J Comput Aided Mol Des (2007) 21:251–267
DOI 10.1007/s10822-007-9112-4

ORIGINAL PAPER

Optimization of biaryl piperidine and 4-amino-2-biarylurea MCH1 receptor antagonists using QSAR modeling, classification techniques and virtual screening

Georgia Melagraki · Antreas Afantitis · Haralambos Sarimveis ·
Panayiotis A. Koutentis · John Markopoulos ·
Olga Igglessi-Markopoulou

Received: 23 June 2006 / Accepted: 6 February 2007 / Published online: 22 March 2007

Abstract This paper presents the results of an optimization study on biaryl piperidine and 4-amino-2-biarylurea MCH1 receptor antagonists, which was accomplished by using quantitative-structure activity relationships (QSARs), classification and virtual screening techniques. First, a linear QSAR model was developed using Multiple Linear Regression (MLR) Analysis, while the Elimination Selection-Stepwise Regression (ES-SWR) method was adopted for selecting the most suitable input variables. The predictive activity of the model was evaluated using an external validation set and the Y-randomization technique. Based on the selected descriptors, the Support Vector Machines (SVM) classification technique was utilized to classify data into two categories: ''actives'' or ''non-actives''. Several attempts were made to optimize the scaffold of most potent compounds by inducing various structural modifications. Potential derivatives with improved activities were identified, as they were classified ''actives'' by the SVM classifier. Their activities were estimated using the produced MLR model. A detailed analysis on the model applicability domain defined the compounds, whose estimations can be accepted with confidence.

Keywords MCH1R · QSAR · Classification · SVM · Virtual screening

Introduction

Melanin-concentrating hormone (MCH) is a cyclic nonadecapeptide which is expressed in the brain of all vertebrates. It has been demonstrated that MCH is involved in feeding and body weight regulation. MCH stimulates food intake in rodents and chronic administration leads to increased body weight. Animals that lack the gene encoding MCH receptor are hypophagic, lean and maintain elevated metabolic rates [1–3].

Two G-protein coupled receptors (GPCRs) have been identified for MCH, namely MCH1R and MCH2R. MCH1R is present in rodents and high mammalian species while MCH2R is expressed only in ferrets, dogs, rhesus monkeys and humans. The pharmacological role of MCH2R in metabolic homeostasis is still undefined whereas the critical role of MCH1R in the regulation of food intake and energy homeostasis has been extensively studied [4–6].

MCH1R has been identified as a key target in MCH regulation, as small molecule antagonists of MCH1R have demonstrated activity in vivo. MCH1R antagonists are potentially interesting agents for treatment of metabolic or obesity-related disorders. This fact has prompted different research groups to design and synthesize MCH1R antagonists which demonstrate in vivo efficiency in the therapy of obesity [7–9]. Alternative techniques could help to decrease the number of animals sacrificed during in vivo

G. Melagraki · A. Afantitis · H. Sarimveis (✉) ·
O. Igglessi-Markopoulou
School of Chemical Engineering, National Technical University
of Athens, Athens, Greece
e-mail: hsarimv@central.ntua.gr

A. Afantitis
Department of ChemoInformatics, NovaMechanics Ltd,
Larnaca, Cyprus

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

J. Markopoulos
Department of Chemistry, University of Athens, Athens, Greece

testing. In this sense QSAR, virtual screening and pattern recognition could be very useful and promising methodologies. To the best of our knowledge, these techniques have never been attempted so far in the general field of MCH1R. It is thus of particular interest to investigate this possibility.

In this paper we show that QSAR modeling, classification and virtual screening can contribute greatly to the modeling, design and optimization of MCH1R antagonists. The first major result is the development of a QSAR model involving seven descriptors that is able to predict successfully MCH1R binding affinity. The model was generated using a database consisting of a series of 63 MCH1 receptor antagonists including biaryl piperidine- and 4-amino-2-biarylurea-based derivatives [10, 11]. About 69 physicochemical and topological descriptors were examined in terms of their efficacy to determine and predict the activity of the investigated derivatives. The effects of various structural modifications on biological activity were investigated next, within a classification pattern. In particular, the popular SVM classification methodology was utilized to afford novel active patterns. Biological activities of novel structures were estimated using the new QSAR model, while the detection of its domain of applicability defined the compounds whose estimations can be accepted with confidence.

Materials and methods

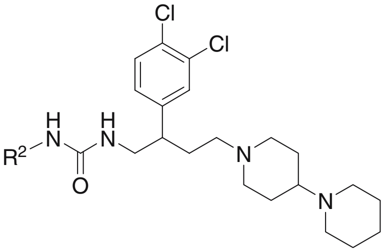
Data set

The database consists of 63 recently discovered biaryl piperidine and 4-amino-2-biarylbutylureas (Tables 1–3) [10, 11]. In order to model and predict the binding affinity of MCH receptor antagonists, 69 physicochemical constants, topological and structural descriptors (Table 4) were considered as possible input candidates to the model. Before the calculation of the descriptors, all structures were fully optimized using CS Mechanics and more specifically MM2 force fields and the Truncated-Newton-Raphson optimizer, which provide a balance between speed and accuracy (Chemoffice Manual). Before calculating the HOMO and LUMO Energies (eV) all the structures were additionally fully optimized using the AM1 basis set. All the descriptors were calculated using ChemSar and Topix [12, 13].

Separation into a training and a validation set

The separation of the dataset into training and validation sets was performed according to the popular Kennard and Stones algorithm [14]. 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.,

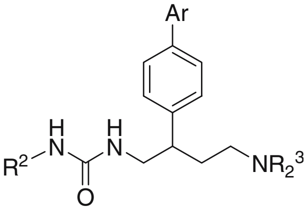
Table 1 Binding biological data of the 1-[4-(1,4′-bipiperidin-1′-yl)-2-(3,4-dichlorophenyl)butyl]-3-arylurea derivatives. Training and test data



ID	R^2	K_i (nM) observed	Log ($1/K_i$) observed	Training data log ($1/K_i$) predicted	Test data log ($1/K_i$) predicted
1[b]	3,5-$Cl_2C_6H_3$	135	–2.1303		–3.7673
2[b]	3,4-$Cl_2C_6H_3$	300	–2.4771		–4.1723
3[b]	3,4-$F_2C_6H_3$	240	–2.3802		–3.0638
4[b]	3-Cl-4-FC_6H_3	100	–2.0000		–3.2225
5	C_6H_5	18,500	–4.2672	–3.9305	
6	3-$MeOC_6H_4$	11,000	–4.0414	–3.4531	
7[b]	3-MeC_6H_4	6,800	–3.8325		–4.1118
8[b]	3-ClC_6H_4	1,641	–3.2151		–3.7187
9	3-$CF_3C_6H_4$	205	–2.3118	–3.4331	
10	4-$CF_3C_6H_4$	1,080	–3.0334	–3.7417	
11	2-$CF_3C_6H_4$	22,500	–4.3522	–3.7048	
12	3,5-$(CF_3)_2C_6H_3$	1,223	–3.0874	–2.9427	

