Journal of Chromatography B, 910 (2012) 84–94

Contents lists available at SciVerse ScienceDirect

Journal of Chromatography B

journal homepage: www.elsevier.com/locate/chromb

QSRR modeling for diverse drugs using different feature selection methods coupled with linear and nonlinear regressions[☆]

Mohammad Goodarzi[a], Richard Jensen[b], Yvan Vander Heyden[a,*]

[a] Department of Analytical Chemistry and Pharmaceutical Technology, Center for Pharmaceutical Research (CePhaR), Vrije Universiteit Brussel (VUB), Laarbeeklaan 103, B-1090 Brussels, Belgium
[b] Department of Computer Science, Aberystwyth University, Aberystwyth, Wales, UK

ARTICLE INFO

Article history:
Received 7 September 2011
Accepted 17 January 2012
Available online 24 January 2012

Keywords:
QSRR
Chromatographic retention
ACO
MLR
SVM
Relief method

ABSTRACT

A Quantitative Structure-Retention Relationship (QSRR) is proposed to estimate the chromatographic retention of 83 diverse drugs on a Unisphere poly butadiene (PBD) column, using isocratic elutions at pH 11.7. Previous work has generated QSRR models for them using Classification And Regression Trees (CART). In this work, Ant Colony Optimization is used as a feature selection method to find the best molecular descriptors from a large pool. In addition, several other selection methods have been applied, such as Genetic Algorithms, Stepwise Regression and the Relief method, not only to evaluate Ant Colony Optimization as a feature selection method but also to investigate its ability to find the important descriptors in QSRR. Multiple Linear Regression (MLR) and Support Vector Machines (SVMs) were applied as linear and nonlinear regression methods, respectively, giving excellent correlation between the experimental, i.e. extrapolated to a mobile phase consisting of pure water, and predicted logarithms of the retention factors of the drugs ($\log k_w$). The overall best model was the SVM one built using descriptors selected by ACO.

1. Introduction

For many years, the separation of drugs has been a critical and important stage in analytical chemistry and pharmaceutical science. One of the most applied techniques is High-Performance Liquid Chromatography (HPLC), which is able to analyze a wide polarity range of acidic, basic and neutral compounds. High-Performance Liquid Chromatography is well recognized as a powerful, fast, selective and highly efficient technique, successfully employed for the separation and determination of many drugs [1]. To perform separations, a broad range of chromatographic stationary phases provide meaningfully different retention and selectivity. However, the mechanisms of retention are not always entirely known [2,3]. The choice of the stationary phase is very important and is based on user knowledge or on chromatographic tests to select columns with similar or dissimilar characteristics (selectivities).

The prediction of the physicochemical behavior of compounds, such as chromatographic retention, is useful for estimating, for instance, how well two similar substances will be distinguished in a given separation system, at the moment standards are not (yet) available in the drug development process. Quantitative Structure-Retention Relationship modeling has been utilized for the prediction of retention and migration behaviors [4–9]. In the resulting models a retention parameters is modeled as a function of molecular descriptors. It should be noted that QSRR is a kind of Quantitative Structure-Property Relationship (QSPR) study. Put et al. [9] have performed Classification And Regression Tree (CART) analysis as a QSRR study of the chromatographic retention of 83 drugs. CART selected three descriptors: a hydrophobicity parameter ($\log P$), the hydrophilic factor (Hy) [10] and the total path count (TPC) [11] from 266 calculated descriptors, to predict chromatographic retention. CART divided the retentions of the 83 molecules into five classes called very low, low, intermediate, high and very high retention [9].

In the present study, we have performed regression instead of classification. One of the most important stages, not only in classification but also in regression, is feature selection. As many pattern recognition and regression techniques were originally not designed to cope with large amounts of irrelevant features (e.g. given molecular descriptors), combining them with feature selection techniques has become a necessity in many applications [12–14]. The application of feature selection methods has several goals: firstly, to avoid overfitting and improve model performance;

[☆] This paper belongs to the Special Issue Chemometrics in Chromatography, Edited by Pedro Araujo and Bjørn Grung.

[*] Corresponding author. Tel.: +32 24774723; fax: +32 24774735.
E-mail address: yvanvdh@vub.ac.be (Y. Vander Heyden).

1570-0232/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.jchromb.2012.01.012

secondly, to provide faster and more cost-effective models; and thirdly, to acquire a deeper insight into the underlying processes that generated the data, and to identify important variables that have an intuitive physical interpretation [15]. In this study we mainly focus on the first two goals and less on the latter.

Recently, Swarm Intelligence has been used in different fields of study for the purpose of feature selection [16]. One interesting method is Ant Colony Optimization (ACO). ACO [16–18] is based on the behavior of real ants that are capable of finding the shortest route between a food source and their nest by means of pheromone deposition, without the use of visual information and hence possessing no global world model, while being able to adapt to changes in the environment. If a sudden environmental change occurs (e.g. a large obstacle appears on the shortest path), the ants can respond to this and will eventually converge to a new path. Based on this idea, artificial ants can be deployed to solve complex optimization problems via the use of artificial pheromone deposition. ACO is particularly attractive for feature selection as there seems to be no heuristic that can guide incremental search to the optimal subset of features. Additionally, it can be the case that ants discover the best feature combinations as they proceed throughout the search space. The ACO-based Fuzzy-Rough Set feature selection method has been applied recently for the first time in QSAR [19], giving excellent results for a class of glycogen synthase kinase-3β inhibitors.

Another important item which affects the prediction ability of any QSRR model is the choice of the regression technique for correlating descriptors with the experimental chromatographic retention. The significance of simple Multiple Linear Regression (MLR) in QSAR and QSRR has received attention from the literature [20,21], while accounting for non-linearity in the building of QSAR and QSRR models has also played an important role in the accuracy of activity and retention predictions, respectively [22,23]. It should be noted that in this study Support Vector Machines were used as nonlinear modeling technique. However, for the evaluation of ACO and to assess the ability of other feature selection methods, we have also used Genetic Algorithms (GAs), the Relief method and Stepwise Regression to select relevant variables in the construction of different QSRR models.

2. Theory

2.1. Feature selection and Ant Colony Optimization

The main aim of feature selection is to determine a minimal feature subset from a problem domain while retaining a suitably high accuracy in representing the original features. In real world problems feature selection is a must because of the abundance of noisy, irrelevant or misleading features. The usefulness of a feature or feature subset is determined by both its relevancy and its redundancy. A feature is said to be relevant if it is predictive for the decision feature(s) (i.e. dependent variable(s) here retention expressed as $\log k_{\mathrm{w}}$), otherwise it is irrelevant. A feature is considered to be redundant if it is highly correlated with other features (for instance, different $\log P$ estimates may be highly correlated). Hence, the search for a good feature subset involves finding those features that are highly correlated with the decision feature(s), but are uncorrelated with each other. However, the complexity of locating such a globally optimal subset of features is usually prohibitive, which motivates the use of more advanced search techniques, such as ACO.

For ACO-based feature selection, the process begins with the generation of a number of ants, placed randomly on a graph that represents every possible combination of features. Here, each node corresponds to a dataset feature and each edge permits the traversal of an ant from one feature to another (Fig. 1). An amount of virtual pheromone (a real number in ([0,1]) is associated with each edge that indicates the popularity of this particular traversal by past ants. Ants then traverse the graph, making probabilistic decisions as to which nodes to visit based on this virtual pheromone and also a heuristic desirability measure, until a traversal stopping criterion is satisfied. This is typically when the heuristic measure has reached a pre-calculated global optimum for the data. If the criterion is not satisfied, the virtual pheromone on the edges is updated based on ant traversals, a new set of ants is created and the process iterates once more. More details and definitions can be found in [17].

