Mol Divers (2010) 14:225–235
DOI 10.1007/s11030-009-9163-7

FULL-LENGTH PAPER

A combined LS-SVM & MLR QSAR workflow for predicting the inhibition of CXCR3 receptor by quinazolinone analogs

Antreas Afantitis · Georgia Melagraki · Haralambos Sarimveis · Panayiotis A. Koutentis · Olga Igglessi-Markopoulou · George Kollias

Received: 5 March 2009 / Accepted: 9 May 2009 / Published online: 30 May 2009

Abstract A novel QSAR workflow is constructed that combines MLR with LS-SVM classification techniques for the identification of quinazolinone analogs as "active" or "non-active" CXCR3 antagonists. The accuracy of the LS-SVM classification technique for the training set and test was 100% and 90%, respectively. For the "active" analogs a validated MLR QSAR model estimates accurately their I-IP10 IC_{50} inhibition values. The accuracy of the QSAR model ($R^2 = 0.80$) is illustrated using various evaluation techniques, such as leave-one-out procedure ($R^2_{\mathrm{LOO}} = 0.67$) and validation through an external test set ($R^2_{\mathrm{pred}} = 0.78$). The key conclusion of this study is that the selected molecular descriptors, Highest Occupied Molecular Orbital energy (HOMO), Principal Moment of Inertia along X and Y axes PMIX and PMIZ, Polar Surface Area (PSA), Presence of triple bond (PTrplBnd), and Kier shape descriptor ($^1\kappa$), demonstrate discriminatory and pharmacophore abilities.

A. Afantitis (✉) · G. Kollias (✉)
Biomedical Sciences Research Center "Alexander Fleming",
Athens, Greece
e-mail: afantitis@fleming.gr

G. Kollias
e-mail: kollias@fleming.gr

A. Afantitis
Department of ChemInformatics, NovaMechanics Ltd, Nicosia, Cyprus
e-mail: afantitis@novamechanics.com

G. Melagraki · H. Sarimveis · O. Igglessi-Markopoulou
School of Chemical Engineering, National Technical University of Athens, Athens, Greece

P. A. Koutentis
Department of Chemistry, University of Cyprus, P.O. Box 20537, 1678 Nicosia, Cyprus

Keywords QSAR · CXCR3 · Molecular modelling · In silico predictions

Introduction

Recent reports from several major pharmaceutical companies indicate that there is significant interest in the identification of small-molecule antagonists of CXCR3 [1,2]. CXCR3 and its ligands Mig (CXCL9), IP-10 (CXCL10), and ITAC (CXCL11) have been involved in a variety of inflammatory diseases such as rheumatoid arthritis, multiple sclerosis, inflammatory bowel diseases [2,3]. Recently the quinazoline analogs [4,5] were identified as promising functional antagonists of CXCR3 that could be developed into new therapeutic agents for the treatment of inflammatory disorders.

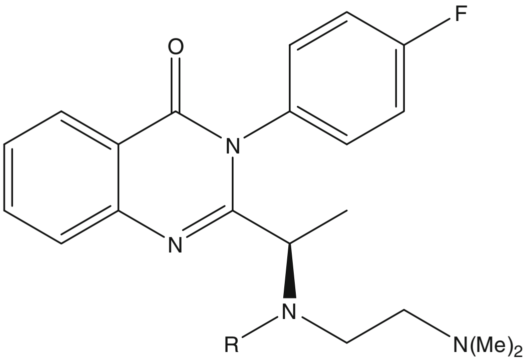
In the past, several attempts have been made to build QSAR models in the general field of chemokine antagonists such as CCR5 [6–8], CCR2 [9,10], CXCR2 [11], and CXCR4 [12]. Previously, we reported the first QSAR study concerning small-molecule antagonists of CXCR3 [13].

In this work we show that a combination of ES-SWR (Forward Selection & Backward Elimination)–Multiple Linear Regression (MLR) QSAR modeling [14] and Least Squares-Support Vector Machines (LS-SVM) classification techniques [15] can contribute greatly to building a workflow with a double role. Firstly, discriminate compounds into two categories (threshold $IC_{50} = 790$ nM), "actives" (IC_{50} range: 0.8–790 nM), and "non-actives" (IC_{50} range: 1,370–25,000 nM), and secondly, accurately estimate the ^{125}I-IP10 IC_{50} values for the active small molecules.

Materials and methods

In this computational study, 55 novel CXCR3 antagonists (quinazolinone analogs) were collected from Medina et al.'s

Table 1 Dataset (amide moiety) and model predictions using LS-SVM



Id	*R*	I-IP10 IC_{50} (nM) (experimental)	I-IP10 $\log(1/IC_{50})$ (experimental)	Class threshold (IC_{50} 790 nM)	Predicted class
1	$–CO \cdot (CH_2)_8CH_3$	146	−2.16	Active	
2	$–CO \cdot (CH_2)_{10}CH_3$	154	−2.19	Active	
3[a]	$–CO \cdot (CH_2)_6CH_3$	375	−2.57	Active	Non-active[b]
4	$–CO \cdot (CH_2)_4CH_3$	10,000	−4.00	Non-active	
5	$–(CH_2)_8CH_3$	710	−2.85	Active	
6	$–(CH_2)_7CH_3$	790	−2.90	Active	
7	$–S(O)_2(CH_2)_9CH_3$	587	−2.77	Active	
8[a]	$–S(O)_2(CH_2)_7CH_3$	1,400	−3.15	Non-active	Active[b]
9	$–CO \cdot CH_2Ph–4–Ph$	75	−1.88	Active	
10	$–CO \cdot Ph–4–Ph$	10,000	−4.00	Non-active	
11[a]	$–CO \cdot CH_2Ph$	10,000	−4.00	Non-active	Non-active
12	$–CO \cdot CH_2Ph–4–CH_3$	10,000	−4.00	Non-active	
13	$–CO \cdot CH_2Ph–4–CF_3$	88	−1.95	Active	
14	$–CO \cdot CH_2Ph–4–OCF_3$	156	−2.19	Active	

