JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

Two New Parameters Based on Distances in a Receiver Operating Characteristic Chart for the Selection of Classification Models

Alfonso Pérez-Garrido,[*,†,‡] Aliuska Morales Helguera,[¶,§,∥] Fernanda Borges,[∥] M. Natália D. S. Cordeiro,[⊥] Virginia Rivero,[#] and Amalio Garrido Escudero[†]

[†]Cátedra de Ingeniería y Toxicología Ambiental, Universidad Cátolica San Antonio, Guadalupe, Murcia, Spain

[‡]Departamento de Tecnologia de la Alimentacion y de la Nutricion, Universidad Cátolica San Antonio, Guadalupe, Murcia, Spain

[¶]Departamento de Química, Universidad Central Marta Abreu de Las Villas, Santa Clara, 54830 Villa Clara, Cuba

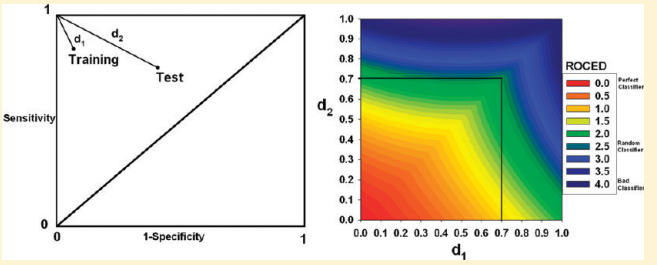
[§]CBQ, Universidad Central Marta Abreu de Las Villas, Santa Clara, 54830 Villa Clara, Cuba

[∥]CIQUP, Departamento de Química e Bioquímica, Faculdade de Ciências, 4169-007, Universidade do Porto, Porto, Portugal

[⊥]REQUIMTE, Departamento de Química e Bioquímica, Faculdade de Ciências, 4169-007, Universidade do Porto, Porto, Portugal

[#]Departmento de Química, Universidad de Vigo, 36310, Spain

ABSTRACT: There are several indices that provide an indication of different types on the performance of QSAR classification models, being the area under a Receiver Operating Characteristic (ROC) curve still the most powerful test to overall assess such performance. All ROC related parameters can be calculated for both the training and test sets, but, nevertheless, neither of them constitutes an absolute indicator of the classification performance by themselves. Moreover, one of the biggest drawbacks is the computing time needed to obtain the area under the ROC curve, which naturally slows down any calculation algorithm. The present study proposes two new parameters based on distances in a ROC curve for the selection of classification models with an appropriate balance in both training and test sets, namely the following: the ROC graph Euclidean distance (ROCED) and the ROC graph Euclidean distance corrected with Fitness Function (FIT(λ)) (ROCFIT). The behavior of these indices was observed through the study on the mutagenicity for four genotoxicity end points of a number of nonaromatic halogenated derivatives. It was found that the ROCED parameter gets a better balance between sensitivity and specificity for both the training and prediction sets than other indices such as the Matthews correlation coefficient, the Wilk's lambda, or parameters like the area under the ROC curve. However, when the ROCED parameter was used, the follow-on linear discriminant models showed the lower statistical significance. But the other parameter, ROCFIT, maintains the ROCED capabilities while improving the significance of the models due to the inclusion of FIT(λ).



INTRODUCTION

Traditionally, the techniques used to measure the performance of a discriminant analysis QSAR model are derived from Wilk's lambda or by indices based on the confusion matrix.[1−5] It is well-known that the choice of the Wilk's lambda-based model involves the same problems of predictive choosing the best regression model based only on the determination coefficient R^2.[6] Among the indices based on confusion matrix, the accuracy, sensitivity, specificity, precision, enrichment factor, and Matthews correlation coefficient should be remarked.[7] In the field of toxicology, Benigni et al.[8] quite successfully introduced the Receiver Operating Characteristic (ROC) chart where the true positive rate (or sensitivity) is plotted against the false positive rate (1-specificity). This chart has the advantage of comparing simultaneously the different aspects of the performance of several systems or models.[8] It has been observed too that ROC curves visually convey the same information as the confusion matrix in a much more intuitive and robust fashion.[9] The Area under the ROC curve (AUC) can be directly computed for any classification model that attaches a probability, like discriminant analysis[10−14] and is also widely used in many disciplines.[15−20] The AUC is the probability of active compounds being ranked earlier than decoy compounds, and it can take values between 1 (perfect classifiers) and 0.5 (useless random classifiers). This AUC metric parameter is not sensitive to *early recognition* (quickly ability to recognize positives). Truchon and Bayly[21] have discussed several methods to address this problem using the parameter named Boltzmann-Enhanced Discrimination of ROC (BEDROC) based on Robust Initial Enhancement (RIE),[22] which provided a good early recognition of actives. Recently, McGaughey et al.[23] used the Enrichment Factor (EF),[24] the AUC, the RIE and the BEDROC parameters to evaluate different virtual screening (VS) methods. The RIE and BEDROC

Received: July 5, 2011
Published: September 17, 2011

dx.doi.org/10.1021/ci2003076 | *J. Chem. Inf. Model.* 2011, 51, 2746–2759

parameters did not lead to dramatically different conclusions about the performance of VS methods; therefore, the AUC remained as the most powerful test.[23] Besides one of the biggest drawbacks is the computing time needed to obtain the AUC, which naturally slows down any calculation algorithm.

In addition, all these parameters can be calculated for both the training and test sets, but, nevertheless, none of them constitutes an absolute indicator of the classification performance by themselves. Also, if we only try to improve the performance for the training set, we might get some bad predictions, and, on the other hand, if all our efforts are used to obtain good predictions, these predictions could come from poorly trained models (unreal situation). In the present study, we propose two new parameters to achieve a good balance of classification for both the training and test sets. Furthermore, the involved procedure is not computational expensive. To assess the performance of the new proposed parameters, we choose the mutagenicity of nonaromatic halogenated derivatives because they represent an environmentally important family of compounds derived from High Production Volume lists (HPV) of chemicals in Europe.[25] To the best of our knowledge, there are no mutagenicity QSAR studies for this family of compounds, employing end points included among recognized regulatory methods like bacterial mutagenesis, *in vivo* micronucleous, *in vitro* chromosome aberration, and *in vitro* mammalian cell gene mutation tests.

MATERIALS AND METHODS

Data Set. The chosen data set contains a number of nonaromatic halogenated derivatives with Bacterial mutagenesis (B), *in vitro* Chromosome Aberration (CA), *in vivo* Micronucleus (M), and *in vitro* Mammalian Cell Gene mutation (MCG) data obtained from the Food and Drug Administration (FDA) SAR Genetox Database (www.leadscope.com - accessed Aug 20, 2011). The substances were classified as positive for each test if the result from a single test (strain, cell, etc.) was positive and were classified as negative if they did not show mutagenicity in the strains. To develop the classification functions, values of 1 and −1 were assigned to mutagenic and nonmutagenic substances, respectively. The structures of the molecules were downloaded from the application in sdf format followed by a fully optimization with the quantum-mechanics semiempirical Parametric Method Number 3 (PM3) method implemented in MOE 8.0.[26] A total of 3314 descriptors were calculated, specifically: 2340 with the DRAGON package[27] (further divided into several families), 655 with the Modeslab software (http://www.modeslab.com - accessed Aug 20, 2011),[28] and 319 with the MOE 8.0[26] and MOPAC 7.1[29] programs. Variables with constant or close to constant values were deleted. The data set included 536 substances (360 mutagenic and 176 nonmutagenic) for bacterial mutagenesis, 88 (66 mutagenic and 22 nonmutagenic) for *in vitro* mammalian mutagenesis, 124 (80 mutagenic and 44 nonmutagenic) for *in vitro* chromosome aberration, and 70 (23 mutagenic and 47 nonmutagenic) for *in vivo* micronucleus. Whenever applying non three-dimensional families of descriptors, stereoisomers were eliminated to avoid duplication of information because they would get identical values. Each group of stereoisomers was represented by a single substance. Therefore, the number of substances pertaining to those non-3D descriptors was 519 (351 mutagenic and 168 nonmutagenic) for bacterial mutagenesis and 123 for *in vitro* chromosome aberration (80 mutagenic and 43 nonmutagenic). As usually, the data should be divided into training and test sets to obtain validated QSAR models. In this work, we have applied the k-Means Cluster Analysis (k-MCA) technique toward designing both training and test sets which were representative of the entire *experimental universe*.

k-Means Cluster Analysis. A cluster analysis was first performed to divide the series of compounds into several statistically representative classes of chemicals and to then select the training and test sets among the members of such classes. In the course of doing so, the training set ended up with 80% of the data set (429/536 compounds and 418/519 for bacterial mutagenesis data dealt with 3D and non-3D descriptors, respectively; 71/88 for the *in vitro* mammalian mutagenesis data; 98/124 and 97/123 for the *in vitro* chromosome aberration data dealt with 3D and non-3D descriptors, respectively; and 57/70 for the *in vivo* micronucleus data), while the test set with the remaining 20% (107/536 compounds and 101/519 for the bacterial mutagenesis data dealt with 3D and non-3D descriptors, respectively; 17/88 for the *in vitro* mammalian mutagenesis data; 26/124 for the *in vitro* chromosome aberration data; and 13/70 for the *in vivo* micronucleus data). This procedure ensures that any kind of substance, as determined by the clusters derived from k-MCA, will be represented in each series of compounds (training and prediction series).

A Principal Component (PC) analysis was used to reduce the list of descriptors into two PCs for each family. That is, we have used the PCs of each family of descriptors which produces the greater separation of clusters, ensuring a statistically acceptable data partition. The number of members in each cluster and the standard deviation of the PCs in the cluster (as low as possible) was taken into account. For the *in vitro* chromosome aberration data, the k-MCA split the mutagenic compounds into five clusters comprising 11, 13, 13, 20, and 23 members with standard deviations of 0.09, 0.07, 0.07, 0.08, and 0.07, respectively, and the nonmutagenic compounds into four clusters comprising 16, 8, 10, and 9 members with standard deviations of 0.19, 0.16, 0.09 and 0.06, respectively. For the *in vitro* mammalian mutagenesis data, the k-MCA split the mutagenic compounds into five clusters comprising 14, 18, 17, 11, and 6 members with standard deviations of 0.05, 0.06, 0.08, 0.05, and 0.03, respectively, and the nonmutagenic compounds into two clusters comprising 8 and 14 members with standard deviations of 0.50 and 0.27, respectively. For the bacterial mutagenesis data, the k-MCA split the mutagenic compounds into five clusters comprising 117, 100, 35, 48, and 60 members with standard deviations of 0.14, 0.14, 0.21, 0.15, and 0.23, respectively, and the nonmutagenic compounds into seven clusters comprising 9, 32, 27, 37, 36, 17, and 18 members with standard deviations of 0.39, 0.20, 0.12, 0.07, 0.06, and 0.06, respectively. For the *in vivo* micronucleus data, the k-MCA split the mutagenic compounds into three clusters comprising 6, 6, and 11 members with standard deviations of 0.16, 0.14 and 0.08, respectively, and the nonmutagenic compounds into four clusters comprising 9, 9, 13, and 16 members with standard deviations of 0.22, 0.12, 0.04, and 0.04, respectively. Selection of the training and prediction sets was then carried out taking, proportionally to size of the cluster, the compounds belonging to each cluster.