[b] Compound included in the test set

Table 2 Binding biological data of 1-[4-amino-2-(biaryl-4-yl)butyl]-3-arylurea derivatives. Training and test data



ID	Ar	R^2	NR_2^3	K_i (nM) observed	Log $(1/K_i)$ observed	Training data log $(1/K_i)$ predicted	Test data log $(1/K_i)$ predicted
13[b]	C_6H_5	3,5-$Cl_2C_6H_3$	*cyclo*-Pentylamino	50	–1.6990		–2.3385
14[b]	3-ClC_6H_4	3,5-$Cl_2C_6H_3$	*cyclo*-Pentylamino	26	–1.4150		–2.0376
15	2-ClC_6H_4	3,5-$Cl_2C_6H_3$	*cyclo*-Pentylamino	130	–2.1140	–2.0740	
16	4-ClC_6H_4	3,5-$Cl_2C_6H_3$	*cyclo*-Pentylamino	280	–2.4472	–2.2969	
17[b]	3-FC_6H_4	3,5-$Cl_2C_6H_3$	*cyclo*-Pentylamino	37	–1.5682		–1.4960
18	3-$CF_3OC_6H_4$	3,5-$Cl_2C_6H_3$	*cyclo*-Pentylamino	20	–1.3010	–0.5212	
19	3-NCC_6H_4	3,5-$Cl_2C_6H_3$	*cyclo*-Pentylamino	3	–0.4771	–1.2844	
20[b]	4-NCC_6H_4	3,5-$Cl_2C_6H_3$	*cyclo*-Pentylamino	46	–1.6628		–1.2995
21[b]	3-Pyridyl	3,5-$Cl_2C_6H_3$	*cyclo*-Pentylamino	24	–1.3802		–1.6171
22[b]	4-Pyridyl	3,5-$Cl_2C_6H_3$	*cyclo*-Pentylamino	122	–2.0864		–2.1144
23[b]	3-NCC_6H_4	3,5-$Cl_2C_6H_3$	Methylamino	2.6	–0.4150		–0.5162
24[b]	3-NCC_6H_4	3,5-$Cl_2C_6H_3$	Dimethylamino	4.7	–0.6721		–0.1203
25[b]	3-NCC_6H_4	3,5-$Cl_2C_6H_3$	Ethylamino	4.3	–0.6335		–0.2554
26	3-NCC_6H_4	3,5-$Cl_2C_6H_3$	Isopropylamino	5.6	–0.7482	0.0417	
27	3-NCC_6H_4	3,5-$Cl_2C_6H_3$	Piperidin-1-yl	5.4	–0.7324	–0.8475	
28	3-NCC_6H_4	C_6H_5	Dimethylamino	2.4	–0.3802	–0.9114	
29	3-NCC_6H_4	3-ClC_6H_4	Dimethylamino	1.1	–0.04140	–0.4782	
30[b]	3-NCC_6H_4	4-ClC_6H_4	Dimethylamino	9.2	–0.9638		–0.7859
31[b]	3-NCC_6H_4	3-FC_6H_4	Dimethylamino	1.2	–0.0792		–0.3362
32[b]	3-NCC_6H_4	4-FC_6H_4	Dimethylamino	7.4	–0.8692		–0.6100
33[b]	3-NCC_6H_4	3-Cl-4-FC_6H_3	Dimethylamino	0.98	0.0088		0.1711
34[b]	3-NCC_6H_4	3,4-$F_2C_6H_3$	Dimethylamino	0.84	0.0757		0.2678
35	3-NCC_6H_4	3,5-$F_2C_6H_3$	Dimethylamino	0.88	0.0555	0.1702	

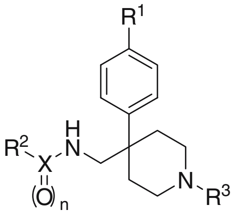
[b] Compound included in the test set

the Euclidean distance. These two samples are removed from the original data set and placed into the calibration data set. This procedure is repeated until the desired number of samples has been reached in the calibration set. The advantages of this algorithm are that the calibration samples map the measured region of the input variable space completely with respect to the induced metric and that the test samples all fall inside the measured region. According to Tropsha [15] and Wu [16], the Kennard and Stones algorithm is one of the best ways to build training and test sets.

MLR model development-variable selection

Our first objective was to determine the best variables which produce the most significant linear QSAR models

Table 3 Binding biological data of aryl and biaryl piperidine analogs. Training and test data



ID	R^1	$R^2X(O)_n$	R^3	K_i (nM) observed	Log (1/K_i) observed	Training data log (1/K_i) predicted	Test data log (1/K_i) predicted
36	3-Pyridyl	3,5-$Cl_2C_6H_3NHCO$	H	39	–1.5911	–1.4857	
37[b]	3-Pyridyl	3,5-$Cl_2C_6H_3NHCO$	$MeSO_2$	2.2	–0.3424		–0.4954
38	3-Pyridyl	3,5-$Cl_2C_6H_3NHCO$	Me	3.1	–0.4914	–0.9034	
39	3-Pyridyl	3,5-$Cl_2C_6H_3NHCO$	Me_2NSO_2	84	–1.9243	–1.8360	
40	Me	3,5-$Cl_2C_6H_3NHCO$	Me	300	–2.4771	–2.1551	
41	5-Indolyl	3,5-$Cl_2C_6H_3NHCO$	Me	1,188	–3.0748	–2.7598	
42	3-Cl-C_6H_4	3,5-$Cl_2C_6H_3NHCO$	Me	2.6	–0.4150	–1.3040	
43	3-AcHN-C_6H_4	3,5-$Cl_2C_6H_3NHCO$	Me	2.1	–0.3222	–0.8970	
44[b]	3-OHN-C_6H_4	3,5-$Cl_2C_6H_3NHCO$	Me	5.5	–0.7404		–0.7806
45[b]	3-NCC_6H_4	3,5-$Cl_2C_6H_3NHCO$	Me	1.4	–0.1461		–0.6108
46	3-NCC_6H_4	3,5-$Cl_2C_6H_3NHCO$	Et	2.6	–0.4150	–0.5641	
47	3-NCC_6H_4	3,5-$Cl_2C_6H_3NHCO$	*n*-Pr	0.31	0.5086	–0.2618	
48	3-NCC_6H_4	3,5-$Cl_2C_6H_3NHCO$	*n*-Bu	11	–1.0414	–0.1373	
49[b]	3-NCC_6H_4	3,5-$Cl_2C_6H_3NHCO$	$MeOCH_2CH_2$	0.41	0.3872		–0.0616
50[b]	3-NCC_6H_4	3,5-$Cl_2C_6H_3NHCO$	*i*-Pr	0.45	0.3468		0.3092
51	3-NCC_6H_4	3,5-$Cl_2C_6H_3NHCO$	*sec*-Bu	13	–1.1139	–0.0678	
52[b]	3-NCC_6H_4	3,5-$Cl_2C_6H_3NHCO$	*cyclo*-Propylmethyl	0.17	0.7696		–0.2959
53	3-NCC_6H_4	3,5-$Cl_2C_6H_3NHCO$	*cyclo*-Butyl	11	–1.0414	–1.1280	
54	3-NCC_6H_4	3,5-$Cl_2C_6H_3NHCO$	*cyclo*-Pentyl	0.63	0.2006	–0.6877	
55	3-NCC_6H_4	3,5-$Cl_2C_6H_3NHCO$	*cyclo*-Hexyl	1.2	–0.0792	–0.5842	
56[b]	3-NCC_6H_4	C_6H_5NHCO	Me	2.2	–0.3424		–1.2445
57	3-NCC_6H_4	3-CF_3-4-ClC_6H_3NHCO	Me	0.98	0.0088	–0.5199	
58	3-NCC_6H_4	3,4-$F_2C_6H_3NHCO$	Me	1.4	–0.1461	–0.2917	
59	3-NCC_6H_4	3-NCC_6H_4NHCO	Me	1.5	–0.1761	–0.7080	
60	3-NCC_6H_4	2,5-$Cl_2C_6H_3NHCO$	Me	140	–2.1461	–0.8839	
61[b]	3-NCC_6H_4	3,5-$Cl_2C_6H_3CH_2CO$	Me	164	–2.2148		–1.9125
62	3-NCC_6H_4	3,5-$Cl_2C_6H_3CO$	Me	155	–2.1903	–1.1466	
63	3-NCC_6H_4	3,5-$Cl_2C_6H_3SO_2$	Me	281	–2.4487	–1.9255	