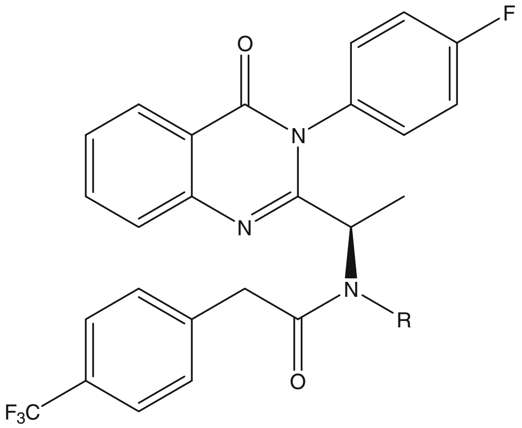
[a] Test set
[b] Misclassified compound

[4,5] recently published work (Tables 1–6). Before the calculation of the molecular descriptors, the chemical structures were fully optimized using PM6 [16,17], which provided a balance between computational speed and accuracy. A recent paper [18] highlighted the quality of QSPR models obtained by PM6 method as similar to that of models based on B3LYP (Density Functional Theory). Then approximately 200 physicochemical constants, topological and structural molecular descriptors, were calculated using Chem3D [19], Topix [20], MOPAC2007 [21], and ROCS & EON [22]. For the development of the workflow the available small molecules were separated (55 quinazolinone analogs) into two independent sets, "actives" and "non-actives." The cutoff value for the discrimination of "actives" and "non-actives" was set to $IC_{50} =$ 790 nM. The separation of the dataset into training and validation sets was performed according to the popular Kennard and Stones algorithm [23–25]. The algorithm starts by finding two samples that are the farthest apart from each other on the basis of the input variables in terms of some metric, e.g., the Euclidean distance. These two samples are removed from the original dataset and put into the calibration dataset. The procedure described is repeated until the desired number of samples has been reached in the calibration set. The advantage of the specific algorithm was that the calibration samples mapped the measured region of the input variable space completely with respect to the induced metric and that the test samples all fell inside the measured region. A commonly used ratio of training to validation objects (75:25) was also adopted in this work [26]. The training set contains 40 compounds (32 "actives" and 8 "non-actives") and the test set 15 compounds (10 "actives" and 5 "non-actives").

The ES-SWR algorithm [14] was used on the training dataset (40 compounds) to select the most appropriate descriptors. ES-SWR combines the advantages of both Forward Selection (FS-SWR) and Backward Elimination (BE-SWR). Forward Selection is computationally efficient for the generation of nested subsets of variables. On the other hand Backward Selection eliminates the most appropriate variable, so that the remaining variables perform best [27]. The objective of the variable selection was to determine the optimum set of descriptors that produce the most significant QSAR models linking and interpreting the chemical structure of the small molecules with their functional activity [27].

Firstly, Least Squares–Support Vector Machines (LS-SVM) classification techniques [15] were applied for the discrimination and the investigation of the pharmaco-

Table 2 Dataset (*N*-substituents) and model predictions using LS-SVM



Id	*R*	I-IP10 IC_{50} (nM) (experimental)	I-IP10 log(1/IC_{50}) (experimental)	Class threshold (IC_{50} 790 nM)	Predicted class
15	–$(CH_2)_2$OMe	300	−2.48	Active	
16	–$(CH_2)_2$OEt	40	−1.60	Active	
17	–$(CH_2)_2CH_3$	10,000	−4.00	Non-active	
18	–CH_2–2-thiazolyl	100	−2.00	Active	
19	–CH_2–2-imidazoyl	230	−2.36	Active	
20[a]	−CH_2–4-imidazoyl	650	−2.81	Active	Active
21	–CH_2–4-(1-methyl-imidazoyl)	240	−2.38	Active	
22[a]	−CH_2–2-pyridyl	73	−1.86	Active	Active
23	–CH_2–3-pyridyl	13	−1.11	Active	

[a] Test set

phore ability of the selected by ES-SWR algorithm molecular descriptors to the above data (55 quinazolinone analogs). For constructing the SVM model the LS-SVM package [15] was used after scaling both training and validation data in the range [0,1], for "non-active" (IC_{50} range: 1,370–25,000 nM) and "active" (IC_{50} range: 0.8–790 nM) molecules, respectively. The threshold was defined to $IC_{50} = 790$ nM in order to have a considerable distance between the nearest lower and higher experimental values. The gap (580 nM) between 790 and 1,370 nM is the largest between two consecutive data in terms of IC_{50} in the available experimental dataset.

The Kernel type that was adopted in the present work was the Radial Basis Function (RBF). The first task was the assignment of each molecule to one class, namely "actives" or "non-actives," based on a cutoff value that was set to $IC_{50} = 790$ nM. The bandwidth σ^2 and the regularization parameter γ in the kernel function were optimized to achieve the best possible discrimination between classes. The optimized values obtained were $\sigma^2 = 5$ and $\gamma = 400$ [15].

For the performance evaluation of the SVM models, several statistical tests such as recall (or sensitivity), specificity, accuracy, precision, and *F*-measure were used [28,29]. Recall (or sensitivity) and Specificity are able to identify the discrimination ability of the SVM model and accuracy presents the ratio of the correctly discriminated classes. *F*-measure is a function of Recall and Precision which indicate the accuracy of real and estimated class, respectively. According to Fawcett et al. [29], for the calculation of the above statistics the Confusion Matrix (Table 7) should be constructed. In Table 7, TA indicates that active compound is correctly classified as active (TA); FA indicates that active compound is wrongly classified as active (FA); FN indicates that non-active compound is wrongly classified as active (FN); and TN indicates that non-active compound is correctly classified as non-active (TN).

$$\text{Recall} = \frac{\text{TA}}{\text{TA} + \text{FN}} \tag{1}$$

$$\text{Precision} = \frac{\text{TN}}{\text{FA} + \text{TN}} \tag{2}$$

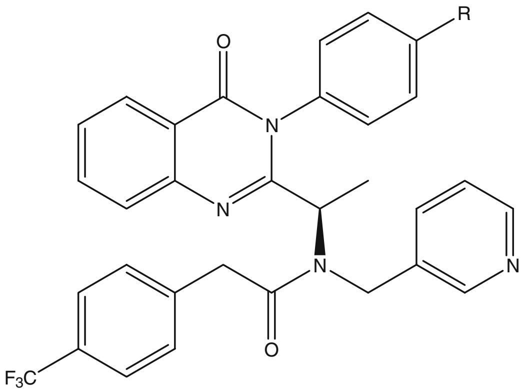
$$\text{Specificity} = \frac{\text{TA}}{\text{TA} + \text{FA}} \tag{3}$$

$$\text{Accuracy} = \frac{\text{TA} + \text{TN}}{\text{TA} + \text{FA} + \text{FN} + \text{TN}} \tag{4}$$

$$F\text{-measure} = \frac{2(\text{Recall})(\text{Precision})}{\text{Recall} + \text{Precision}} \tag{5}$$