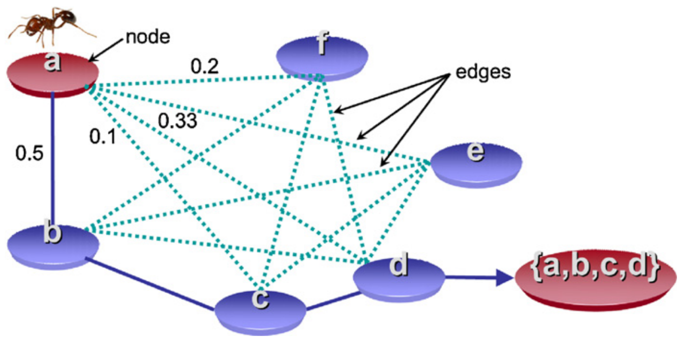

Fig. 1. ACO representation of feature selection. Nodes *a* to *f* represent features. The path highlighted (in blue) indicates the path taken by one ant and the resulting feature subset. The numbers on the edges are example of virtual pheromone. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

2.2. The Relief method

In the Relief method [24,25], each feature is given a relevance weighting that reflects its ability to discern between decision class labels. It thus first was applied on classification problems. A user-specified threshold determines the number of sampled objects used for constructing the weights. For each sampling, an object **x** is randomly chosen, and its nearest neighbor of the same class and nearest neighbor of a different class are calculated. Based on these neighbors, the feature weights are updated such that more weight is given to features that discriminate the object from neighbors of different classes. The user must supply a threshold which determines the level of relevance that feature weights must surpass in order to be finally chosen. The method has been extended to enable it to handle inconsistency, noisy and multi-class datasets [26]. Relief has also been extended to handle continuous decision variables (e.g. retention parameters). Instead of requiring the exact knowledge whether two objects belong to the same class or not, which is not applicable in regression problems, the relative distance between the predicted values of two objects (compounds) is used in order to calculate feature weightings.

2.3. Support Vector Machine regression

Support Vector Machine (SVM) is a new and very promising classification and regression technique developed by Vapnik [27]. Here, we give only a brief introduction to its main principle. Given a training data set of compounds $\{(\mathbf{x}_1, y_1, \ldots, (\mathbf{x}_l, y_l)\}_{i=1}^{l}$, where $\mathbf{x}_i \in \mathbf{X} \subseteq R$ is the ith input data point in input space (a descriptor) and $y_i \in \mathbf{y} \subseteq R$ is the associated output value of $\mathbf{x}_i$ (retention parameters). Initially SVM considered classification problems of two classes. An SVM model is a representation of the samples as points in space, mapped in such a way that the two classes are separated by a gap that is as wide as possible. Because the classes are not always linearly separable in the initial data space (of the descriptors), the technique

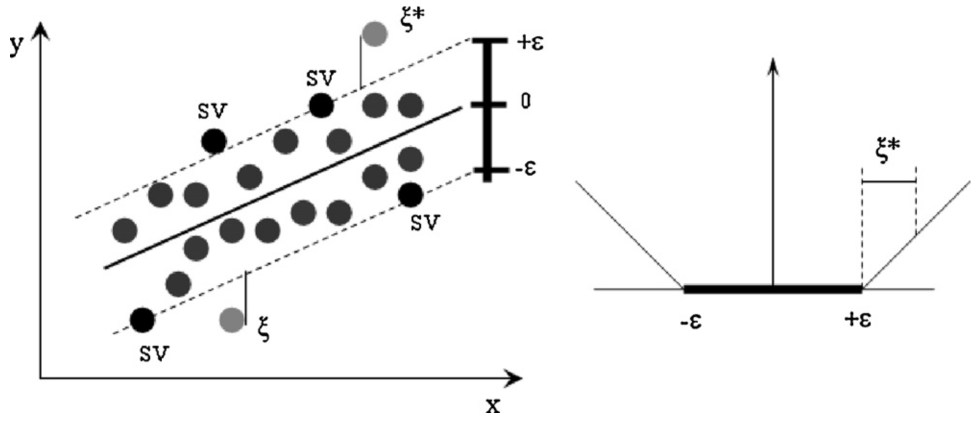

Fig. 2. The soft margin (ε-insensitive) loss function, consisting of support vectors (SVs), ξ, and ε for a linear SVR.

constructs a hyperplane in a high-dimensional space which allows a linear separation of the classes (see further, kernel function).

Later an SVM version for regression was proposed, called Support Vector Regression. When applied to regression problems a loss function is introduced. For example, quadratic, Laplace, Huber, and ε-insensitive functions are four possible loss functions [27].

In Support Vector Regression, the goal is to find a function $f(\mathbf{x})$ that has at the most ε deviation from the actually obtained targets y_i for all the training data, and at the same time is a linear function to link the nonlinear relationship between input and output data. Deviation larger than ε is not accepted. The function $f(\mathbf{x})$, the SVR function, can be represented as follows

$$f(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b \tag{1}$$

where $\phi(\mathbf{x})$ represents the nonlinear mapping of the training data in the high-dimensional space, $\mathbf{w}$ are the coefficients and b the bias term (b can be dropped if the mean is zero). For more detailed information on the estimation of $\mathbf{w}$ we refer to [27].

Fig. 2 shows the ε-insensitive loss function graphically. Three parameters determine the quality: C, the parameter controlling the trade-off between a large margin and less constrained violation, ε which is a precision parameter representing the radius of the tube located around the regression function $f(\mathbf{x})$, and the kernel function.

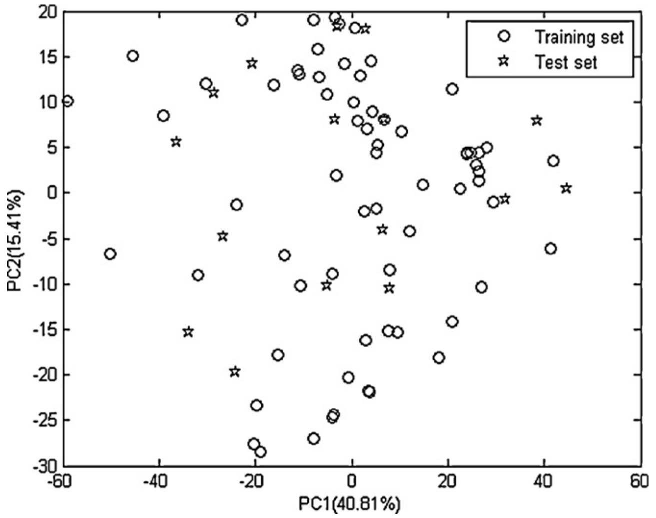

Fig. 3. The PCA score plot (PC1–PC2) of entire descriptor data set.

The kernel function used here is the Gaussian Radial Basis Function (RBF) kernel:

$$k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \tag{2}$$

where σ^2 denotes the width of the Gaussian kernel.

The parameters of SVM, i.e. C, σ^2, and ε were optimized by systematically changing their values in the training step and calculating the Mean Squared Error (MSE) of the model using 5-fold cross-validation.

2.4. Model building

The 2D structures of the molecules of Table 1 were drawn using HyperChem 7 software (Hypercube, Gainesville, Florida, United States). Then the structures were first pre-optimized with the Molecular Mechanics Force Field (MM+) procedure. Final geometries were obtained with the semi-empirical Austin Model 1 (AM1) method in Hyperchem, applying the Polak–Ribiere algorithm until the root mean square gradient reached 0.001 kcal mol^{-1} [28]. The resulting geometry was transferred into the Dragon program (Talete srl, DRAGON for Windows – Software for molecular descriptors calculation, Milano, Italy, 2007) in order to obtain 1497 descriptors, grouped in constitutional, Topological, Geometrical, Charge, GETAWAY (Geometry, Topology and Atoms-Weighted AssemblY), WHIM (Weighted Holistic Invariant Molecular descriptors), 3D-MoRSE (3DMolecular Representation of Structure based on Electron diffraction), Molecular Walk Count, BCUT, 2D-Autocorrelation, Aromaticity Index, Randic molecular profile, Radial Distribution Function, Functional group and Atom-Centred Fragment classes [29].

The calculated descriptors were first analyzed for the existence of constant or near constant variables, which were then removed. In addition, to decrease the redundancy in the descriptor data matrix, the descriptors' mutual correlation and that with the retention of the molecules was determined. The collinear descriptors (i.e. $r > 0.9$) were detected and those having the highest correlation with the retention were retained, while the others were removed from the data matrix.

The chromatographic data used were obtained from [9]. The QSRR models were validated through leave-one-out and leave-25%-out cross-validation, and external validation (with a test set), as well as by a y-randomization test, in which the y-block was shuffled, while the descriptors block was kept unaltered. The models were statistically evaluated by the squared correlation coefficient of the experimental versus predicted logarithms of the retention

Table 1
Compound names and their experimental (Exp) and predicted log k_w values from the ACO/MLR, GA/MLR, Relief/MLR, SR/MLR, ACO/SVM, GA/SVM, Relief/SVM, SR/SVM-based QSRR modeling.