[b] Compound included in the test set, * outlier

linking the structure of compounds with their binding affinity. The ES-SWR algorithm was used on the training data set to select the most appropriate descriptors. ES-SWR is a popular stepwise technique [17] that combines Forward Selection (FS-SWR) and Backward Elimination (BE-SWR).

Model validation

The accuracy of the proposed MLR model was illustrated using validation through an external test set and Y-randomization. The leave-one-out and leave-five-out cross-validation procedures were used to illustrate the robustness

Table 4 Calculated descriptors

ID	Description	Notation	ID	Description	Notation
1	Molar Refractivity	MR	2	Diameter	Diam
3	Partition Coefficient (Octanol Water)	ClogP	4	Molecular Topological Index	TIndx
5	Principal Moment of Inertia Z	PMIZ	6	Number of Rotatable Bonds	NRBo
7	Principal Moment of Inertia Y	PMIY	8	Polar Surface Area	PSAr
9	Principal Moment of Inertia X	PMIX	10	Radius	Rad
11	LUMO Energy	LUMO	12	Shape attribute	ShpA
13	HOMO Energy	HOMO	14	Shape coefficient	ShpC
15	Balaban Index	BIndx	16	Sum of Valence Degrees	SVDe
17	Cluster Count	ClsC	18	Total Connectivity	TCon
19	Wiener Index	WIndx	20	Total Valence Connectivity	TVCon
21	DistEqTotal	DistEqTotal	22	Randic 0	Chi0
23	Randic 1	Chi1	24	Randic 2	Chi2
25	Randic 3	Chi3	26	Randic 4	Chi4
27	Randic Information 0	ChiInf0	28	Randic Information 1	ChiInf1
29	Randic Information 2	ChiInf2	30	Randic Information 3	ChiInf3
31	Randic Information 4	ChiInf4	32	Molecular Weight	MW
33	Randic Mod	ChiMod	34	Xu1	Xu1
35	Xu2	Xu2	36	Xu3	Xu3
37	Balaban Topological	TopoJ	38	Number of Branches	NBranch
39	Number of Rings	NRings	40	Wiener Dim	Wiener Dim
41	Bertz	Bertz	42	AtomCompMean	AtomCompMean
43	AtomCompTot	AtomCompTot	44	Zagreb1	Zagreb1
45	Zagreb2	Zagreb2	46	Kappa1	Kappa1
47	Kappa2	Kappa2	48	Kappa3	Kappa3
49	Wiener Distance	WienerDistCode	50	Polarity	Polarity
51	DistEqMean	DistEqMean	52	Quadratic	Quadr
53	InfMagnitDistTot	InfMagnitDistTot	54	ScHultz	ScHultz
55	Gordon	Gordon	56	Kier-Hall 0	Ki0
57	Kier-Hall 1	Ki1	58	Kier-Hall 2	Ki2
59	Kier-Hall 3	Ki3	60	Kier-Hall 4	Ki4
61	Kier-Hall Information 0	KiInf0	62	Kier-Hall Information 1	KiInf1
63	Kier-Hall Information 2	KiInf2	64	Kier-Hall Information 3	KiInf3
65	Kier-Hall Information 4	KiInf4	66	Randic Cluster 3	ChiCl3
67	Randic Cluster 4	ChiCl4	68	Wiener Information	InfWiener
69	Wiener Index	WIndx			

of the MLR modeling technique in our particular QSAR study.

Cross-validation test

Cross-validation is popular technique used to explore the reliability of statistical models. Based on this technique, a number of modified data sets are created by deleting in each case one or a small group (leave-some-out) of objects. For each data set, an input–output model is developed, based on the utilized modeling technique. The model is evaluated by measuring its accuracy in predicting the responses of the remaining data (the ones that have not been utilized in the development of the model) [18].

Validation through the external validation set

According to Tropsha' group [15, 19] a QSAR model is considered predictive, if the following conditions are satisfied:

$$R^2_{\text{pred}} > 0.6 \tag{1}$$

$$\frac{(R^2 - R^2_{\text{o}})}{R^2} \text{ or } \frac{(R^2 - R'^2_{\text{o}})}{R^2} \text{ is less than } 0.1 \tag{2}$$

$$k \text{ or } k' \text{ is close to } 1. \tag{3}$$

In Eqs. 2 and 3, R^2 is the coefficient of determination between experimental values and model prediction on the training 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 the opposite (regression of the predicted activities against observed activities). The definitions are presented clearly in [20] and are not repeated here for brevity.

Y-randomization test

This technique ensures the robustness of a QSAR model [15, 21]. The dependent variable vector (biological action) is randomly shuffled and a new QSAR model is developed, using the given modeling algorithm. The procedure is repeated several times and the new QSAR models are expected to have low R^2 and Q^2 values. If the opposite happens then an acceptable QSAR model cannot be obtained for the specific modeling method and data.

Defining model applicability domain

In order for a QSAR model to be used for screening new compounds, its domain of application [15, 20] must be defined and predictions for only those compounds that fall into this domain may be considered reliable. *Extent of Extrapolation* [15] is one simple approach to define the applicability of the domain. It is based on the calculation of the leverage h_i [22] for each chemical, where the QSAR model is used to predict its activity

$$h_i = x_i(X^T X)^{-1} x_i^T. \tag{4}$$

In Eq. 4 x_i is the row vector containing the k model parameters of the query compound and X is the $n \times k$ matrix containing the k model parameters for each one of the n training compounds. A leverage value greater than $3k/n$ is considered large. It means that the predicted response is the result of a substantial extrapolation of the model and may be not reliable.