In the second stage, a MLR QSAR model was developed by applying the selected molecular descriptors to the 32 "actives" small molecules of the dataset. The QSAR model was

Table 3 Dataset (4-phenyl substitutions) and model predictions using LS-SVM



Id	*R*	I-IP10 IC_{50} (nM) (experimental)	I-IP10 $\log(1/IC_{50})$ (experimental)	Class threshold IC_{50} 790 nM	Predicted class
24	–H	299	−2.48	Active	
25	–F	22	−1.34	Active	
26	–Cl	25	−1.40	Active	
27[a]	–Me	14	−1.15	Active	Active
28	–OEt	6	−0.78	Active	
29	$-C\equiv CCH_3$	4	−0.60	Active	
30	$-NO_2$	7	−0.85	Active	
31[a]	$-C\equiv N$	11	−1.04	Active	Active
32[a]	$-SO_2Me$	10,000	−4.00	Non-active	Non-active
33	$-CO_2H$	1,370	−3.14	Non-active	
34	–NHAc	25,000	−4.40	Non-active	

[a] Test set

evaluated for its robustness, accuracy, and reliability (Table 9, Fig. 1).

To illustrate this, the following evaluation techniques were used: the leave-one-out (LOO) cross-validation procedure, validation through an external test set, and Y-randomization [26,30,31].

Specifically for the external validation based on the validation set, the following criteria were used:

$$R^2_{\text{pred}} > 0.6 \tag{6}$$

$$\frac{(R^2_{\text{pred}} - R^2_o)}{R^2_{\text{pred}}} \quad \text{or} \quad \frac{(R^2_{\text{pred}} - R'^2_o)}{R^2_{\text{pred}}} \leq 0.1 \tag{7}$$

$$k \text{ or } k' \approx 1 \tag{8}$$

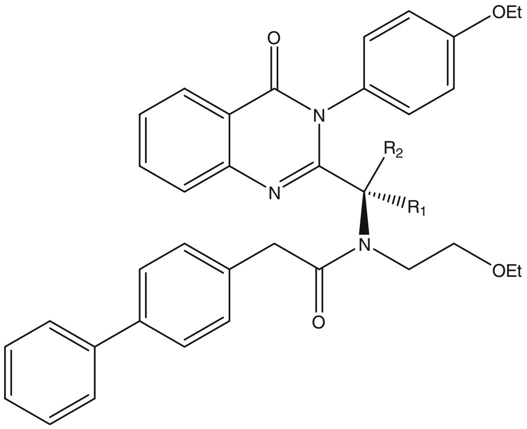
In Eqs. 6 and 7, R^2_{pred} is the coefficient of determination between experimental values and model predictions on the validation set. Mathematical definitions of R^2_o, R'^2_o, k, and k' are based on regression of the observed activities against predicted activities and regression of the predicted activities against observed activities. The definitions of the aforementioned statistical indices are presented in detail in reference [32,33].

In order for a QSAR model to be used for screening new compounds, its domain of application [30,32] must be defined and predictions for only those compounds that fall into this domain may be considered reliable. The *Extent of Extrapolation* method was adopted for defining the domain of applicability of the produced method, based on the calculation of the leverages for the components in the available dataset [34–36].

Results and discussion

The results support that the molecular descriptors selected by the ES-SWR algorithm combined with the LS-SVM modeling methodology have discrimination and pharmacophore ability and could be used as a filter with great sensitivity and specificity in the range of 1,370 to 25,000 nM. The selected molecular descriptors are the following: Highest Occupied Molecular Orbital energy (HOMO), Principal Moment of Inertia along *X* and *Z* axes (PMIX and PMIZ), Polar Surface Area (PSA), Presence of triple bond (PTrplBnd), and Kier shape descriptor (${}^1\kappa$). Furthermore, the proposed MLR model (Eq. 9) has the ability to predict accurately the I-IP10 IC_{50} inhibition in the range of 0.8 to 790 nM using as inputs the same molecular descriptors used with the LS-SVM.

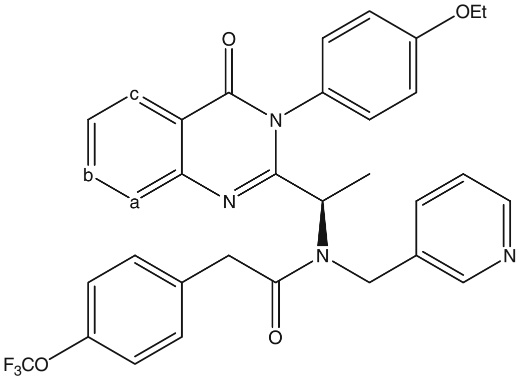
Table 4 Dataset (stereocenter) and model predictions using LS-SVM



Id	R_1	R_3	I-IP10 IC_{50} (nM) (experimental)	I-IP10 log(1/IC_{50}) (experimental)	Class threshold (IC_{50} 790 nM)	Predicted class
35[a]	H	H	2,300	−3.36	Non-active	Non-active
36	H	Me	75	−1.88	Active	
37	Me	Me	10,000	−4.00	Non-active	
38	H	Et	9	−0.95	Active	
39[a]	H	Ph	4,000	−3.60	Non-active	Non-active