An inspection of the standard deviation between and within the clusters, the respective Fisher ratio and their p level of significance (forced to be lower than 0.05) was also done.[30,31]

The ROC Graph Euclidean Distance (ROCED). The predictive value of any QSAR in toxicology is usually measured by their representation in a ROC chart,[8] which has the advantage of comparing simultaneously different aspects of the model's

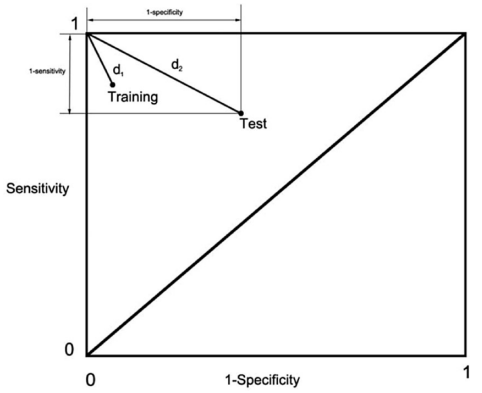


Figure 1. Representation of the distances in a ROC graph for the training (d_1) and for the test set (d_2).

performance. For example, the accuracy index alone does not distinguish between positive and negative predictions, and it is influenced by the performance on the most populated class, whereas the axes of the ROC graphs display independently the information relative to the prediction of positive and negative chemicals. In a ROC plot, the true positive rate (or sensitivity) is plotted against the false positive rate (1-specificity). According to the ROC curve theory, the diagonal line corresponds to random responses, whereas the top left corner is obviously the ideal performance. Thus, the most finely tuned models are those whose points fall in the left upper triangle, as close as possible to the corner.[8] Considering this, we decided to develop a new potentially useful parameter for optimizing the predictive models.

If the best model is one whose representation is located as close as possible to the upper left corner in the ROC chart, a good indicator would be a measure of this distance. The Euclidean distance between the perfect and a real classifier (d_i) expressed as a function of their respective values of sensitivity and specificity is

$$d_i = \sqrt{(Se_p - Se_r)^2 + (Sp_p - Sp_r)^2} \quad (1)$$

where Se_p and Se_r are the respective sensitivity values of the perfect and the real classifier, while Sp_p and Sp_r represent the specificity values of the perfect and real classifier, respectively. Since the sensitivity and specificity for a perfect classifier takes values of 1, then the euclidean distance can be expressed as

$$d_i = \sqrt{(1 - Se_r)^2 + (1 - Sp_r)^2} \quad (2)$$

where $i = 1$ stands for the training set, and $i = 2$ for the test set (see Figure 1).

Since we are concerned that these two distances corresponding to the training and test sets are as small as possible, we can define a parameter that relates as follows

$$\text{ROCED} = (|d_1 - d_2| + 1)(d_1 + d_2)(d_2 + 1) \quad (3)$$

where representation of the distances in a ROC graph for the training (d_1) and for the test set (d_2). Therefore minimizing the value of the latter parameter, we achieve three goals:

1 The obtained model has a similar accuracy for the training and test series, first factor;
2 Both training and test sets have ratings close to perfection (AUC = 1), second factor;

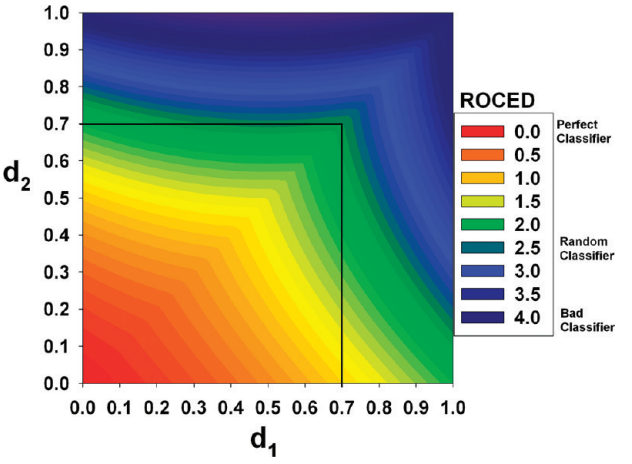


Figure 2. Contour graph representation of distances d_1, d_2, and ROCED values. ROCED = 0 represent a perfect classifier, and ROCED > 2.5 represent a random classifier.

3 A maximum accuracy on the test set (less distance to a perfect classifier), last factor.

This parameter can take values between 0 (perfect classifier for both training and test set) and 4.5 ($d_1 = 0.5$ random classifier and $d_2 = 1$). Values greater than 2.5 should not be considered as this means that some of the two distances (training or test) has a value greater than 0.7 which places it closed to the diagonal line in a ROC graph, which indicates that these models have random responses (Figure 2).

It would be very interesting to have a predictive model for the vast majority of chemicals, and so we have penalized the distance d_2 for substances outside of the applicability domain of the model. As we know that for the test set

$$1 - \text{Sensitivity} = \frac{FN}{P} \quad (4)$$

$$1 - \text{Specificity} = \frac{FP}{N} \quad (5)$$

where FN stands for false negatives, FP stands for false positives, P stands for the number of positive substances, and N stands for the number of negative substances. In an attempt to include more substances within the applicability domain of the model, we have added to the false positive substances the positive compounds outside the applicability domain and to the false negative substances negative compounds outside the applicability domain. In so doing, we recalculate the distance d_2 as

$$d_{2-rec} = \sqrt{\left(\frac{FN_{ED} + EP}{P}\right)^2 + \left(\frac{FP_{ED} + EN}{N}\right)^2} \quad (6)$$

where FN_{ED} are positive substances classified as negative in the domain (false negatives), EP are the positive substances outside of the applicability domain (excluded positives), P are the total positive substances; FP_{ED} are negative substances classified as positive in the domain (false positives), EN are the negative substances outside of the applicability domain (excluded negatives), and N are the total negative substances. In all cases the calculation is based on a priori probability of 0.5. Details about the construction of the applicability domain will be seen in the next subsection.

Table 1. Best Six Models Sorted by Wilk's Lambda for Every Family of Descriptors and Mutagenic End Points

								training									test											
end point	ID	size	ROCED	FIT	ROCFIT	N[a]	λ	AUC	MCC	Se	Sp	Ac	EF	Pr[b]	RIE	BEDROC	AUC	MCC	Se	Sp	Ac	Se[c]	Sp[c]	Ac[c]	E[d]	EF	Pr[b]	RIE	BEDROC
CA	DRAGON	10	2.66	0.51	5.21	0	0.46	0.94	0.62	0.81	0.83	0.82	1.39	0.81	1.56	1.00	0.65	−0.02	0.65	0.33	0.54	0.65	0.33	0.54	0	0.99	0.65	1.52	1.00
	MOE+MOPAC	10	1.95	0.43	4.56	1	0.51	0.93	0.69	0.89	0.80	0.86	1.38	0.89	1.55	1.00	0.64	0.26	0.71	0.56	0.65	0.71	0.50	0.64	1	1.15	0.75	1.13	0.74
	binary fingerprints	10	1.34	0.42	3.16	0	0.51	0.91	0.64	0.81	0.85	0.82	1.40	0.81	1.54	1.00	0.60	0.30	0.65	0.67	0.65	0.69	0.67	0.68	1	1.20	0.65	0.69	0.45
	GETAWAY	10	1.87	0.42	4.46	1	0.51	0.92	0.62	0.86	0.77	0.83	1.35	0.87	1.56	1.00	0.67	0.22	0.76	0.44	0.65	0.81	0.44	0.68	1	1.10	0.72	0.71	0.46
	3DMoRSE	10	2.66	0.35	7.50	0	0.55	0.89	0.63	0.84	0.80	0.83	1.37	0.84	1.50	0.96	0.73	−0.02	0.65	0.33	0.54	0.65	0.33	0.54	0	0.99	0.65	1.52	0.99
	atom-centered fragments	10	1.92	0.29	6.70	3	0.61	0.85	0.61	0.90	0.68	0.82	1.29	0.90	1.53	0.99	0.58	0.20	0.65	0.56	0.62	0.67	0.56	0.63	2	1.12	0.65	1.47	0.96
B	DRAGON	10	1.72	0.54	3.18	0	0.59	0.87	0.59	0.89	0.70	0.82	1.27	0.89	1.42	0.95	0.75	0.40	0.86	0.51	0.75	0.87	0.50	0.75	5	1.17	0.86	1.27	0.85
	MOE+MOPAC	10	1.79	0.50	3.62	0	0.61	0.85	0.58	0.88	0.69	0.82	1.27	0.85	1.40	0.94	0.71	0.25	0.74	0.51	0.66	0.78	0.50	0.69	5	1.13	0.76	1.38	0.93
	TOPS-MODE	10	2.09	0.48	4.37	0	0.62	0.84	0.57	0.91	0.64	0.82	1.24	0.84	1.41	0.94	0.64	0.30	0.86	0.42	0.72	0.85	0.43	0.73	5	1.11	0.77	0.99	0.68
	atom-centered	10	2.18	0.45	4.83	1	0.64	0.83	0.53	0.88	0.63	0.80	1.23	0.83	1.43	0.96	0.66	0.39	0.91	0.42	0.76	0.91	0.41	0.77	5	1.13	0.78	0.97	0.67
	functional groups count	10	2.13	0.41	5.25	1	0.66	0.82	0.53	0.91	0.57	0.80	1.21	0.81	1.41	0.95	0.65	0.28	0.89	0.35	0.72	0.88	0.32	0.72	4	1.09	0.76	1.05	0.73
	frequency fingerprints	10	1.71	0.40	4.23	0	0.66	0.81	0.51	0.89	0.59	0.79	1.22	0.82	1.43	0.96	0.71	0.44	0.87	0.55	0.77	0.89	0.54	0.78	10	1.17	0.81	1.23	0.85
MCG	DRAGON	10	1.27	1.23	1.04	0	0.22	0.99	0.93	1.00	0.89	0.97	1.29	1.00	1.34	1.00	0.73	0.66	1.00	0.50	0.88	1.00	0.50	0.88	0	1.13	1.00	1.03	0.78
	MOE+MOPAC	10	2.56	1.00	2.57	0	0.26	0.94	0.89	1.00	0.83	0.96	1.27	1.00	1.34	1.00	0.63	0.23	0.92	0.25	0.76	0.92	0.25	0.76	0	1.05	0.92	1.31	1.00
	GETAWAY	10	2.56	0.90	2.84	0	0.28	0.94	0.89	0.98	0.89	0.96	1.29	0.98	1.34	1.00	0.64	0.11	0.85	0.25	0.71	0.85	0.25	0.71	0	1.03	0.85	1.31	1.00
	frequency fingerprints	10	0.56	0.87	0.65	0	0.29	0.93	0.89	1.00	0.83	0.96	1.27	1.00	1.34	1.00	0.75	0.83	1.00	0.75	0.94	1.00	0.75	0.94	0	1.21	1.00	1.31	1.00
	3DMoRSE	10	2.62	0.86	3.05	1	0.29	0.94	0.85	1.00	0.78	0.94	1.25	1.00	1.34	1.00	0.70	0.23	0.92	0.25	0.76	0.92	0.25	0.76	0	1.05	0.92	1.31	1.00
	atom-centered fragments	10	1.48	0.78	1.88	2	0.31	0.92	0.81	0.98	0.78	0.93	1.24	0.98	1.34	1.00	0.67	0.55	0.85	0.75	0.82	0.85	0.67	0.81	1	1.20	0.85	1.03	0.78
M	DRAGON	10	1.16	0.95	1.23	0	0.24	1.00	1.00	1.00	1.00	1.00	3.00	1.00	3.00	1.00	0.67	0.28	0.75	0.56	0.62	0.75	0.56	0.62	0	1.39	0.75	0.67	0.21
	MOE+MOPAC	10	1.53	0.44	3.48	1	0.40	0.94	0.74	0.89	0.87	0.88	2.32	0.89	2.97	0.99	0.60	0.28	0.50	0.78	0.69	0.50	0.78	0.69	0	1.63	0.50	0.14	0.04
	GETAWAY	10	0.76	0.41	1.88	2	0.42	0.96	0.77	0.89	0.89	0.89	2.43	0.89	2.97	0.99	0.72	0.50	0.75	0.78	0.77	0.75	0.78	0.77	0	1.95	0.75	3.13	0.96
	3DMoRSE	10	0.87	0.40	2.14	3	0.42	0.93	0.64	0.74	0.89	0.84	2.33	0.74	2.97	0.99	0.71	0.50	0.75	0.78	0.77	0.75	0.78	0.77	0	1.95	0.75	3.11	0.96
	atom-centered fragments	10	3.06	0.37	8.33	1	0.44	0.95	0.74	0.89	0.87	0.88	2.32	0.89	2.99	1.00	0.50	−0.05	0.50	0.44	0.46	0.50	0.38	0.42	1	0.93	0.50	0.67	0.21
	geometrical	10	1.39	0.36	3.88	1	0.45	0.95	0.73	0.84	0.89	0.88	2.40	0.84	2.95	0.98	0.75	0.39	0.75	0.67	0.69	0.75	0.63	0.67	1	1.63	0.75	3.13	0.96

