Chemometrics and Intelligent Laboratory Systems 98 (2009) 65–77

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Critical comparative analysis, validation and interpretation of SVM and PLS regression models in a QSAR study on HIV-1 protease inhibitors

Noslen Hernández [a], Rudolf Kiralj [b], Márcia M.C. Ferreira [b,*], Isneri Talavera [a]

[a] Advanced Technologies Application Center, Havana, 12200, Cuba
[b] Institute of Chemistry, University of Campinas, Campinas, SP 13083-970, Brazil

ARTICLE INFO

Article history:
Received 21 April 2009
Accepted 30 April 2009
Available online 15 May 2009

Keywords:
Peptidic protease inhibitors
Molecular descriptors
Regression models
Validation
Statistics

ABSTRACT

Four Quantitative Structure–Activity Relationship (QSAR) models were constructed for a set of 32 and 16 HIV-1 protease inhibitors in the training and external validation sets, respectively, using the biological activity and molecular descriptors from the literature. Two QSAR models were based on Support Vector Machines methods (SVM): Support Vector Regression (SVR) and Least-Squares Support Vector Machines (LS-SVM) models. The other two models were an ordinary Partial Least Squares (PLS) and Ordered Predictors Selection-based PLS (OPS-PLS). The SVR and LS-SVM models showed to be somewhat better than the PLS model in external validation and leave-*N*-out crossvalidation. SVR and LS-SVM were better than OPS-PLS in external validation, but showed equal performance in leave-*N*-out crossvalidation. However, despite of their high predictive ability, the SVM models failed in **y**-randomization, which did not happen with the PLS and OPS-PLS models. The OPS-PLS model was the only one that undoubtedly showed satisfactory performance both in prediction and all validations. The selection of inhibitors by the SVM-based models and variable selection by the OPS-PLS model were rationalized by means of Hierarchical Cluster Analysis (HCA) and Principal Component Analysis (PCA). Lagrange multipliers from the SVR and LS-SVM models were explained for the first time in terms of molecular structures, descriptors, biological activity and principal components. Some unresolved difficulties in practical usage of SVM in QSAR and QSPR were pointed out. The presented validation and interpretation of SVR and LS-SVM models is a proposal for future investigations about SVM applications in QSAR and QSPR, valid for any modeling and validation condition of the final regression equations.

1. Introduction

Quantitative Structure–Activity Relationships (QSAR) [1] are mathematical equations for calculation of biological activity from molecular descriptors (physico-chemical properties or molecular features). Nowadays, with fast computer programs, large descriptor sets are commonly generated, which frequently introduce significant descriptor intercorrelations. Latent variable methods like Principal Component Regression (PCR) and Partial Least Squares (PLS) regression [1–5] have shown to be adequate for solving this problem. Besides, PLS allows limited modeling of non-linear relations by using more latent variables or descriptor transforms.

Support Vector Machines method (SVM) is a relatively new alternative to the existing linear and non-linear multivariate calibration approaches in chemometrics [6,7]. SVM was originally proposed by Vapnik [8,9] within the area of statistical learning theory and structural risk minimization. SVM is able to treat both, linear and non-linear data sets and control or even reduce overfitting. The non-linearity is achieved by applying the kernel trick, *i.e.*, every dot product in linear SVM is replaced by a non-linear kernel which satisfies the Mercer's theorem [9]. Compared to Artificial Neural Networks (ANN), SVM always finds a global, usually unique solution. SVM-based regressions can solve ill-posed problems leading to models that are often unique and exhibit good prediction power.

Least-Squares Support Vector Machines method (LS-SVM) [10,11] appears as a newer variant of the SVM formulation in which equality instead of inequality constraints are applied and a sum squared error cost function is used. LS-SVM is computationally more efficient and simpler than its predecessor Support Vector Regression method (SVR). Although LS-SVM suffers from the lack of sparsity, which can be achieved only using pruning techniques applied to Lagrange multipliers [11], and is less robust to the presence of outliers and non-Gaussian noise, it outperforms SVR in many cases, probably because its optimization procedure is more accurate (less parameters to optimize).

SVR methods have appeared in QSAR and QSPR (Quantitative Structure–Property Relationship) in 2002 [12–14], and LS-SVM methods only in 2005 [15,16]. SVM-based methods are still modestly used in QSAR and QSPR, accounting to more than two hundred research articles, majority of which (95%) prefers SVR. Nowadays, several numerical methods in QSAR and QSPR compete in proposing

* Corresponding author. Tel.: +55 19 3788 3102; fax: +55 19 3788 3023.
E-mail address: marcia@iqm.unicamp.br (M.M.C. Ferreira).

0169-7439/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2009.04.012

regression models for prediction of biological activities, chemical or physical properties by selecting a few molecular descriptors among hundreds or thousands of measured or calculated variables. This increasing complexity of regression models, possibility for chance [17] and non-causal [18] correlations encoded in final models, and danger of models' non-interpretability [19], impose rather rigorous criteria for model validations [20–25], regardless of the modeling methodology applied. Leave-*N*-out crossvalidation (LNO) [22,26] and **y**-randomization [22,27] are among the most rigorous validation procedures. Up to our knowledge, there are only a few articles including LNO or **y**-randomization of SVM models in the QSAR literature.

All these facts motivated the authors of this work to compare SVR and LS-SVM approaches with PLS when applied to a particular QSAR data set with linear structure, but unlike similar comparisons in the literature, the models are rigorously validated. The performance of SVR and LS-SVM methods is analyzed parallel to that of an ordinary PLS and Ordered Predictors Selection-based PLS (OPS-PLS) [28], and is rationalized by means of Principal Component (PCA) and Hierarchical Cluster (HCA) Analyses [1,3,5]. The data set is from a QSAR study [29,30] on peptidic inhibitors of HIV-1 integrase (Fig. 1), with *in vitro* biological activity expressed as the negative logarithm of molar concentration IC_{50} (inhibitory concentration for 50% viral inhibition). Fourteen *a priori* molecular descriptors [29,31,32] of various natures (compositional, electronic, steric, hydrophobic, topological and mixed), were calculated. In this work, a new HCA-based data split, has been applied. The data set has been shown useful for validation of various chemometric approaches [28,33]. It should be stressed that the purpose of this work is not to propose new, more potent HIV-1 protease inhibitors, but to offer methodologies for rigorous validation and chemical interpretation of SVM-based models as it is the usual practice for conventional regression models in QSAR and QSPR.

2. Methods

2.1. The data set

The QSAR data set [29,30] was in the form of a matrix 48×14, where 48 peptidic inhibitors with four substituents at the main peptide-like chain (Fig. 1) were described by 14 molecular descriptors (named originally as $X_1, X_2, \ldots, X_{14}$), resulting from a variable selection on more than fifty *a priori* descriptors [29]. The anti-HIV activity was in the form $pIC_{50} = -\log IC_{50}$. Original IC_{50} was determined and rationalized *via* inhibitors docking to HIV-1 protease by Holloway et al. [34]. The former data split [29] consisted of compounds **1**–**32** in the training and **33**–**48** in the external validation set, according to Holloway et al. [34]. Detailed inspection of this split by means of HCA for the complete data set and the corresponding QSAR model (published previously [29]) has revealed a more consistent split based on inhibitor clustering. Therefore, inhibitors selected for the new external validation set were: **1**, **3**, **8**, **12**–**14**, **22**, **24**, **26**, **29**, **32**, **35**, **40**, **41**, **44** and **47**. When carefully observing the molecular structures in Fig. 1, the reader can find out that this new external validation set is a better representative of the training set.

2.2. The SVR method

The main idea of SVR [35,36] is to find the "flattest" (*i.e.* less complex) linear function that approximates the given data with ε precision in a kernel-induced feature space [36]. This is reached using the ε-insensitive loss function, which penalizes errors greater that ε. The trade between the flatness of the estimate and the amount up to which deviations greater than ε are tolerated, is determined by the regularization constant $C \geq 0$. This setting is transformed into a constrained optimization problem, in which the Wolfe dual is computed, resulting in a convex programming problem. The solution of this problem is sparse: a subset of the resulting Lagrange multipliers will be nonzero [35–37] and the associated samples will be support vectors (SV). Only these vectors contribute to the regression function. In this setting, the regression vector cannot be given explicitly, only for linear SVR it can be described as a linear combination of the training patterns. Consequently, the information regarding the original input variables is vanished in most cases, and a direct interpretation of the SVR model is more complicated.

2.3. The LS-SVM method

The least squares version of SVM, the LS-SVM method [10,11], requires the solution of a set of linear equations instead of the long and computationally hard quadratic programming problem involved by the standard SVM. LS-SVM solves a constrained optimization problem [11], where the values of Lagrange multipliers for each sample are obtained as solutions of a set of linear equations. In this procedure, sparseness is lost, every data point is a SV, but some points contribute more than others, as follows from optimality conditions [11]. Some pruning techniques have been introduced in order to achieve sparsity [11,38,39]. Besides, the use of a sum of squared errors cost function might lead to estimates which are less robust, but several variants have been developed to overcome this drawback, such as the incorporation of methods from robust statistics [38].

2.4. Software and optimization procedures for the SVR and LS-SVM models

A Gaussian (radial basis function, RBF) kernel was used for both SVR and LS-SVM models. The LS-SVM pruning algorithm was that of Suykens et al. [11]. For pruning, 15% of the training objects with the lowest absolute values of Lagrange multiplier were removed, after which the model was reconstructed and the pruning was repeated until the validation performance degraded up to 80% of that of the original model. Optimization of the SVR (C, ε and kernel width σ) and LS-SVM (C and kernel width σ) hyperparameters was performed by a grid search based on leave-one-out crossvalidation. For the SVR model, the ranges of parameter values used to tune the values of parameters C, ε and σ were 100–1000, 0.05–0.5 and 1–50, respectively. For the LS-SVM model, the range supply for tuning parameters C and σ^2 was exp(2.5)–exp(5) and exp(3.5)–exp(5.5), respectively. The SVR calculations were performed by using the e1071 package in R [40]. For the LS-SVM model, calculations employing the Matlab/C toolbox [41] were carried out.

2.5. The PLS model

The ordinary PLS model (denominated as the PLS model) was constructed with 14 autoscalled molecular descriptors *via* leave-one-out crossvalidation, analogously to the published model [29].

2.6. The OPS-PLS model

All data analyses were performed using home-built functions written for Matlab [42]. The OPS® Toolbox routines, implemented in Matlab, are registered and are available online [43]. The core of OPS [28] is to sort the variables from informative vectors (regression vector, correlation vector, residual vector, variable influence on projection, net analyte signal, covariance procedures vector, signal-to-noise ratios, and their combinations) and investigate the PLS models systematically, to find the most relevant set of interpretable variables by comparing the crossvalidation parameters of the models. The OPS procedure using autoscaled descriptors and leave-one-out crossvalidation resulted in the final model with reduced number of variables, denominated as the OPS-PLS model.

