Method as Novel Variable Selection in QSPR. 2. Soil Sorption Coefficients. *Chemom. Intell. Lab. Syst.* **2007**, *88*, 197–203.

(19) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **1953**, *21*, 1087–1092.

(20) Kirkpatrick, S.; Gelatt, C. D., Jr.; Vecchi, M. P. Optimization by Simulated Annealing. *Science* **1983**, *220*, 671–680.

(21) Mercader, A. G.; Duchowicz, P. R.; Fernández, F. M.; Castro, E. A.; Bennardi, D. O.; Autino, J. C.; Romanelli, G. P. QSAR prediction of inhibition of aldose reductase for flavonoids. *Bioorg. Med. Chem.* **2008**, *16*, 7470–7476.

(22) Mercader, A. G.; Duchowicz, P. R.; Fernández, F. M.; Castro, E. A.; Cabrerizo, F. M.; Thomas, A. H. Predictive Modeling of the Total Deactivation Rate Constant of Singlet Oxygen by Heterocyclic Compounds. *J. Mol. Graphics Modell.* **2009**, *28*, 12–19.

(23) Mercader, A. G.; Duchowicz, P. R.; Fernández, F. M.; Castro, E. A.; Wolcan, E. QSPR Study of solvent quenching of the $^5D_0$ -> $^7F_2$ emission of Eu(6,6,7,7,8,8,8-heptafluoro-2,2-dimethyl-3,5-octanedionate)$_3$. *Chem. Phys. Lett.* **2008**, *462*, 352–357.

(24) Abdi, H. Partial least squares regression and projection on latent structure regression (PLS Regression). *Wiley Interdisciplinary Reviews: Computational Statistics* **2010**, *2*, 97–106.

(25) *Matlab 5.0*; The MathWorks Inc.: Natick, MA; http://www.mathworks.com/. Accessed May 23, 2011.

(26) Hawkins, D. M.; Basak, S. C.; Mills, D. Assessing model fit by cross-validation. *J. Chem. Inf. Model.* **2003**, *43*, 579–586.

(27) Edgell, S. E.; Noon, S. M. Effect of violation of normality on the t test of the correlation coefficient. *Psych. Bull.* **1984**, *95*, 576–583.

(28) Duchowicz, P. R.; Fernández, M.; Caballero, J.; Castro, E. A.; Fernández, F. M. QSAR of Non-Nucleoside Inhibitors of HIV-1 Reverse Transcriptase. *Bioorg. Med. Chem.* **2006**, *14*, 5876–5889.

(29) Duchowicz, P. R.; Mercader, A. G.; Fernández, F. M.; Castro, E. A. Prediction of aqueous toxicity for heterogeneous phenol derivatives by QSAR. *Chemom. Intell. Lab. Syst.* **2008**, *90*, 97–107.

(30) Duchowicz, P. R.; Vitale, M. G.; Castro, E. A.; Autino, J. C.; Romanelli, G. P.; Bennardi, D. O. QSAR Modeling of the Interaction of Flavonoids with GABA(A) Receptor. *Eur. J. Med. Chem.* **2007**, *43*, 1593–1602.

(31) Duchowicz, P. R. F.; Caballero, M.; Castro, J.; Fernández, E. A. F. M. QSAR of Non-Nucleoside Inhibitors of HIV-1 Reverse Transcriptase. *Bioorg. Med. Chem.* **2006**, *14*, 5876–5889.

(32) Garro Martinez, J. C.; Duchowicz, P. R.; Estrada, M. R.; Zamarbide, G. N.; Castro, E. A. Anticonvulsant Activity of Ringed Enaminones: A QSAR Study. *QSAR Comb. Sci.* **2009**, *28*, 1376–1385.

(33) Allinger, N. L.; Yuh, Y. H.; Lii, J. H. Molecular mechanics. The MM3 force field for hydrocarbons. *J. Am. Chem. Soc.* **1989**, *111*, 8551–8566.

(34) *HYPERCHEM 6.03*; Hypercube, Inc.: Gainesville, FL; http://www.hyper.com. Accessed May 23, 2011.

(35) Stewart, J. J. P. Optimization of parameters for semiempirical methods I. *J. Comput. Chem.* **1989**, *10*, 209–220.

(36) *DRAGON*, release 5.0 Evaluation Version; Milano Chemometrics and QSAR Research Group: Milano, Italy; http://michem.disat.unimib.it/chm. Accessed May 23, 2011.