[a] Test set

Table 5 Dataset (quinazolinone) and model predictions using LS-SVM

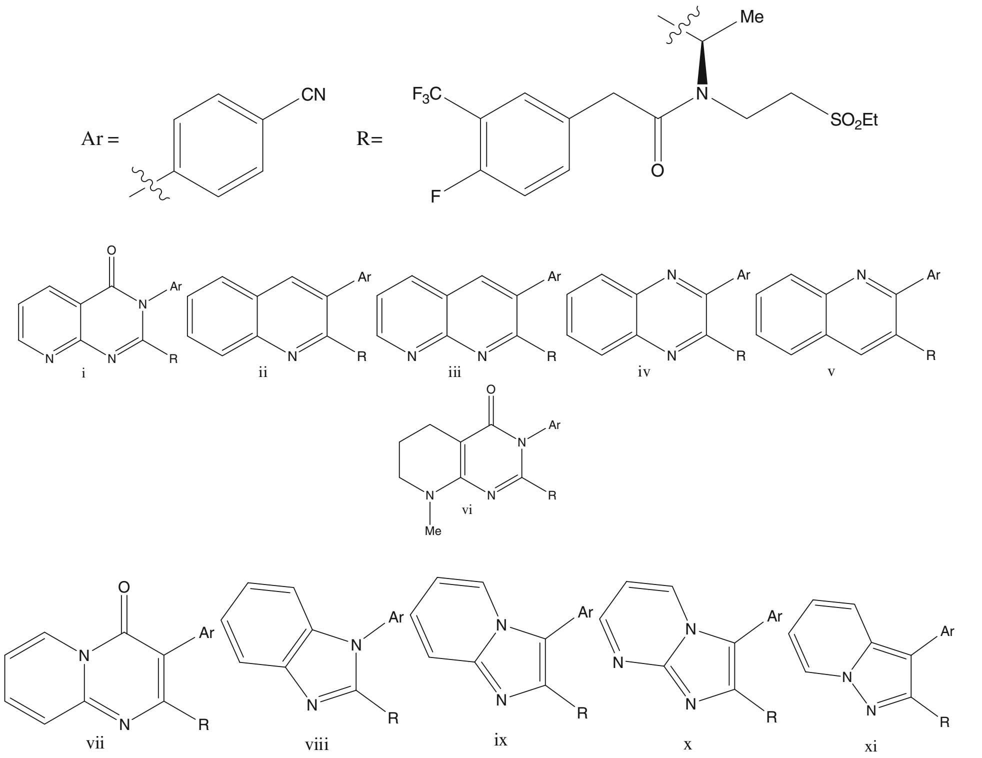


Id	*a*	*b*	*c*	I-IP10 IC_{50} (nM) (experimental)	I-IP10 log(1/IC_{50}) (experimental)	Class threshold (IC_{50} 790 nM)	Predicted class
40[a]	C	C	C	6	−0.78	Active	Active
41	N	C	C	8	−0.90	Active	
42	C	N	C	144	−2.16	Active	
43	C	C	N	1, 400	−3.15	Non-active	
44	N	C	N	480	−2.68	Active	

[a] Test set

The accuracy when LS-SVM classification technique was applied to the training set (Tables 1–6) was 100%. The LS-SVM model also demonstrated good performance in the separate test set of 15 analogs. It accurately identified the 90% of the small molecules of medium or high inhibitory activities (“actives,” IC_{50} range: 0.8–790 nM) and misclassified only one of the analogs of low or no inhibitory activity (“non-actives,” IC_{50} range: 1,370–25,000 nM). The model produced by the LS-SVM method was also validated by applying the Y-randomization test. In all random shuffles of the Y vector

Table 6 Dataset (six–six fused heterocyclic ring systems) and model predictions using LS-SVM



Id	*R*	I-IP10 IC_{50} (nM) (experimental)	I-IP10 $\log(1/IC_{50})$ (experimental)	Class threshold (IC_{50} 790 nM)	Predicted class
45	i	11	−1.04	Active	
46[a]	ii	0.8	0.10	Active	Active
47[a]	iii	2	−0.30	Active	Active
48	iv	9	−0.95	Active	
49[a]	v	5	−0.70	Active	Active
50	vi	4	−0.60	Active	
51	vii	2	−0.30	Active	
52	vii	6	−0.78	Active	
53	ix	3	−0.48	Active	
54[a]	x	10	−1.00	Active	Active
55	xi	18	−1.26	Active	

[a] Test set

we tried, the performance of the produced model was significantly reduced (20–40%). This illustrates that the accuracy of the LS-SVM model is not due to a chance correlation. LS-SVM model also passed successfully the tests for performance evaluation [28,29]:

$$\text{Recall} = \frac{\text{TA}}{\text{TA} + \text{FN}} = 0.9$$

$$\text{Precision} = \frac{\text{TN}}{\text{FA} + \text{TN}} = 0.9$$

$$\text{Specificity} = \frac{\text{TA}}{\text{TA} + \text{FA}} = 0.8$$

$$\text{Accuracy} = \frac{\text{TA} + \text{TN}}{\text{TA} + \text{FA} + \text{FN} + \text{TN}} = 0.87$$

$$F\text{-measure} = \frac{2(\text{Recall})(\text{Precision})}{\text{Recall} + \text{Precision}} = 0.90$$

In the next stage an MLR QSAR model was developed by applying the selected molecular descriptors to the "active compounds" of the training data (32 compounds). The R^2 statistic for the training set is equal to 0.80 as shown below, while for the validation set the R^2_{pred} statistic is 0.78. The MLR QSAR equation was the following:

$$\begin{aligned}\log(1/\text{IC50}) = {} & -18.3\ (\pm 3.51) - 0.40\ (\pm 0.22)\ \text{HOMO} \\ & - 0.384 \times 10^{-3}\ (\pm 0.104 \times 10^{-4})\ \text{PMIX} \\ & - 0.217 \times 10^{-3}\ (\pm 0.591 \times 10^{-4})\ \text{PMIZ} \\ & - 0.021\ (\pm 0.007)\ \text{PSA} \\ & + 1.19\ (\pm 0.213)\ \text{PTrplBnd} \\ & + 0.550\ (\pm 0.09)\,^{1}\kappa \qquad (9)\end{aligned}$$

Table 7 Confusion matrix

Actual predict	Active	Non-active
Active	TA = 9	FA = 1
Non-active	FN = 1	TN = 4

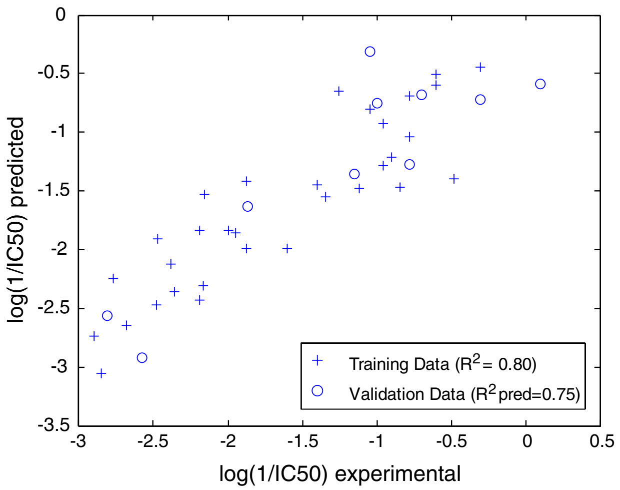


Fig. 1 Experimental versus predicted values log(1/IC50) for the training and validation data (Eq. 9)

$R^2 = 0.80$, RMS $= 0.39$, $F = 16.5$, $R^2_{\mathrm{LOO}} = 0.67$, PRESS $=$ 6.37, $n = 32$ (training set), $R^2_{\mathrm{pred}} = 0.78$, RMS$_{\mathrm{pred}}$ =0.42, $n = 10$ (test set).