2.7. Validations of the QSAR models

The final regression models (PLS, OPS-PLS, SVR and LS-SVM) were validated by leave-*N*-out crossvalidation and **y**-randomization on the

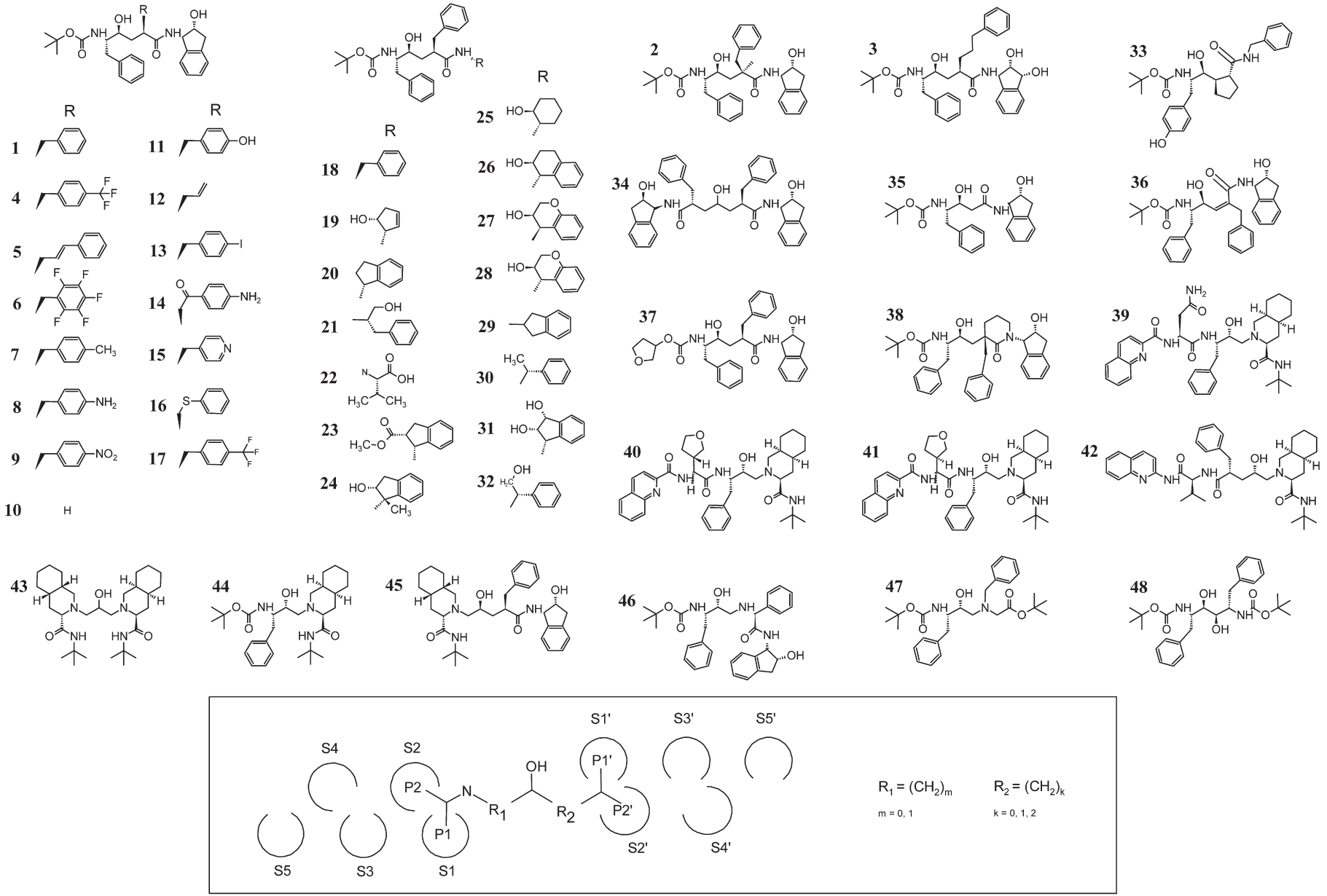

Fig. 1. Molecular structures of HIV-1 inhibitors **1**–**48** and their general structure in the box, consisting of a chain that varies in fragments R_1 and R_2, and four substituents P1, P2, P1′ and P2′ in four protease pockets S1, S2, S1′ and S2′. Six unoccupied pockets S3, S4, S5, S3′, S4′ and S5′ are shown also.

previously randomized training data set, and by external validation, according to recent recommendations for model validations in QSAR and QSPR [20–27]. Ten randomizations of the **y** vector were performed according to Wold and Eriksson [44], as well as additional 1000 randomizations. The same data randomizations and validation procedures were carried out by Matlab and Pirouette [45] for all the four models, with the purpose of consistent comparison.

2.8. Exploratory analysis

PCA and HCA with incremental linkage were applied to the autoscaled training data set using Pirouette. The analyses were performed to rationalize inhibitor selections made by the SVR and LS-SVM and variable selection by OPS-PLS in terms of classes of **y**, molecular structures and molecular features. Exploratory analysis also aided in the interpretation of the regression models.

3. Results and discussion

3.1. Statistical comparison of the QSAR models

Comparative statistics of the final PLS, OPS-PLS, SVR and LS-SVM regression models, is presented in Table 1 and Fig. 2. PLS and OPS-PLS use the same number of latent variables with rather different contents of the original information (OPS uses 5 from 14 molecular descriptors). The SVM-based models use half of the inhibitor set: SVR selects 14 (44%) and LS-SVM 18 (56%) inhibitors. Values of optimized parameters C, σ and ε for SVR and LS-SVM are reported in Table 1.

With the new split, the PLS model becomes somewhat weaker in the training statistics ($R^2=0.84$, $Q^2=0.64$ and SEV = 0.92) than the analogue model for the old split [29], but it improved the external statistics significantly ($SEV_{ext}=0.61$ instead of 1.12). OPS-PLS is superior to PLS in the training statistics, with a slight improvement in prediction, whilst the external predictions are similar for both models. Considering the basic statistical requirements ($R^2>0.6$ and $Q^2>0.5$ [20,21,24]), all the quality indices in Table 1 and predictions in Fig. 2, both PLS and OPS-PLS are good models.

SVR and LS-SVM are very similar in terms of all statistics. This is probably due to equilibrium between the conditions under which the models were obtained: 14 inhibitors and 3 optimized parameters (SVR) against 18 inhibitors and 2 parameters (LS-SVM). SVR and LS-SVM exhibit improvement with respect to OPS-PLS and especially to PLS in terms of training and external statistics, which is visible from Table 1 and Fig. 2. This is a typical situation in the literature, reported when non-linear SVR or LS-SVM is compared to linear methods traditionally used in QSAR and QSPR, such as PLS, PCR and MLR (Multiple Linear Regression).

Robustness of the four regression models was tested by leave-N-out crossvalidation (known also as leave-many-out, LMO), where N took values from 1 to 7 as the region within which all the four models were considered robust (Table 2 and Fig. 3). For satisfactory LNO validation, Q^2_{LNO} is greater than 0.5 and stable, with the maximum N still allowing construction of meaningful models [23], knowing that the maximum N is problem-dependent [26] and cannot be large for small and medium data sets [23]. Therefore, the four models can be considered robust. However, the models are differentiated in terms of average and oscillation degree (standard deviation) of Q^2_{LNO}, well visible in Fig. 3. LNO gives more reliable comparison among the models than leave-one-out crossvalidation: PLS is distinguished from the other models, having the lowest average Q^2_{LNO} and two times greater oscillation degree (Table 2). LS-SVM is at the mid-way between OPS-PLS and SVR. Small differences between average Q^2_{LNO} (below 0.05), oscillation degrees and almost coinciding values of Q^2_{L2O} and Q^2_{L6O} (Fig. 3) make the three models statistically indistinguishable. This is clearer when confidence levels from normal distribution [46] of the differences between average values of Q^2_{LNO} for the models (Table 2) are calculated.

The essence of **y**-randomization is to detect and quantify chance correlations in M runs (randomizations) for K samples in the training sets with scrambled or randomized vectors **y**. The basic statistics of randomization models (Q^2_{yrand} and R^2_{yrand}) should be poor and not in the range of that for acceptable regression models. Otherwise, each resulting model may be considered as a chance correlation. The chance correlation's frequency depends primarily on two statistical factors [17,27]: it strongly increases with the decrease of K and moderately with greater M. Chemical factors such as the nature of compounds and their structural similarity, data quality, distribution profile of each variable, variable intercorrelations, among others, modify at a certain extent these statistical dependences. There are two main issues about **y**-randomization: how large should be M, and how exactly the chance correlation degree should be qualified or quantified that one could characterize a model as having or not having chance correlation. The simple approach of Wold and Eriksson [44] consisting of ten randomizations ($M=10$) for any K, known in QSAR and QSPR literature as a fast and effective procedure [24,47], was also adopted in this work. According to our experience, a model possessing real chance correlations is detected for any number of randomization runs $M\geq 10$, but the question is what criteria are reliable at a certain number M. Applying additional 1000 randomization, this issue may be analyzed and discussed in details.

Table 1
Statistics of PLS, OPS-PLS, SVR and LS-SVM regression models.

Parameter	PLS	OPS-PLS	SVR	LS-SVM
No. of inhibitors	32	32	14[a]	18[b]
No. of descriptors	14	5	14	14
No. of LV (%Var)[c]	3 (69.48%)	3 (82.90%)	–	–
No. of optimized parameters	–	–	3[d]	2[e]
$Q2$[f]	0.641	0.759	0.844	0.819
R^{2}[f]	0.836	0.852	0.935	0.938
Q^2_{ext}[f]	0.841	0.818	0.891	0.899
SEV[g]	0.921	0.754	0.606	0.653
SEC[g]	0.666	0.631	0.417	0.401
SEV_{ext}[g]	0.612	0.654	0.506	0.487
No. of residuals >10% (training set)[h]	8	8	0	2
No. of residuals >10% (exter. valid. set)[h]	3	3	2	2
Average residual (training set)[i]	6.5	6.4	4.7	4.1
Average residual (exter. valid. set)[i]	6.3	7.4	4.8	5.1
PLS regression vector: X_1: 0.037, X_2: −0.120, X_3: −0.103, X_4: −0.008, X_5: 0.036, X_6: −0.103, X_7: −0.017, X_8: −0.007, X_9: 0.503, X_{10}: 0.208, X_{11}: 0.496, X_{12}: 0.097, X_{13}: 0.055, and X_{14}: −0.016				
OPS-PLS regression vector: X_3: −0.155, X_6: −0.548, X_9: 0.976, X_{10}: 0.404, and X_{13}: 0.312				
SVR Lagrange multipliers and b: $\alpha_2=100.00$, $\alpha_{10}=-25.58$, $\alpha_{17}=27.55$, $\alpha_{18}=32.43$, $\alpha_{19}=40.56$, $\alpha_{20}=-37.73$, $\alpha_{21}=-100.00$, $\alpha_{25}=-6.39$, $\alpha_{34}=-2.18$, $\alpha_{37}=13.04$, $\alpha_{38}=-53.65$, $\alpha_{42}=-3.28$, $\alpha_{43}=1.95$, $\alpha_{45}=13.27$, and $b=1.250$				
LS-SVM Lagrange multipliers and b: $\alpha_2=18.97$, $\alpha_5=7.83$, $\alpha_{10}=-9.02$, $\alpha_{11}=-6.31$, $\alpha_{16}=5.29$, $\alpha_{17}=9.42$, $\alpha_{18}=10.07$, $\alpha_{19}=17.54$, $\alpha_{20}=-15.85$, $\alpha_{21}=-21.44$, $\alpha_{23}=-9.22$, $\alpha_{25}=-6.77$, $\alpha_{27}=8.39$, $\alpha_{28}=-12.01$, $\alpha_{37}=6.44$, $\alpha_{38}=-16.66$, $\alpha_{45}=5.24$, $\alpha_{46}=8.08$, and $b=-0.773$				

[a] Inhibitors used to build the SVR model: **2**, **10**, **17**, **18**, **19**, **20**, **21**, **25**, **34**, **37**, **38**, **42**, **43** and **45**.