Support vector machines

The Support Vector Machine (SVM) method is a new and very promising supervised machine learning technique, originally developed by Vapnik and co-workers while working on Structural Risk Minimization [23–25]. The SVM method is quite different from empirical risk minimization algorithms and is gaining popularity due to the many attractive features it possesses and its promising empirical performance. SVM is a classification approach that has been suggested as being particularly appropriate for chemical applications. In particular, the popular Library for Support Vector Machines (LIBSVM) [26] was utilized in this work. A detailed presentation of the theory behind the SVM technique can be found in several books and tutorials [27]. Here we briefly summarize the main principles of SVMs, when they are used for classification purposes.

A typical application of the SVM technology in chemoinformatics consists of defining two classes of molecules, determining a set of descriptors that characterize each molecule and using the SVM algorithm to develop a classification model. If we assume that a number of training compounds have been classified using experimental data into an active class or an inactive class and that a set of significant descriptors has been calculated for each training compound, the procedure for developing an SVM classification model can be summarized as follows: every molecule has its own image in the multidimensional space, where each dimension corresponds to a different descriptor. The coordinates of the image are obviously the values of the various descriptors. The SVM model seeks to find an optimal hyperplane that best separates the two sets of classes corresponding to the active and non-active compounds in the multidimensional space. There are numerous hyperplanes that may separate the data in this manner. The optimal hyperplane is the one which maximizes the margin, defined as the closest distance from any point to the separating hyperplane. The points used to define the optimal hyperplane are often a small fraction of the entire data and, as such, they allow a SVM model to be less prone to overtraining while maintaining an excellent degree of generalizability. Thus, the produced SVM model can be used to classify other than the training molecules as active or inactive. The predicted class of a molecule that is not included in the training set depends on which side of the separating hyperplane the image of the molecule is located.

Results and discussion

First, the data set of 63 derivatives was partitioned into a training set of 35 compounds, and a validation set of 28 compounds according to the Kennard and Stones [14] algorithm. The algorithm was applied on the complete database consisting of all 69 available descriptors. The validation examples are marked with [b] in Tables 1–3. The validation data were not involved by any means in the process of selecting the most appropriate descriptors or in the development of the QSAR model. They were considered as a completely unknown external set of data, which was used only to test the accuracy of the produced model.

The MLR QSAR model was thus developed by applying the ES-SWR algorithm on the set of training data. The result was the following 8-parameter (seven descriptors and the intercept) equation:

$$\begin{aligned}\log(1/K_i) = &- 19.3\,(t\text{ - value}: -4.68)\\ &- 1.87 * \text{LUMO}\,(t\text{ - value}: -2.64)\\ &+ 0.516 * \text{CLogP}\,(t\text{ - value}: 3.69)\\ &+ 0.0253 * \text{PSAr}\,(t\text{ - value}: 1.81)\\ &+ 2.85 * \text{KiInf2}\,(t\text{ - value}: 2.78)\\ &- 1.95 * \text{Ki3}\,(t\text{ - value}: -6.40)\\ &+ 0.142 * \text{Xu3}\,(t\text{ - value}: 5.47)\\ &- 3.99 * \text{ChiCl4}\,(t\text{ - value}: -2.95)\end{aligned} \quad (5)$$

$n = 35$; $R^2 = 0.78$; $R^2_{\text{adj}} = 0.72$; $F = 13.72$; RMSE = 0.6256; $Q^2 = 0.65$; $S_{\text{PRESS}} = 0.7959$.

The above equation shows that the most significant descriptors according to the ES-SWR algorithm are Lipophilicity (ClogP), LUMO energy, PSAr, Kier&Hall information index order 2 (KiInf2), Kier&Hall index order 3 (Ki3), Xu3 index and Chicluster4 (ChiCl4). Table 5 presents the correlation matrix and the variance inflation factors (VIF), for the seven descriptors. These statistics indicate that the selected descriptors are not highly correlated. The chemical meaning of the seven descriptors is briefly described next.

Lipophilicity is known to be important for absorption, permeability, and in vivo distribution of organic compounds [28] and has been used as a physicochemical descriptor in QSARs with great success. From the derived QSAR equation we can conclude that lipophilic groups favor the biological action under study.

Polar surface area (PSAr) is defined as the part of the surface area of the module associated with oxygens, nitrogens, sulfurs and the hydrogens bonded to any of these atoms. PSAr has proven to be a very useful parameter for QSAR studies [17].

LUMO energy in particular has been identified as being of significant value to QSAR studies [17, 29]. Molecules with low LUMO values are more able to accept electrons than molecules with high LUMO energy values. The LUMO energy value is 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. From the derived QSAR equation we can conclude that EWGs favor the biological action under study.

In addition to the aforementioned indices, four topological indices were found to significantly influence the activity [17, 30]. Topological indices give information not only about the atomic constitution of a compound but also about the presence and character of chemical bonds by which the atoms are connected to each other.

Equation 5 was used to predict the binding affinity for the validation examples. The results are presented in the last columns of Tables 1–3 and correspond to the following statistics: $R^2_{\text{pred}} = 0.83$, RMSE = 0.6549. In Fig. 1, the experimental vs. predicted values are plotted for both the training and validation sets. The results illustrate that the linear MLR technique combined with a successful variable selection procedure are adequate to generate an efficient QSAR model for predicting the binding affinity of different compounds.

The proposed model (Eq. 5) passed all the tests related to the predictive ability (Eqs. 1–3)

$$R^2_{\text{pred}} = 0.83 > 0.6$$

$$\frac{(R^2 - R'^2_{\text{o}})}{R^2} = -0.22 < 0.1, \quad k' = 1.14.$$

For a more exhaustive testing of the model building technique that was followed in this work, the LOO and L5O cross-validation techniques were applied on the training set of compounds. The L5O method was imple-

Table 5 Correlation matrix and VIF values for the seven selected descriptors

	LUMO	ClogP	PSAr	KiInf2	Ki3	Xu3	ChiCl4	VIF*
LUMO	1							1.4
ClogP	–0.011	1						1.9
PSAr	–0.442	–0.307	1					1.9
KiInf2	0.012	0.059	–0.210	1				4.4
Ki3	–0.095	0.441	–0.016	0.669	1			6.7
Xu3	0.184	0.344	0.272	0.206	0.706	1		4.0
ChiCl4	0.208	0.214	–0.080	0.566	0.507	0.478	1	2.3