No.	Compounds	Exp	ACO/MLR	GA/MLR	Relief/MLR	SR/MLR	ACO/SVM	GA/SVM	Relief/SVM	SR/SVM
1	Acebutolol	0.35	0.67	0.76	−0.07	0.64	0.25	0.25	0.25	0.45
2	Acetopromazine	2.93	3.06	3.15	3.32	3.15	2.97	2.95	3.03	2.9
3	2-Acetylphenothiazine	3.06	2.84	3.15	2.55	3.28	3.17	3.16	2.96	2.96
4[a]	Alprenolol	1.72	0.90	0.96	0.14	1.20	1.38	1.02	0.48	1.25
5	Antazoline	1.89	1.67	1.47	1.01	1.53	1.79	1.74	1.79	0.89
6	Astemizole	3.51	3.06	2.79	3.69	3.36	3.41	3.61	3.61	3.41
7	Atenolol	−1.05	−0.85	−0.38	−1.32	−0.92	−0.99	−0.41	−0.98	−0.95
8	Betaxolol	1.77	1.08	1.01	1.40	0.79	1.67	0.99	1.67	1.67
9	Bisoprolol	0.09	0.57	0.54	0.47	0.63	0.20	0.19	0.02	0.19
10	Brimonidine	0.17	0.06	0.17	1.11	0.25	0.14	0.07	0.27	0.27
11	Bupranolol	2.05	1.67	1.54	1.80	1.87	1.96	2.01	1.95	1.95
12	Carbamazepine	0.93	0.90	1.91	0.85	0.51	1.03	2.18	0.98	0.83
13[a]	Carteolol	0.23	−0.13	−0.03	−0.29	0.27	−0.16	−1.14	−0.52	−0.26
14[a]	Celiprolol	0.23	0.70	0.63	0.36	1.17	0.01	−0.37	0.95	0.16
15	Chloropyramine	2.77	2.19	2.16	2.52	2.37	2.62	2.51	2.67	2.81
16	Chlorpheniramine(+)	1.90	2.37	2.40	1.89	2.44	2.20	2.01	1.94	1.99
17[a]	Chlorpheniramine(+/−)	2.04	2.37	2.40	1.89	2.44	2.20	1.88	1.94	1.99
18	Chlorpromazine	4.08	3.75	3.84	3.62	3.85	3.78	4.16	3.97	3.98
19	Chlorprothixene	4.23	4.52	4.66	4.79	4.10	4.33	4.33	4.13	4.33
20	Cicloprolol	0.57	1.09	0.85	0.54	0.77	0.67	0.67	0.47	0.47
21	Cimetidine	0.72	1.14	1.08	0.84	1.16	0.62	0.8	0.62	0.64
22[a]	Cinnarizine	4.66	3.43	3.47	2.61	3.14	4.91	3.97	3.17	3.88
23	Cirazoline	1.58	1.33	0.45	1.16	0.98	1.48	1.48	1.48	1.68
24	Clomipramine	3.91	3.68	3.50	3.67	3.25	3.69	3.81	3.81	3.81
25	Clonidine	1.28	0.63	1.35	0.98	0.56	1.11	1.18	1.34	1.18
26	Desipramine	2.89	2.77	2.79	2.29	2.39	2.79	2.79	2.78	2.78
27[a]	Detomidine	1.63	1.52	0.97	1.19	1.33	1.66	1.43	1.43	2.17
28	Dilevalol	−1.26	−0.57	0.33	−0.21	−0.33	−0.24	−1.16	−1.16	−1.16
29	Dimethindene	2.24	3.12	2.89	2.64	3.23	2.34	2.61	2.34	2.14
30	Diphenhydramine	2.11	2.11	2.03	2.47	2.29	1.85	2.01	2.01	2.21
31	Doxazosin	2.82	1.63	1.97	2.16	1.84	2.72	2.72	2.72	2.72
32	Esmolol	0.92	0.83	1.16	1.05	0.81	1.02	0.82	0.98	1.02
33	Ethopropazine	4.18	3.49	3.74	4.35	4.19	3.70	4.08	4.08	4.08
34	Famotidine	0.19	−0.48	0.03	−0.08	−0.36	0.29	0.29	0.29	0.29
35	Fluphenazine	3.35	2.96	2.95	3.14	2.88	3.33	3.25	3.25	3.25
36	Imipramine	3.02	3.05	2.91	2.87	2.83	3.12	3.12	3.12	3.12
37	Indoramin	2.30	2.14	2.21	2.03	2.07	2.39	2.19	2.19	2.19
38	Isothipendyl	2.53	2.11	2.48	2.93	2.14	2.43	2.43	2.63	2.43
39	Ketotifen	1.95	2.98	2.79	2.29	2.80	1.85	2.19	2.05	2.05
40	Lofexidine	1.41	1.91	1.79	1.65	1.56	1.90	1.51	1.51	1.51
41	Medetomidine	2.52	1.82	1.23	2.30	1.61	1.90	2.12	2.41	2.41
42	Mepyramine	2.05	1.74	1.38	1.99	2.03	1.33	1.6	1.95	1.95
43	2-Methoxyphenothiazine	3.40	3.12	3.04	2.94	3.28	3.30	3.31	3.34	3.29
44	Metiamide	0.04	0.32	0.30	0.10	0.78	0.14	0.14	−0.05	0.14
45	Metoprolol	−0.55	0.09	0.21	−0.18	0.12	−0.45	0.21	−0.45	−0.45
46	Moxonidine	−1.12	−0.51	−0.56	−1.06	−0.49	−1.03	−1.02	−1.02	−1.02
47	Nadolol	−0.64	−1.29	−0.93	−0.60	−0.94	−0.54	−0.74	−0.54	−0.54
48	Naphazoline	1.48	1.82	1.93	1.81	1.74	1.58	1.58	1.57	1.57
49	Nifenalol	0.07	0.39	0.28	−0.39	0.70	0.17	0.17	−0.02	0.16
50	Nizatidine	−0.57	−0.35	−0.34	0.40	−0.76	−0.47	−0.47	−0.47	−0.47
51	Oxprenolol	1.22	0.95	0.82	0.68	1.19	0.15	1.12	1.12	1.11
52	Oxymetazoline	1.27	0.89	0.98	1.80	0.89	1.37	1.37	1.37	1.37
53[a]	Perphenazine	3.07	2.18	2.54	2.91	2.35	2.94	2.17	2.86	3.01
54	Pheniramine	1.27	1.97	1.91	1.25	1.92	1.37	1.38	1.37	1.17
55[a]	Phenothiazine	3.37	2.36	3.18	3.34	1.78	3.24	1.74	3.86	3.16
56	Phentolamine	−0.83	1.44	1.34	0.20	0.87	1.63	1.60	−0.73	0.11
57	Pindolol	0.33	−0.13	−0.10	−0.52	−0.16	0.43	0.43	0.43	0.23
58	Pizotifen	3.46	3.81	3.48	3.46	3.71	3.36	3.36	3.36	3.55
59	Practolol	−0.63	−0.30	−0.25	−0.79	−0.16	−0.53	−0.73	−0.64	−0.54
60[a]	Prazosin	1.17	1.31	1.09	1.54	1.32	0.45	0.40	0.95	1.69
61[a]	Prochlorperazine	3.52	3.02	3.19	3.28	3.32	3.48	2.89	3.29	3.58
62	Promazine	3.29	3.17	3.31	3.48	3.24	3.52	3.39	3.39	3.39
63[a]	Promethazine	3.22	3.47	3.69	3.73	3.87	3.52	3.58	2.97	3.56
64	Propiomazine	3.49	3.24	3.41	3.19	3.11	3.39	3.39	3.39	3.39
65[a]	Propranolol	2.04	1.36	1.78	1.59	1.99	1.63	2.07	1.84	1.64
66	Ranitidine	1.78	0.88	0.69	0.65	0.57	1.68	1.68	1.68	1.88
67[a]	Roxatidine acetate	1.15	1.66	1.61	1.20	1.91	1.50	1.19	0.30	2.62
68	Sotalol	−1.60	−0.87	−1.36	−1.06	−1.25	−1.50	−1.61	−1.50	−1.50
69	Terazosin	0.17	1.53	1.29	1.75	1.46	0.27	0.27	0.28	0.26
70	Tetryzoline	0.68	0.64	1.11	1.01	0.74	1.29	0.78	0.78	1.25
71	Thioridazine	4.65	4.98	5.39	4.06	5.11	4.55	4.75	4.55	4.55
72	Thiothixene-cis	2.77	3.22	2.59	2.62	2.79	2.67	2.67	2.67	2.67
73[a]	Tiamenidine	−0.23	−0.17	0.30	−0.04	−0.21	1.08	−0.27	−0.26	0.78
74	Timolol	0.17	−0.59	−1.24	−0.35	−0.46	0.27	0.07	0.07	0.27
75	Tolazoline	−0.06	−0.19	0.08	0.14	−0.15	0.04	0.04	0.04	−0.16