[b] Inhibitors used to build the LS-SVM model: **2**, **5**, **10**, **11**, **16**, **17**, **18**, **19**, **20**, **21**, **23**, **25**, **27**, **28**, **37**, **38**, **45** and **46**.

[c] Number of latent variables and the corresponding percentage of the total variance (in brackets).

[d] Optimized parameters of the SVR model: $C=100$, $\sigma=15$ and $\varepsilon=0.3$.

[e] Optimized parameters of the LS-SVM model: $C=44.902$ and $\sigma=9.78$.

[f] Correlation coefficients calculated: Q^2 — correlation coefficient of crossvalidation, R^2 — correlation coefficient of multiple determination (calibration) and Q^2_{ext} — correlation coefficient of external validation.

[g] Standard errors calculated: SEV — standard error of crossvalidation, SEC — standard error of calibration and SEV_{ext} — standard error of external validation.

[h] Number of inhibitors with absolute values of relative residuals greater than 10%, for the training and external validation sets.

[i] Average of absolute values of relative residuals, for the training and external validation sets, expressed in %.

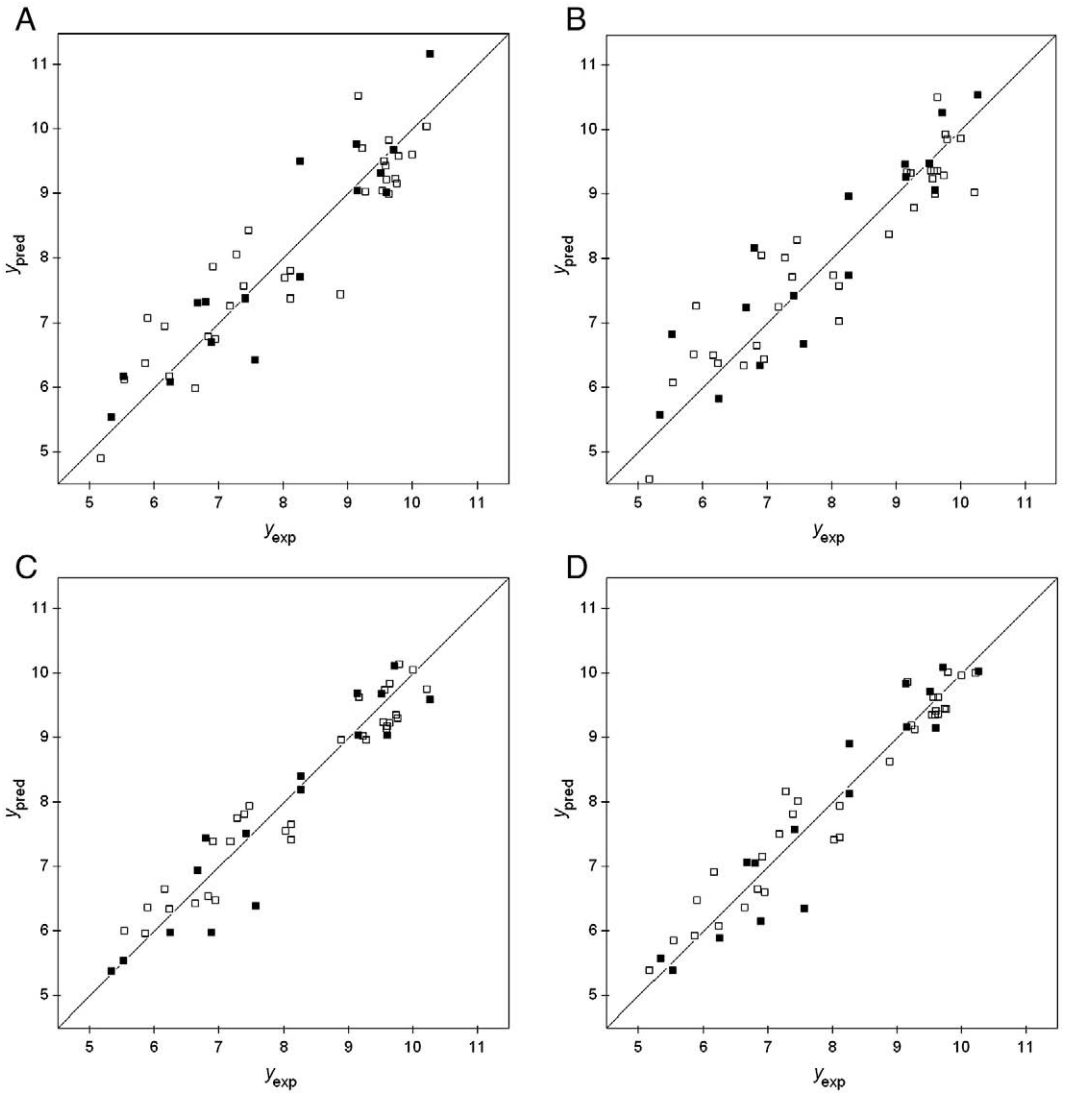

Fig. 2. Experimental against predicted biological activity for A) PLS, B) OPS-PLS, C) SVR and D) LS-SVM regression model. Training and external validation sets are differentiated by white (□) and solid (■) squares, respectively.

Results from 10 and 1000 **y**-randomization validations for the four models are presented in Table 3 and Fig. 4, and also in Supplementary data (Figs. S1, S2 and S3, and Table S1). Four approaches to qualify and quantify chance correlations were applied. To understand their usefulness, the results for 10 randomizations are analyzed and discussed first. The qualitative approach of Wold and Eriksson [44] is based on detecting the randomization runs with Q^2_{yrand} or R^2_{yrand} above

Table 2
Comparative statistics of leave-*N*-out crossvalidation for PLS, OPS-PLS, SVR and LS-SVM regression models.

Parameter[a]	PLS	OPS-PLS	SVR	LS-SVM
Average Q^2_{LNO}	0.617	0.769	0.845	0.803
Standard deviation (Q^2_{LNO})	0.038	0.018	0.019	0.020
Confidence levels matrix[b]				
PLS		0.0003	<0.0001	<0.0001
OPS-PLS			**0.0037**	**0.2077**
SVR				**0.1285**

[a] Statistical parameters are calculated from Q^2 from leave-*N*-out crossvalidation (Q^2_{LNO}).
[b] Confidence levels for normal distribution of the differences between average Q^2_{LNO} values obtained for different models, taking into account respective standard deviations. Bold values indicate that the models are not distinguishable.

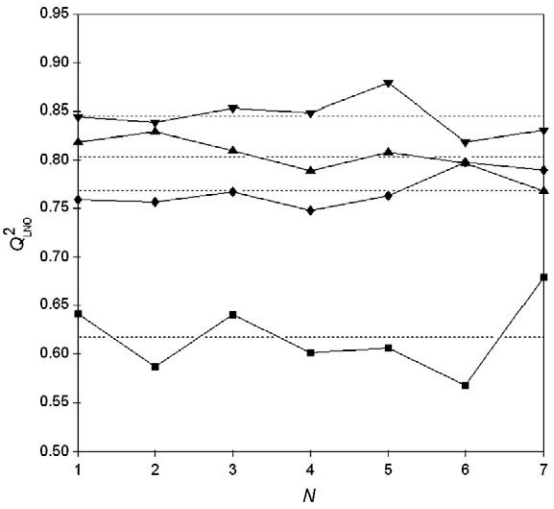

Fig. 3. Leave-*N*-out crossvalidation ($N = 1, 2, \ldots, 7$) for PLS (■), OPS-PLS (♦), SVR (▼), and LS-SVM (▲) models. Average Q^2_{LNO} for the models are represented by dashed and oscillations by solid lines.

Table 3
Comparative statistics of **y**-randomization for PLS, OPS-PLS, SVR and LS-SVM regression models.

Parameter[a]	PLS		OPS-PLS		SVR		LS-SVM	
	$M=10$	$M=1000$	$M=10$	$M=1000$	$M=10$	$M=1000$	$M=10$	$M=1000$
Maximum (Q^2_{yrand})	−0.388	0.367	−0.185	0.320	−0.410	0.280	−0.272	0.330
Maximum (R^2_{yrand})	0.310	**0.614**	0.189	**0.494**	**0.639**	**0.826**	**0.685**	**0.863**
Average (Q^2_{yrand})	−0.594	−0.455	−0.383	−0.306	−1.020	−1.002	−0.784	−0.643
Average (R^2_{yrand})	0.227	0.274	0.108	0.152	**0.435**	**0.471**	**0.526**	**0.572**
Standard deviation (Q^2_{yrand})	0.177	0.311	0.147	0.206	0.363	0.616	0.263	0.357
Standard deviation (R^2_{yrand})	0.053	0.092	0.047	0.089	0.118	0.105	0.095	0.083
Minimum model-random. diff. (Q^2_{yrand})[b]	5.81	**0.88**	6.42	2.13	3.45	0.92	4.15	1.37
Minimum model-random. diff. (R^2_{yrand})[b]	9.92	2.41	14.11	4.04	2.51	1.04	2.66	0.90
Confidence level for min. diff. (Q^2_{yrand})[c]	<0.0001	**0.3788**	<0.0001	**0.0332**	**0.0006**	**0.3576**	<0.0001	**0.1707**
Confidence level for min. diff. (R^2_{yrand})[c]	<0.0001	**0.0160**	<0.0001	<0.0001	**0.0121**	**0.2983**	**0.0078**	**0.3681**
Randomizations %, conf. level >0.0001 (Q^2_{yrand})[d]	0	**68%**	0	10%	10%	**83%**	0	**43%**
Randomizations %, conf. level >0.0001 (R^2_{yrand})[d]	0	2%	0	0	**40%**	**29%**	**40%**	**30%**
y-Randomization intercept (r_{yrand} *vs.* Q^2_{yrand})[e]	−0.8(1)	−0.52(2)	−0.53(8)	−0.37(1)	−1.3 (1)	−1.08(3)	−1.03(9)	−0.71(2)
y-Randomization intercept (r_{yrand} *vs.* R^2_{yrand})[e]	0.15(4)	0.241(5)	0.01(4)	0.110(4)	**0.37(5)**	**0.44(1)**	**0.47(4)**	**0.548(4)**

[a] Statistical parameters are based on Q^2 from **y**-randomization (Q^2_{yrand}) and R^2 from **y**-randomization (R^2_{yrand}), with the number of randomizations $M=10$ and $M=1000$. Values typed bold represent critical cases.