*VIF less than 10 indicates that the model contains no multicollinearity

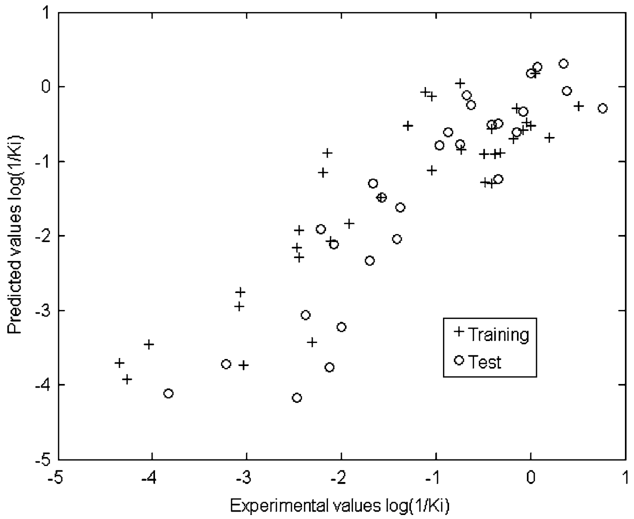


Fig. 1 Predicted vs. experimental values for the training and test sets

mented by selecting randomly groups of five compounds from the available training data. Each group was left out and that group was predicted by the model developed from the remaining observations. 3000 random groups of five compounds were selected for the implementation of the L5O cross-validation test. It should be emphasized that the procedure for developing the QSAR models included the selection of the best descriptors. Therefore, each time one (LOO) or five (L5O) compounds were excluded from the training set, the modeling procedure selected the best descriptors and developed an MLR model based only on the remaining observations. The excluded compounds were not involved by any means in the development of the model. It was important that the model was stable to the inclusion–exclusion of compounds. The results produced by the LOO ($Q^2 = 0.65$) and the L5O ($Q^2_{\mathrm{L5O}} = 0.66$) cross-validation tests illustrated the validity of the modeling approach.

The model was further validated by applying the Y-randomization. Several random shuffles of the Y vector were performed and the low R^2 and Q^2 values that were obtained show that the good results in our original model are not due to a chance correlation or structural dependency of the training set. It should be noted that for each random permutation of the Y vector, the complete training procedure was followed for developing the new QSAR model, including the selection of the most appropriate descriptors. The results of the Y-randomization test are presented in Table 6.

The extrapolation method was applied to the compounds that constitute the validation set. The leverages for all 28 compounds were computed (Table 7). All 28 compounds in the test set fall inside the domain of the model (the warning leverage limit is $3k/n$ 3*8/35 = 0.686).

Table 6 R^2 and Q^2 values after several Y-randomization test

Iteration	R^2	Q^2
1	0.28	0.00
2	0.08	0.00
3	0.36	0.25
4	0.13	0.05
5	0.47	0.23
6	0.07	0.00
7	0.35	0.17
8	0.37	0.24
9	0.35	0.09
10	0.30	0.23

After the pre-selection of the descriptors the next step was to build the classification model by using SVM. The SVM classification approach has been suggested as being particularly appropriate for chemical applications and well-suited for virtual screening purposes. A successful SVM model is expected to efficiently discriminate between

Table 7 Leverages for the test set

Compound	Leverage
1	0.4167
2	0.3112
3	0.4059
4	0.4025
7	0.3897
8	0.4152
13	0.4847
14	0.4487
17	0.5390
20	0.3972
21	0.5416
22	0.5250
23	0.4962
24	0.5345
25	0.5074
30	0.5170
31	0.4842
32	0.4854
33	0.4386
34	0.4278
37	0.0635
44	0.6324
45	0.6295
49	0.5732
50	0.5127
52	0.5815
56	0.5249
61	0.4590

Warning leverage limit = 0.686

important and unimportant features. For constructing the SVM model, the LIBSVM package was used [26] after scaling both the training and validation data in the range [–1,+1]. The Kernel type that was adopted in the present work was the Radial Basis Function (RBF). The first task is the assignment of each compound to one class, namely ''active'' or ''non-active'' based on a cut-off value that was set to 60 for binding affinity K_i. For classification purposes, active compounds are assigned a +1 value whereas non-active compounds are assigned a –1 value. These classes are defined a priori by groups of objects in the training set belonging to these classes. The cost parameter C and the gamma parameter γ in the kernel function were optimized to achieve the best possible discrimination between classes. The optimized values obtained were $C = 10$ and $\gamma = 1$. The predictive ability of the model was tested on the validation set of compounds. The total accuracy of the SVM model was 91%, meaning that the model assigned the correct class to 91% of the compounds. Accuracy for the training and validations sets was 100 and 82%, respectively. The results for the validation set are listed in Table 8. The misclassified samples (marked with an asterisk) are clearly indicated.

Table 8 SVM classification results for the test set

Compound	Class	Predicted class
1	–1	–1
2	–1	–1
3	–1	–1
4	–1	–1
7	–1	–1
8	–1	–1
13*	+1	–1
14*	+1	–1
17*	+1	–1
20	+1	+1
21	+1	+1
22*	–1	+1
23	+1	+1
24	+1	+1
25	+1	+1
30	+1	+1
31	+1	+1
32	+1	+1
33	+1	+1
34	+1	+1
37*	+1	–1
44	+1	+1
45	+1	+1
49	+1	+1
50	+1	+1
52	+1	+1
56	+1	+1
61	–1	–1

*Misclassified compound

Our final objective was to be able to classify compounds that are not involved in the training procedure. These compounds were derived from virtual optimization of the lead compounds by insertions, substitutions, and deletions of pharmacophoric substituents of the main building block scaffolds. More specifically, based on the produced SVM classification model, a group of new derivatives, previously not tested for the specific biological action, was subjected to virtual screening. The aim was, starting from a primary hit and using both pharmacophore-based and substructure-based modifications to discover a structurally diverse set of potent leads [31, 32].

A variety of modifications of the initial compounds were introduced and the representative modifications that led to ''active'' compounds are shown in Tables 9–14. Biological activities of the compounds characterized as ''actives'' were estimated using the developed MLR equation. The activity values together with the leverages are shown in Tables 9–14. More precisely, the last column in Tables 9–14 shows the difference between the warning limit 0.686 and the leverage calculated for each compound. A negative value means that the respective compound falls outside the domain of applicability of the model. The initial study focused on the substitution of the biaryl moiety and indicated that the biphenyl analogs (id 1v–4v, bearing either 2° or 3° alkylamine side chains while falling well within the domain of applicability were not predicted to be significantly active compounds. The fused analogs (id 5v–9v however, gave good predicted activity but the most active structures (id 7v, 9v) fell outside (or marginally inside) of the applicability domain. The tetrahydroquinoline structure 6v was considered for further modification and the nitrogen heterocyclic ring was converted to the isomeric tetrahydroisoquinolines (id 10v and 11v). Structure id 10v (pred. 1.2616, domain 0.2071) which gave improved predicted activity and remained within the applicability domain was further modified by investigating the nature of the *N*-substitution (id 12v–18v). Generally the addition of larger branched alkyl chains helped improve the activity but several structures fell outside or marginally inside of the domain of applicability. This was in agreement with the model which emphasizes the lipophilicity descriptor. Interestingly, moving the alkyl groups from the ring nitrogen to the neighboring *peri* position (id 19v–27v) greatly improved the structures fit within the applicability domain and structure id 25v bearing an *iso*-butyl substituent showed good activity (1.6211) well within the domain (0.3642). Structure id 25v was therefore considered for further modification. *N*-Alkylation (id 28v) greatly improved the activity (3.7658) but the structure fell outside

Table 9 Virtual screening results (id 1v–11v)