Table 1 (*Continued*)

No.	Compounds	Exp	ACO/MLR	GA/MLR	Relief/MLR	SR/MLR	ACO/SVM	GA/SVM	Relief/SVM	SR/SVM
76	Trifluoperazine	3.63	3.94	3.76	3.78	3.92	3.83	3.73	3.73	3.73
77	2-Trifluoromethylphenothiazine	4.80	4.17	3.89	4.59	4.44	4.70	4.70	4.70	4.70
78	Triflupromazine	4.12	4.32	4.09	4.50	4.39	4.02	4.02	4.11	4.22
79[a]	Trimeprazine	3.51	3.48	3.72	3.44	3.63	3.67	3.78	3.52	3.75
80	Tripelennamine	1.81	1.68	1.40	2.01	1.83	1.71	1.91	1.91	1.91
81[a]	Triprolidine	2.62	2.55	2.22	2.16	2.20	2.38	1.75	3.07	2.00
82	Tymazoline	2.01	1.24	0.82	1.97	1.46	1.91	1.83	1.91	1.88
83	Xylometazoline	2.38	1.68	1.82	2.14	1.79	1.82	1.71	2.28	2.23

[a] Compounds used in test set (external set).

Table 2
Correlation matrix and multicollinearity parameters for the descriptors selected by the Genetic Algorithm.

Descriptors	Descriptors							Multicollinearity parameters	
	BEHm5	ATS6e	Qneg	Mor30m	H4m	PSA	MLOGP	Tolerance	VIF
BEHm5	1							0.232	4.31
ATS6e	0.531	1						0.325	3.07
Qneg	0.319	0.210	1					0.357	2.80
Mor30m	0.004	0.000	0.038	1				0.948	1.05
H4m	0.400	0.187	0.428	0.006	1			0.413	2.42
PSA	0.153	0.001	0.404	0.026	0.352	1		0.310	3.23
MLOGP	0.103	0.150	0.051	0.014	0.015	0.219	1	0.445	2.24

BEHm5(*BCUT descriptors*): highest eigenvalue n.5 of Burden matrix/weighted by atomic masses; ATS6e (*2D autocorrelation*): Broto-Moreau autocorrelation of a topological structure −lag 6/weighted by atomic Sanderson electronegativities; Qneg (*Charge descriptors*): total negative charge; Mor30m (*3D-MoRSE descriptors*): 3D-MoRSE-signal 30/weighted by atomic masses; H4m (*GETAWAY descriptors*): H autocorrelation of lag 4/weighted by atomic masses; PSA (*Properties*): fragment-based polar surface area; MLOGP (*Properties*): Moriguchi octanol-water partition coefficient (log *P*)

Table 3
Correlation matrix and multicollinearity parameters for the descriptors selected by Stepwise Regression.

Descriptors	Descriptors									Multicollinearity parameters	
	MATS1v	GATS8e	Qpos	FDI	Mor30m	E3p	H4m	PSA	MLOGP	Tolerance	VIF
MATS1v	1									0.626	1.59
GATS8e	0.047	1								0.714	1.40
Qpos	0.001	0.005	1							0.377	2.65
FDI	0.016	0.003	0.091	1						0.741	1.35
Mor30m	0.005	0.059	0.038	0.027	1					0.780	1.28
E3p	0.155	0.049	0.002	0.104	0.068	1				0.572	1.75
H4m	0.028	0.011	0.429	0.024	0.006	0.019	1			0.468	2.14
PSA	0.026	0.012	0.404	0.000	0.026	0.000	0.352	1		0.361	2.77
MLOGP	0.061	0.076	0.051	0.000	0.014	0.034	0.015	0.218	1	0.524	1.91

MATS1v (*2D autocorrelation*): Moran autocorrelation −lag1/weighted by atomic van der Waals volumes; GATS8e (*2D autocorrelation*): Geary autocorrelation −lag8/weighted by atomic Sanderson electronegativities; Qpos (*Charge descriptors*): total positive charge; FDI (*Geometrical descriptors*): folding degree index; Mor30m (*3D-MoRSE descriptors*): 3D-MoRSE −signal 30/weighted by atomic masses; E3p (*WHIM descriptors*): 3rd component accessibility directional WHIM index/weighted by atomic polarizabilities; H4m (*GETAWAY descriptors*): H autocorrelation of lag 4/weighted by atomic masses; PSA (*Properties*): fragment-based polar surface area. MLOGP (*Properties*): Moriguchi octanol-water partition coefficient (log *P*).

Table 4
Correlation matrix and multicollinearity parameters for the descriptors selected by Relief.

Descriptors	Descriptors									Multicollinearity parameters	
	X3Av	BEHe5	MATS1p	GATS5e	Mor17m	Mor27m	Mor14e	HATS5e	MLOGP	Tolerance	VIF
X3Av	1									0.699	1.49
BEHe5	0.035	1								0.491	2.04
MATS1p	0.066	0.030	1							0.57	1.75
GATS5e	0.014	0.049	0.015	1						0.783	1.28
Mor17m	0.065	0.001	0.000	0.028	1					0.683	1.46
Mor27m	0.103	0.040	0.042	0.001	0.033	1				0.594	1.68
Mor14e	0.036	0.306	0.000	0.011	0.087	0.176	1			0.393	2.54
HATS5e	0.003	0.357	0.161	0.019	0.000	0.010	0.159	1		0.501	1.99
MLOGP	0.009	0.038	0.072	0.159	0.084	0.041	0.090	0.029	1	0.519	1.93

X3Av (*Topological descriptors*): Average valence connectivity index chi-3; BEHe5 (*BCUT descriptors*): highest eigenvalue n.5 of Burden matrix/weighted by atomic Sanderson electronegativities; MATS1p (*2D autocorrelation*): Moran autocorrelation −lag1/weighted by atomic polarizabilities; GATS5e (*2D autocorrelation*): Geary autocorrelation −lag5/weighted by atomic Sanderson electronegativities; Mor17m (*3D-MoRSE descriptors*): 3D-MoRSE signal 17/weighted by atomic masses. Mor27m (*3D-MoRSE descriptors*): 3D-MoRSE signal 27/weighted by atomic masses; Mor14e (*3D-MoRSE descriptors*):3D-MoRSE signal 14/weighted by atomic Sanderson electronegativities; HATS5e (*GETAWAY descriptors*): leverage-weighted autocorrelation of lag 5/weighted by atomic Sanderson electronegativities; MLOGP (*Properties*): Moriguchi octanol-water partition coefficient (log *P*).

Table 5
Correlation matrix and multicollinearity parameters for the descriptors selected by ACO.

Descriptors	Descriptors							Multicollinearity parameters	
	SRW07	MATS1v	GATS8e	Mor30m	H4m	PSA	MLOGP	Tolerance	VIF
SRW07	1							0.827	1.21
MATS1v	0.000	1						0.864	1.16
GATS8e	0.018	0.047	1					0.751	1.33
Mor30m	0.110	0.005	0.059	1				0.789	1.27
H4m	0.003	0.028	0.011	0.006	1			0.604	1.65
PSA	0.003	0.026	0.012	0.026	0.352	1		0.445	2.25
MLOGP	0.064	0.061	0.076	0.014	0.015	0.218	1	0.593	1.69

SRW07 (*Molecular walk counts*): self-retuning walk count of order 07; MATS1v (*2D autocorrelation*): Moran autocorrelation −lag1/weighted by atomic van der Waals volumes. GATS8e (*2D autocorrelation*): Geary autocorrelation −lag 8/weighted by atomic Sanderson electronegativities; Mor30m (*3D-MoRSE descriptors*): 3D-MoRSE signal 30/weigthed by atomic masses; H4m (*GETAWAY descriptors*): H autocorrelation of lag 4/weighted by atomic masses; PSA (*Properties*): fragment-based polar surface area; MLOGP (*Properties*): Moriguchi octanol-water partition coefficient (log *P*).

factors (log k_w) for the calibration and test sets (r^2 and q^2), the Root Mean Square Errors of calibration and external validation (RMSE), the Relative Standard Error of Prediction (RSEP), the Mean Absolute Error (MAE), Fischer test (*F*), *t*-test, Predicted REsidual Sum of Squares in **y** (PRESS), and the Total Sum of Squares (SST). It should be mentioned that RMSE is used for all the feature selection methods.

Table 6
Statistical parameters for the GA, SR, Relief, and ACO feature selection methods.