[a] Number of variables with a significance less than 0.05. [b] Attained precision. [c] Results obtained taking into account only those substances that were within the applicability domain of the models. [d] Number of substances outside the applicability domain of the models.

dx.doi.org/10.1021/ci2003076 |J. Chem. Inf. Model. 2011, 51, 2746–2759

Table 2. Best Six Model of Every Family of Descriptors and Every Mutagenic End Point Sorted by MCCtest

								training									test											
end point	ID	size	ROCED	FIT	ROCFIT	N[a]	λ	AUC	MCC	Se	Sp	Ac	EF	Pr[b]	RIE	BEDROC	AUC	MCC	Se	Sp	Ac	Se[c]	Sp[c]	Ac[c]	E[d]	EF	Pr[b]	RIE	BEDROC
CA	geometrical	6	1.20	0.02	68.71	6	0.97	0.57	0.09	0.75	0.34	0.60	1.04	0.75	1.00	0.64	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0	1.53	1.00	1.53	1.00
	topologics	8	1.31	0.06	22.34	8	0.90	0.66	0.18	0.60	0.59	0.60	1.13	0.60	1.37	0.89	0.85	0.83	1.00	0.78	0.92	1.00	0.78	0.92	0	1.37	1.00	1.45	0.95
	MOE+MOPAC	7	1.04	0.08	12.77	7	0.88	0.70	0.24	0.65	0.60	0.63	1.16	0.75	1.47	0.95	0.85	0.83	0.94	0.89	0.92	0.94	0.89	0.92	0	1.44	0.94	0.71	0.46
	TOPS-MODE	10	1.26	0.09	13.75	8	0.83	0.73	0.32	0.81	0.50	0.70	1.15	0.75	1.40	0.91	0.88	0.75	0.88	0.89	0.88	0.88	0.88	0.88	1	1.43	0.94	1.49	0.98
	binary fingerprints	9	0.72	0.19	3.87	6	0.73	0.81	0.48	0.73	0.76	0.74	1.31	0.73	1.50	0.98	0.83	0.75	0.88	0.89	0.88	0.88	0.89	0.88	0	1.43	0.88	1.15	0.75
	GETAWAY	7	1.36	0.08	16.52	7	0.88	0.65	0.29	0.75	0.54	0.67	1.16	0.75	0.91	0.58	0.71	0.75	1.00	0.67	0.88	1.00	0.67	0.88	0	1.30	0.85	0.96	0.63
B	binary fingerprints	9	1.34	0.23	5.83	4	0.78	0.74	0.42	0.85	0.55	0.75	1.18	0.80	1.37	0.92	0.74	0.56	0.93	0.58	0.82	0.93	0.58	0.82	1	1.20	0.83	1.41	0.97
	topologics	9	1.02	0.22	4.64	4	0.79	0.76	0.48	0.82	0.66	0.77	1.24	0.83	1.25	0.84	0.78	0.55	0.83	0.74	0.80	0.84	0.73	0.81	2	1.27	0.88	1.39	0.97
	GETAWAY	6	1.14	0.21	5.47	1	0.81	0.75	0.37	0.77	0.60	0.72	1.19	0.80	1.28	0.86	0.77	0.55	0.78	0.80	0.79	0.80	0.80	0.80	2	1.32	0.89	1.18	0.79
	2 Dautocorrelations	10	1.21	0.12	10.25	6	0.87	0.70	0.29	0.67	0.64	0.66	1.18	0.79	1.23	0.83	0.77	0.55	0.80	0.77	0.79	0.80	0.77	0.79	1	1.28	0.89	1.35	0.93
	atom-centered fragments	7	2.50	0.31	8.14	1	0.74	0.75	0.55	0.95	0.52	0.81	1.19	0.80	1.41	0.94	0.65	0.54	0.99	0.42	0.81	0.99	0.33	0.80	5	1.14	0.79	1.10	0.76
	MOE+MOPAC	9	1.06	0.38	2.83	1	0.69	0.81	0.48	0.83	0.65	0.77	1.24	0.83	1.42	0.95	0.74	0.54	0.83	0.71	0.79	0.83	0.71	0.79	1	1.27	0.83	1.23	0.83
MCG	TOPS-MODE	10	0.15	0.38	0.40	2	0.48	0.93	0.79	0.92	0.89	0.92	1.29	0.92	1.31	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0	1.31	1.00	1.31	1.00
	MOE+MOPAC	6	0.30	0.33	0.91	0	0.64	0.84	0.67	0.91	0.78	0.87	1.24	0.91	1.29	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0	1.31	1.00	1.31	1.00
	3DMoRSE	9	0.47	0.24	1.94	5	0.63	0.87	0.58	0.91	0.67	0.85	1.19	0.91	1.34	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0	1.31	1.00	1.31	1.00
	information	9	0.40	0.15	2.70	6	0.73	0.84	0.52	0.79	0.78	0.79	1.22	0.79	1.33	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0	1.31	1.00	1.31	1.00
	topologics	10	0.47	0.23	2.01	5	0.60	0.89	0.58	0.91	0.67	0.85	1.19	0.91	1.34	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0	1.31	1.00	1.31	1.00
	RDF	10	0.66	0.15	4.35	8	0.70	0.78	0.49	0.91	0.56	0.82	1.20	0.96	1.34	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0	1.13	1.00	1.31	1.00
M	MOE+MOPAC	8	0.48	0.13	3.59	7	0.75	0.82	0.53	0.68	0.84	0.79	2.05	0.68	2.73	0.91	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0	3.25	1.00	3.24	1.00
	GETAWAY	9	0.40	0.09	4.27	8	0.78	0.78	0.57	0.74	0.84	0.81	2.10	0.74	2.35	0.79	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0	3.25	1.00	3.24	1.00
	3DMoRSE	8	0.36	0.13	2.73	6	0.75	0.81	0.59	0.79	0.82	0.81	2.05	0.79	2.48	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0	3.25	1.00	3.24	1.00
	DRAGON	8	0.33	0.44	0.77	2	0.48	0.96	0.76	0.74	0.97	0.89	2.80	0.74	2.95	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0	3.25	1.00	3.24	1.00
	TOPS-MODE	7	0.97	0.09	10.88	6	0.84	0.73	0.28	0.58	0.71	0.67	1.50	0.58	2.40	0.80	0.93	0.84	1.00	0.89	0.92	1.00	0.89	0.92	0	2.60	1.00	2.70	0.83
	information	7	0.54	0.26	2.04	3	0.64	0.87	0.56	0.79	0.79	0.79	1.96	0.79	2.81	0.94	0.89	0.84	1.00	0.89	0.92	1.00	0.89	0.92	0	2.60	1.00	0.70	0.21

[a] Number of variables with a significance less than 0.05. [b] Attained precision. [c] Results obtained taking into account only those substances that were within the applicability domain of the models. [d] Number of substances outside the applicability domain of the models.

dx.doi.org/10.1021/ci2003076 |J. Chem. Inf. Model. 2011, 51, 2746–2759

Table 3. Best Six Models Sorted by the AUC of the Training Set for Every Family of Descriptors and Mutagenic End Points