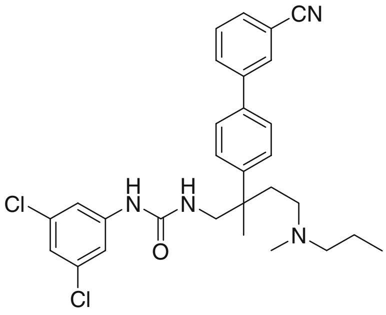
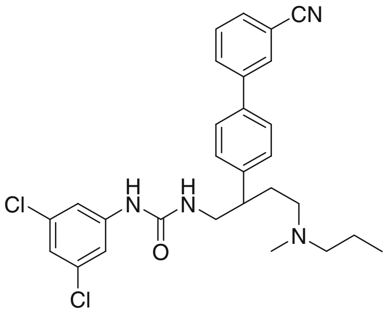
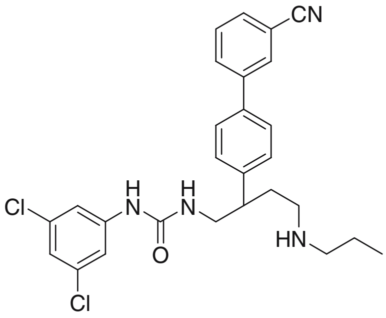
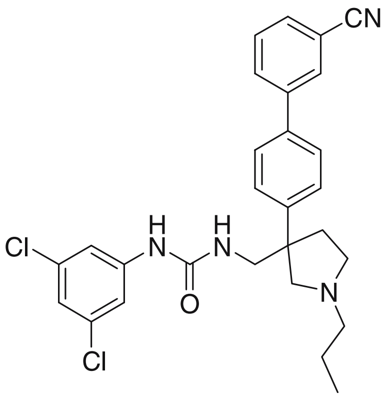
Id	Structure	Predicted activity	Limit-leverage
1v		–0.0632	0.5439
2v		0.2263	0.4835
3v		–0.0044	0.4921
4v		–0.6181	0.6390

Table 9 continued

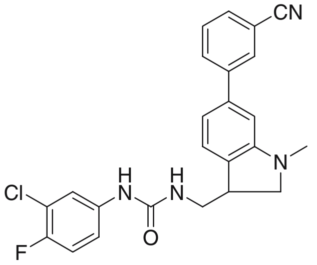
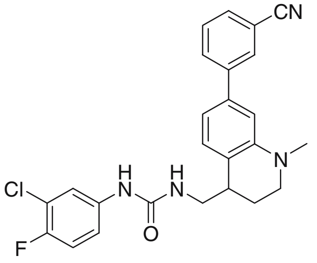
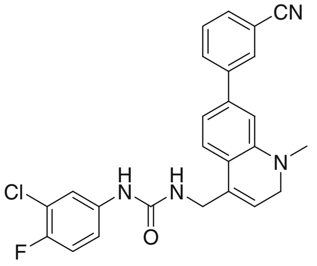
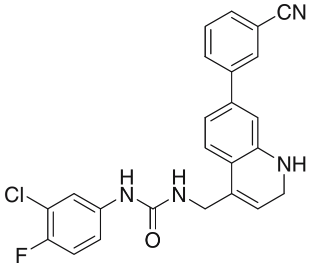
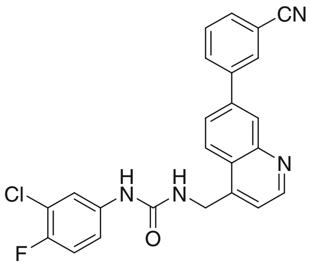
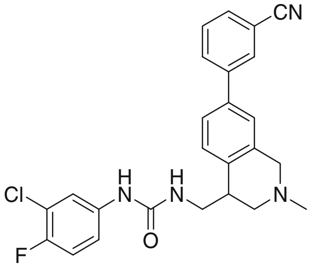
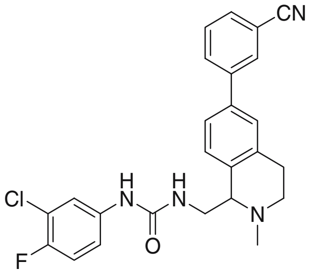
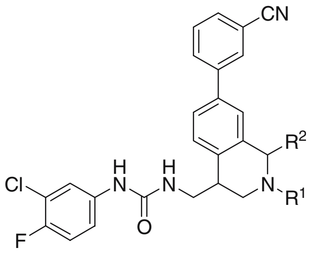
Id	Structure	Predicted activity	Limit-leverage
5v		1.0114	0.1900
6v		1.3332	0.1817
7v		2.3485	–0.0757
8v		0.9605	0.4548

Table 9 continued

Id	Structure	Predicted activity	Limit-leverage
9v		2.3206	0.0318
10v		1.2616	0.2071
11v		1.3288	0.1397

the applicability domain (–1.0082). Introducing a further degree of unsaturation (id 29v) however, significantly improved activity (2.3716) within the applicability domain (0.2389). The model predicts that lipophilicity and LUMO are important descriptors and as such structure id 29v was further modified by introducing fluorine substituents on the *iso*-butyl side chain (id 30v–32v). Structures id 31v and 32v were predicted to have very good activities (2.8676 and 2.6800, respectively) and were within the applicability domain. The urea moiety of structure id 32v was modified to afford *N*-methylurea, carbamate, carbonate or thiocarbamate structures (id 33v–37v) but this did not afford structures with improved activity within the desired domain of applicability. The phenyl urea substituent of structure id 29v was also investigated (id 38v–50v). The 3,4-difluorobenzene substitution pattern gave the best activity comfortably within the applicability domain. The phenyl substituent of the dihydroisoquinoline moiety of this structure (id 44v) was then modified (id 51v–57v) Structure id 51v showed excellent predicted activity (3.1284) within the domain of applicability (0.1105) and introduction of an additional fluorine on the *iso*-butyl group (id 58v) forced the structure outside the acceptable domain. In general it has been demonstrated that several potentially active structures can be predicted via virtual screening which fall within the models domain of applicability. The introduction of branched alkyl chains and also the use of fluorine substitution are in agreement with the descriptor model which showed a preference for increased lipophilicity and a more negative LUMO energy. Compounds id

Table 10 Modifications of 1,2,3,4-tetrahydroisoquinoline id 10v

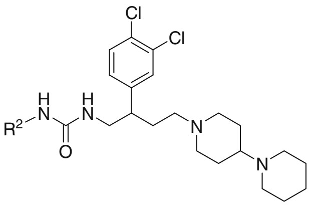
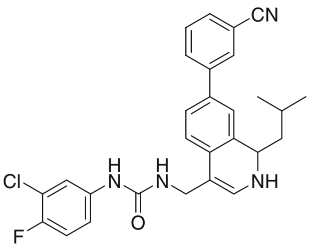


Id	R^1	R^2	Predicted activity	Limit-leverage
12v	Et	H	1.2009	0.2342
13v	n-Pr	H	1.4846	0.1507
14v	i-Pr	H	1.7995	–0.0442
15v	c-Pr	H	1.1890	0.1677
16v	n-Bu	H	1.7116	0.0121
17v	i-Bu	H	2.2567	–0.2531
18v	s-Bu	H	1.3205	0.0690
19v	H	Me	0.2551	0.5669
20v	H	Et	0.5929	0.5630
21v	H	n-Pr	1.1016	0.5126
22v	H	i-Pr	1.2560	0.4831
23v	H	c-Pr	0.5392	0.5088
24v	H	n-Bu	1.1302	0.4767
25v	H	i-Bu	1.6211	0.3642
26v	H	s-Bu	1.0449	0.4725
27v	H	t-Bu	0.6620	0.4373

31v, 32v, 33v, 47v, 51v, 53v are predicted to exhibit an increased biological activity and simultaneously fall inside the domain of applicability of the model.