Table 8, which presents the correlation matrix and the VIF test, clearly supports that the six selected descriptors are not highly correlated. Furthermore, the ratio of the objects in the training set to the number of descriptors is (5:1), which is typical of many QSAR studies [37,38].

The model was stable to the inclusion-exclusion of compounds measured by the Leave-one-out (LOO) and Leave Five Out (L5O) cross-validation procedures. This was supported by the following statistics: $R^2_{\mathrm{LOO}} = 0.67$ and $R^2_{\mathrm{L5O}} =$ 0.65. Calculation of the R^2_{LOO} statistic was performed using all 32 models that are produced by excluding one compound each time from the training examples, while calculation of the R^2_{L5O} statistic was based on 500 random exclusions of 5-member groups of examples.

The model (Eq. 9) also passed successfully Tropsha's [32] recommended tests for predictive ability:

$$R^2_{\mathrm{ext}} = 0.82 > 0.5$$
$$R^2_{\mathrm{pred}} = 0.78 > 0.6$$
$$\frac{(R^2_{\mathrm{pred}} - R^2_o)}{R^2_{\mathrm{pred}}} = -0.25 < 0.1$$
$$k = 0.95 (0.85 \le k \le 1.15)$$

The MLR QSAR model was additionally validated by applying the Y-randomization test [37–39] . In particular, 10 random shuffles of the Y vector gave R^2 and R^2_{LOO} values in the ranges of 0.15 to 0.33 and 0.03 to 0.28, respectively. These low R^2 and R^2_{LOO} values showed that the results from our original model were not due to a chance correlation or structural dependency of the training set.

It needs to be emphasized however that no matter how robust, significant, and validated a QSAR model may be, it cannot be expected to predict reliably the modeled activity for the entire universe of chemicals [30]. The domain of applicability of the model was defined using the extent of extrapolation method as discussed above. According to this method, we considered as reliable only the predictions of the compounds whose leverages lie within the domain of applicability. In Table 9 all leverages for active test set are presented. The warning leverage limit is 0.65, and it can be concluded from the leverage values in Table 9 that the predictions of the QSAR model for test set small molecules are considered reliable. An additional validation test has been carried out in order to further assess the predictability and the applicability of the model. The available data were divided randomly for five times into a ratio 80:20 for training and test set, respectively. The results are presented in Table 10.

The chemical meaning of the six descriptors used in the produced LS-SVM and MLR QSAR workflow are briefly described next:

Polar Surface Area is defined as the part of the surface area of the module associated with nitrogens, oxygens, sulfurs, and the hydrogens bonded to any of these atoms [40]. Polar Surface Area is a descriptor that correlates well the passive molecular transport through membranes and allows the prediction of transport properties of drugs. Molecules with a PSA of greater than 140 Å^2 are usually believed to be poor at permeating cell membranes [40].

Molecular orbital (MO) surfaces visually represent the various stable electron distributions of a molecule. According to Frontier Orbital Theory, the shapes and symmetries of the highest-occupied and lowest-unoccupied molecular orbitals (HOMO and LUMO) are crucial in predicting the reactivity of a species and the stereo- and regiochemical outcome of a chemical reaction [14].

Presence of Triple Bonds (PTrplBnd) is a dummy variable (indicator variable) which is used to distinguish different treatment groups. The presence of a triple bond in a molecule is responsible for many peculiar chemical and physiochemical properties [41].

The principal moments of inertia (PMI) (g/mol Å^2) is a physical quantity which is related to the rotational dynamics of a module [14]. The PMIs are defined by the diagonal elements of the inertia tensor matrix when the Cartesian coordinate axes are the principal axes of the module, with the origin located at the center of mass of the module [43]. In this case

Table 8 Correlation matrix for the five selected descriptors

	HOMO	PMIX	PMIZ	PSA	PTrplBnd	$^1\kappa$	VIF[a]
HOMO	1						1.7
PMIX	−0.17	1					2.2
PMIZ	−0.02	0.33	1				1.6
PSA	−0.46	0.66	−0.32	1			5.5
PTrplBnd	−0.42	0.51	0.35	0.56	1		1.8
$^1\kappa$	−0.18	0.71	0.53	0.84	0.48	1	5.8

[a] VIF less than 10 indicates that the model contains no multicollinearity

Table 9 Model predictions using MLR QSAR equation 9 (only 'active' molecules)

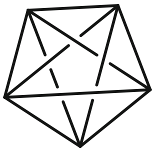
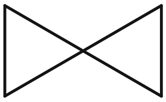
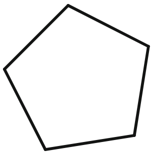
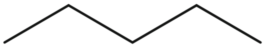
Id	I-IP10 IC_{50} (nM) (experimental)	I-IP10 $\log(1/IC_{50})$ (experimental)	Training data $\log(1/IC_{50})$ (predicted) ($R^2 = 0.80$, $R^2_{LOO} = 0.67$)	Validation data $\log(1/IC_50)$ (predicted) ($R^2_{pred} = 0.78$)	Leverages (limit = 0.65)
1	146	−2.16	−2.31		
2	154	−2.19	−2.43		
3[a]	375	−2.57		−2.92	0.37
5	710	−2.85	−3.05		
6	790	−2.90	−2.74		
7	587	−2.77	−2.25		
9	75	−1.88	−1.99		
13	88	−1.95	−1.86		
14	156	−2.19	−1.84		
15	300	−2.48	−2.47		
16	40	−1.60	−1.99		
18	100	−2.00	−1.84		
19	230	−2.36	−2.36		
20[a]	650	−2.81		−2.56	0.20
21	240	−2.38	−2.12		
22[a]	73	−1.86		−1.63	0.08
23	13	−1.11	−1.47		
24	299	−2.48	−1.91		
25	22	−1.34	−1.55		
26	25	−1.40	−1.45		
27[a]	14	−1.15		−1.36	0.10
28	6	−0.78	−1.04		
29	4	−0.60	−0.50		
30	7	−0.85	−1.47		
31[a]	11	−1.04		−0.31	0.30
36	75	−1.88	−1.42		
38	9	−0.95	−1.28		
40[a]	6	−0.78		−1.27	0.21
41	8	−0.90	−1.21		
42	144	−2.16	−1.53		
44	480	−2.68	−2.64		
45	11	−1.04	−0.81		
46[a]	0.8	0.10		−0.59	0.21
47[a]	2	−0.30		−0.72	0.22
48	9	−0.95	−0.92		
49[a]	5	−0.70		−0.68	0.16
50	4	−0.60	−0.60		
51	2	−0.30	−0.45		
52	6	−0.78	−0.69		
53	3	−0.48	−1.39		
54[a]	10	−1.00		−0.75	0.49
55	18	−1.26	−0.65		

