

## Investigation of substituent effect of 1-(3,3-diphenylpropyl)-piperidinyl phenylacetamides on CCR5 binding affinity using QSAR and virtual screening techniques

Antreas Afantitis<sup>a,b</sup>, Georgia Melagraki<sup>a</sup>, Haralambos Sarimveis<sup>a,\*</sup>, Panayiotis A. Koutentis<sup>c</sup>, John Markopoulos<sup>d</sup> & Olga Igglessi-Markopoulou<sup>a</sup>

<sup>a</sup>School of Chemical Engineering, National Technical University of Athens, Athens, Greece; <sup>b</sup>Department of ChemoInformatics, NovaMechanics Ltd, Larnaca, Cyprus; <sup>c</sup>Department of Chemistry, University of Cyprus, P.O. Box 20537 1678, Nicosia, Cyprus; <sup>d</sup>Department of Chemistry, University of Athens, Athens, Greece

Received 1 December 2005; accepted 26 January 2006  
© Springer 2006

**Key words:** CCR5, binding affinity, QSAR, virtual screening

### Summary

A linear quantitative–structure activity relationship model is developed in this work using Multiple Linear Regression Analysis as applied to a series of 51 1-(3,3-diphenylpropyl)-piperidinyl phenylacetamides derivatives with CCR5 binding affinity. For the selection of the best variables the Elimination Selection-Stepwise Regression Method (ES-SWR) is utilized. The predictive ability of the model is evaluated against a set of 13 compounds. Based on the produced QSAR model and an analysis on the domain of its applicability, the effects of various structural modifications on biological activity are investigated. The study leads to a number of guanidine derivatives with significantly improved predicted activities.

### Introduction

The chemokine receptor CCR5 is expressed on T-lymphocytes, monocytes, macrophages, dendritic cells, microglia and other cell types [1]. These receptors detect and respond to several chemokines principally “regulated on activation normal T-cell expressed and secreted” (RANTES) and macrophage inflammatory proteins (MIP) MIP-1 $\alpha$  and MIP-1 $\beta$ , resulting in the recruitment of cells of the immune system to sites of disease. CCR5 is also co-receptor for HIV-1 and other viruses, enabling these viruses to enter cells. Individuals, who are homozygous for 32-bp deletion in the gene encoding CCR5, whilst otherwise healthy, are strongly protected against infection

[2]. Many studies indicate different roles for CCR5 and its ligands in disorders such as rheumatoid arthritis [3], multiple sclerosis [4], transplant rejection [5] and inflammatory bowel disease [6]. These observations suggest that molecules that modulate the CCR5 receptor would have potential benefit in a wide range of diseases.

In the past, several attempts have been made to build QSAR models in the general field of CCR5 binding affinity. Debnath [7] presented predictive pharmacophore models for CCR5 antagonists using piperidine and piperazine derivatives. Song et al. [8] presented a 3D Quantitative Structure Activity Relationship (QSAR) study using piperidine derivatives. Xu et al. [9] using Molecular Docking and 3D QSAR presented a study based on 1-amino-2-phenyl-4-(piperidin-1-yl)-butane derivatives. Finally, Roy and Leonard presented two validated QSAR studies using

\*To whom correspondence should be addressed. E-mail: hsarimv@central.ntua.gr

substituted benzylpyrazole [10] and 3-(4-benzylpiperidin-1-yl)-*N*-phenylpropylamine [11] derivatives. With the latter derivatives [11] 3D-QSAR and more specifically Molecular Shape Analysis (MSA), Receptor Surface Analysis (RSA) and Molecular Field Analysis (MFA) were applied.

In this work, we selected a series of 51 1-(3,3-diphenylpropyl)-piperidinyl phenylacetamides [1] to investigate their role as CCR5 receptor modulators. Sixty-one physicochemical and topological descriptors were examined in terms of their efficacy to determinate and predict the activity of the investigated derivatives. The descriptors were calculated using Topix ([www.lohninger.com/topix.html](http://www.lohninger.com/topix.html)) and ChemSar which is included in the ChemOffice (Cambridge-Soft Corporation) suite of programs. Among them, the most statistically significant descriptors were selected, using a rigorous variable selection method. The result of this study was the development of a new linear QSAR model containing 7 descriptors. In order to validate the proposed methodology, we used two strategies: Y-randomization and external validation using division of the entire data set into training and test sets. Based on the new QSAR model and the detection of its domain of applicability, the effects of various structural modifications on the biological activity were investigated.

## Materials and methods

### Data set

In this QSAR study 52 biological data from Burrows's et al. [1] work were used. The biological activities of these 52 compounds were reported in the same papers [1]. In order to model and predict the specific activity (CCR5 binding affinity), 61 physicochemical constants, topological and structural descriptors (Table 1) were considered as possible input candidates to the model. All the descriptors were calculated using ChemSar and Topix.

### Stepwise multiple regression

The ES-SWR algorithm was used to select the most appropriate descriptors. ES-SWR is a popular stepwise technique [12] that combines

Forward Selection (FS-SWR) and Backward Elimination (BE-SWR). It is essentially a forward selection approach, but at each step it considers the possibility of deleting a variable as in the backward elimination approach, provided that the number of model variables is greater than two. The two basic elements of the ES-SWR method are described below in more details.

### Forward selection

The variable considered for inclusion at any step is the one yielding the largest single degree of freedom  $F$ -ratio among the variables that are eligible for inclusion. The variable is included only if the corresponding  $F$ -ratio is larger than a fixed value  $F_{in}$ . Consequently, at each step, the  $j$ th variable is added to a  $k$ -size model if

$$F_j = \max_j \left( \frac{RSS_k - RSS_{k+j}}{s_{k+j}^2} \right) > F_{in} \quad (1)$$

In the above in equality RSS is the *residual sum of squares* and  $s$  is the *mean square error*. The subscript  $k+j$  refers to quantities computed when the  $j$ th variable is added to the  $k$  variables that are already included in the model.