								training									test											
end point	id	size	roced	fit	rocfit	n[a]	λ	AUC	MCC	Se	Sp	Ac	EF	Pr[b]	RIE	BEDROC	AUC	MCC	Se	Sp	Ac	Se[c]	Sp[c]	Ac[c]	E[d]	EF	Pr[b]	RIE	BEDROC
CA	DRAGON	10	3.29	0.51	6.46	0	0.46	0.93	0.63	0.84	0.80	0.83	1.37	0.84	1.56	1.00	0.65	−0.02	0.65	0.33	0.54	0.65	0.25	0.52	1	0.99	0.65	1.52	1.00
	binary fingerprints	10	1.34	0.42	3.16	0	0.51	0.91	0.64	0.81	0.85	0.82	1.40	0.81	1.54	1.00	0.60	0.30	0.65	0.67	0.65	0.69	0.67	0.68	1	1.20	0.65	0.69	0.45
	3DMoRSE	10	2.36	0.33	7.11	1	0.57	0.89	0.69	0.89	0.80	0.86	1.38	0.89	1.45	0.93	0.66	0.11	0.76	0.33	0.62	0.76	0.33	0.62	0	1.05	0.76	1.52	1.00
	MOE+MOPAC	8	0.76	0.32	2.35	0	0.63	0.89	0.74	0.89	0.86	0.88	1.43	0.89	1.47	0.95	0.74	0.52	0.76	0.78	0.77	0.76	0.78	0.77	0	1.33	0.76	1.13	0.74
	GETAWAY	10	2.64	0.34	7.75	0	0.56	0.89	0.67	0.87	0.80	0.85	1.38	0.89	1.54	0.99	0.59	0.09	0.65	0.44	0.58	0.69	0.38	0.58	2	1.05	0.69	1.35	0.88
	frequency fingerprints	10	2.90	0.28	10.51	2	0.61	0.88	0.57	0.62	0.97	0.74	1.50	0.62	1.54	1.00	0.54	0.13	0.35	0.78	0.50	0.31	0.75	0.46	2	1.15	0.35	1.43	0.94
B	MOE+MOPAC	10	1.56	0.48	3.29	0	0.62	0.86	0.59	0.88	0.70	0.82	1.28	0.86	1.33	0.89	0.67	0.30	0.76	0.54	0.69	0.80	0.56	0.72	7	1.15	0.77	1.12	0.75
	3DMoRSE	10	2.83	0.36	7.85	0	0.69	0.81	0.54	0.89	0.61	0.79	1.23	0.82	1.33	0.89	0.67	0.26	0.86	0.37	0.70	0.87	0.29	0.68	9	1.10	0.74	1.24	0.84
	frequency fingerprints	10	2.10	0.40	5.23	0	0.66	0.81	0.56	0.92	0.59	0.81	1.22	0.82	1.42	0.95	0.69	0.36	0.86	0.48	0.74	0.87	0.43	0.74	6	1.14	0.79	1.20	0.83
	2 Dautocorrelations	9	1.81	0.30	5.96	0	0.73	0.81	0.46	0.80	0.66	0.76	1.24	0.83	1.42	0.96	0.70	0.31	0.77	0.55	0.70	0.76	0.52	0.69	4	1.15	0.79	1.32	0.91
	constitutional	10	2.43	0.34	7.22	0	0.70	0.80	0.55	0.93	0.57	0.81	1.21	0.82	1.42	0.96	0.63	0.28	0.84	0.42	0.71	0.87	0.37	0.72	7	1.11	0.77	1.17	0.81
	information	8	1.55	0.33	4.76	0	0.72	0.80	0.51	0.83	0.68	0.78	1.25	0.84	1.34	0.90	0.71	0.28	0.74	0.55	0.68	0.74	0.55	0.68	0	1.14	0.79	1.38	0.96
MCG	2 Dautocorrelations	10	1.37	0.63	2.19	2	0.36	0.98	0.85	0.96	0.89	0.94	1.19	0.60	1.34	1.00	0.66	0.35	0.85	0.50	0.76	0.85	0.50	0.76	0	1.31	0.62	0.40	0.30
	MOE+MOPAC	9	0.54	0.99	0.55	0	0.29	0.97	0.93	1.00	0.89	0.97	1.29	1.00	1.34	1.00	0.73	0.67	0.92	0.75	0.88	0.92	0.75	0.88	0	1.21	0.92	1.03	0.78
	information	10	1.56	0.48	3.24	1	0.42	0.97	0.71	0.91	0.83	0.89	1.26	0.91	1.34	1.00	0.56	0.25	0.77	0.50	0.71	0.77	0.50	0.71	0	1.09	0.77	1.22	0.93
	DRAGON	10	0.49	0.78	0.62	1	0.31	0.97	0.93	0.98	0.94	0.97	1.31	0.98	1.34	1.00	0.83	0.67	0.92	0.75	0.88	0.92	0.75	0.88	0	1.21	0.92	1.31	1.00
	WHIM	9	1.65	0.50	3.28	2	0.44	0.97	0.82	0.94	0.89	0.93	1.18	0.83	1.34	1.00	0.68	0.17	0.69	0.50	0.65	0.69	0.50	0.65	0	0.95	0.62	1.31	1.00
	molecular properties	10	2.64	0.42	6.25	3	0.45	0.96	0.78	0.94	0.83	0.92	1.12	0.89	1.34	1.00	0.47	0.23	0.92	0.25	0.76	1.00	0.25	0.80	2	1.12	0.92	1.20	0.91
M	DRAGON	10	0.84	0.91	0.92	0	0.24	1.00	1.00	1.00	1.00	1.00	3.00	1.00	3.00	1.00	0.67	0.39	0.75	0.67	0.69	0.75	0.67	0.69	0	1.63	0.75	0.67	0.21
	MOE+MOPAC	9	0.23	0.30	0.78	2	0.53	0.97	0.85	0.95	0.92	0.93	2.57	0.95	2.94	0.98	0.94	0.84	1.00	0.89	0.92	1.00	0.89	0.92	0	2.60	1.00	3.13	0.97
	GETAWAY	9	0.98	0.45	2.20	1	0.43	0.96	0.82	0.95	0.89	0.91	2.45	0.95	2.96	0.99	0.71	0.39	0.75	0.67	0.69	0.75	0.67	0.69	0	1.63	0.75	3.11	0.96
	geometrical	8	1.02	0.38	2.64	1	0.51	0.96	0.77	0.89	0.89	0.89	2.43	0.89	2.97	0.99	0.78	0.50	0.75	0.78	0.77	0.75	0.75	0.75	1	1.95	0.75	3.13	0.96
	WHIM	10	1.35	0.32	4.26	1	0.48	0.96	0.77	0.89	0.89	0.89	2.43	0.89	2.91	0.97	0.58	0.28	0.75	0.56	0.62	0.75	0.56	0.62	0	1.39	0.75	0.14	0.04
	atom-centered fragments	10	3.06	0.37	8.33	1	0.44	0.95	0.74	0.89	0.87	0.88	2.32	0.89	2.99	1.00	0.50	−0.05	0.50	0.44	0.46	0.50	0.38	0.42	1	0.93	0.50	0.67	0.21

[a] Number of variables with a significance less than 0.05. [b] Attained precision. [c] Results obtained taking into account only those substances that were within the applicability domain of the models. [d] Number of substances outside the applicability domain of the models.

dx.doi.org/10.1021/ci2003076 |J. Chem. Inf. Model. 2011, 51, 2746–2759

Table 4. Best Six Model of Every Family of Descriptors and Every Mutagenic End Point Sorted by ROCED

								training									test											
end point	ID	size	ROCED	FIT	ROCFIT	N[a]	λ	AUC	MCC	Se	Sp	Ac	EF	Pr[b]	RIE	BEDROC	AUC	MCC	Se	Sp	Ac	Se[c]	Sp[c]	Ac[c]	E[d]	EF	Pr[b]	RIE	BEDROC
CA	MOE+MOPAC	8	0.59	0.30	1.96	2	0.65	0.89	0.71	0.84	0.89	0.86	1.45	0.84	1.40	0.90	0.78	0.66	0.88	0.78	0.85	0.88	0.78	0.85	0	1.35	0.88	1.14	0.75
	binary fingerprints	9	0.72	0.19	3.87	6	0.73	0.81	0.48	0.73	0.76	0.74	1.31	0.73	1.50	0.98	0.83	0.75	0.88	0.89	0.88	0.88	0.89	0.88	0	1.43	0.88	1.15	0.75
	DRAGON	6	0.72	0.33	2.21	3	0.67	0.85	0.59	0.81	0.80	0.81	1.37	0.81	1.53	0.98	0.79	0.66	0.88	0.78	0.85	0.88	0.78	0.84	1	1.35	0.88	1.49	0.97
	GETAWAY	9	0.79	0.13	6.03	8	0.79	0.79	0.52	0.71	0.83	0.76	1.37	0.88	1.49	0.96	0.74	0.66	0.88	0.78	0.85	0.88	0.78	0.85	0	1.35	0.88	0.70	0.46
	walk and path counts	10	0.82	0.12	6.71	7	0.78	0.79	0.49	0.75	0.76	0.75	1.32	0.85	1.49	0.97	0.83	0.66	0.88	0.78	0.85	0.88	0.78	0.85	0	1.35	0.88	1.53	1.00
	TOPS-MODE	8	0.85	0.22	3.91	4	0.72	0.83	0.54	0.76	0.79	0.77	1.34	0.76	1.51	0.98	0.77	0.52	0.76	0.78	0.77	0.76	0.78	0.77	0	1.33	0.76	0.70	0.46
B	topologics	9	0.98	0.22	4.53	3	0.79	0.77	0.50	0.83	0.67	0.78	1.25	0.84	1.29	0.87	0.78	0.52	0.80	0.74	0.78	0.81	0.74	0.79	3	1.26	0.88	1.42	0.98
	information	10	1.01	0.23	4.32	5	0.77	0.78	0.47	0.79	0.69	0.76	1.25	0.84	1.27	0.86	0.71	0.46	0.74	0.74	0.74	0.76	0.74	0.76	2	1.25	0.87	1.38	0.96
	MOE+MOPAC	8	1.06	0.17	6.19	2	0.83	0.76	0.43	0.73	0.72	0.73	1.26	0.84	1.25	0.84	0.75	0.47	0.75	0.74	0.75	0.76	0.74	0.76	4	1.27	0.86	1.20	0.81
	WHIM	8	1.10	0.14	7.71	3	0.86	0.74	0.41	0.72	0.72	0.72	1.25	0.84	1.34	0.90	0.70	0.47	0.72	0.77	0.74	0.73	0.76	0.74	2	1.29	0.87	0.99	0.66
	walk and path counts	10	1.12	0.24	4.69	4	0.77	0.76	0.44	0.79	0.66	0.75	1.23	0.83	1.29	0.87	0.71	0.44	0.73	0.74	0.73	0.74	0.73	0.74	2	1.25	0.86	1.32	0.91
	binary fingerprints	9	1.13	0.25	4.51	2	0.77	0.77	0.47	0.83	0.64	0.77	1.22	0.82	1.38	0.93	0.76	0.54	0.90	0.61	0.81	0.90	0.61	0.81	1	1.21	0.84	1.42	0.98
MCG	TOPS-MODE	10	0.15	0.38	0.40	2	0.48	0.93	0.79	0.92	0.89	0.92	1.29	0.92	1.31	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0	1.31	1.00	1.31	1.00
	MOE+MOPAC	6	0.30	0.33	0.91	0	0.64	0.84	0.67	0.91	0.78	0.87	1.24	0.91	1.29	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0	1.31	1.00	1.31	1.00
	DRAGON	5	0.33	0.60	0.54	0	0.53	0.91	0.62	0.87	0.78	0.85	1.23	0.87	1.34	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0	1.31	1.00	1.31	1.00
	information	9	0.40	0.15	2.70	6	0.73	0.84	0.52	0.79	0.78	0.79	1.22	0.79	1.33	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0	1.31	1.00	1.31	1.00
	3DMoRSE	5	0.43	0.24	1.77	3	0.74	0.83	0.52	0.83	0.72	0.80	1.20	0.83	1.34	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0	1.31	1.00	1.31	1.00
	topologics	10	0.47	0.23	2.01	5	0.60	0.89	0.58	0.91	0.67	0.85	1.19	0.91	1.34	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0	1.31	1.00	1.31	1.00
M	MOE+MOPAC	9	0.23	0.30	0.78	2	0.53	0.97	0.85	0.95	0.92	0.93	2.57	0.95	2.94	0.98	0.94	0.84	1.00	0.89	0.92	1.00	0.89	0.92	0	2.60	1.00	3.13	0.97
	DRAGON	8	0.33	0.44	0.77	2	0.48	0.96	0.76	0.74	0.97	0.89	2.80	0.74	2.95	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0	3.25	1.00	3.24	1.00
	3DMoRSE	8	0.36	0.13	2.73	6	0.75	0.81	0.59	0.79	0.82	0.81	2.05	0.79	2.48	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0	3.25	1.00	3.24	1.00
	GETAWAY	9	0.40	0.09	4.27	8	0.78	0.78	0.57	0.74	0.84	0.81	2.10	0.74	2.35	0.79	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0	3.25	1.00	3.24	1.00
	information	7	0.54	0.26	2.04	3	0.64	0.87	0.56	0.79	0.79	0.79	1.96	0.79	2.81	0.94	0.89	0.84	1.00	0.89	0.92	1.00	0.89	0.92	0	2.60	1.00	0.70	0.21
	burden eiguenvalues	10	0.57	0.16	3.56	5	0.65	0.89	0.65	0.89	0.79	0.82	2.04	0.89	2.71	0.91	0.86	0.72	1.00	0.78	0.85	1.00	0.78	0.85	0	2.17	1.00	2.68	0.83

[a] Number of variables with a significance less than 0.05. [b] Attained precision. [c] Results obtained taking into account only those substances that were within the applicability domain of the models. [d] Number of substances outside the applicability domain of the models.