[a] Test set

Table 10 LS-SVM and MLR results from random split (80% for training and 20% for the test set)

Random iteration	LS-SVM					MLR				
	Sensitivity	Specificity	Total accuracy	Recall	Precision	F-measure	R^2	R^2_{LOO}	R^2_{pred}	Domain of applicability
1	0.86	0.75	0.82	0.86	0.86	0.86	0.81	0.66	0.78	All within
2	0.88	1.00	0.91	0.88	1.00	0.93	0.78	0.68	0.75	All within
3	1.00	1.00	1.00	1.00	1.00	1.00	0.77	0.68	0.74	All within
4	1.00	0.80	0.91	1.00	0.86	0.92	0.82	0.65	0.80	All within
5	0.88	1.00	0.91	0.88	1.00	0.93	0.70	0.67	0.79	All within

Table 11 $^1\kappa$ values and corresponding shapes

$A=5$				
$^1\kappa$	0.800	2.222	3.200	5.000

the off-diagonal elements of the inertia tensor matrix are zero and the three diagonal elements I_{xx}, I_{yy}, and I_{zz} correspond to the moments of inertia about the X-, Y-, and Z-axes of the module.

Kier and Hall shape descriptors [42] encode information about several attributes of molecular shape. The specific indices are based on the atom count and the path count of various orders. More specifically $^1\kappa$ (first order shape attribute) quantifies molecular complexity based on cyclicity [40].

According to the produced LS-SVM and MLR QSAR workflow, high values of Kier and Hall shape descriptor ($^1\kappa$) contribute positively to the activity. Thus, an improvement on the activity can be expected by designing small molecules with high $^1\kappa$ values. Kier and Hall shape descriptors are convenient numerical delineators of potential value in molecular series where conformational states are limited or similar. They are also of potential value in searching databases for molecules with a prescribed shape. According to Kier and Hall [42,44], when combined with indices reflecting electronic (in our study HOMO energy) and topological structure, they may be of great value in exploiting the information from combinatorial libraries and data from high throughput screening. In this study $^1\kappa$ descriptor contributes greatly to the discrimination power of the LS-SVM model. $^1\kappa$ is the first order shape attribute which is described by the $^1P_{\max}$, $^1P_{\min}$, P, and A. P is the number of paths in the H-depleted molecular graph, $^1P_{\min}$ is the linear graph, $^1P_{\max}$ is the complete graph in which all atoms are bonded to each other, and A is the number of atoms [44].

$$^1\kappa = \frac{^1P_{\max}\,^1P_{\min}}{(^1P)^2} \quad ^1P_{\min} = A - 1 \quad ^1P_{\max} = \frac{A(A-1)}{2}$$

In Table 11 a range of $^1\kappa$ values is shown for structures (real and hypothetical) ranging from a linear structure to one with a maximum number of cycles [44]. The structural information encoded in $^1\kappa$ is related to the complexity and the cyclicity of a molecule which is not favorable (cyclicity) for the design of novel and potent quinazolinone antagonists of CXCR3. The remarks from the computational study agree also with the experimental results (the compared compounds should have the same number of atoms), for example, small molecules with id. **8** (IC_{50} = 154 nM) versus id. **11** (IC_{50} = 10,000 nM) or id. **10** (IC_{50} = 710 nM) versus id. **12** (IC_{50} = 10,000 nM).

While interpreting the physical meaning of the descriptors, we have noticed that only "actives" compounds contain triple bonds. More specifically 13 out of 42 active compounds contain a triple bond at the 4-phenyl substitution. The majority of the analogs contain cyano group. As is shown from the experimental result, the introduction of a triple bond at the 4-phenyl substitution with a cyano group will improve the activity. Nitrile groups are susceptible to nucleophilic attack at carbon, while electrophilic agents attack the nitrile nitrogen [45].

Polar surface area is related to the hydrogen-bonding ability of the compounds, and the study has showed that values between 65 and 110 Å contribute positively to the activity. The presence of nitrogens, oxygens, sulfurs, and the hydrogens bonded to any of these atoms increases Polar Surface Area value with a specific weight which depends on the atomic contributions of each group [46]. Once again the in silico study confirms the experimental results since it is clearly stated in Johnson et al. [4,5] that substitution from other polar groups may cause significant loss of the activity and therefore there is not a clear direction. Computational analysis was used to interpret this phenomenon using quantitatively polar groups through PSA.

On the other hand, molecules with high HOMO (highest occupied molecular orbital energy) values are more able to donate electron density more easily than molecules with low HOMO energy values [14]. The HOMO energy value can be increased with the presence of electron-donating groups (EDG) such us NMe_2, NH_2, NHEt, and OMe and decreased with the presence of electron-withdrawing groups (EWG) such as halogens and cyano and nitro groups. From the derived LS-SVM and MLR QSAR workflow we can conclude that EWGs favor the biological action under study. It is important to emphasize that the most potent analogs which are those with id. 45–55, have strong electron-withdrawing groups in both the sides of the molecule (cyano and fluorine).

Large values of principal moments of inertia (along *X* and *Z* axes PMIX and PMIZ) correspond to lower inhibition activity. PMIX and PMIZ give information about how the product of mass and distance influence the investigated activity along the *X* and *Z* [35] axes of the quinazolines analogs.