[b] Minimum model-randomization difference: the difference between the proposed model (Table 1) and the best **y**-randomization in terms of correlation coefficients Q^2_{yrand} or R^2_{yrand}, expressed in units of the standard deviations of Q^2_{yrand} or R^2_{yrand}, respectively. The best **y**-randomization is defined by the highest Q^2_{rand} or R^2_{rand}.

[c] Confidence level for normal distribution of the minimum model-randomization difference.

[d] Percentage of randomizations characterized by model-randomization difference (in terms of Q^2_{yrand} or R^2_{yrand}) at confidence levels >0.0001.

[e] Intercepts obtained from two **y**-randomization plots for each regression model proposed, with statistical errors in brackets for significant digits. Q^2_{yrand} or R^2_{yrand} is the vertical axis, whilst the horizontal axis is the absolute correlation coefficient r_{yrand} between the original and randomized vectors **y**. The randomization plots are completed with the data for the proposed model ($r_{yrand}=1.000$, Q^2 or R^2).

0.4 and considering them as serious cases of chance correlation. It is noticeable that the four models in Fig. 4A are discriminated in the R^2_{yrand}–Q^2_{yrand} space by the distribution of the randomizations (see also the maximum and average values and standard deviations of R^2_{yrand} and Q^2_{yrand} in Table 3). Although randomizations for all models are placed at negative Q^2_{yrand} and thus are far from the models along the vertical axis, the same is not valid with respect to the horizontal axis. As a cumulative effect, randomizations for OPS-PLS are well clustered, those for PLS are moderately spread, whilst randomizations for SVR and LS-SVM are rather dispersed in the central part of the R^2_{yrand}–Q^2_{yrand} space and exceed 0.4 at the horizontal axis, characterizing the presence of chance correlation in the SVM-based models. The second, a semi-quantitative criterion for **y**-randomizations detection is based on the work of Rücker et al. [27], where the smallest difference between the model and randomizations in terms of Q^2_{yrand} or R^2_{yrand}, were expressed in units of the corresponding standard deviations. Since differences are normally distributed, the smallest differences can be easily expressed in terms of confidence levels [46], as shown in Table 3. It can be noticed that SVR has somewhat critical confidence level for Q^2_{yrand}, but the analogue statistics for R^2_{yrand} of the SVM-based models is rather unfavorable. The third approach, more quantitative than the previous ones, takes into account frequencies of all randomizations with distances from the respective models at confidence level >0.0001. This way, 40% of randomizations are critical in R^2_{yrand} for the SVM-based models, but only 10% are unfavorable in Q^2_{yrand} for SVR. The last, most quantitative and rigorous approach, is based on Eriksson et al. [48], in which the absolute Pearson correlation coefficient r_{yrand} between the original and randomized vectors **y** is

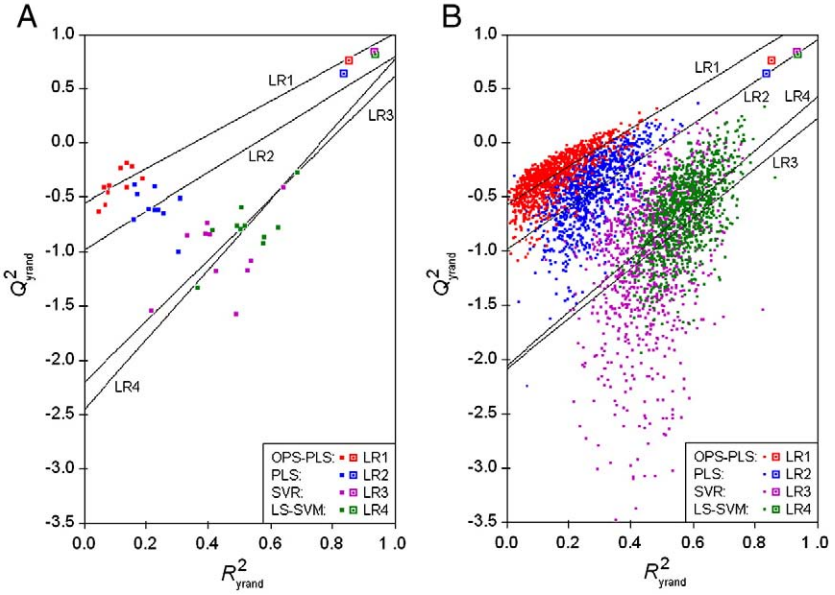

Fig. 4. The Q^2_{yrand} against R^2_{yrand} plots for 10 (A) and 1000 (B) randomizations of the four QSAR models, with the corresponding linear regression (LR) lines. The real models are placed at the right upper corner, whilst the respective randomizations are spread over the central and left parts of the plots.

plotted against Q^2_{yrand} and R^2_{yrand}. The plots (Figs. S1 and S2 left in supplementary data) are extended to the model's data (*i.e.*, $r=1$, $Q^2_{yrand}=Q^2$ and $R^2_{yrand}=R^2$), lines from linear regression are drawn, and the intercepts are compared with recommended limits of 0.05 and 0.3 for Q^2_{yrand} and R^2_{yrand}, respectively. Taking into account these criteria (Table 3), the intercepts on the axis R^2_{yrand} for SVR and LS-SVM exceed the limits. The three previous approaches to analyze **y**-randomizations seem to be suitable for small *M*, whilst this one including r_{yrand} shows uncertainties not in the graphical presentation (Figs. S1 and S2 left) but in numerical analysis, due to possible small intercept-to-error ratios (see the respective values for OPS-PLS in Table 3). Concluding, 10 **y**-randomizations pointed out that SVR and LS-SVM have failed in this validation, OPS-PLS has shown to be certainly free of chance correlations, whilst PLS is probably still inside the acceptable limits.

By increasing *M* from 10 to 1000 several effects, from minor to significant, can be observed graphically (Figs. 4, S1 and S2) and numerically (Table 3 and Table S1). The Q^2_{yrand}–R^2_{yrand} plots for 10 and 1000 randomizations are very similar at qualitative level. Dispersion of data points is increasing in the order OPS-PLS → PLS → LS-SVM → SVR (see standard deviations in Table 3), the data distributions are of similar basic shapes although more dispersed for SVR, and are centered at very similar coordinates (average Q^2_{yrand} and R^2_{yrand} increase slightly to moderately by applying large *M*, see Table 3). Such similarities can be also observed in the plots Q^2_{yrand}–r_{yrand} (Fig. S1) and R^2_{yrand}–r_{yrand} (Fig. S2). The corresponding linear regression equations for each QSAR model (Table S1), regardless of somewhat larger values of intercepts and significantly smaller respective errors at large *M* (Table 3), are mostly statistically indistinguishable when the strict criteria of normal confidence level <0.0001 are applied (Table S1). However, high standard deviations and large maximum values of Q^2_{yrand} and R^2_{yrand} are due to large *M*. This means that two first criteria presented, namely the rule of limit 0.4 for Q^2_{yrand} and R^2_{yrand}, and the criterion of the smallest difference between the proposed model and randomizations, are not applicable for large *M*. The other two criteria, frequency of chance correlations over 25%, and the **y**-randomization intercepts are statistically more adequate for large *M*. In this sense, one can see from Table 3 that SVR and LS-SVM have obviously failed in the 1000 **y**-randomizations test, whilst PLS can be characterized as a still tolerable case but close to the limits. PLS has high chance correlation frequency in Q^2_{yrand} due to relatively low Q^2, and relatively high intercept on R^2_{yrand} with substantial number of data points above the intercept for r_{rand} close to zero. The presented **y**-randomization validations illustrate rather clearly that there is no need to apply large *M* in QSAR, once sufficient information at small *M* is obtained to judge validated models. It is interesting to note for $M=1000$ that, whilst the four QSAR models are distinguished in the R^2_{yrand}–r_{yrand} plot (Fig. S1), they are rather overlapped in the Q^2_{yrand}–r_{yrand} plot (Fig. S2). The highest concentration of the data points is situated around a R^2_{yrand}–r_{yrand} plane at about $Q^2_{yrand}=-0.5$, which is well visible in the three-dimensional r_{yrand}–R^2_{yrand}–Q^2_{yrand} plot (Fig. S3), describing **y**-randomization by three parameters of different sensitivity with respect to chance correlation.

Why SVM is so promising in the predictions and validations except for **y**-randomization at low and high *M*? Possible explanation is the high susceptibility of SVM to overtraining when not good model selection is obtained. Extensive literature search and inspection of 188 QSAR and QSPR research articles employing SVM has revealed that the SVM performance in LNO and **y**-randomization with respect to other methods has never been explored systematically and in details. Only routine comparison of SVM to other regression methods in terms of LNO or **y**-randomization has been found in eight articles. The other possible reason for the unfavorable behavior of SVM is that non-linear SVM should be used in non-linear modeling, and not as a competitive approach to simple calibration methods such as PLS in linear problems. Reader should notice that the choice of SVM for calibration and classification purposes in QSAR and QSPR literature is frequently not justified properly. SVM is being computationally more complex and time-consuming than PLS even when used in its linear variant. Non-linear SVM is a good choice when statistical tests like residuals analysis for linear models, analysis of descriptor–**y** scatterplots etc., strongly indicate the presence of non-linearities. In the present study, the linear data set is used with the purpose to raise questions about various issues and difficulties in using non-linear SVM in typical QSAR works.

3.2. Inhibitor selection by the SVM-based models

The history of SVM in QSAR and QSPR is characterized by difficulties in interpretation of regression equations, up to the point to consider SVM as "a state-of-the-art black-box modeling technique" [49]. A large majority of QSAR and QSPR articles employing SVM (over 90%) interpreted regression models indirectly, by discussing SVM prediction, validation performances and meaning of molecular descriptors. Some articles dealt with linear SVM models by interpreting SVM regression equations [50,51], and using star plots as quantification of descriptor importance over bootstrap folds [12] (in total 10 articles). Such a situation is partially due to insufficient efforts in chemical interpretation of SVM equations and the fact that it is much easier to interpret relationships among descriptors than among compounds. A general QSAR or QSPR has to be accompanied with "a mechanistic interpretation, if possible" [25], and this section shows that it is indeed possible to interpret SVM-based models directly.

The 14 descriptors were interpreted in details with respect to their use in drug design and understanding intermolecular interactions some years ago [29,31]. Understanding interactions between **1**–**48** and the protease at molecular level is important for interpretation of the SVM-based models. The previous exploratory analysis for the whole data set [29] and the analogue analysis for the training set (Fig. 5) may give some insight into the inhibitor selections by SVR and LS-SVM, *i.e.*, discrimination of support vectors from non support vectors. **1**–**48** possess a central peptidic chain and four substituents (rings, aliphatic fragments or combinations), as shown in the box in Fig. 1. There are two inner substituents (drawn upwards or downwards of the chain) P1 and P1′, and two terminal substituents (drawn left and right of the chain) P2 and P2′ which occupy corresponding pockets of the HIV-1 protease [29,31]. The protein possesses six more pockets, named as S3, S4, S5, and S3′, S4′ and S5′ in accordance to the distance from the catalytic active site [52]. Larger peptidic inhibitors occupy six or more pockets [53], but even in such cases a good complementarity between substituents P1, P1′, P2 and P2′ and the pockets is necessary. These facts and previous observations [29,31] indicate the importance of molecular size for anti-HIV activity of **1**–**48** rather clearly.