dx.doi.org/10.1021/ci2003076 |J. Chem. Inf. Model. 2011, 51, 2746–2759

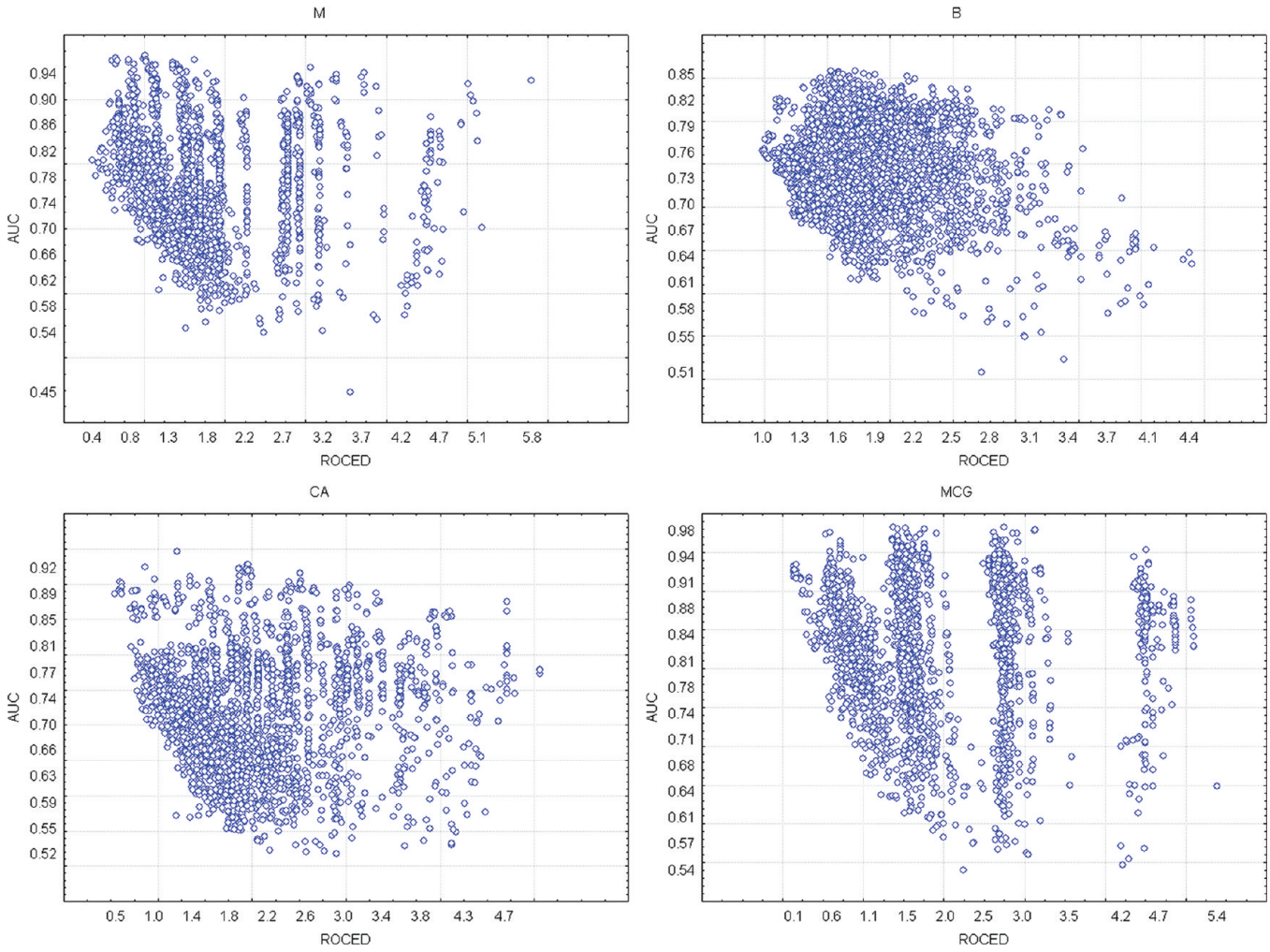


Figure 3. Scatterplot of AUC of the training set against ROCED for the Micronucleous (M), Bacterial (B), Chromosome Aberration (CA), and Mammalian Cell Gene mutation (MCG) tests.

To avoid possible loss of significance in the variables of the models obtained by linear discriminant analysis using only eq 3, we have defined a new parameter: ROCFIT -ROC graph Euclidean Distance corrected with FIT(λ).[1,2] ROCFIT is thus defined as follows

$$\text{ROCFIT} = \frac{\text{ROCED}}{\text{FIT}(\lambda)} \qquad (7)$$

For those end points where we have different numbers of substances due to the dimensionality of descriptors -bacterial mutagenesis and *in vitro* chromosomal aberration, the value of FIT was computed taking $n = 429$ (bacterial mutagenesis) and $n = 98$ (*in vitro* chromosomal aberration).

Applicability Domain of the Models. Given that the real utility of a QSAR/QSPR model relies on its ability to accurately predict the modeled activity for new chemicals, careful assessment of the model's true predictive power is a must. This includes the model validation but also the definition of the applicability domain of the model in the space of molecular descriptors used for deriving the model. There are several methods for assessing the applicability domain of QSAR/QSPR models,[32,33] but the most common one encompasses determining the leverage values for each compound.[34] A Williams plot, i.e. the plot of standardized residuals versus leverage values (h), can then be used for an immediate and simple graphical detection of both the response outliers and structurally influential chemicals in the model. In this plot, the applicability domain is established inside a squared area within $\pm x$ standard deviations and a leverage threshold h^* (h^* is generally fixed at $3(k+1)/n$, where n is the number of training compounds and k the number of model parameters, whereas $x = 2$ or 3), lying outside this area (vertical lines) the outliers and (horizontal lines) the influential chemicals. For future predictions, only predicted mutagenicity for chemicals belonging to the chemical domain of the model should be proposed and used.[35]

Variable Selection. Nowadays, there is a vast amount and a wide range of molecular descriptors with which one can model the activity of interest. This makes the search for gathering the most suitable subset quite complicated and time-consuming because of the many possible combinations, especially if one tries to define an accurate, robust, and (above all) interpretable model. That is the reason why we have applied here the replacement method (RM)[36] for selecting the variables, implemented by us in the STATISTICA Visual Basic.[37] Following this method, we observed the variation in the performance of the model selected by choosing any of the parameters below:

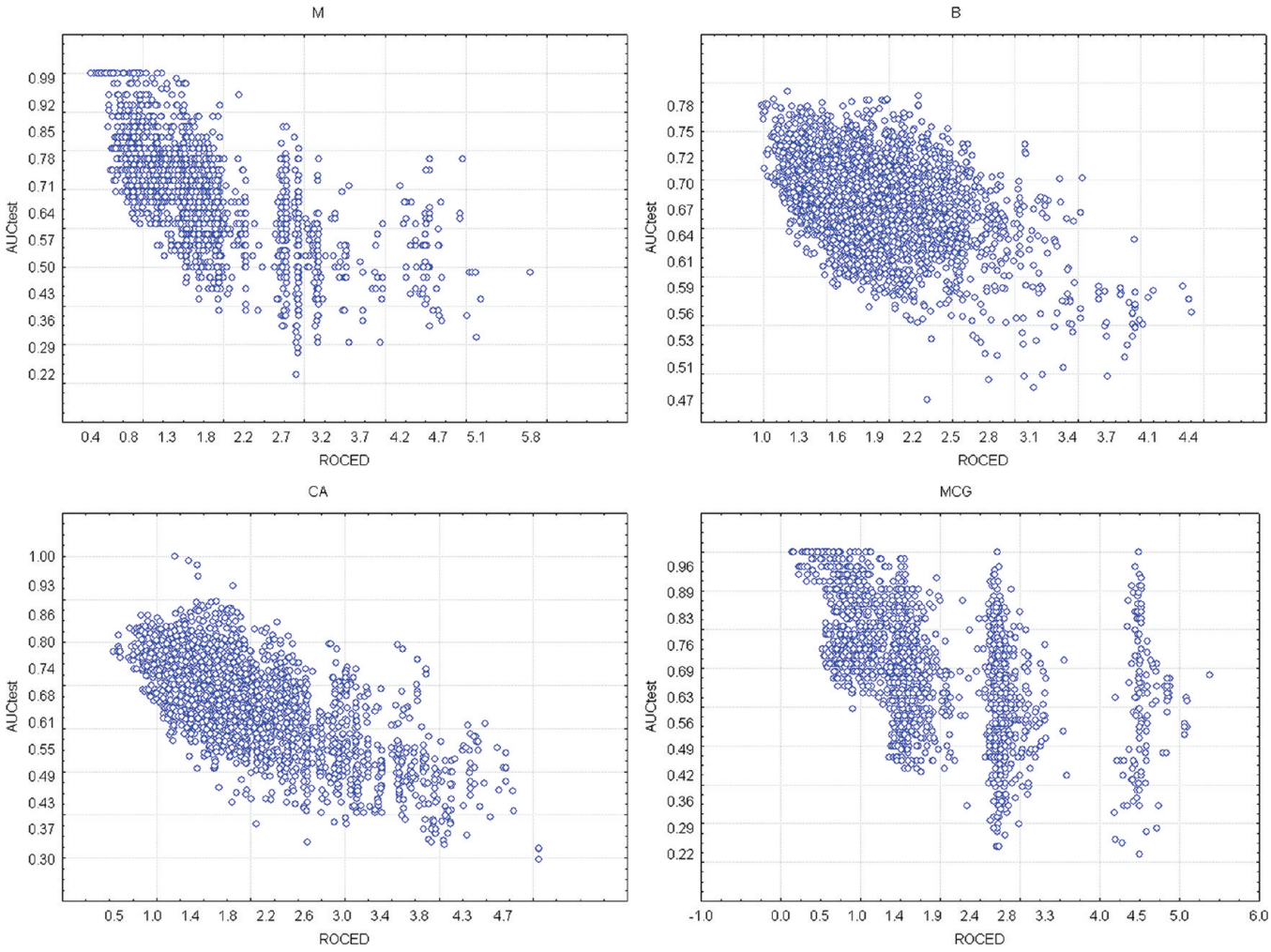


Figure 4. Scatterplot of AUC of the test set against ROCED for the Micronucleous (M), Bacterial (B), Chromosome Aberration (CA), and Mammalian Cell Gene mutation (MCG) tests.

- minimizing Wilk's lambda (λ),
- minimizing the Matthews correlation coeficcient of the test set (MCCtest),
- maximizing the area under the ROC curve of the training set (AUCtraining),
- minimizing the parameter ROCED,
- minimizing the parameter ROCFIT.

In all cases, the calculations were performed for a number of variables between 3 and 10. For the evaluation of the classification performance several indices were calculated for each model (sensibility, specificity, accuracy, precision, enrichment factor,[24] Matthews correlation coefficient,[7] RIE,[22] and BEDROC[21]) based on a prior probability of 0.5 and the area under the ROC curve for both the training and test sets. We also included the statistical significance (p-level) of the model computed by χ-square tests. All calculations were performed with the STATISTICA software.[37]

RESULTS AND DISCUSSION

Wilk's Lambda. Table 1 summarizes the results obtained based on the Wilk's lambda parameter. As can be seen, in general, models with better Wilk's lambda values afford a good separation of classes but a low predictivity. In turn, indices such as the Matthews's correlation coefficient, accuracy, and especially specificity have very low values for the prediction set with respect to the training set. Further, there was an *early recognition* of active substances for the training set, judging by the BEDROC values close to 1, but that also ensure us about their fair results for the prediction set (see the results for the micronucleous test in Table 1). Therefore, the parameter Wilk's lambda provides a good classification for the training set, but it does not guarantee a reasonable prediction. Moreover, comparing with the values of the ROCED parameter, one can see as a smaller value can well achieve a better balance between sensitivity and specificity for both the training and prediction sets (see Table 1).