Parameters	Methods			
	GA	SR	Relief	ACO
S	0.691	0.596	0.526	0.620
R	0.915	0.940	0.954	0.932
S_{loo}	0.783	0.716	0.636	0.718
R_{loo}	0.890	0.913	0.932	0.909
$R_{l\text{-}25\%\text{-}o}$	0.791	0.837	0.875	0.850
$S_{l\text{-}25\%\text{-}o}$	1.083	0.990	0.862	0.952
S_{val}	0.726	1.158	1.163	0.832
R_{val}	0.932	0.858	0.882	0.921
S_{rand}	1.269	1.202	1.239	1.310
$\bar{S}_{rand}$	1.620	1.623	1.623	1.624
Fit	2.632	2.927	3.867	3.40

3. Results and discussion

3.1. Linear models

It is common practice to calculate large numbers of molecular descriptors to construct regression models for (bio)activity, retention or other properties. However, this makes the use of the well known MLR modeling impracticable, interpretation difficult, and does not avoid descriptor collinearity. Thus, variable selection is an indispensable task for the building of simple, predictive QSPR models. The recently implemented ACO method has demonstrated to be a valuable tool for this purpose [19], and was applied here to reduce the number of calculated Dragon descriptors (1497 descriptors) to just a few representatives. In addition, other feature selection techniques have been applied in order to evaluate the efficacy of ACO.

When building a QSRR model for retention two approaches are possible. Either one includes only descriptors with well known physicochemical properties to be able to explain the models and the importance of the descriptors in it. Else one can built models starting by selecting the theoretical descriptors from a large pool. The theoretical descriptors can not always evidently be linked to given properties but the quality of the obtained models is evaluated by their model-fit and predictive-properties describing parameters. This latter approach was followed in this paper.

Firstly, the dataset was split into two, a training set and test set. The training set comprised 67 compounds and the test set 16, i.e. 80% and 20% of the full data set, respectively. In fact, it was also taken into account that the training set covers the test set domain, which is checked for the descriptors by principal component analysis (PCA, score plot) and by the log k_w range for the retention data. The ranges of log k_w are from −1.6 to 4.8 and from −0.26 to 3.88 for the training and test sets, respectively. Fig. 3 shows the PC1–PC2 score plot, from PCA performed on the autoscaled descriptors. The results indicate that the training and the test sets cover the entire data set.

Table 7
Statistical measures obtained for the different QSRR models. The SVM parameters used in the models also are given.

Measure	Set	ACO		GA		SR		Relief	
		MLR	SVM	MLR	SVM	MLR	SVM	MLR	SVM
RMSE	Training set	0.58	0.41	0.65	0.40	0.55	0.21	0.48	0.09
	Test set	0.59	0.44	0.51	0.74	0.71	0.59	0.71	0.62
RSEP(%)	Training set	25.02	17.57	27.89	17.22	23.64	8.89	20.86	4.09
	Test set	23.38	17.55	20.41	29.29	28.17	23.54	28.31	24.88
MAE(%)	Training set	8.28	5.70	8.62	5.53	8.05	4.44	7.49	3.73
	Test set	17.05	14.19	16.65	19.03	18.11	16.94	16.98	17.02
R^2	Training set	0.87	0.94	0.84	0.94	0.88	0.98	0.91	0.99
	Test set	0.85	0.89	0.87	0.86	0.74	0.81	0.78	0.83
PRESS	Training set	22.68	11.19	28.21	10.74	20.26	2.87	15.77	0.61
	Test set	5.54	3.12	4.22	8.69	8.04	5.61	8.12	6.27
SVM parameters									
C			100.00		78.15		80.15		110.87
γ			0.001		0.07		0.09		1.00
ε			0.10		0.10		0.10		0.10
#SVs			63		61		60		58

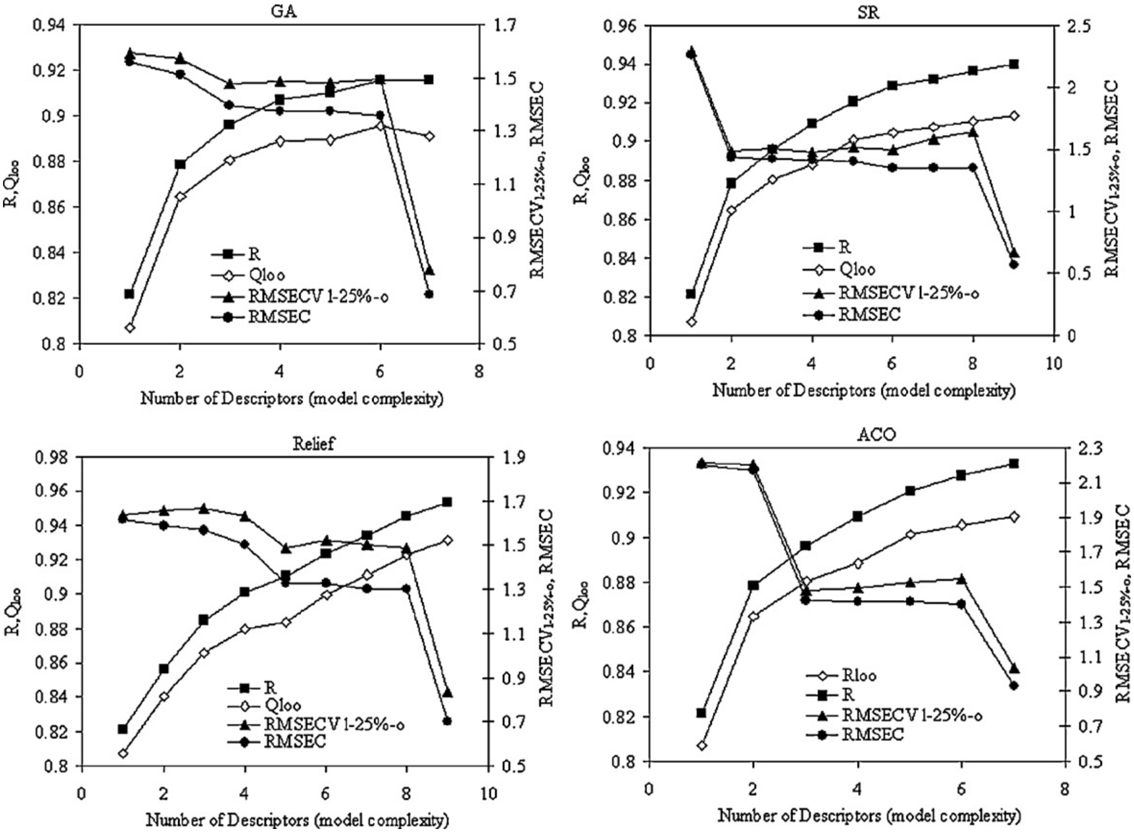

Fig. 4. The correlation coefficients of calibration (R) and leave-one-out (Q_{loo}), the root mean squared error of calibration (RMSEC) and leave-25%-out ($\mathrm{RMSECV}_{1\text{-}25\%\text{-}o}$) versus the model complexity.

Then, feature selection was performed based on the training set data to find the most important features. In this study we also used Genetic Algorithms, Stepwise Regression and the Relief method for feature selection. There is no general rationale for a given feature selection method as being better for all datasets. Different datasets have different properties, such as linearity/nonlinearity, and noise, which may require different selections to describe given properties.

With the GA-based method, seven descriptors were selected and the following MLR equation obtained:

$$\begin{aligned}\log k_{\mathrm{w}}\,(\mathrm{GA/MLR}) = {} & -4.73(\pm 1.9) + 1.23(\pm 0.79) \times \mathrm{BEHm5} \\ & + 0.00053(\pm 0.005) \times \mathrm{ATS6e} - 0.45(\pm 0.20) \\ & \times \mathrm{Qneg} + 2.44(\pm 0.79) \times \mathrm{Mor30m} \\ & - 1.43(\pm 0.70) \times \mathrm{H4m} + 0.03(\pm 0.01) \\ & \times \mathrm{PSA} + 1.09(\pm 0.09) \times \mathrm{MLOGP} \qquad (3)\end{aligned}$$

Table 2 shows the correlation matrix of the selected descriptors, their tolerance and Variance Inflation Factor (VIF), which show that multicollinearity is not exhibited by the selected descriptors. If VIF or tolerance assume values >10 or below 0.10, respectively, then multicollinearity is a problem. For VIF <5, no significant collinearity is present [30].