Three biological activity classes [29], inhibitor selections by SVR and LS-SVM, and clustering patterns for the inhibitors were inspected in the exploratory analyses for the training data (Fig. 5). Classes for slightly ($pIC_{50}=5.158$ to 6.246), moderately ($pIC_{50}=6.640$ to 8.268) and highly ($pIC_{50}=8.886$ to 10.267) active compounds are denominated class I, class II and class III, respectively. Class I contains 6 inhibitors (**10**, **21**, **33**, **38**, **43** and **48**), class II has 11 (**2**, **18**–**20**, **23**, **25**, **28**, **30**, **42**, **45** and **46**) and class III 14 (**4**–**7**, **9**, **11**, **15**–**17**, **27**, **31**, **34**, **37** and **39**) compounds. SVR selects 4 (**10**, **21**, **38** and **43**) inhibitors from class I, 7 (**2**, **18**, **19**, **20**, **25**, **42** and **45**) from class II and only 3 (**17**, **34** and **37**) from class III. LS-SVM uses 3 (**10**, **21** and **38**) inhibitors from class I, 9 (**2**, **18**, **19**, **20**, **23**, **25**, **28**, **45** and **46**) from class II and 6 (**5**, **11**, **16**, **17**, **27** and **37**) from class III. Fractions of classes I, II and III captured by SVR are 67%, 64% and 21%, respectively, and by LS-SVM are 50%, 82% and 43%, respectively. Moderately active compounds are well represented in both models, slightly active have weaker participation in LS-SVM than in SVR, whilst highly active compounds are not well presented, especially in SVR (Fig. 5).

HCA analysis for 48 inhibitors [29] resulted in four clusters (G1, G2, G3 and G4) which reflected well systematic variations in biological activity, molecular size and molecular structures. Basically the same

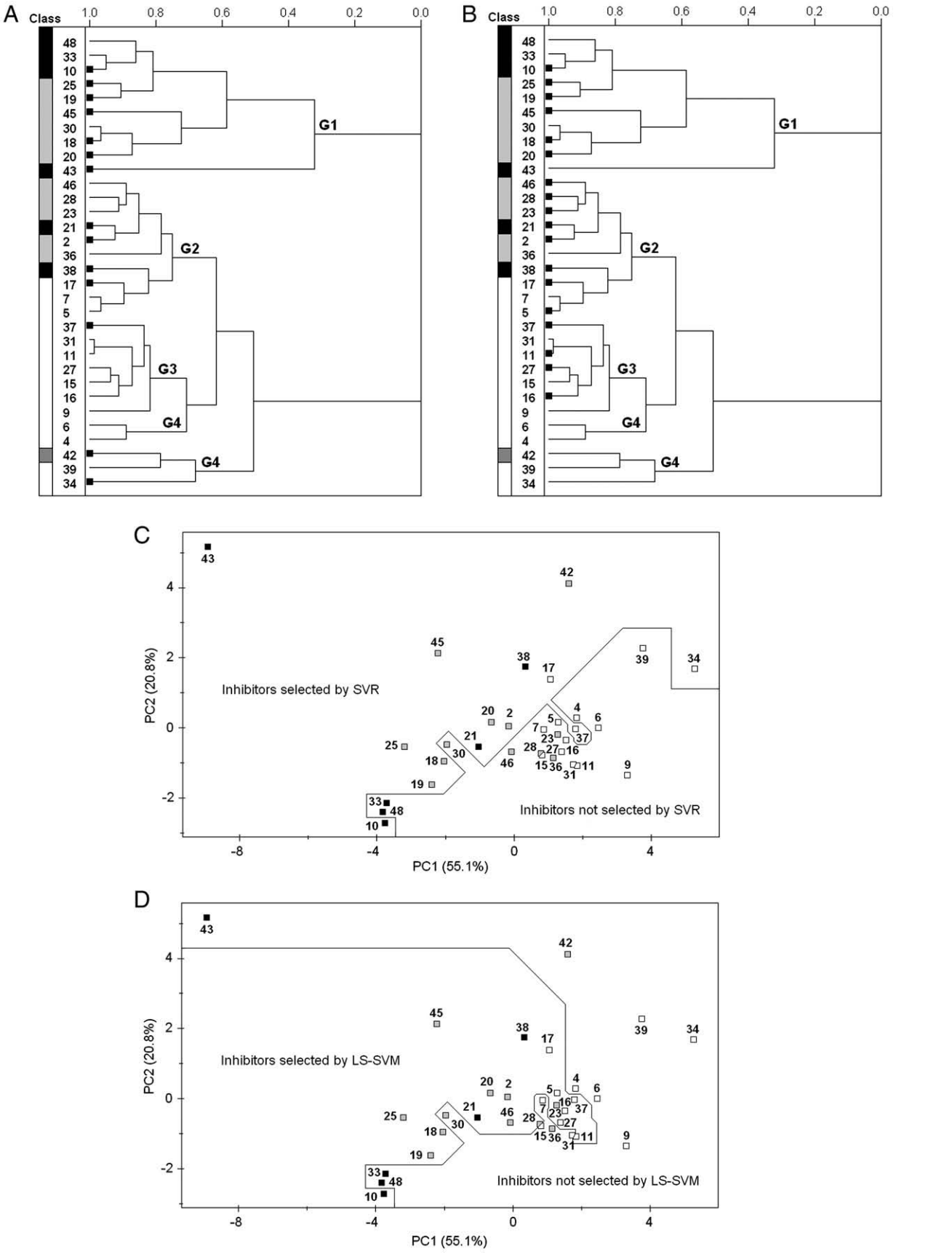

Fig. 5. HCA dendrogram with incremental linkage for the A) SVR model and B) LS-SVM model, showing the clustering pattern of inhibitors (clusters denominated G1, G2, G3 and G4 as from the literature [29]), inhibitor selection by these models (■), and activity classes column (black: class I, gray: class II, white: class III). PC1–PC2 scores plot the C) SVR model and D) LS-SVM model, showing the inhibitor selection by the models and activity classes (black squares: class I, gray squares: class II, white squares: class III). Selected and not selected inhibitors are separated by a set of arbitrary lines.

trends can be observed for the training set (Fig. 5A and B). Biological activity and molecular mass roughly increase in the direction G1 → G2 → G3 → G4. The SVM-based models capture the cluster G1 equally well. LS-SVM takes inhibitors from G2 and G3 more frequently than SVR. However, whilst G3 participates in the SVR model very poorly, G4 (*i.e.*, its part containing **34**, **39** and **42**) is well represented by this model. LS-SVM does not use G4 at all. Probably this equilibrium between SVR and LS-SVM is responsible for rather similar performances of the two models.

PCA analysis for the training set shows that the first three principal components (PCs) capture 84.4% of the total variance, similarly to the analysis for the complete set (85.9% [29]). Both analyses exhibit rather similar distribution of inhibitors and activity classes in scores plots. PC1–PC2 scores plot is sufficient to illustrate that inhibitors selected by the SVM-based models can be discriminated from those not selected (broken lines in Fig. 5C and D). The inhibitors not selected by SVR are at the right-central and right-bottom regions. The analogue area for LS-SVM is somewhat larger, including also the narrow top region of the PC1–PC2 space, but excluding a part of the central-right region. PC1 correlates with the biological activity and extensive features like molecular size and content of valence electrons. PC2 is related to molecular topology, especially branchness. The right-central region has the highest concentration of moderately and highly active inhibitors which are of elevated molecular size and modest branchness.

Why the SVM-based models exclude a large portion of highly active inhibitors (more than 50%), which are of the greatest interest for drug design? What does it mean chemically to select inhibitors by SVM? The answer lies in the obtained Lagrange multipliers $\alpha_i^{(*)}$ (Table 1) and molecular structures (Fig. 1). According to the SVM theory [9,36,54], instead of finding the regression function that fits best to the data, the flattest hyperplane (LS-SVM), or the flattest hyperplane with ε-precision (insensitive zone) considered as a hypertube (SVR) that best fits the data, is constructed in a high-dimensional feature space. Geometrically speaking, all samples can be characterized by two Lagrange multipliers in SVR: α_i and α_i^* related to the position regarding the hypertube. Samples within the hypertube are non support vectors ($\alpha_i^{(*)}=0$), whilst samples above ($\alpha_i>0$ and $\alpha_i^*=0$) and below the hypertube ($\alpha_i=0$ and $\alpha_i^*>0$) are support vectors. Only one Lagrange multiplier per sample is active ($\alpha_i\alpha_i^*=0$). Depending on which of them is active, it appears as a positive (α_i) or negative ($-\alpha_i^*$) coefficient in the regression equation, shown by α values in Table 1. LS-SVM, on the contrary, uses all the training data, so each sample is associated to a nonzero Lagrange multiplier α_i ($\alpha_i = C\, e_i$) related to its distance and position (orientation) with respect to the estimated function in the variables space. The sparse approximation is obtained, in which samples associated to smaller absolute values of Lagrange multipliers are eliminated (non support vectors).

Table 1 shows that the SVR regression equation is based on 7 samples above ($\alpha>0$) and 7 below the hypertube ($\alpha<0$). The LS-SVM regression equation includes 9 samples above ($\alpha>0$) and 9 below the hypercurve ($\alpha<0$). Detailed inspection of molecular structures (Fig. 1) was combined with the exploratory analysis (Fig. 5) and inspection of Lagrange multipliers (Table 1) expressed as percentages of the maximum multiplier (Table 4). This analysis has indicated that

Table 4
Structural interpretation of nonzero Lagrange multipliers in the SVR and LS-SVM regression equations.