Finally a cautionary note should be included dealing with the biological activity scales. While the data from the experimental and virtual studies has been recorded with the same units it must be noted that the predicted activities produced by the virtual model are significantly higher. It would be truly remarkable if the model was able to accurately predict such activities quantitatively but this is unlikely. The synthesis and study of these compounds would be required to truly validate the virtual model and as such is a worthy pursuit but this is outside the scope of this present paper. It must therefore be noted that the virtual screening study acts only as an aid in proposing structural modifications to assist ongoing SAR studies. The high biological activities predicted are only indicative of which structures should be targeted for synthesis on the basis that they meet or approach the optimal values for the chosen descriptors for the given model.

Table 11 Modifications of compound id 25v

Id	Structure	Predicted activity	Limit-leverage
28v		3.7658	–1.0082
29v		2.3716	0.2389

Conclusion

In the present study seven descriptors (namely LUMO energy, PSAr, ClogP and four topological descriptors: Ki3, KiInf2, ChiCl4, Xu3) were found to be important for describing biological activity of potent MCH receptor antagonists. The seven-descriptor set contains electronic, topological and physicochemical information about molecules, and describes and models successfully the binding affinity of these small molecules.

An SVM classifier was developed based on the partition of the initial dataset into training and validation compounds. The SVM model was then used to classify novel compounds that were derived by inducing structural modification to the initial compounds of the database. Biological activities of novel compounds were estimated by the produced MLR model. The detailed validation procedure that was followed (separation of the data into two independent sets, cross-validation, Y-randomization) illustrated the accuracy and robustness of the produced model not only by calculating its fitness on the training data, but also by testing its predicting ability. The applicability domain served as a valuable tool to filter out ''dissimilar'' compounds.

Due to its predictive ability, the proposed model could be a useful aid to the costly and time consuming experiments for determining binding affinity of the MCH receptor antagonists.

Table 12 Modifications of the 1,2-dihydroisoquinoline id 29v (introduction of fluorine substituents and modification of the urea moiety)

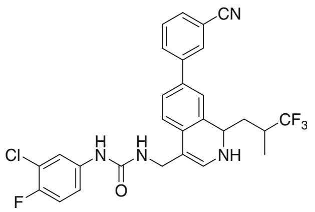
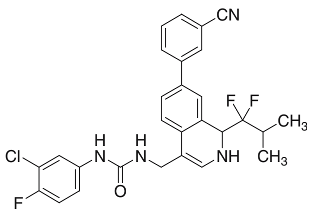
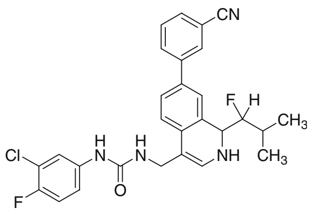
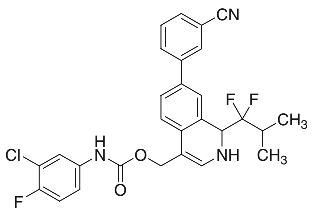
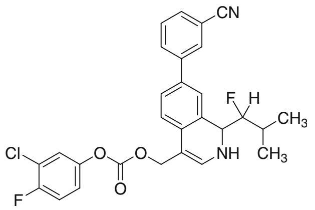
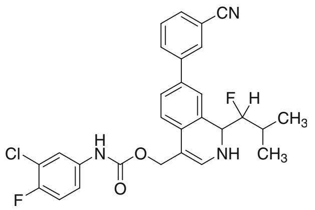
Id	Structure	Predicted Activity	limit-leverage
30v		1.8484	0.2434
31v	 CN F F CH$_3$ CH$_3$ Cl F H N H N O NH	2.8676	0.1321
32v	 CN F H CH$_3$ CH$_3$ Cl F H N H N O NH	2.6800	0.1784
33v	 CN F F CH$_3$ CH$_3$ Cl F H N O O NH	2.8841	0.1338
34v	 CN F H CH$_3$ CH$_3$ Cl F O O O NH	2.9348	–0.0314
35v	 CN F H CH$_3$ CH$_3$ Cl F H N O O NH	2.6045	0.2058

Table 12 continued

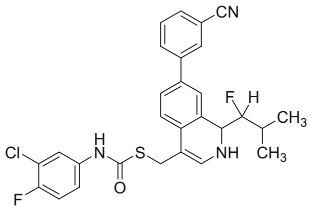
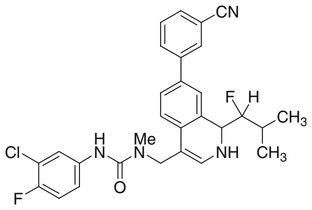
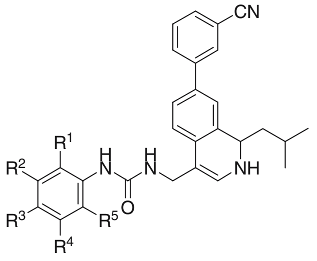
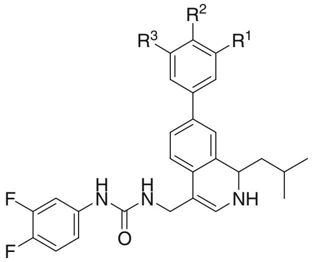
Id	Structure	Predicted Activity	limit-leverage
36v	 CN F H CH3 H N S Cl NH CH3 O F	1.9578	0.1470
37v	 CN F H CH3 H Me N N Cl NH CH3 O F	3.0980	–0.3189

Table 13 Modifications of the 1,2-dihydroisoquinoline id 29v (variation of the urea phenyl substituents)



Id	R^1	R^2	R^3	R^4	R^5	Predicted activity	Limit-leverage
38v	H	F	Cl	H	H	2.3194	0.2490
39v	H	Cl	Cl	H	H	1.4953	0.4141
40v	Cl	H	Cl	H	H	1.6983	0.3908
41v	F	H	Cl	H	H	2.2156	0.2723
42v	F	H	F	H	H	2.1562	0.3046
43v	Cl	H	F	H	H	2.0106	0.3211
44v	H	H	F	F	H	2.3314	0.2970
45v	F	H	H	F	H	2.1661	0.3023
46v	H	F	H	F	H	2.2490	0.3463
47v	F	F	H	H	H	2.4724	0.1615
48v	F	H	H	H	F	2.3694	0.1587
49v	F	F	H	H	F	3.7355	0.0225
50v	F	F	F	H	H	2.8719	0.0081

Table 14 Modifications of the (3,4-difluorophenyl)urea id 44v (variation of the isoquinoline aryl substituent)



Id	R^1	R^2	R^3	Predicted activity	Limit-leverage
51v	CN	F	H	3.1284	0.1105
52v	F	CN	H	3.3585	0.0123
53v	F	H	CN	2.8806	0.2040
54v	CF_3	H	H	2.1193	0.0523
55v	F	H	H	1.9889	0.1910
56v	F	F	H	2.4967	0.0855
57v	CN	Cl	H	2.9882	0.0906
58v	CN	F	H	3.6585	–0.1387

Acknowledgments G.M. thanks the Greek State Scholarship Foundation for a doctoral assistantship. A.A. wishes to thank Cyprus Research Promotion Foundation (Grant No. PENEK/ENISX/0603/05) and the Committee of Research of the National Technical University of Athens, Greece for a doctoral assistantship.