Stepwise Regression which is a combination of Forward Selection and Backward Elimination was also performed for feature selection. The algorithm starts by selecting the independent variable (descriptors) which has largest correlation with the dependent variable ($\log k_{\mathrm{w}}$); then that with the next highest correlation, and later incorporates a mechanism for eliminating earlier selected variables in the backward elimination phase in case these latter are not significant anymore. Each epoch of the selection procedure comprises an inclusion phase followed by an exclusion phase. The number of variables retained in the model is based on the levels of significance assumed for inclusion and exclusion of variables. Nine descriptors were selected by Stepwise Regression and the following equation was obtained.

$$\begin{aligned}\log k_{\mathrm{w}}(\mathrm{SR/MLR}) = {} & 5.49(\pm 2.91) - 6.04(\pm 1.61) \times \mathrm{MATS1v} \\ & + 0.50(\pm 0.21) \times \mathrm{GATS8e} - 0.32(\pm 0.172) \\ & \times \mathrm{Qpos} - 7.05(\pm 2.81) \times \mathrm{FDI} + 1.34(\pm 0.75) \\ & \times \mathrm{Mor30m} - 4.24(\pm 1.68) \times \mathrm{E3p} - 1.21(\pm 0.57) \\ & \times \mathrm{H4m} + 0.03(\pm 0.005) \times \mathrm{PSA} + 1.26(\pm 0.08) \\ & \times \mathrm{MLOGP} \qquad (4)\end{aligned}$$

Table 3 shows details of the nine descriptors selected using Stepwise Regression. This table also indicates that there is no high correlation or multicollinearity (VIF <5 and Tolerance >0.1) between the selected descriptors [31].

When the Relief method was applied, nine descriptors were selected. Table 4 shows the description of the selected descriptors and indicates that there is also no multicollinearity problem for these descriptors (VIF <5 and Tolerance >0.1). The following MLR equation was obtained.

$$\begin{aligned}\log k_{\mathrm{w}}(\mathrm{Relief/MLR}) = {} & -12.66(\pm 1.60) + 27.18(\pm 4.87) \times \mathrm{X3Av} \\ & + 2.50(\pm 0.45) \times \mathrm{BEHe5} - 4.77(\pm 1.16) \\ & \times \mathrm{MATS1p} + 0.76(\pm 0.20) \times \mathrm{GATS5e} \\ & + 1.30(\pm 0.29) \times \mathrm{Mor17m} + 1.47(\pm 0.47) \\ & \times \mathrm{Mor27m} - 0.91(\pm 0.17) \times \mathrm{Mor14e}\end{aligned}$$

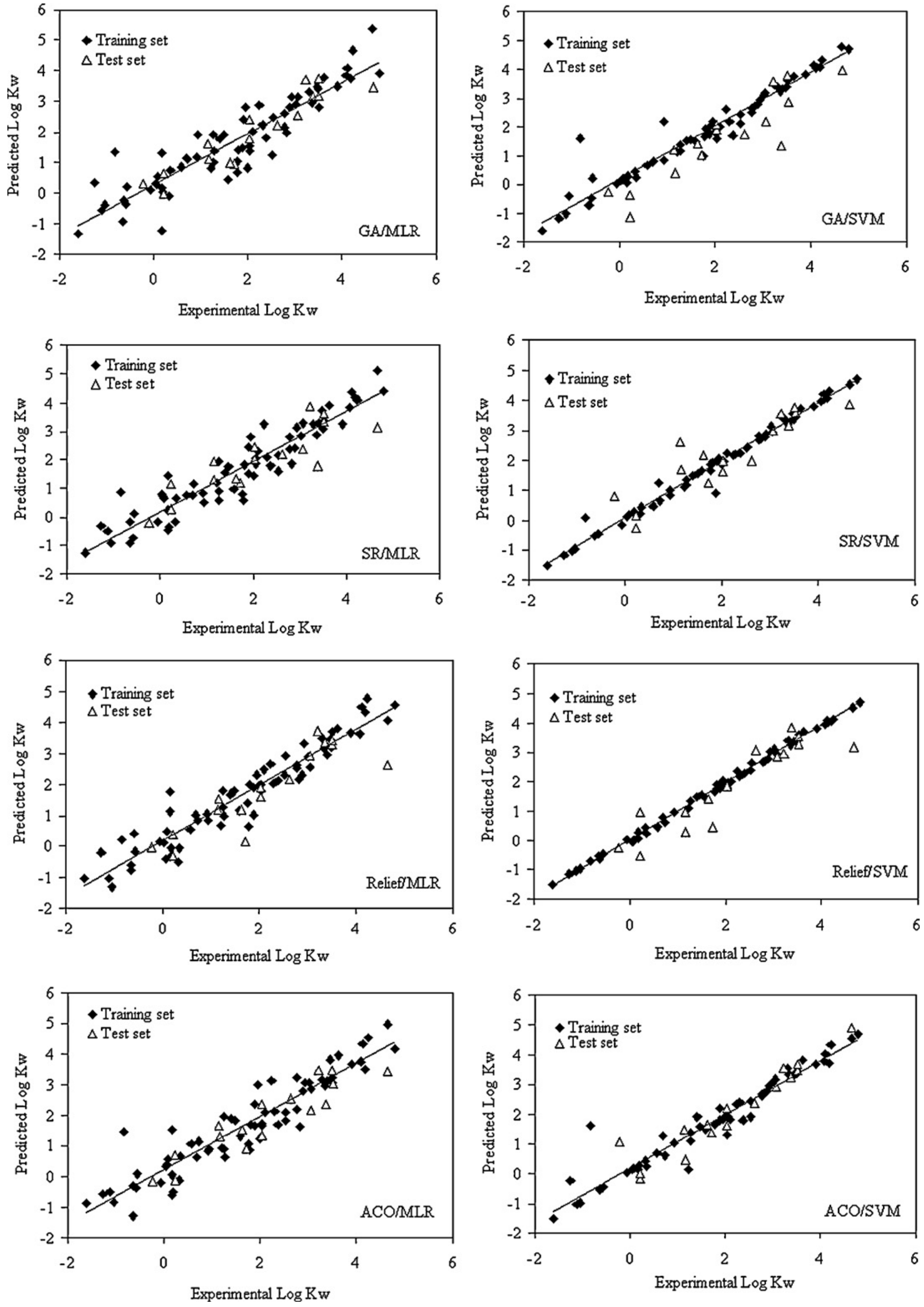

Fig. 5. Plot of experimental vs. predicted chromatographic retention $\log k_\mathrm{w}$ for the different models.

$$+2.58(\pm0.68)\times \mathrm{HATS5e}+1.22(\pm0.07)\times \mathrm{MLOGP} \qquad (5)$$

Finally, ACO-based feature selection was performed and the following MLR equation obtained;

$$\begin{aligned}\log k_\mathrm{w}(\mathrm{ACO/MLR}) = {} & -2.93(\pm0.31)+0.00263(\pm0.001)\times \mathrm{SRW07} \\ & -5.37(\pm1.43)\times \mathrm{MATS1v}+0.63(\pm0.21) \\ & \times \mathrm{GATS8e}+2.26(\pm0.77)\times \mathrm{Mor30m} \\ & -1.54(\pm0.52)\times \mathrm{H4m}+0.0298(\pm0.005) \\ & \times \mathrm{PSA}+1.22(\pm0.07)\times \mathrm{MLOGP} \qquad (6)\end{aligned}$$

The ACO optimization process converged to the seven descriptors depicted in Table 5.