Matthews Correlation Coefficient of the Test Set. In Table 2, the best models obtained based on the MCCtest for each end point and family of descriptors can be judged. In contrast to the Wilk's lambda results, one can see that high quality predictive models are obtained but whole coming from an unreal poor training. In general, the values of the indices for the prediction test reveal better ratings as well as a better performance and an *early recognition* of actives with respect to the training group. This fact is more accentuated in the micronucleous and chromosomal aberrations end points due to the smaller number of substances present, which in turn most likely led to a

dx.doi.org/10.1021/ci2003076 |J. Chem. Inf. Model. 2011, 51, 2746–2759

Table 5. Best Six Models Sorted by ROCFIT for Every Family of Descriptors and Mutagenic End Points

								training									test											
end point	ID	size	ROCED	FIT	ROCFIT	N[a]	λ	AUC	MCC	Se	Sp	Ac	EF	Pr[b]	RIE	BEDROC	AUC	MCC	Se	Sp	Ac	Se[c]	Sp[c]	Ac[c]	E[d]	EF	Pr[b]	RIE	BEDROC
CA	MOE+MOPAC	8	0.60	0.37	1.62	0	0.60	0.89	0.70	0.86	0.86	0.86	1.42	0.86	1.40	0.90	0.82	0.66	0.88	0.78	0.85	0.88	0.78	0.85	0	1.35	0.88	1.44	0.94
	DRAGON	6	0.72	0.33	2.21	3	0.67	0.85	0.59	0.81	0.80	0.81	1.37	0.81	1.53	0.98	0.79	0.66	0.88	0.78	0.85	0.88	0.78	0.84	1	1.35	0.88	1.49	0.97
	binary fingerprints	8	0.86	0.32	2.67	1	0.63	0.92	0.57	0.86	0.71	0.80	1.30	0.86	1.53	0.99	0.75	0.52	0.76	0.78	0.77	0.81	0.78	0.80	1	1.33	0.76	0.70	0.46
	TOPS-MODE	5	0.91	0.24	3.80	2	0.76	0.80	0.47	0.78	0.71	0.75	1.28	0.78	1.51	0.98	0.77	0.59	0.82	0.78	0.81	0.82	0.78	0.81	0	1.34	0.82	1.15	0.75
	GETAWAY	8	1.87	0.44	4.29	0	0.56	0.90	0.61	0.84	0.77	0.82	1.35	0.87	1.55	1.00	0.57	0.22	0.76	0.44	0.65	0.81	0.44	0.68	1	1.10	0.72	0.70	0.46
	2 Dautocorrelations	10	1.19	0.22	5.34	2	0.66	0.86	0.65	0.79	0.88	0.82	1.43	0.79	1.50	0.97	0.60	0.36	0.71	0.67	0.69	0.71	0.67	0.69	0	1.22	0.71	1.45	0.94
B	MOE+MOPAC	9	1.09	0.40	2.77	0	0.68	0.82	0.50	0.86	0.63	0.78	1.23	0.86	1.42	0.96	0.73	0.49	0.83	0.66	0.78	0.83	0.66	0.77	1	1.24	0.83	1.28	0.86
	DRAGON	8	1.29	0.46	2.84	0	0.65	0.85	0.57	0.89	0.67	0.82	1.26	0.89	1.39	0.93	0.73	0.50	0.88	0.60	0.79	0.88	0.59	0.79	4	1.22	0.88	1.20	0.81
	TOPS-MODE	10	1.30	0.45	2.87	0	0.64	0.82	0.57	0.91	0.64	0.82	1.24	0.91	1.35	0.91	0.74	0.51	0.86	0.65	0.79	0.89	0.62	0.81	7	1.22	0.86	1.38	0.95
	information	8	1.10	0.33	3.35	0	0.72	0.80	0.47	0.81	0.67	0.76	1.24	0.83	1.45	0.97	0.73	0.46	0.77	0.71	0.75	0.79	0.70	0.77	3	1.24	0.86	1.41	0.97
	topologics	8	1.34	0.39	3.47	0	0.69	0.81	0.50	0.84	0.65	0.78	1.24	0.84	1.38	0.93	0.70	0.45	0.81	0.65	0.76	0.82	0.64	0.77	5	1.21	0.81	1.06	0.73
	frequency fingerprints	9	1.54	0.39	3.93	0	0.68	0.81	0.50	0.86	0.62	0.78	1.23	0.82	1.40	0.94	0.66	0.37	0.84	0.52	0.74	0.87	0.52	0.76	3	1.15	0.80	0.90	0.62
MCG	TOPS-MODE	9	0.17	0.42	0.40	0	0.49	0.92	0.76	0.91	0.89	0.90	1.29	0.91	1.29	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0	1.31	1.00	1.31	1.00
	DRAGON	5	0.33	0.60	0.54	0	0.53	0.91	0.62	0.87	0.78	0.85	1.23	0.87	1.34	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0	1.31	1.00	1.31	1.00
	MOE+MOPAC	9	0.54	0.99	0.55	0	0.29	0.97	0.93	1.00	0.89	0.97	1.29	1.00	1.34	1.00	0.73	0.67	0.92	0.75	0.88	0.92	0.75	0.88	0	1.21	0.92	1.03	0.78
	frequency fingerprints	10	0.56	0.87	0.65	0	0.29	0.93	0.89	1.00	0.83	0.96	1.27	1.00	1.34	1.00	0.75	0.83	1.00	0.75	0.94	1.00	0.75	0.94	0	1.21	1.00	1.31	1.00
	atom-centered	10	0.72	0.72	0.99	1	0.33	0.96	0.81	0.98	0.78	0.93	1.24	0.98	1.34	1.00	0.67	0.55	0.85	0.75	0.82	0.85	0.75	0.82	0	1.20	0.85	1.03	0.78
	2 Dautocorrelations	6	0.72	0.70	1.02	0	0.46	0.92	0.77	0.96	0.78	0.92	1.24	0.96	1.34	1.00	0.65	0.55	0.85	0.75	0.82	0.85	0.75	0.82	0	1.19	0.77	1.22	0.93
M	DRAGON	8	0.53	0.75	0.70	0	0.34	0.98	0.88	0.95	0.95	0.95	2.70	0.95	2.99	1.00	0.81	0.64	0.75	0.89	0.85	0.75	0.89	0.85	0	2.44	0.75	3.22	0.99
	MOE+MOPAC	9	0.28	0.38	0.72	2	0.47	0.95	0.81	0.89	0.92	0.91	2.55	0.89	2.98	0.99	0.89	0.84	1.00	0.89	0.92	1.00	0.89	0.92	0	2.60	1.00	0.70	0.21
	GETAWAY	9	0.61	0.42	1.43	1	0.45	0.96	0.77	0.89	0.89	0.89	2.43	0.89	2.97	0.99	0.75	0.64	0.75	0.89	0.85	0.75	0.89	0.85	0	2.44	0.75	3.22	0.99
	3DMoRSE	9	0.70	0.41	1.72	2	0.46	0.92	0.68	0.74	0.92	0.86	2.47	0.74	2.97	0.99	0.72	0.64	0.75	0.89	0.85	0.75	0.89	0.85	0	2.44	0.75	3.13	0.96
	burden eiguenvalues	9	0.66	0.33	2.04	3	0.51	0.92	0.66	0.84	0.84	0.84	2.18	0.84	2.76	0.92	0.72	0.64	0.75	0.89	0.85	0.75	0.89	0.85	0	2.44	0.75	2.70	0.83
	information	7	0.54	0.26	2.04	3	0.64	0.87	0.56	0.79	0.79	0.79	1.96	0.79	2.81	0.94	0.89	0.84	1.00	0.89	0.92	1.00	0.89	0.92	0	2.60	1.00	0.70	0.21

[a] Number of variables with a significance less than 0.05. [b] Attained precision. [c] Results obtained taking into account only those substances that were within the applicability domain of the models. [d] Number of substances outside the applicability domain of the models.

dx.doi.org/10.1021/ci2003076 |J. Chem. Inf. Model. 2011, 51, 2746–2759

Table 6. Best Models Found on the Work by Rizzi *et al.*[14] Together with the Corresponding ROCED Value

	training					test							
classifiers	recall	precision	EF	MCC	AUC	recall	precision	EF	MCC	AUC	d_1	d_2	ROCED
DRAGON all[a]	0.9	0.65	7.17	0.74	0.98	0.88	0.49	5.38	0.61	0.96	0.1109	0.1510	0.3135
DRAGON no 3D[b]	0.98	0.68	7.49	0.8	0.99	0.9	0.54	5.89	0.66	0.95	0.0502	0.1260	0.2134
DRAGON shape[c]	0.88	0.5	5.5	0.62	0.96	0.9	0.4	4.38	0.54	0.94	0.1488	0.1680	0.3771
QikProp all[d]	0.8	0.29	3.21	0.4	0.87	0.74	0.22	2.37	0.29	0.82	0.2799	0.3694	0.9687
QikProp no ADME[e]	0.68	0.3	3.34	0.37	0.85	0.62	0.23	2.58	0.28	0.82	0.3569	0.4330	1.2180
EVA[f]	0.74	0.36	3.91	0.44	0.89	0.7	0.36	4.01	0.44	0.83	0.2914	0.3248	0.8436
3D-pharm AAR1[g]	0.74	0.32	3.57	0.42	0.84	0.6	0.39	4.29	0.42	0.8	0.3039	0.4109	1.1163
3D-pharm AHR11[h]	0.8	0.34	3.73	0.45	0.86	0.64	0.42	4.57	0.46	0.78	0.2532	0.3707	0.9556
3D-pharm AHR15[i]	0.8	0.37	4.04	0.48	0.88	0.66	0.34	3.7	0.4	0.78	0.2419	0.3633	0.9253
SP docking[j]	0.7	0.22	2.39	0.28	0.77	0.62	0.2	2.17	0.23	0.77	0.3892	0.4538	1.3046
XP docking[k]	0.74	0.22	2.44	0.3	0.75	0.68	0.19	2.1	0.24	0.73	0.3691	0.4318	1.2186
MASC[l]	0.64	0.31	3.45	0.37	0.78	0.56	0.28	3.08	0.31	0.79	0.3870	0.4630	1.3379
PLS-DA[m]	0.58	0.39	4.25	0.41	0.84	0.62	0.34	3.75	0.39	0.84	0.4295	0.3986	1.1941

[a] Model using all types of DRAGON descriptors. [b] Model using only non-3D DRAGON descriptors. [c] Model using only DRAGON descriptors related to the molecular shape. [d] Model using all types of QikProp descriptors. [e] Model using only QikProp descriptors nonrelated to the ADME profile. [f] Model using all types of EVA descriptors. [g] 3D Pharmacophore models using only AAR1 docking procedure hypotheses. [h] 3D Pharmacophore models using only AHR11 docking procedure hypotheses. [i] 3D Pharmacophore models using only AHR15 docking procedure hypotheses. [j] Model using standard precision (SP) docking score. [k] Model using extra precision (XP) docking score. [l] Model using multiple active site corrections (MASC) docking score. [m] Model using PLS-DA classifier.