Conclusions

The proposed method, due to the high predictive ability [30] and simplicity, could be a useful aid to the costly and time-consuming experiments for determining the CXCR3 functional antagonism effect of quinazolinone analogs. The two-stage approach that is proposed in this work increases the accuracy of the produced QSAR model, since it covers a narrower chemical space, compared to a model that uses all the available data [47,48]. A virtual screening procedure [49] could be based on the proposed QSAR model. The design of novel active molecules by the insertion, deletion, or modification of substituents on different sites of the molecule and at different positions could therefore be guided by the proposed model. The method [50,51] can also be used to screen existing databases or virtual combinations to identify derivatives with desired activity. In this scenario, the classification model will be used to screen out inactive compounds, while the applicability domain will serve as a valuable tool to filter out “dissimilar” combinations. The molecular descriptors used in QSAR workflow encode information about the structure, branching, electronic effects, and polarity of the modules and thus implicitly account for cooperative effects between functional groups. The proposed QSAR workflow aims to help researchers to design novel chemistry driven molecules with desired biological activity.

Acknowledgments This work was supported by funding under the Sixth Research Framework Programme of the European Union, Project MUGEN (MUGEN LSHG-CT-2005-005203). A.A would like to acknowledge support from OpenEye Scientific Software, Inc. software license provided at no cost.

References

1. Neote K (2007) Chemokine biology: basic research and clinical application: vol 2: pathophysiology of chemokines (Progress in Inflammation Research). Birkhäuser, Basel
2. Wijtmans M, Verzijl D, Leurs R, de Esch IJ, Smit MJ (2008) Towards small-molecule CXCR3 ligands with clinical potential. ChemMedChem 3:861–872. doi:10.1002/cmdc.200700365
3. Cole AG, Stroke IL, Brescia MR, Simhadri S, Zhang JJ, Hussain Z et al (2006) Identification and initial evaluation of 4-*N*-aryl-[1,4]diazepane ureas as potent CXCR3 antagonists. Bioorg Med Chem Lett 16:200–203. doi:10.1016/j.bmcl.2005.09.020
4. Johnson M, Li AR, Liu J, Fu Z, Zhu L, Miao S et al (2007) Discovery and optimization of a series of quinazolinone-derived antagonists of CXCR3. Bioorg Med Chem Lett 17:3339–3343. doi:10.1016/j.bmcl.2007.03.106
5. Du X, Chen X, Mihalic JT, Deignan J, Duquette J, Li AR et al (2008) Design and optimization of imidazole derivatives as potent CXCR3 antagonists. Bioorg Med Chem Lett 18:608–613. doi:10.1016/j.bmcl.2007.11.072
6. Roy K, Mandal AS (2009) Predictive QSAR modeling of CCR5 antagonist piperidine derivatives using chemometric tools. J Enzyme Inhib Med Chem 24:205–223. doi:10.1080/14756360802051297
7. Aher YD, Agrawal A, Bharatam PV, Garg P (2007) 3D-QSAR studies of substituted 1-(3,3-diphenylpropyl)-piperidinyl amides and ureas as CCR5 receptor antagonists. J Mol Model 13:519–529. doi:10.1007/s00894-007-0173-z
8. Afantitis A, Melagraki G, Sarimveis H, Koutentis PA, Markopoulos J, Igglessi-Markopoulou O (2006) Investigation of substituent effect of 1-(3,3-diphenylpropyl)-piperidinyl phenylacetamides on CCR5 binding affinity using QSAR and virtual screening techniques. J Comput Aided Mol Des 20:83–95. doi:10.1007/s10822-006-9038-2
9. Nair PC, Srikanth K, Sobhia ME (2008) QSAR studies on CCR2 antagonists with chiral sensitive hologram descriptors. Bioorg Med Chem Lett 18:1323–1330. doi:10.1016/j.bmcl.2008.01.023
10. Srikanth K, Nair PC, Sobhia ME (2008) Probing the structural and topological requirements for CCR2 antagonism: holographic QSAR for indolopiperidine derivatives. Bioorg Med Chem Lett 18:1450–1456. doi:10.1016/j.bmcl.2007.12.072
11. Khlebnikov AI, Schepetkin IA, Quinn MT (2006) Quantitative structure-activity relationships for small non-peptide antagonists of CXCR2: indirect 3D approach using the frontal polygon method. Bioorg Med Chem 14:352–365. doi:10.1016/j.bmc.2005.08.026
12. Bhonsle JB, Wang Z, Tamamura H, Fujii N, Peiper SC, Trent JO (2005) A simple, automated quasi-4D-QSAR, quasi-multi way PLS approach to develop highly predictive QSAR models for highly flexible CXCR4 inhibitor cyclic pentapeptide ligands using scripted common molecular modeling tools. QSAR Comb Sci 24:620–630. doi:10.1002/qsar.200430912
13. Afantitis A, Melagraki G, Sarimveis H, Igglessi-Markopoulou O, Kollias G (2009) A novel QSAR model for predicting the inhibition of CXCR3 receptor by 4-*N*-aryl-[1,4] diazepane ureas. Eur J Med Chem 44:877–884. doi:10.1016/j.ejmech.2008.05.028
14. Todeschini R, Consonni V, Mannhold R (2000) In: Kubinyi H, Timmerman H (eds) Handbook of molecular descriptors. Wiley-VCH, Weinheim