Inhibitors, clusters[a]	Clustering pattern			%Lagrange multiplier[b]		Structural peculiarity of selected inhibitors with respect to others in the same cluster G[c]
	Clustering[d]	Structural basis[e]	G[f]	SVR	LS-SVM	
Support vectors: inhibitors selected by both SVR and LS-SVM						
{**2**, **21**}	Cluster (0.924)	Small difference in P1′/P2′ (–CH_3)	G2	**2**: 100 **21**: −100	**2**: 88 **21**: −100	**2**: P1′ consists of two substituents **21**: P2′ is a branched substituent
{**19**, **25**}	Cluster (0.905)	Small difference in P2′ (–CH_2–, 2H)	G1	**19**: 41 **25**: −6	**19**: 82 **25**: −32	**19**: P2′ is a cycloalkene group (C_5) **25**: P2′ is a cycloalkane group (C_6)
{**18**, **20**}	Cluster (0.872)	Small difference in P2′ (–CH_2CH_2–)	G1	**18**: 32 **20**: −38	**18**: 47 **20**: −74	**18**: P2′ is an unsubstited benzyl group **20**: P2′ is an unsubstituted indanyl group
{**17**, **38**}	Cluster (0.821)	Small difference in P1′ (–CH_2–, 2H)	G2	**17**: 28 **38**: −54	**17**: 44 **38**: −78	**17**: P1′ ends in a highly hydrophobic *t*-Bu **38**: P1′ consists of two substituents; one is a hydrophobic ring fragment at the chain
10	Close to {19, 25} (0.808)	P1′ or P2′ is small or non-existing	G1	−26	−42	P1′ is non-existing (H atom)
45	Close to {18, 20} (0.724)	P1 and P2′ are rather hydrophobic	G1	13	24	P1 is hydrophobic, aliphatic bicyclic system; P2 is in unusual conformation
37	Isolated (SVR); close to {16, 27} and 11 (LS-SVM; 0.837)	P1′ and P2′ are rather hydrophobic, one of them has 1–2 heteroatoms (O, S)	G3	13	30	P2 contains a furane ring
Support vectors: inhibitors selected by either SVR or LS-SVM						
{**23**, **28**}	Cluster (0.913)	Small difference in P2′ (–CH_2–)	G2	**23**: 0 **28**: 0	**23**: −43 **28**:−56	**23**: P2′ contains a methylformyl group **28**: P2′ contains an endocyclic O atom
{**16**, **27**}	Cluster (0.911)	P1′ and P2′ contain 2 heteroatoms (O, S)	G3	**16**: 0 **27**:0	**16**: 25 **27**: 39	**16**: P1′ contains an exocyclic S atom **27**: P2′ contains an endocyclic O atom
46	Close to {23, 28} (0.889)	P2′ contains 1–2 heteroatoms (O)	G2	0	38	P1′ is a phenyl and not a benzyl group
5[g]	Close to 17 (0.966)	P1′ is longer than benzyl by two C atoms	G2	0	37	P1′ is a long and rigid conjugated p system
11	Close to {16, 27} (0.872)	P1′ and P2′ contain 2 heteroatoms (O, S)	G3	0	−29	P1′ ends in a strong sp^3 hydrogen bonding group (OH)
42	Close to 34 (0.681)	P2 and P2′ are bicyclic systems; P1′ is a benzyl; the chain around OH and P1′ is the same	G4	−3	0	P1 is a small group (*i*-Pr) and P2 is a bicyclic heteroaromatic system
34	Close to 42 (0.681)	P2 and P2′ are bicyclic systems; P1′ is a benzyl; the chain around OH and P1′ is the same	G4	−2	0	Highly symmetric; all substituents are aromatic rings
43[h]	Isolated	G1: rather hydrophobic substituents	G1	2	0	Highly symmetric; P1, P1′ are bicyclic aliphatic system; P2, P2′ contain *t*-Bu

[a] Two-membered inhibitor clusters, inhibitors close to some cluster or inhibitors far from any cluster. It is understood by inhibitor or support vector every compound which was selected by the SVR or LS-SVM model as important for the final regression equation.
[b] Lagrange multipliers for inhibitors expressed as percentage of the maximum absolute value obtained from the SVR or LS-SVM regression equation.
[c] Main structural characteristics which distinguish an inhibitor from other inhibitors in the same cluster G.
[d] Position of a selected inhibitor with respect to other selected inhibitors. Similarity index for a cluster or between a cluster and an inhibitor is given in brackets.
[e] Structural similarity or small difference between close inhibitors.
[f] One of the clusters G1, G2, G3 and G4 to which the inhibitor or cluster belongs (HCA analysis in Fig. 5A and B).
[g] Not treated in cluster with **17** because {**17**, **38**} is common for SVR and LS-SVM.
[h] At similarity index 0.321 with respect to other inhibitors in the cluster G1.

there are three types of inhibitors: 1) support vectors for both models SVR and LS-SVM; 2) support vectors in either SVR or LS-SVM; and 3) non support vectors, (always zero Lagrange multipliers). Clusters {**2**, **21**}, {**17**, **38**}, {**18**, **20**}, and {**19**, **25**} at similarity index >0.8 and three inhibitors (**10**, **37** and **45**) participate in both SVM equations. Clusters {**16**, **27**} and {**23**, **28**} at similarity index >0.9 and 6 inhibitors (**5**, **11**, **34**, **42**, **43** and **46**) are included in one of the SVM models. The clusters are connected with other selected inhibitors into larger clusters within one of the clusters G (G1, G2, G3 and G4, Fig. 5A and B). The rationale for these patterns is a high molecular similarity in substituents and in chains, frequently including small group rearrangements (specified in Table 4). The percentage Lagrange multipliers and their absolute values for the whole training set do not correlate clearly with the biological activity, molecular descriptors or scores in Fig. 5C and D. HCA dendrogram (Fig. 5A and B) also does not show noticeable relationship between the Lagrange multipliers and positions of selected inhibitors. However, molecular diversity is the feature which greatly affects the selection of inhibitors by the final models. Fig. 1 and Table 4 show that the more peculiar the molecule within its cluster G or the whole training set, the more distant it is from the group of inhibitors along an imagined high-dimensional line or curve in the feature space. Consequently, inhibitors with small structural variations are close to each other (highly concentrated as in Fig. 5C and D), defining a rather well geometrical pattern in the feature space, and are well captured by a hypertube or hyperplane (well predicted by the estimate function), so, they are not support vectors. Other inhibitors, due to their peculiar molecular structure, deviate from such trends, so the final hypertube or hypercurve has to be adjusted to incorporate them. Small clusters and inhibitors in Table 4 are arranged in decreasing order of the absolute Lagrange multipliers, coinciding well with the degree of structural diversity relative to that of **1**. {**2**, **21**} is characterized by branching at a substituent site (P1′ or P2′) and not by one substituent, which makes this cluster unique in this aspect. {**19**, **25**} possesses a small hydrophobic and aliphatic ring instead of an aromatic at P1′, which also makes it unique, but this molecular diversity is somewhat smaller. {**18**, **20**} contains unsubstituted aromatic rings at P1′, meaning that there is no change in branching or size, but no hydrogen bonding or polar group present is present. {**17**, **38**} is characterized by increasing hydrophobicity of the benzyl at P1′ by placing a new group at the benzene ring (**17**) or introducing a ring

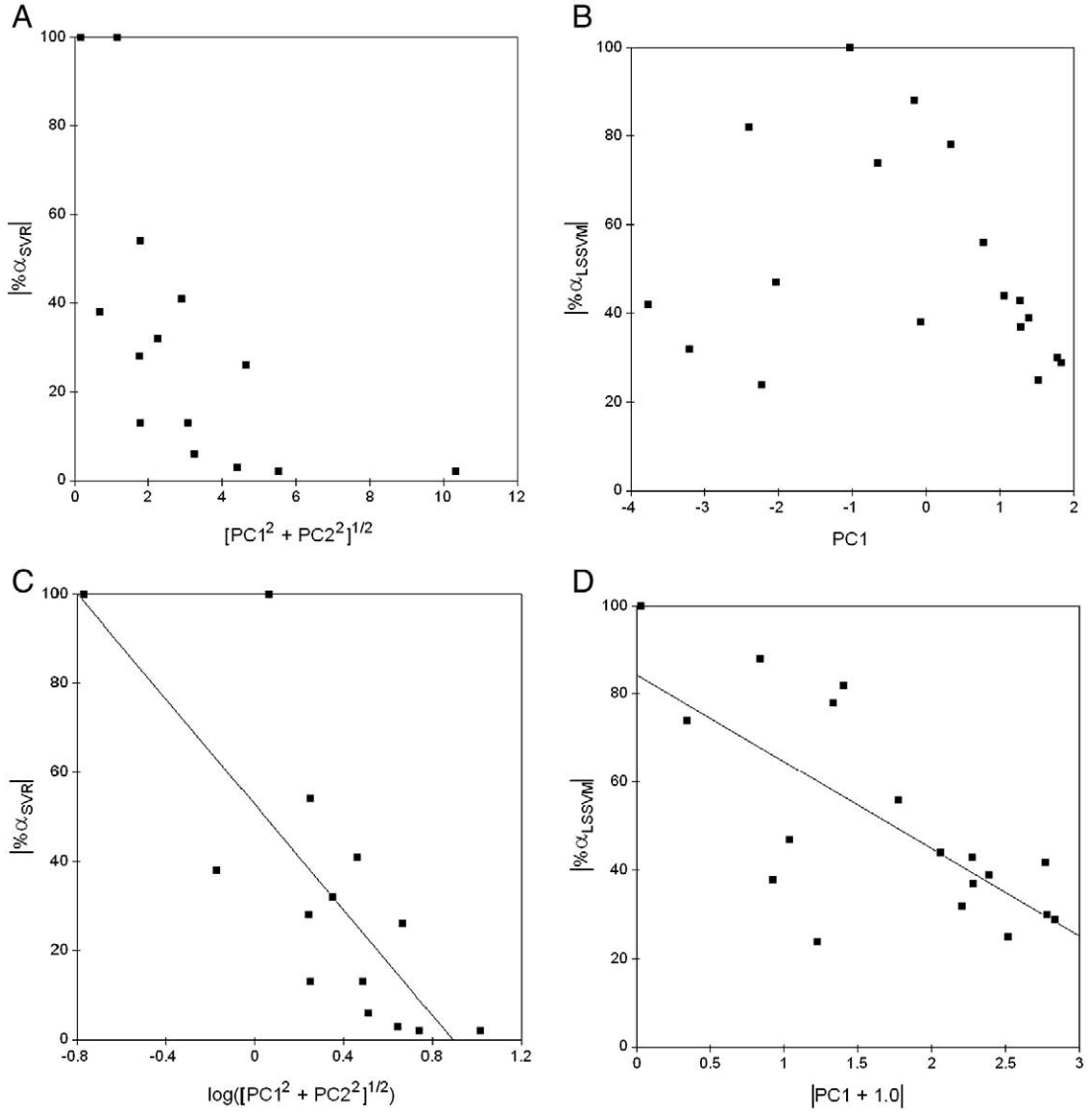

Fig. 6. Relationships between absolute values of relative Lagrange multipliers (Table 3) and principal components (PCs) for the SVR (left plots) and LS-SVM (right plots) models. The original relationships (plots A and B) are linearized (plots C and D) by means of linear regression. The obtained equations have the following statistics: a) for SVR: 14 data points, $|\%\alpha_{SVR}| = 53(7) - 59(13)\ \log([PC1^2 + PC2^2]^{1/2})$ with $p<0.001$ for t-statistics, $r = -0.788$, $F_{1,12} = 20$ with $p<0.001$; and b) for LS-SVM: 18 data points, $|\%\alpha_{LSSVM}| = 84(9) - 20(5)\ |PC1 + 1.0|$ with $p<0.001$ for t-statistics, $r = -0.723$, $F_{1,16} = 17$ with $p<0.001$.

at the same substitution site within the chain. Table 4 continues with smaller variations and ends with structures having larger substituents among which one is small (**42**) or which are positioned in some symmetrical way (**34** and **43**). These structural variations give origin to changes in molecular descriptors, intermolecular interactions with the protease, and finally, in the biological activity.