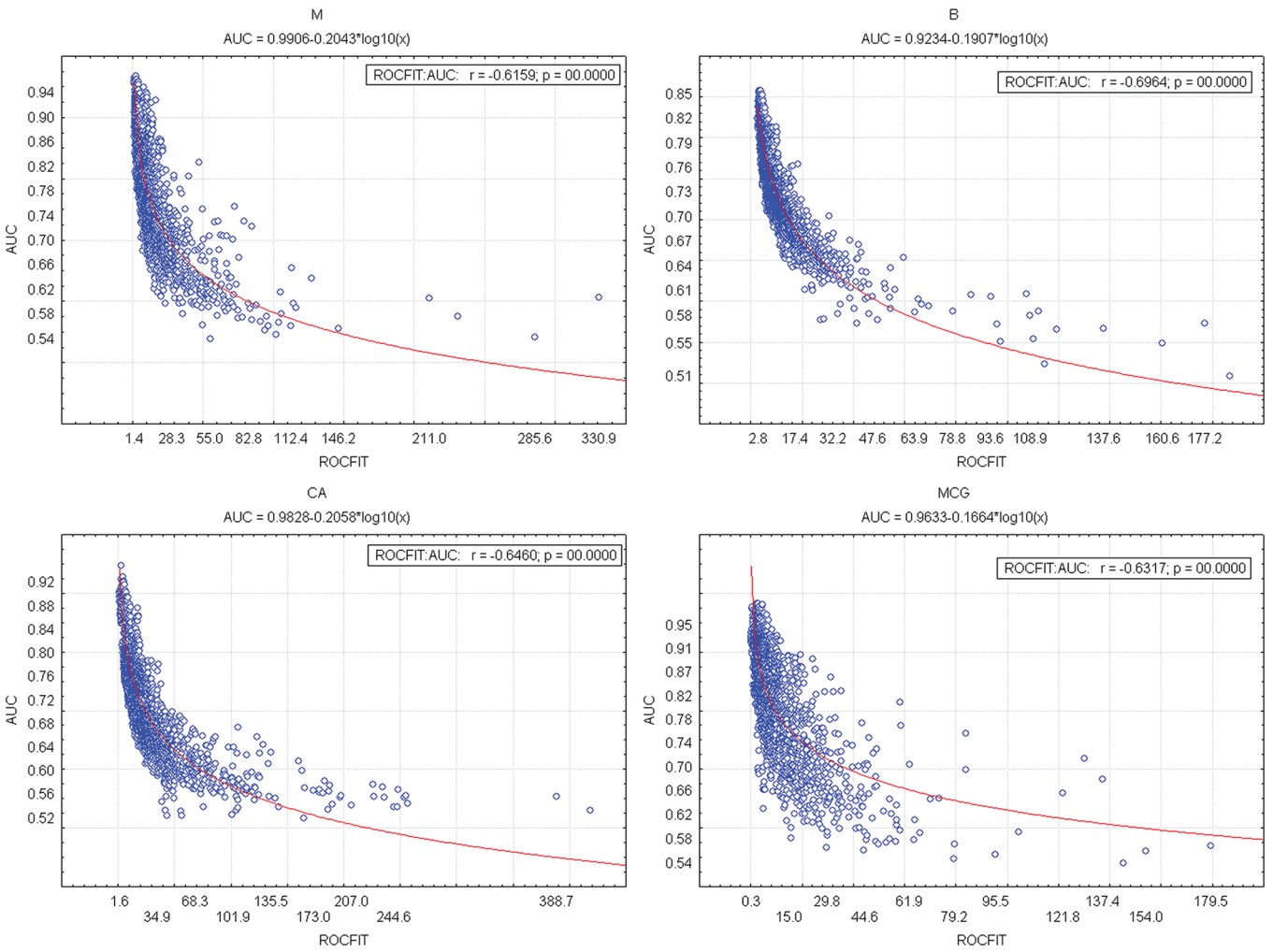


Figure 5. Scatterplot of AUC of the training set against ROCFIT for the Micronucleous (M), Bacterial (B), Chromosome Aberration (CA), and Mammalian Cell Gene mutation (MCG) tests.

dx.doi.org/10.1021/ci2003076 |J. Chem. Inf. Model. 2011, 51, 2746–2759

Table 7. Best Models[a] Found on the Work by Pérez-Garrido *et al.*[1] Together with the Corresponding ROCED and ROCFIT Values

							training					test								
	ID	size	ROCED	FIT	ROCFIT	λ	AUC	MCC	Se	Sp	Ac	AUC	MCC	Se	Sp	Ac	Se[b]	Sp[b]	Ac[b]	excluded
AMES	funcional groups count	8	0.46	0.71	0.65	0.49	0.89	0.73	0.81	0.91	0.86	0.93	0.77	85.71	91.30	88.64	85.71	91.30	88.64	0
	atom-centered fragments	8	0.99	0.70	1.42	0.50	0.91	0.71	0.80	90.32	0.85	0.80	0.64	76.19	86.96	81.82	75.00	85.00	80.00	4
	3DMoRSE	8	0.74	0.61	1.22	0.53	0.89	0.65	80.72	83.87	0.82	0.87	0.69	90.48	78.26	84.09	90.00	77.27	83.33	2
	GETAWAY	7	0.81	0.62	1.29	0.54	0.89	0.64	74.70	88.17	0.82	0.80	0.64	76.19	86.96	81.82	75.00	86.96	81.40	1
	2 Dautocorrelations	7	0.74	0.54	1.37	0.58	0.88	0.67	78.31	88.17	0.84	0.84	0.59	76.19	82.61	79.55	76.19	82.61	79.55	0
	geometrical	6	0.94	0.57	1.66	0.58	0.88	0.59	79.52	79.57	0.80	0.86	0.59	76.19	82.61	79.55	75.00	81.82	78.57	2
	RDF	8	0.87	0.42	2.08	0.62	0.88	0.59	80.72	78.49	0.80	0.81	0.60	85.71	73.91	79.55	85.71	80.00	82.93	3
	WHIM	7	0.96	0.44	2.18	0.63	0.85	0.58	71.08	86.02	0.79	0.90	0.55	80.95	73.91	77.27	80.95	72.73	76.74	1
	constitutional	5	1.37	0.47	2.90	0.64	0.86	0.60	72.29	87.10	0.80	0.78	0.41	71.43	69.57	70.45	70.00	68.18	69.05	2
	burden eiguenvalues	7	0.91	0.40	2.26	0.65	0.85	0.52	71.08	80.65	0.76	0.92	0.55	80.95	73.91	77.27	80.95	73.91	77.27	0
	information	4	0.84	0.47	1.80	0.64	0.84	0.54	72.29	81.72	0.77	0.83	0.59	76.19	82.61	79.55	76.19	82.61	79.55	0
	topologics	3	1.42	0.33	4.26	0.74	0.82	0.51	71.08	79.57	0.76	0.78	0.36	66.67	69.57	68.18	65.00	69.57	67.44	1
	eiguenvalue-based	2	1.20	0.20	6.00	0.83	0.76	0.45	63.86	80.65	0.73	0.76	0.45	66.67	78.26	72.73	66.67	77.27	72.09	1
	walk and path counts	4	1.34	0.16	8.17	0.84	0.74	0.41	55.42	83.87	0.70	0.81	0.59	76.19	82.61	79.55	73.68	80.95	77.50	4
	conectivity	2	1.34	0.12	10.97	0.89	0.74	0.40	54.22	83.87	0.70	0.78	0.50	71.43	78.26	75.00	70.00	78.26	74.42	1
	Galvez topological charge	2	1.56	0.09	18.19	0.92	0.69	0.37	68.67	68.82	0.69	0.58	0.27	61.90	65.22	63.64	61.90	65.22	63.64	0
	molecular properties	2	3.35	0.08	40.37	0.92	0.67	0.32	85.54	44.09	0.64	0.51	0.03	80.95	21.74	50.00	85.00	18.18	50.00	2
	Randić molecular profiles	1	2.29	0.05	43.04	0.95	0.63	0.14	57.83	55.91	0.57	0.55	0.09	52.38	56.52	54.55	52.38	54.55	53.49	1
MCGN	atom centered fragments	4	0.41	1.10	0.38	0.36	0.97	0.71	88.89	83.33	0.87	0.86	0.76	85.71	100.00	88.89	85.71	100.00	88.89	0
	functional groups count	4	0.65	0.94	0.69	0.40	0.91	0.76	100.00	66.67	0.90	0.86	0.76	85.71	100.00	88.89	85.71	100.00	88.89	0
	topologics	4	0.50	0.92	0.54	0.40	0.96	0.75	96.30	75.00	0.90	0.86	0.76	85.71	100.00	88.89	85.71	100.00	88.89	0
	conectivity	4	0.50	0.83	0.60	0.42	0.94	0.82	100.00	75.00	0.92	0.86	0.76	85.71	100.00	88.89	85.71	100.00	88.89	0
	GETAWAY	4	4.38	0.82	5.35	0.43	0.94	0.76	92.59	83.33	0.90	0.79	−0.19	85.71	0.00	66.67	85.71	0.00	66.67	0
	RDF	4	1.43	0.77	1.86	0.43	0.89	0.76	92.59	83.33	0.90	0.79	0.36	85.71	50.00	77.78	85.71	50.00	77.78	0
	burden eiguenvalues	4	0.63	0.75	0.84	0.45	0.97	0.80	85.19	100.00	0.90	0.86	0.76	85.71	100.00	88.89	83.33	100.00	87.50	1
	2 Dautocorrelations	4	1.67	0.71	2.36	0.45	0.94	0.76	92.59	83.33	0.90	0.93	0.60	71.43	100.00	77.78	71.43	100.00	75.00	1
	constitutional	4	0.65	0.69	0.95	0.47	0.89	0.69	96.30	66.67	0.87	0.86	0.76	85.71	100.00	88.89	85.71	100.00	88.89	0
	walk and path counts	4	0.72	0.64	1.13	0.47	0.95	0.69	92.59	75.00	0.87	1.00	0.60	71.43	100.00	77.78	71.43	100.00	77.78	0
	eiguenvalue-based	4	0.54	0.61	0.89	0.49	0.90	0.64	88.89	75.00	0.85	0.93	0.76	85.71	100.00	88.89	85.71	100.00	88.89	0
	geometrical	4	4.79	0.78	6.10	0.51	0.90	0.63	92.59	66.67	0.85	0.43	−0.29	71.43	0.00	55.56	71.43	0.00	55.56	0
	3DMoRSE	2	0.66	0.58	1.15	0.52	0.91	0.63	92.59	66.67	0.85	0.86	0.76	85.71	100.00	88.89	85.71	100.00	88.89	0
	Galvez topological charge	3	1.74	0.54	3.22	0.57	0.85	0.64	88.89	75.00	0.85	0.57	0.19	71.43	50.00	66.67	71.43	50.00	66.67	0

[a] All models have been derived using DRAGON descriptors. [b] Results taking into account only those substances that were within the applicability domain of the models.

dx.doi.org/10.1021/ci2003076 |J. Chem. Inf. Model. 2011, 51, 2746–2759

quickly convergence of the algorithm. For the end point of mutagenicity in bacteria, one can also appreciate that the best model is obtained for the descriptors binary fingerprints, i.e. the well-known limitation of this index.[4] That is to say, for a case where we have very few or no false negatives, but at the same time very few true negatives, or what is the same, high sensitivity and low specificity, the model's correlation coefficient is relatively high. In this case, the parameter ROCED has a higher value than the following model derived from topological descriptors, which has a better balance between sensitivity and specificity for both training and test sets as well as slightly higher AUC values.

AUC. The results for the models derived based on the AUC parameter are presented in Table 3. Whenever an improvement of the AUC values was reached (i.e.: values close to 1), that also meant getting better values of accuracy, precision, enrichment factor, Matthews correlation coefficient, BEDROC, and a high number of significant variables for the training set but, nevertheless, a low prediction. For instance, for AUC values slightly above 0.5, the accuracy is near 0.7 and a greater number of substances are outside of the applicability domain (especially for bacterial mutagenesis; see Table 3). Moreover, in all end points, the test set showed an imbalance between sensitivity and specificity.

ROCED. Table 4 shows the results obtained by classifying our models for each end point based on the ROCED parameter. Here, one should again emphasize that, in general, both the AUC and the balance between sensitivity and specificity for the training and prediction sets is better when the value of ROCED is lower. This is to be expected since the value of this parameter is derived from the values of selectivity and specificity represented in the ROC graph. Also, as expected, the number of substances outside of the applicability domain is significantly lower than those obtained with the AUC and Wilk's lambda parameters save for the MCCtest.