References

1. Kowalski TJ, Spar BD, Weig B, Farley C, Cook J, Ghibaudi L, Fried S, O'Neill K, Del Vecchio RA, McBriar M, Guzik H, Clader J, Hawes BE, Hwa J (2006) Eur J Pharmacol 535:182
2. McBriar M, Guzik H, Shapiro S, Paruchova J, Xu R, Palani A, Clader JW, Cox K, Greenlee WJ, Hawes BE, Kowalski TJ, O'Neill K, Spar BD, Weig B, Weston DJ, Farley C, Cook J (2006) J Med Chem 49:2294
3. (a) Palani A, Shapiro S, McBriar MD, Clader JW, Greenlee WJ, Spar B, Kowalski TJ, Farley C, Cook J, van Heek M, Weig B, O'Neill K, Graziano M, Hawes B (2005) J Med Chem 48:4746. (b) McBriar MD, Guzik H, Xu R, Paruchova J, Li S, Palani A, Clader JW, Greenlee WJ, Hawes BE, Kowalski TJ, O'Neill K, Spar B, Weig B (2005) J Med Chem 48:2274
4. Receveur JM, Bjurling E, Ulven T, Little PB, Norregaard PK, Hogberg T (2004) Bioorg Med Chem Lett 14:5075
5. Rowbottom MW, Vickers TD, Dyck B, Taminiya J, Zhang M, Zhao L, Grey J, Provencal D, Schwarz D, Heise CE, Mistry M, Fisher A, Dong T, Hu T, Saunders J, Goodfellow VS (2005) Bioorg Med Chem Lett 15:3439
6. Vasudevan A, Wodka D, Verzal MK, Souers AJ, Gao J, Brodjian S, Fry D, Dayton B, Marsh KC, Hernandez LE, Ogiela CA, Collins CA, Kym PR (2004) Bioorg Med Chem Lett 14:4879
7. (a) Xu R, Li S, Paruchova J, McBriar MD, Guzik H, Palani A, Clader JW, Cox K, Greenlee WJ, Hawes BE, Kowalski TJ, O'Neill K, Spar BD, Weig B, Weston DJ (2006) Bioorg Med Chem 14:3285. (b) Su J, McKittrick BA, Tang H, Czarniecki M, Greenlee WJ, Hawes BE, O'Neill K (2005) Bioorg Med Chem 13:1829
8. (a) Kanuma K, Omodera K, Nishiguchi M, Funakoshi T, Chaki S, Semple G, Tran T-A, Kramer B, Hsu D, Casper M, Thomsen B, Beeley N, Sekiguchi Y (2005) Bioorg Med Chem Lett 15:2565. (b) Kanuma K, Omodera K, Nishiguchi M, Funakoshi T, Chaki S, Semple G, Tran T-A, Kramer B, Hsu D, Casper M, Thomsen B, Sekiguchi Y (2005) Bioorg Med Chem Lett 15:3853. (c) Kanuma K, Omodera K, Nishiguchi M, Funakoshi T, Chaki S, Nagase Y, Iida I, Yamaguchi J-I, Semple G, Tran T-A, Sekiguchi Y (2006) Bioorg Med Chem 14:3307
9. Vasudevan A, Wodka D, Verzal MK, Souers AJ, Gao J, Brodjian S, Fry D, Dayton B, Marsh KC, Hernandez LE, Ogiela CA, Collins CA, Kym PR (2004) Bioorg Med Chem Lett 14:4879
10. Guo T, Hunter RC, Gu H, Shao Y, Rokosz LL, Stauffer TM, Hobbs DW (2005) Bioorg Med Chem Lett 15:3691
11. Guo T, Shao Y, Qian G, Rokosz LL, Stauffer TM, Hunter RC, Babu SD, Gu H, Hobbs DW (2005) Bioorg Med Chem Lett 15:3696
12. CambridgeSoft Corporation http://www.cambridgesoft.com
13. http://www.lohninger.com/topix.html
14. Kennard RW, Stone LA (1969) Technometrics 11:137
15. Tropsha A, Gramatica P, Gombar VK (2003) QSAR Comb Sci 22:69
16. Wu W, Walczak B, Massart DL, Heuerding S, Erni F, Last IR, Prebble KA (1996) Chemometr Intell Lab Syst 33:35
17. Todeschini R, Consonni V, Mannhold R, Kubinyi H, Timmerman H (Series Editor) (2000) Handbook of molecular descriptors. Wiley-VCH, Weinheim
18. (a) Efron B (1983) J Am Stat Assoc 78:316. (b) Osten DW (1998) J Chemom 2:39
19. Shen M, Beguin C, Golbraikh A, Stables J, Kohn H, Tropsha A (2004) J Med Chem 47:2356
20. Golbraikh A, Tropsha A (2002) J Mol Graph Mod 20:269

21. Wold S, Eriksson L (1995) In: Van de Waterbeemd H (ed) Chemometrics methods in molecular design, VCH Weinheim, Germany
22. Atkinson A (1985) Plots, transformations and regression. Clarendon Press, Oxford (UK)
23. Cortes C, Vapnik V (1995) Mach Learning 20:273
24. Jorissen RN, Gilson MK (2005) J Chem Inf Model 45:549
25. Wilton D, Willet P, Lawson K, Mullier G (2003) J Chem Inf Comput Sci 43:469
26. Chang CC, Lin CJ LIBSVM: http://www.csie.ntu.edu.tw/~cjlin/libsvm
27. Burges CJC (1998) Data Min Knowl Discov 2:127
28. Walters WPA, Murcko MA (1999) Curr Opin Chem Biol 3:384
29. Karelson M (2000) Molecular descriptors in QSAR/QSPR. Wiley, NY
30. Kier LB, Hall LB (1986) Molecular connectivity in structure activity analysis. Wiley, Chichester
31. Afantitis A, Melagraki G, Sarimveis H, Koutentis PA, Markopoulos J, Igglessi-Markopoulou O (2006) J Comput Aid Design 20:83
32. Afantitis A, Melagraki G, Sarimveis H, Koutentis PA, Markopoulos J, Igglessi-Markopoulou O (2006) Mol Div. 10:405