Two-membered clusters and other associated inhibitors in Table 4 can be recognized in the dendrogram and scores plot (Fig. 5) and also in the feature space. It means that a hypertube or hypercurve passes through a cluster whose members have different signs of Lagrange multipliers (the signs are the same for SVR and LS-SVM, see Table 1), and passes by a cluster with members of the same sign of α_i or α_i^*. Besides this information, one more fact can be extracted for selected inhibitors which are in common for SVR and LS-SVM. Lagrange multipliers are positive for **2**, **17**–**19**, **37** and **45**, and negative for **10**, **20**, **21**, **25** and **38**. Careful inspection of respective molecular structures using **1** as a standard reveals that structural changes result in mass shift from right to left (from P2′ to P1′ or from P1′ to P1) for positive Lagrange multipliers, whilst negative α_i and α_i^* are related to the mass concentration in the opposite direction (from P1′ to P2′). The two-membered clusters nicely show these trends: $-CH_3$ is placed at P1′ site in **2** ($\alpha_2>0$) whilst $-CH_2-$ is a part of P2′ in **21** ($\alpha_{21}<0$). P2′ is smaller in **18** or **19** ($\alpha_i>0$) than in **20** or **25** ($\alpha_i<0$). More complicated are {**17**, **38**} where an alkyl (*t*-Bu) is at P1′ in **17** ($\alpha_{17}>0$), and another alkyl ($-CH_2CH_2-$) is in the ring between P1′ and P2′ in **38** ($\alpha_{38}<0$). **10** is unsubstituted in P1′ ($\alpha_{10}>0$), whilst there is a large substituent P1 in **45** ($\alpha_{45}>0$) and a ring (instead of *t*-Bu) in P2 of **37** ($\alpha_{37}>0$).

Very small Lagrange multipliers for **34**, **42** and **43** in SVR minimize the differences between the two SVM-based model selections of highly active inhibitors from the cluster G4 (Fig. 5). Now it becomes clear why SVM does not select a representative subset of inhibitors from class III. There are four possible reasons for this: the tight clustering of highly active compounds with respect to the biological activity (Fig. 2) and molecular descriptors (Fig. 5), intrinsic insensitivity of SVM to molecular diversity of larger molecules, and initial data transformation which reduces molecular diversity. Variations in structures of larger molecules (mainly from class III), become small when considered relative to molecular size. Inhibitors with larger absolute Lagrange multipliers are from clusters G1 and G2 (Table 4), and regularly smaller are from classes I and II. Variation in structures of highly active compounds is important for drug design to find and optimize the lead compound, so this is an issue to analyze carefully in future SVM applications in QSAR.

The absolute values of Lagrange multipliers from Table 4 (SVR and LS-SVM data) show interesting correlations with principal components (PCs), some of which are statistically very significant (Fig. 6). The absolute Lagrange multipliers of the SVR model are non-linearly correlated to $[PC1^2+PC2^2]^{1/2}$ (Fig. 6A), *i.e.*, the diagonal of PC1 and PC2 along which the support vectors are mainly concentrated (Fig. 5C), with the maximum absolute multipliers around the plot's origin. The multipliers of the LS-SVM model are also non-linearly correlated to the absolute values of PC1 along which the inhibitors are mainly concentrated (Fig. 6B), with absolute maxima at small PC1. It is likely that the non-linear character of these relationships between PCs (or more precisely, functions of PCs) and Lagrange multipliers are problem- and method-dependent. Physical meaning of both scores and multipliers is the position of a sample point with respect to some reference in the feature space. When taking into account these observations, the connection between SVM and PCA methodologies may be better understood in a particular QSAR study. The two relationships can be linearized by logarithmic (Fig. 6C) and absolute value (Fig. 6D) transformations followed by linear regression. The resulting regression equations are statistically significant, which can be seen from correlation coefficients (absolute values >0.7) and *t*- and *F*-statistics at confidence level <0.001, as obtained by using the QuickCalcs software [55]. When extended to all non support vectors, the most significant correlations respect to descriptors X_1–X_{14}, activity **y** and principal components are modest (correlation coefficients from −0.20 to −0.48), and relationships are visible in the scatterplots (not shown).

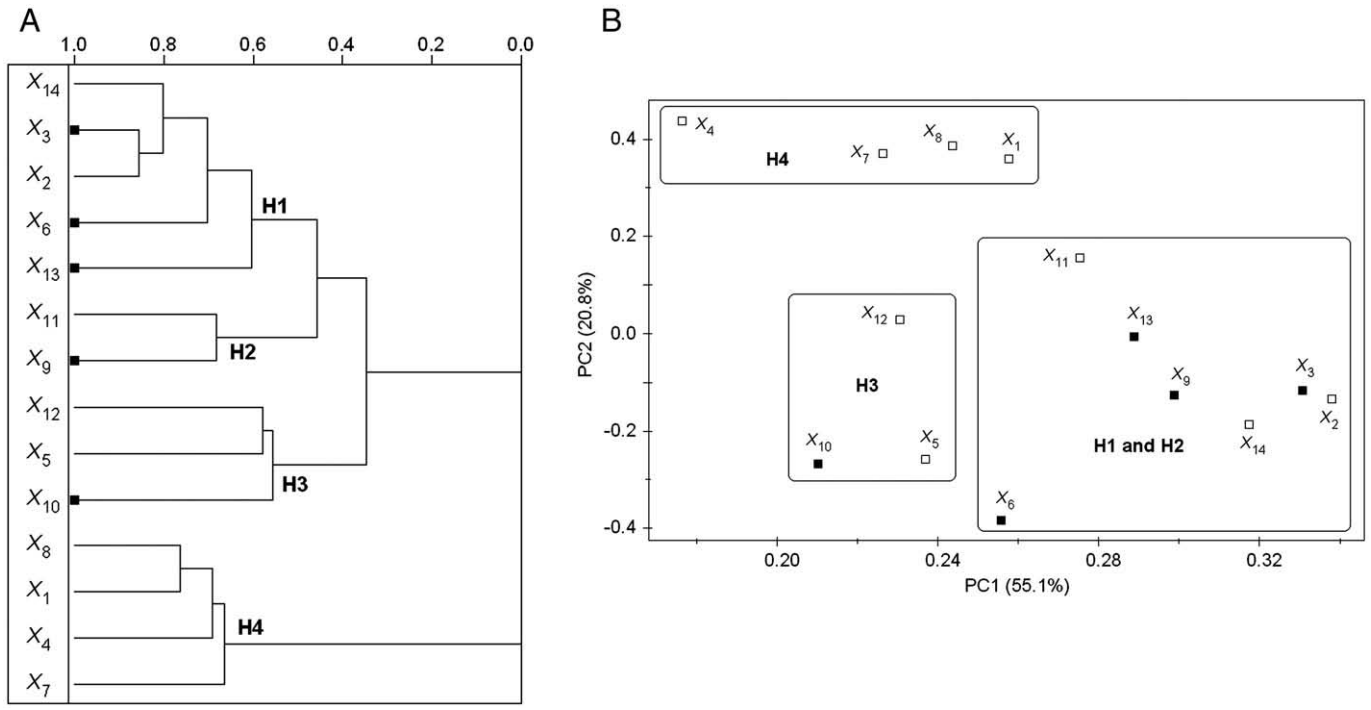

Fig. 7. Exploratory analysis explaining the variable selection by OPS-PLS. A) HCA dendrogram with incremental linkage showing the clustering pattern of molecular descriptors (clusters denominated H1, H2, H3 and H4 as from the literature [29]) and their selection (■). B) PC1–PC2 loadings plot showing the inhibitor selection (■) and clustering (mixed H1–H2 and separated H3 and H4 clusters).

3.3. Variable selection by the OPS-PLS model

OPS-PLS [28] is a new PLS approach to QSAR, and hence, its performance in terms of variable selection is rather unknown and worth to analyze. Fig. 7 presents exploratory analysis for the training data set (using 14 descriptors). These results are very similar to those for the full data set (*i.e.*, for 48 descriptors [29]). Inhibitors are grouped in two main clusters: the larger consisting of H1, H2 and H3, and the smaller H4. The two clusters differ in the profile of descriptor–activity scatterplots, according to which all descriptors in H4 (X_1, X_4, X_7 and X_8) have more dispersed data in these plots than those for H1–H3. This difference is more obvious when the regression coefficients of the PLS model (Table 1) and respective correlation coefficients are taken into account. A good visualization of this differentiation can be obtained when the descriptors are plotted against the Mahalanobis distance (Fig. S4 in Supplementary data), where the scatterplots for descriptors from cluster H4 are better structurally defined than those for other descriptors. By other words, due to high correlation between Mahalanobis distance and the scores from significant principal components, these four descriptors do not bring new information to the model. Therefore, since the OPS algorithm uses various information vectors, exclusion of H4 and the corresponding loadings space (Fig. 7) during variable selection becomes more obvious. Five selected descriptors are well spread in the remaining clusters: X_3, X_6 and X_{13} are from H1, X_9 is from H2, and X_{10} is from H3. A rather good distribution of these descriptors is noticeable in the loadings plot. X_3 is from {X_2, X_3, X_{14}}, so the complete structure of H1 is well preserved.

Descriptors X_1–X_{14} have been classified [29] in four ways. With respect to their dimensionality, they are 1D (one-dimensional) when originated from physical and chemical constants, 2D when based only on molecular topology, and 2D with some 3D information when created from molecular structures encoding some stereochemical information as in Fig. 1. The 1D and 2D classes are not present in the OPS-PLS model, which means that the most important descriptor do include certain 3D information. X_1–X_{14} can be classified according to real phenomena they describe: 1D (chemical composition), 2D (molecular topology, chemical bonding and main intramolecular interactions), and 3D (intermolecular interactions in the 3D space). The 3D class is abundantly presented in OPS-PLS (X_3, X_6, X_9 and X_{10}) and the 2D class has one descriptor (X_{13}). The descriptors can be divided into extensive (depending on molecular size) and intensive (not depending on molecular size). Among the selected descriptors, only X_6, the electron density for π- and lone pair electrons, is intensive. X_1–X_{14} can be discriminated with respect to their natures as electronic, steric-geometrical, electronic-geometrical, compositional, hydrophobic and topological descriptors. Most classes, namely electronic (X_6 and X_{13}), topological (X_3), steric-geometrical (X_9) and electronic-geometrical (X_{10}) participate in the OPS-PLS model. These classes show that two aspects of molecular structure, the geometrical and electronic structures, are inseparable for **1**–**48**, as has been noticed previously [29]. How to interpret an OPS-PLS model in terms of molecular descriptors? Since the OPS algorithm, like any other computational approach, deals with mathematical and computational aspects and not chemistry, the way to ensure that the final model will be interpretable is to use informative descriptors, understandable from the chemical point of view whenever possible.

4. Conclusions

The common practice about SVM-based methods in QSAR and QSPR is to consider only prediction power, leave-one-out cross-validation and external validation of final regression models. Conventional linear models in QSAR and QSPR are required to be validated by more tests, and mechanistic interpretation must be given whenever the action of the studied compounds is known. Hence, it has become a common taboo that SVM models are always good, they do not need to be additionally validated, there is no need and sense or is not possible to give direct mechanistic interpretation of these models. Another frequent approach is not to test the data properly for non-linearity, but to go directly into parallel construction of non-linear SVM and linear models and conclude from some comparative tests the superiority of SVM, ignoring other validations that could or could not confirm the validity of SVM. These viewpoints presume that there is no necessity to investigate the addressed issues about SVM, especially because of several bootstrapping validations during the selection of the best SVM model.