Table 6 shows the statistical parameters calculated for the different feature selections. *R* and *S* are the correlation coefficient

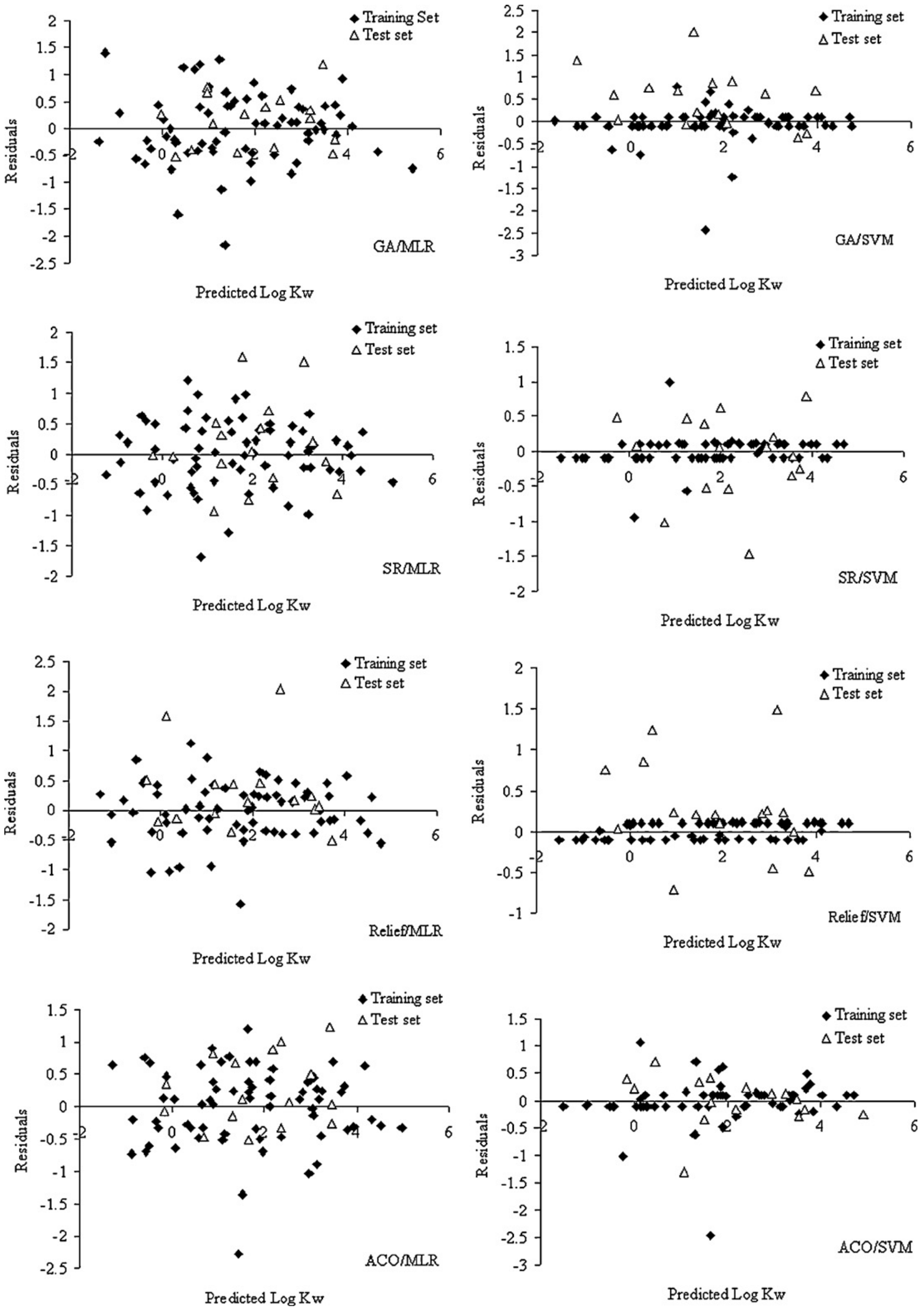

Fig. 6. Residuals plot for the different models.

and standard deviation, respectively, of the training set, p is the significance of the model, and Fit is the Kubiny function [32], for which the larger the value indicates the better the fit to the linear equation. It should be noted that leave-one-out (LOO) and leave-25%-out (L-25%-o) cross-validation techniques measure the internal validation of the developed QSRR upon inclusion/exclusion of compounds. We have also performed y-randomization, where S_{Rand} is the smallest standard deviation from 100,000 cases of randomization for the model. The $\bar{S}_{\text{Rand}}$ is the average of the 100,000 y-randomization cases. Thus, as $S_{\text{Rand}} > S$ and $\bar{S}_{\text{Rand}} > S$ this indicates that the obtained correlation is not fortuitous and results in a real structure-retention relationship, i.e. it is not a chance correlation.

When we compare the statistical parameters of the four feature selection methods employed in this study, all methods were able to find relevant features.

In order to find the optimum number of descriptors from those selected, for all feature selection techniques, models with different complexities were built. First the model with one descriptor was constructed and then descriptors were progressively added until all were used, evaluating the model each time. Fig. 4 shows the Root Mean Squared Error of Calibration (RMSEC), the Root Mean Squared Error of leave-25%-out ($\text{RMSECV}_{\text{l-25\%-o}}$), the correlation coefficient (R), and the correlation coefficient of leave-one-out cross-validation (R_{loo}) for all feature selection methods and all models. The higher the numbers of descriptors used in the models the more suitable models are obtained, i.e. the best fitting models with the best predictive properties. This is seen from the facts that R and Q_{loo} are continuously increasing, while the RMSE-values continue decreasing. A notable observation is that for all methods, a large improvement in prediction (large drop in RMSE-values) is obtained, after which the stopping criterion of the method has been met (since no more complex models were built). The figures also show that the built models do not seem to overfit the data (no increase obtained in the RMSE-values for the more complex models).

In order to find the most important descriptors in any of the models 1 to 4, standardization of the regression coefficients [33] was performed.

One should take into account the size of the regression coefficients for their comparison in a modeling. This is difficult when the variables are measured in different units. On the other word, we cannot compare the size of the various coefficients because the independent variables are measured on different scales and thus modeled. Therefore we need either to scale the descriptors before modeling resulting in a comparable coefficient or use a metric to compare the different coefficients to each other, which is the case here. Using standardized coefficients then helps to overcome this problem.

Standardized regression coefficients can be calculated as following:

$$B_i = \hat{\beta}_i \left(\frac{s_i}{s_y} \right) \quad i = 1, \ldots, k \tag{7}$$

where B_i is the standardized regression coefficient, $\hat{\beta}_i$ and s_i the regression coefficient and the standard deviation of the ith independent variable, respectively, and s_y is the standard deviation of the dependent variable.

As a result the following ranking of the contributions to $\log k_{\text{w}}$ is achieved:

$$\text{Model}\ \ 1(\text{GA/MLR});\quad \underset{(0.913)}{\text{MLOGP}} > \underset{(0.499)}{\text{PSA}} > \underset{(-0.192)}{\text{Qneg}} > \underset{(0.170)}{\text{BEHm5}} > \underset{(-0.167)}{\text{H4m}} > \underset{(0.167)}{\text{Mor30m}} > \underset{(0.010)}{\text{ATS6e}}$$

$$\text{Model}\ \ 2(\text{SR/MLR});\quad \underset{(1.053)}{\text{MLOGP}} > \underset{(0.520)}{\text{PSA}} > \underset{(-0.213)}{\text{MATS1v}} > \underset{(-0.151)}{\text{E3p}} > \underset{(-0.140)}{\text{H4m}} > \underset{(-0.138)}{\text{Qpos}} > \underset{(-0.132)}{\text{FDI}} > \underset{(0.128)}{\text{GATS8e}} > \underset{(0.092)}{\text{Mor30m}}$$

$$\text{Model}\ \ 3(\text{Relief/MLR});\quad \underset{(1.020)}{\text{MLOGP}} > \underset{(-0.341)}{\text{Mor14e}} > \underset{(0.312)}{\text{BEHe5}} > \underset{(0.272)}{\text{X3Av}} > \underset{(0.218)}{\text{Mor17m}} > \underset{(-0.216)}{\text{MATS1p}} > \underset{(0.215)}{\text{HATS5e}} > \underset{(0.168)}{\text{GATS5e}} > \underset{(0.163)}{\text{Mor27m}}$$

$$\text{Model}\ \ 4(\text{ACO/MLR});\quad \underset{(1.018)}{\text{MLOGP}} > \underset{(0.445)}{\text{PSA}} > \underset{(-0.189)}{\text{MATS1v}} > \underset{(-0.179)}{\text{H4m}} > \underset{(0.160)}{\text{GATS8e}} > \underset{(0.154)}{\text{Mor30m}} > \underset{(0.105)}{\text{SRW07}}$$

MLOGP (Moriguchi octanol-water partition coefficient ($\log P$)) [34], selected by all methods, is the most important descriptor in all models for the prediction of $\log k_{\text{w}}$. Its selection is not a surprise as was already discussed earlier [9]. PSA also seems of a considerable importance. It is selected by three of the four methods. Again, its selection could be expected since it is a descriptor frequently used in QSPR/QSRR modeling. The other descriptors are usually less important and are fine-tuning the modeling and prediction. This may explain their diversity in the different models. Even though these selectors are different in the different models, they may stand for similar properties and be thus related. However, as mentioned already higher, it is not our purpose to study the individual descriptors. The Relief method, which did not select PSA, required the selection of the highest number of descriptors to model the retention. A notable observation is that the feature selection methods only select PSA and MLOGP as the two last variables before stopping. However, though PSA is important, this is not seen from Fig. 4. The measured and predicted values for all compounds from the different models are shown in Table 1.