By plotting the AUC values against the ROCED values for the 11889 models obtained separately for the four mutagenicity end points (see Figure 3 and Figure 4), one can observe that as the values of ROCED become lower the AUC parameter usually presents higher values. In other words, in these working models that have high levels of ROCED, the AUC falls between 0.5 and 1, but models with low values of ROCED (between 0.6 and 1.3, depending on the group of substances) often have AUC values higher than 0.7 for both the training and test sets.

In addition, one can notice that the number of nonsignificant variables ($p > 0.05$) is much higher than in the models derived from the Wilk's lambda and AUC parameters. This is because the ROCED parameter only takes into account the classification obtained and not the significance of the model. Therefore, for other nonlinear methods (i.e: Neural Network, Machine Learning techniques, etc.), this capability of summarizing the information into a single numerical value, resulting from the classification of both the training and test sets, could be exploited to simplify the selection of algorithms and models based on the values of ROCED.

Also considering the reach of this parameter in other models in the literature, Rizzi and Fyn[14] used PLS-Discriminant Analysis to study a series of PDE4 inhibitors. Table 6 shows the best models obtained by the authors, where they have chosen the model derived from no-3D descriptors (taken from DRAGON) as it presented a better classification and a low difference in performance between the training and the prediction sets. Since the authors did not define an applicability domain for that model, the calculation of ROCED was done with eq 3. Logically, the minimum value for this parameter was obtained for the model chosen by the authors, since this parameter summarizes the classification and performance contained in the model.

ROCFIT. As expected and as the results in Table 5 show, though the scope of the ROCFIT classifications with respect to the ROCED parameter (Table 4) decreases on the prediction group, an increase in significance is achieved meanwhile keeping the balance between specificity and selectivity.

In addition, one can see that slightly higher AUC values are attained for the training set, but however that does not happen for the prediction group. Plotting the AUC values against the values of ROCFIT (Figure 5) for the 11,889 models obtained for the four mutagenicity end points, one can notice that these two parameters are related logarithmically, and so a decrease in AUC produces an increase exponentially in ROCFIT.

In another linear discriminant analysis work done by some of us on bacterial and mammalian cell mutagenesis for a series of alpha, beta-unsaturated carbonyl compounds,[1] the best model has been chosen based on the values of Wilk's lambda, FIT(λ), and the ratings for both the training and prediction sets. As can be seen in Table 7, the models chosen are those with a lower value for both ROCED and ROCFIT parameters. Therefore, these indexes are interesting in assessing the models if the main goal is to obtain a good performance and a balance between training and test set classifications. If the goal is to use the model for mechanistic interpretation or molecular design, then ROCFIT should render better results.

CONCLUSIONS

Herein, we proved that the combination of the distances given by the representation of the classifications obtained for the training and test sets in an ROC graph with parameters ROCFIT and ROCED is a quick and useful tool for choosing classification models. The final models displayed lower values of these parameters, especially of ROCFIT, achieved a better balance between sensitivity and specificity, better classifications for both the training and test sets along with higher AUC values. Overall, that turned out to only improve parameters such as the Wilk's lambda, the AUC values for the training set, and the Matthews's correlation coefficients for the test set.

Due to the low computational resources required for their calculation, such parameters are interesting to minimize the computational costs involved on obtaining the AUC in each iteration. For discriminant analysis studies, the ROCED parameter is not enough because the models lose significance, even though it would be interesting to study the reach of this last parameter in nonlinear analysis. On the other hand, the ROCFIT parameter maintains the balance between sensitivity and specificity of ROCED but improves the significance of the model since it includes the FIT(λ) parameter.

AUTHOR INFORMATION

Corresponding Author
*E-mail: Aperez@pdi.ucam.edu.

ACKNOWLEDGMENT

Foundation for Science and Technology (FCT), Portugal and COMPETE/QREN/EU (projects PTDC/QUI/70359/2006 and PTDC/QUI-QUI/113687/2009) are acknowledged for

financial support. The authors acknowledge the FCT and Social European Found grant (SFRH/BPD/63946/2009).

REFERENCES

(1) Pérez-Garrido, A.; Helguera, A. M.; Girón-Rodríguez, F.; Cordeiro, M. N. D. S. QSAR Models to Predict Mutagenicity of Acrylates, Methacrylates and α,β-unsaturated Carbonyl Compounds. *Dent. Mater.* **2010**, *26*, 397–415.

(2) Pérez-Garrido, A.; Helguera, A. M.; Caravaca, G.; Cordeiro, M. N. D. S.; Escudero, A. G. A TOPological Substructural MOlecular Design Approach for Predicting Mutagenesis End-points of α, β-unsaturated Carbonyl Compounds. *Toxicology* **2010**, *268*, 64–77.

(3) Estrada, E.; Molina, E. Automatic Extraction of Structural Alerts for Predicting Chromosome Aberrations of Organic Compounds. *J. Mol. Graphics Modell.* **2006**, *25*, 275–288.

(4) Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A.; Nielsen, H. Assessing the Accuracy of Prediction Algorithms for Classification: an Overview. *Bioinformatics* **2000**, *16*, 412–424.

(5) Yang, X.-G.; Lv, W.; Chen, Y.-Z.; Xue, Y. In Silico Prediction and Screening of gamma-Secretase Inhibitors by Molecular Descriptors and Machine Learning Methods. *J. Comput. Chem.* **2010**, *31*, 1249–1258.

(6) Golbraikh, A.; Tropsha, A. Beware of q2!. *J. Mol. Graphics Modell.* **2002**, *20*, 269–276.

(7) Matthews, B. W. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozime. *Biochim. Biophys. Acta* **1975**, *405*, 442–451.

(8) Benigni, R. Structure-activity Relationship Studies of Chemical Mutagens and Carcinogens: Mechanistic Investigations and Prediction Approaches. *Chem. Rev.* **2005**, *105*, 1767–1800.

(9) Swets, J. A.; Dawes, R. M.; Monahan, J. Better Decisions Through Science. *Sci. Am.* **2000**, *283*, 82–87.

(10) Diamond, G. A. ROC steady: a Receiver Operating Characteristic Curve that is Invariant Relative to Selection Bias. *Med. Decis. Making* **1987**, *7*, 238–243.

(11) Hanley, J. A. Receiver Operating Characteristic (ROC) Methodology: the State of the Art. *Crit. Rev. Diagn. Imaging* **1989**, *29*, 307–335.

(12) Mann, F. A.; Hildebolt, C. F.; Wilson, A. J. Statistical Snalysis with Receiver Operating Characteristic Curves. *Radiology* **1992**, *184*, 37–38.

(13) Metz, C. E.; Goodenough, D. J.; Rossmann, K. Evaluation of Receiver Operating Characteristic Curve Data in Terms of Information Theory, with Applications in Radiography. *Radiology* **1973**, *109*, 297–303.

(14) Rizzi, A.; Fioni, A. Virtual Screening using PLS Discriminant Analysis and ROC Curve Approach: An Application Study on PDE4 Inhibitors. *J. Chem. Inf. Model.* **2008**, *41*, 1686–1692.

(15) Klon, A. E.; Glick, M.; Thoma, M.; Acklin, P.; Davies, J. W. Finding More Needles in the Haystack: A Simple and Efficient Method for Improving High-Throughput Docking Results. *J. Med. Chem.* **2004**, *47*, 2743–2749.

(16) Triballeau, N.; Acher, F.; Brabet, I.; Pin, J. P.; Bertrand, H. O. Virtual Screening Workflow Development Guided by the Receiver Operating Characteristic Curve Approach. Application to High-Throughput Docking on Metabotropic Glutamate Receptor Subtype 4. *J. Med. Chem.* **2005**, *48*, 2534–2547.

(17) Cleves, A. E.; Jain, A. N. Robust Ligand-Based Modeling of the Biological Targets of Known Drugs. *J. Med. Chem.* **2006**, *49*, 2921–2938.

(18) Pham, T. A.; Jain, A. N. Parameter Estimation for Scoring Protein-Ligand Interactions Using Negative Training Data. *J. Med. Chem.* **2005**, *47*, 6128–6136.

(19) Klon, A. E.; Glick, M.; Davies, J. W. Application of Machine Learning to Improve the Results of High-Throughput Docking Against the HIV-1 Protease. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2216–2224.

(20) Kumar, R.; Antony, G. M. A Review of Methods and Applications of the ROC Curve in Clinical Trials. *Drug Inf. J.* **2010**, *44*, 659–671.

(21) Truchon, J.-F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the *Early Recognition* Problem. *J. Chem. Inf. Model.* **2007**, *47*, 488–508.

(22) Sheridan, R. P.; Singh, S. B.; Fluder, E. M.; Kearsley, S. K. Protocols for Bridging the Peptide to Nonpeptide gap in Topological Similarity Searches. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1395–1406.

(23) McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J.-F.; Cornell, W. D. Comparison of Topological, Shape, and Docking Methods in Virtual Screening. *J. Chem. Inf. Model.* **2007**, *47*, 1504–1519.

(24) Pearlman, D. A.; Charifson, P. S. Improved Scoring of Ligand-protein Interactions Using OWFEG Free Energy Grids. *J. Med. Chem.* **2001**, *44*, 502–511.

(25) Benigni, R.; Bossa, C. Predictivity and Reliability of QSAR Models: The Case of Mutagens and Carcinogens. *Toxicol. Mech. Methods* **2008**, *18*, 137–147.

(26) MOE Molecular Operating Environment, version 2008.10. Chemical Computing Group, Inc.: 2008.

(27) Todeschini, R.; Consonni, V.; Pavan, M. DRAGON for Windows, version 5.4. TALETE srl: Milano, Italy, 2004.

(28) Gutierrez, Y.; Estrada, E. MODesLab MOlecular Design Laboratory, version 1.5. 2002.

(29) Frank, J. MOPAC, version 7.1. Seiler Research Laboratory: Colorado, Springs CO, 2007.

(30) McFarland, J. W.; Gans, D. J. On Identifying Likely Determinants of Biological Activity in High Dimensional QSAR Problems. *Quant. Struct-Act. Relat.* **1994**, *13*, 11–17.

(31) Johnson, R. A.; Wichern, D. W. *Applied MultiVariate Statistical Analysis*; Prentice-Hall: New York, 1988; pp 696−703.

(32) Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P. Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification- and Regression-Based QSARs. *Environ. Health Perspect.* **2003**, *111*, 1361–1375.

(33) Netzeva, T. I.; Worth, A. P.; Aldenberg, T.; Benigni, R.; Cronin, M. T. D.; Gramatica, P.; Jaworska, J. S.; Kahn, S.; Klopman, P.; Marchant, C. A.; Myatt, G.; Nikolova-Jeliazkova, N.; Patlewicz, G. Y.; Perkins, R.; Roberts, D. W.; Schultz, T. W.; Stanton, D. T.; van de Sandt, J. J. H.; Tong, W.; Veith, G.; Yang, C. Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure-Activity Relationships. *ATLA* **2005**, *33*, 155–173.

(34) Gramatica, P. Principles of QSAR Models Validation: Internal and External. *QSAR Comb. Sci.* **2007**, *26*, 694–701.

(35) Vighi, M.; Gramatica, P.; Consolaro, F.; Todeschini, R. QSAR and Chemometrics Approaches for Setting Water Quality Objetives for Dangerous Chemicals. *Ecotoxicol. Environ. Saf.* **2001**, *49*, 206–220.

(36) 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.

(37) Statistica, version 8.0. StatSoft, Inc.: Tulsa, USA, 2002.