15. Suykens JAK, Van Gestel T, De Brabanter J, De Moor B, Vandewalle J (2002) Least squares support vector machines. World Scientific Pub Co., Singapore
16. Stewart JJP (2007) Optimization of parameters for semiempirical methods V: modification of NDDO approximations and application to 70 elements. J Mol Model 13:1173–1213. doi:10.1007/s00894-007-0233-4
17. Stewart JJP (2008) Application of the PM6 method to modeling the solid state. J Mol Model 14:499–535. doi:10.1007/s00894-008-0299-7
18. Puzyn T, Suzuki N, Haranczyk M, Rak J (2008) Calculation of quantum-mechanical descriptors for QSPR at the dft level: is it necessary? J Chem Inf Model 48:1174–1180. doi:10.1021/ci800021p
19. Chem 3D. CambridgeSoft Corporation, 100 CambridgePark Drive Cambridge, MA 02140, USA. http://www.cambridgesoft.com
20. Topix. Epina GmbH, Am Wienerwald 15, 3013 Pressbaum, Austria. http://www.lohninger.com/topix.html
21. MOPAC2007. Stewart Computational Chemisitry (SCC), 15210 Paddington Circle Colorado Springs, CO80921, USA, http://openmopac.net/home.html
22. ROCS & EON. OpenEye Scientific Software Inc, 9 Bisbee Court, Suite D Santa Fe, NM 87508, USA. http://www.eyesopen.com
23. Kennard RW, Stone LA (1969) Computer aided design of experiments. Technometrics 11:137–148. doi:10.2307/1266770
24. Ghosh P, Thanadath M, Bagchi MC (2006) On an aspect of calculated molecular descriptors in QSAR studies of quinolone antibacterials. Mol Divers 10:415–427. doi:10.1007/s11030-006-9018-4
25. Melagraki G, Afantitis A, Sarimveis H, Koutentis PA, Markopoulos J, Igglessi-Markopoulou O (2007) Optimization of biaryl piperidine and 4-amino-2-biarylurea MCH1 receptor antagonists using QSAR modeling, classification techniques and virtual screening. J Comput Aided Mol Des 21:251–267. doi:10.1007/s10822-007-9112-4
26. Afantitis A, Melagraki G, Sarimveis H, Koutentis PA, Markopoulos J, Igglessi-Markopoulou O (2006) A novel simple QSAR model for the prediction of anti-HIV activity using multiple linear regression analysis. Mol Divers 10:405–414. doi:10.1007/s11030-005-9012-2
27. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182. doi:10.1162/153244303322753616
28. Hung YH, Liao YS (2008) Applying PCA and fixed size LS-SVM method for large scale classification problems. Inf Technol J 7:890–896. doi:10.3923/itj.2008.890.896
29. Fawcett T (2006) An introduction to ROC analysis. Pattern Recognit Lett 27:861–874. doi:10.1016/j.patrec.2005.10.010
30. Tropsha A, Golbraikh A (2007) Predictive QSAR modeling workflow, model applicability domains, and virtual screening. Curr Pharm Des 13:3494–3504. doi:10.2174/138161207782794257
31. Melagraki G, Afantitis A, Sarimveis H, Koutentis PA, Markopoulos J, Igglessi-Markopoulou O (2007) A novel QSPR model for predicting θ (lower critical solution temperature) in polymer solutions using molecular descriptors. J Mol Model 13:55–64. doi:10.1007/s00894-006-0125-z
32. Golbraikh A, Tropsha A (2002) Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. Mol Divers 5:231–243. doi:10.1023/A:1021372108686
33. Melagraki G, Afantitis A, Sarimveis H, Igglessi-Markopoulou O, Alexandridis A (2006) A novel RBF neural network training methodology to predict toxicity to Vibrio fischeri. Mol Divers 10:213–221. doi:10.1007/s11030-005-9008-y
34. Melagraki G, Afantitis A, Sarimveis H, Koutentis PA, Markopoulos J, Igglessi-Markopoulou O (2007) Identification of a series of novel derivatives as potent HCV inhibitors by a ligand-based virtual screening optimized procedure. Bioorg Med Chem 15:7237–7247. doi:10.1016/j.bmc.2007.08.036
35. Afantitis A, Melagraki G, Sarimveis H, Koutentis PA, Markopoulos J, Igglessi-Markopoulou O (2008) Development and evaluation of a QSPR model for the prediction of diamagnetic susceptibility. QSAR Comb Sci 27:432–436. doi:10.1002/qsar.200730083
36. Toropov AA, Benfenati E (2008) Additive SMILES-based optimal descriptors in QSAR modelling bee toxicity: using rare SMILES attributes to define the applicability domain. Bioorg Med Chem 16:4801–4809. doi:10.1016/j.bmc.2008.03.048
37. Todeschini R, Consonni V, Mauri A, Pavan M (2004) Detecting "bad" regression models: multicriteria fitness functions in regression analysis. Anal Chim Acta 515:199–208. doi:10.1016/j.aca.2003.12.010
38. Afantitis A, Melagraki G, Sarimveis H, Koutentis PA, Markopoulos J, Igglessi-Markopoulou O (2006) A novel QSAR model for evaluating and predicting the inhibition activity of dipeptidyl aspartyl fluoromethylketones. QSAR Comb Sci 25:928–935. doi:10.1002/qsar.200530208
39. Jalali-Heravi M, Asadollahi-Baboli M, Shahbazikhah P (2008) QSAR study of heparanase inhibitors activity using artificial neural networks and Levenberg–Marquardt algorithm. Eur J Med Chem 43:548–556. doi:10.1016/j.ejmech.2007.04.014
40. Ertl P, Rohde B, Selzer P (2000) Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. J Med Chem 43:3714–3717. doi:10.1021/jm000942e
41. Patai S (1992) Patai's 1992 guide to the chemistry of functional groups. Wiley, Chichester
42. McQuarrie DA, Simon JD (1997) Physical chemistry: a molecular approach. University Science Books, CA
43. Kier LB (1986) Molecular connectivity in structure–activity analysis (chemometrics series). Wiley, New York
44. Devillers J, Balaban AT (1999) Topological indices and related descriptors in QSAR and QSPAR. Taylor & Francis Inc, New York
45. Colombo A, Benfenati E, Karelson M, Maran U (2008) The proposal of architecture for chemical splitting to optimize QSAR models for aquatic toxicity. Chemosphere 72:772–780. doi:10.1016/j.chemosphere.2008.03.016
46. Baumann K (2003) Cross-validation as the objective function for variable-selection techniques. Trends Analyt Chem 22:395–406. doi:10.1016/S0165-9936(03)00607-1
47. Agrafiotis DK, Bandyopadhyay D, Wegner JK, Vlijmen H (2007) Recent advances in chemoinformatics. J Chem Inf Model 47:1279–1293. doi:10.1021/ci700059g
48. Muegge I, Oloff S (2006) Advances in virtual screening. Drug Discov Today Technol 3:405–411. doi:10.1016/j.ddtec.2006.12.002
49. Melagraki G, Afantitis A, Sarimveis H, Koutentis PA, Kollias G, Igglessi-Markopoulou O (2009) Predictive QSAR workflow for the in silico identification and screening of novel HDAC inhibitors. Mol Divers. doi:10.1007/s11030-009-9115-2
50. Salum LB, Andricopulo AD (2009) Fragment-based QSAR: perspectives in drug design. Mol Divers 2009. doi:10.1007/s11030-009-9112-5
51. Guido RV, Oliva G, Andricopulo AD (2008) Virtual screening and its integration with modern drug design technologies. Curr Med Chem 15:37–46. doi:10.2174/092986708783330683