3.2. Non-linear models

As mentioned before, several parameters need to be optimally set for SVMs: controller of trade off C, γ and ε-insensitive loss function. The models were built using the descriptors that were selected by the different feature selection methods when building the linear models. In this study we used 5-fold cross-validation and optimized the values based on the accuracy (MSE) of the resulting model. Different values of C ranging from 1 to 500, of γ ranging from 0.001 to 5, and for ε ranging from 0.1 to 5 are evaluated following a grid search. When the values of C, γ, and ε increased the mean squared error of 5-fold cross-validation decreased. The best parameters for γ, C, ε, and the number of SVs obtained are presented in Table 7. The results of the predictions by the nonlinear models constructed by SVM using the different features selected by ACO, Relief, GA and SR are shown in Table 7.

A crucial aspect for QSRR model development is validation. Generally, the most conclusive proof of the predictive capacity of a QSRR model is from external validation. In this study both external and internal validations are considered. For the evaluation of the multivariate calibration models, several statistical measures, described previously, were used. Table 7 shows the parameters to evaluate the quality of the constructed models. The following conclusions can, for instance, be made based on the RMSE results. This table shows that all techniques are suitable for feature selection since they result in similar values for the different models, both for the training and test sets. The descriptors selected by ACO combined with SVM as modeling technique gave the best model (the RMSE-values for both training and test sets are smallest and similar). It is also better than the linear models since RMSE is smaller than for the MLR models.

Although Relief/SVM has the lowest RMSE value (0.09) for the training set, for the prediction it is 0.62, which is not as good as ACO/SVM with 0.41 and 0.44 for training and test sets, respectively. Similar conclusions as from RMSE can be drawn from the other parameters from Table 7.

Predictions for the external validation (test) set were carried out using the calibration models, resulting in accurate predicted $\log k_{\text{w}}$ values, especially for the ACO/SVM model (r^2 of 0.899 and RMSEP of 0.441), as illustrated in Figs. 5 and 6. The residuals distributions in Fig. 6 indicate the absence of systematic error. Both MLR and SVM regression models gave very high correlation between "experimental" and fitted/predicted $\log k_{\mathrm{w}}$, with SVM notably superior, mainly for the training, less for the test set.

From Fig. 5 we also see that the calibration errors for the SVM models are always smaller than for the MLR models. The very low RMSE errors when using the Relief and SR descriptors in SVM can also be seen. These are the techniques that selected the highest number of descriptors. It is possible that these models start to overfit given their very low RMSE for the training set and the considerably higher values for the test set. Also, the SVM models built after the elimination of the least important descriptors may perform generally better than the actual. For the models built with the ACO and GA descriptors the calibration errors are intermediate and for the ACO/SVM model the predictions are best. For the GA/SVM model the prediction is worse, because many test set compounds are predicted too low.

4. Conclusion

In this study, Multiple Linear Regression and Support Vector Regression were used to build Quantitative Structure-Retention Relationships. Ant Colony Optimization, a Genetic Algorithm, Stepwise Regression and the Relief method were used to select the most important descriptors. The results indicated that ACO/SVM was the best. For the linear models, the four feature selection methods lead to similar results, although the selected descriptors were different. For some SVM models, some feature selection techniques (Relief and Stepwise) seem to select too many descriptors possibly leading to an over-fitting to the training set data. From the selected descriptors, MLOGP (Moriguchi octanol-water partition coefficient ($\log P$)) is the most important for the prediction of chromatographic retention $\log (k_{\mathrm{w}})$ of drug compounds. This observation was also to be expected from a chromatographic point of view.

Overall, all models behaved rather similarly regarding prediction, so both linear (MLR equations) and nonlinear models (SVMR) can be built to predict chromatographic retention of drug compounds. A slight preference may go to the ACO/SVMR model. The ACO method showed to be valuable for feature selection mainly for SVM modeling, since few variables were selected from a pool of descriptors, and models with good predictive properties were obtained.

References

[1] P. Jandera, in: K. Valko (Ed.), Handbook of Analytical Separations, vol. 1, Elsevier, Amsterdam, 2000, p. 1 (Chapter 1).

[2] R. Kaliszan, J. Chromatogr. B 715 (1998) 229.

[3] L.A. Lopez, S.C. Rutan, J. Chromatogr. A 965 (2002) 301.

[4] B. Lučić, N. Trinajstić, S. Sild, M. Karelson, A.R. Katritzky, J. Chem. Inf. Comput. Sci. 39 (1999) 610.

[5] X. Gironés, R. Carbó-Dorca, J. Chem. Inf. Comput. Sci. 42 (2002) 317.

[6] E.R. Collantes, W. Tong, W.J. Weish, W.L. Zielinski, Anal. Chem. 68 (1996) 2038.

[7] F.A.L. Ribeiro, M.M.C. Ferreira, J. Mol. Struct. 663 (2003) 109.

[8] A.R. Timerbaev, O.P. Semenova, O.M. Petrukhin, Electrophoresis 23 (2002) 1786.

[9] R. Put, C. Perrin, F. Questier, D. Coomans, D.L. Massart, Y. Vander Heyden, J. Chromatogr. A 988 (2003) 261.

[10] R. Todeschini, P. Gramatica, Quant. Struct. Act. Rel. 16 (1997) 120.

[11] M. Randic, G.M. Brissey, R.B. Spencer, C.L. Wilkins, Comput. Chem. 3 (1979) 5.

[12] I. Guyon, A. Elisseeff, J. Mach. Learn. 3 (2003) 1157.

[13] H. Liu, H. Motoda, Feature Selection for Knowledge Discovery and Data Mining, Kluwer Academic Publishers, Norwell, MA, USA, 1998.

[14] H. Liu, L. Yu, IEEE Trans. Knowl. Data Eng. 17 (2005) 491.

[15] M.H. Nguyen, F.D.L. Torre, Pattern Recogn. 43 (2010) 584.

[16] E. Bonabeau, M. Dorigo, G. Theraulez, Swarm Intelligence: From Natural to Artificial Systems, Oxford University Press, New York, 1999.

[17] R. Jensen, in: A. Abraham, C. Grosan, V. Ramos (Eds.), Swarm Intelligence and Data Mining, Springer, Heidelberg, 2006, p. 45.

[18] M. Dorigo, C. Blum, Theor. Comput. Sci. 344 (2005) 243.

[19] M. Goodarzi, M.P. Freitas, R. Jensen, J. Chem. Inf. Model. 49 (2009) 824.

[20] D.J. Livingstone, D.W. Salt, J. Med. Chem. 48 (2005) 661.

[21] K. Bodzioch, A. Durand, R. Kaliszan, T. Bączek, Y. Vander Heyden, Talanta 81 (2010) 1711.

[22] E. Deconinck, Q.S. Xu, R. Put, D. Coomans, D.L. Massart, Y. Vander Heyden, J. Pharm. Biomed. Anal. 39 (2005) 1021.

[23] R. Put, Q.S. Xu, D.L. Massart, Y. Vander Heyden, J. Chromatogr. A 1055 (2004) 11.

[24] K. Kira, L.A. Rendell, Proceedings AAAI-92, San Jose, CA, MIT Press, Cambridge, MA, 1992, p. 129.

[25] K. Kira, L.A. Rendell, Proceedings 9th International Conference on Machine Learning, Aberdeen, Scotland, Morgan Kaufmann, Los Altos, CA, 1992, p. 249.

[26] I. Kononenko, in: F. Bergadano, L. De Raedt (Eds.), European Conference on Machine Learning, LNCS 784, Springer-Verlag New York, Secaucus, NJ, USA, 1994, p. 171.

[27] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.

[28] D.C. Young, Computational Chemistry: A Practical Guide for Applying Techniques to Real-World Problems, John Wiley & Sons, New York, 2001.

[29] V. Consonni, R. Todeschini, Handbook of Molecular Descriptors, Wiley-VCH, Weinheim, 2000.

[30] S. Chatterjee, B. Price, Regression Analysis by Example, 2nd ed., John Wiley & Sons, New York, 1991.

[31] K. Kira, L. Rendell, Proceedings of International Workshop on Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1992.

[32] H. Kubiny, Quant. Struct. Act. Rel. 13 (1994) 285.

[33] N.R. Draper, H. Smith, Applied Regression Analysis, John Wiley & Sons, New York, 1981.

[34] I. Moriguchi, S. Hirono, Q. Liu, I. Nakagome, Y. Matsushita, Chem. Pharm. Bull. 40 (1992) 127.