The present work is an initial study with the aim to break these taboos, question the validation performance of SVM and propose its direct mechanistic (chemical) interpretation. In this study case, SVR and LS-SVM models were compared to PLS and OPS-PLS models. The results showed that SVM was superior to PLS and OPS-PLS in prediction, leave-one-out crossvalidation and external validation. However, SVM was not superior to OPS-PLS in leave-*N*-out cross-validation, and failed in **y**-randomization, which can be a consequence of SVM overtraining and the linear character of the data set used. OPS-PLS has shown to be the only one with undoubtedly satisfactory performance both in prediction and all validations. The OPS-PLS is a promising methodology which should be used in conjunction with informative and interpretable descriptors.

The regression models were interpreted in terms of selected descriptors and inhibitors *via* exploratory analysis. For the first time up to our knowledge, Lagrange multipliers, their absolute values and in some cases their signs, were interpreted in terms of molecular structure, descriptors and biological activity. Certain relationships between absolute Lagrange multipliers and principal components were detected as statistically significant. Elevated molecular diversity significantly contributed to nonzero Lagrange multipliers, and systematic differences in molecular structures determined signs of the multipliers. SVM showed to be insensitive to structural variations of highly active inhibitors.

The principles of validation and chemical interpretation of SVR and LS-SVM models given in this work are proposals for future investigations about SVM models in QSAR and QSPR, valid for any modeling and validation conditions. It has been shown that a small number of **y**-randomization runs is sufficient to detect the presence of chance correlation. Definite conclusions about the SVM behavior should be obtained by more extensive studies where various data sets with several training-external splits are explored, and different modeling procedures are tested.

Acknowledgement

The authors acknowledge The State of São Paulo Funding Agency (FAPESP) for financial support (MMCF and RK).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.chemolab.2009.04.012.

References

[1] M.M.C. Ferreira, J. Braz. Chem. Soc. 13 (2002) 742–753.
[2] J.O. Rawlings, Applied Regression Analysis: a Research Tool, Cole Statistics and Probability Series, Wadsworth & Brooks, *Wadsworth*, Belmont, CA, 1988.
[3] K.R. Beebe, R. Pell, M.B. Seasholtz, Chemometrics: a Practical Guide, Wiley, New York, NY, 1998.
[4] M.M.C. Ferreira, A.M. Antunes, M.S. Melgo, P.L.O. Volpe, Quim. Nova 22 (1999) 724–731.
[5] H. Martens, T. Naes, Multivariate Calibration, 2nd ed.Wiley, New York, NY, 1989.
[6] A.I. Belousov, S.A. Verzakov, J. von Frese, Chemom. Intell. Lab. Syst. 64 (2002) 15–25.
[7] U. Thissen, R. van Brakel, A.P. de Weijer, W.J. Melssen, L.M.C. Buydens, Chemom. Intell. Lab. Syst. 69 (2003) 35–49.
[8] V. Vapnik, A. Chervonenkis, Theory of Pattern Recognition, Nauka, Moscow, 1974.

[9] V. Vapnik, Statistical Learning Theory, Wiley, New York, NY, 1998.
[10] J.A.K. Suykens, J. Vandewalle, Neural Process. Lett. 9 (1999) 293–300.
[11] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, Least Squares Support Vector Machines, World Scientific, Singapore, 2002.
[12] M. Song, C.M. Breneman, J. Bi, N. Sukumar, K.P. Bennett, S. Cramer, N. Tugcu, J. Chem. Inf. Comput. Sci. 42 (2002) 1347–1357.
[13] M. Momma, K.P. Bennett, in: R.L. Grossman, J. Han, V. Kumar, H. Mannila, R. Motwani (Eds.), Proceedings of the 2nd SIAM International Conference on Data Mining, 11–13 April 2002, Arlington, VA, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2002, pp. 261–274.
[14] M.J. Embrechts, F. Arciniegas, M. Ozdemir, M. Momma, C.M. Breneman, L. Lockwood, K.P. Bennett, R.H. Kewley, Proceedings of the 2002 International Joint Conference on Neural Networks, Honolulu, HI, 12–17 May 2002, Vol. 1, IEEE, Piscataway, NJ, 2002, pp. 305–310.
[15] H.X. Liu, X.J. Yao, R.S. Zhang, M.C. Liu, Z.D. Hu, B.T. Fan, J. Comput.-Aided Mol. Des. 19 (2005) 499–508.
[16] H.X. Liu, X.J. Yao, R.S. Zhang, M.C. Liu, Z.D. Hu, B.T. Fan, J. Phys. Chem. B 109 (2005) 20565–20571.
[17] J.G. Topliss, R.P. Edwards, J. Med. Chem. 22 (1979) 1238–1244.
[18] S.R. Johnson, J. Chem. Inf. Model. 48 (2008) 25–26.
[19] R. Guha, J. Comput.-Aided Mol. Des. 22 (2008) 857–871.
[20] A. Golbraikh, A. Tropsha, J. Mol. Graph. Model. 20 (2002) 269–276.
[21] A. Tropsha, P. Gramatica, V.K. Gombar, QSAR Comb. Sci. 22 (2003) 69–77.
[22] K. Baumann, N. Stiefl, J. Comput.-Aided Mol. Des. 18 (2004) 549–562.
[23] P. Gramatica, QSAR Comb. Sci. 26 (2007) 694–701.
[24] R. Kiralj and M.M.C. Ferreira, J. Braz. Chem. Soc., R. Kiralj, M.M.C. Ferreira, J. Braz, Chem. Soc. 20 (2009) 770–787.
[25] Guidance Document on the Validation of (Quantitative) Structure–Activity Relationship [(Q)SAR] Models, OECD Environment Health and Safety Publications Series on Testing and Assessment No. 69, Paris, 2007. http://www.oecd.org/dataoecd/55/35/38130292.pdf [last access on May 8, 2008].
[26] K. Baumann, Trends Anal. Chem. 22 (2003) 395–406.
[27] C. Rücker, G. Rücker, M. Meringer, J. Chem. Inf. Model. 47 (2007) 2345–2357.
[28] R.F. Teófilo, J.P.A. Martins, M.M.C. Ferreira, J. Chemom. 23 (2009) 32–48.
[29] R. Kiralj, M.M.C. Ferreira, J. Mol. Graph. Model. 21 (2003) 435–448.
[30] HIV-1 QSAR data set. Available online at: http://chipre.iqm.unicamp.br/marcia/hiv1qsardata.html.
[31] R. Kiralj, M.M.C. Ferreira, J. Mol. Graph. Model. 21 (2003) 499–515.
[32] R. Kiralj, Y. Takahata, M.M.C. Ferreira, QSAR Comb. Sci. 22 (2003) 430–448.
[33] B.D. Hudson, D.C. Whitley, A. Browne, M.G. Ford, Croat. Chem. Acta 78 (2005) 557–561.
[34] M.K. Holloway, J.M. Wai, T.A. Halgren, P.M.D. Fitzgerald, J.P. Vacca, B.D. Dorsey, R.B. Levin, W.J. Thompson, L.J. Chen, S.J. Desolms, N. Gaffin, A.K. Ghosh, E.A. Giuliani, S.L. Graham, J.P. Guare, R.W. Hungate, T.A. Lyle, W.M. Sanders, T.J. Tucker, M. Wiggins, C.M. Wiscount, O.W. Woltersdorf, S.D. Young, P.L. Darke, J.A. Zugay, J. Med. Chem. 38 (1995) 305–317.
[35] A.J. Smola, B. Schölkopf, Stat. Comput. 14 (2004) 199–222.
[36] B. Schölkopf, A.J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond, MIT Press, Cambridge, MA, 2002.
[37] H.W. Kuhn, A.W. Tucker, in: J. Neyman (Ed.), Proceeding of the Second Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, University of California Press, Los Angeles, 1951, pp. 481–492.
[38] J.A.K. Suykens, J. De Brabanter, L. Lukas, J. Vandewalle, Neurocomput. 48 (2002) 85–105.
[39] L.Z. Gan, H.K. Liu, Y.X. Sun, in: J. Wang, Z. Yi, J.M. Zurada, B.-L. Lu, Y. Hujun (Eds.), Advances in Neural Networks, Lecture Notes in Computer Science, Vol. 3971, Springer, Berlin, 2006, p. 1016.
[40] E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, A. Weingessel, Misc Functions of the Department of Statistics (e1071), R Package Version 1, 5–17, Technical University of Vienna, Vienna, 2007.
[41] LS-SVMlab: a Matlab/C Toolbox for Least Squares Support Vector Machines, 2002. Available at http://www.esat.kuleuven.ac.be/sista/lssvmlab/.
[42] Matlab 7.3.0. MathWorks, Inc., Natick, MA, 2006.
[43] OPS® Toolbox routines for Matlab. Available online at: http://lqta.iqm.unicamp.br.
[44] S. Wold, L. Eriksson, in: H. van de Waterbeemd (Ed.), Chemometric Methods in Molecular Design, VCH, Weinheim, 1995, p. 309.
[45] Pirouette 3.11, Infometrix, Inc., Woodinville, WA, 2003.
[46] M. Abramowitz, I.A. Stegun (Eds.), Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables, Dover Publications, New York, NY, 1965.
[47] M. Jalali-Heravi, M. Asadollahi-Baboli, QSAR Comb. Sci. 27 (2008) 750–757.
[48] L. Eriksson, J. Jaworska, A.P. Worth, M.T.D. Cronin, R.M. McDowell, P. Gramatica, Environ. Health Perspect. 111 (2003) 1361–1375.
[49] I.S. Han, Ind. Eng. Chem. Res. 45 (2006) 670–680.
[50] S.S. Yang, W.C. Lu, N.N. Chen, Q.N. Hu, J. Mol. Struct., Theochem 719 (2005) 119–127.
[51] A.A. Ghaibeh, M. Sasaki, H. Chuman, Proceedings of the 2006 IEEE Symposium on Computational Intelligence in Bioinformatics Computational Biology, Toronto, ON, 28–29 September 2006, IEEE, Piscataway, NJ, 2006, pp. 1–6.
[52] B. Bhhatarai, R. Garg, Bioorg. Med. Chem. 13 (2005) 4078–4084.
[53] M. Sanczes, S. Krauchenco, N.H. Martins, A. Gustchina, A. Wlodawer, I. Polikarpov, J. Mol. Biol. 369 (2007) 1029–1040.
[54] O. Ivanciuc, in: K.B. Lipkowitz, T.R. Cundari (Eds.), Reviews in Computational Chemistry, Vol. 23, Wiley-VCH, Weinheim, 2007, p. 291.
[55] QuickCalcs Online Calculator for Scientists, GraphPad Software, Inc., La Jolla, CA, 2005. http://www.graphpad.com/quickcalcs/DistMenu.cfm [last access on February 4, 2009].