### Backward elimination

The variable considered for elimination at any step is the one yielding the minimum single degree of freedom  $F$ -ratio among the variables that are included in the model. The variable is eliminated only if the corresponding  $F$ -ratio does not exceed a specified value  $F_{out}$ . Consequently, at each step, the  $j$ th variable is eliminated from the  $k$ -size model if

$$F_j = \min_j \left( \frac{RSS_{k-j} - RSS_k}{s_k^2} \right) < F_{out} \quad (2)$$

The subscript  $k-j$  refers to quantities computed when the  $j$ th variable is eliminated from the  $k$  variables that have been included in the model so far.

### Cross-validation technique

The reliability of the proposed method was explored using the cross-validation method. Based on this technique, a number of modified data sets are created by deleting in each case one (LOO,

Table 1. Physicochemical constants, topological and structural 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	Connolly Accessible Area	SAS	12	Shape attribute	ShpA
13	Connolly Molecular Area	MS	14	Shape coefficient	ShpC
15	Total Energy	TotE	16	Sum of Valence Degrees	SVDe
17	LUMO Energy	LUMO	18	Total Connectivity	TCon
19	HOMO Energy	HOMO	20	Total Valence Connectivity	TVCon
21	Balaban Index	BIndx	22	Wiener Index	WIndx
23	Cluster Count	ClsC	24	Randic 0	Chi0
25	Randic 1	Chi1	26	Randic 2	Chi2
27	Randic 3	Chi3	28	Randic 4	Chi4
29	Randic Information 0	ChiInf0	30	Randic Information 1	ChiInf1
31	Randic Information 2	ChiInf2	32	Randic Information 3	ChiInf3
33	Randic Information 4	ChiInf4	34	Kier-Hall 0	Ki0
35	Randic Mod	ChiMod	36	Xu1	Xu1
37	Xu2	Xu2	38	Xu3	Xu3
39	Balaban Topological	TopoJ	40	Topological Radius	TopoRad
41	Topological Diameter	TopoDia	42	Number of Branches	NBranch
43	Number of Rings	Nrings	44	WienerDim	WienerDim
45	Bertz	Bertz	46	AtomCompMean	AtomCompMean
47	AtomCompTot	AtomCompTot	48	Zagreb1	Zagreb1
49	Zagreb2	Zagreb2	50	Quadratic	Quadr
51	ScHultz	ScHultz	52	Kappa1	Kappa1
53	Kappa3	Kappa3	54	Kappa2	Kappa2
55	Wiener Distance	WienerDistCode	56	Wiener Information	InfWiener
57	DistEqMean	DistEqMean	58	DistEqTotal	DistEqTotal
59	InfMagnitDistTot	InfMagnitDistTot	60	Polarity	Polarity
61	Gordon	Gordon			

leave one out) or a small group (leave some out) of objects [13–15], thus leading to the development of multiple input–output models. Each 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). In particular, the LOO procedure was utilized in this study, where the total number of produced models is equal to the number of available examples. More precisely, a different model is produced by deleting each time one object from the training set. Prediction error sum of squares (PRESS) is a standard index to measure the accuracy of a modeling method based on the cross-validation technique. Using the PRESS and SSY (sum of squares of deviations of the

experimental values from their mean) statistics, the  $R_{CV}^2$  and  $S_{PRESS}$  values can be easily calculated. The formulae that calculate the aforementioned statistics are presented below (Equations 3 and 4):

$$R_{CV}^2 = 1 - \frac{\text{PRESS}}{\text{SSY}} = 1 - \frac{\sum_{i=1}^n (Y_{\text{exp}} - Y_{\text{pred}})^2}{\sum_{i=1}^n (Y_{\text{exp}} - \bar{Y})^2} \quad (3)$$

$$S_{PRESS} = \sqrt{\frac{\text{PRESS}}{n}} \quad (4)$$

*Y-randomization test*

This technique ensures the robustness of a QSAR model [16, 17]. The dependent variable vector (biological action) is randomly shuffled and a new QSAR model is developed using the original independent variable matrix. The new QSAR models (after several repetitions) are expected to have low  $R^2$  and  $R_{CV}^2$  values. If the opposite happens then an acceptable QSAR model cannot be obtained for the specific modeling method and data.

*Estimation of the predictive ability of a QSAR model*

According to Tropsha [17] the predictive power of a QSAR model can be conveniently estimated by an external  $R_{CV,ext}^2$  (Equation 5).

$$R_{CV,ext}^2 = 1 - \frac{\sum_{i=1}^{test} (Y_{exp} - Y_{pred})^2}{\sum_{i=1}^{test} (Y_{exp} - \bar{Y}_{tr})^2} \quad (5)$$

where  $\bar{Y}_{tr}$  is the averaged value for the dependent variable for the training set.

Furthermore Tropsha's research group [17, 18] considered a QSAR model predictive, if the following conditions are satisfied:

$$R_{CV,ext}^2 > 0.5 \quad (6)$$

$$R^2 > 0.6 \quad (7)$$

$$\frac{(R^2 - R_o^2)}{R^2} < 0.1 \quad \text{or} \quad \frac{(R^2 - R_o'^2)}{R^2} < 0.1 \quad (8)$$

$$0.85 \leq k \leq 1.15 \quad \text{or} \quad 0.85 \leq k' \leq 1.15 \quad (9)$$

Mathematical definitions of  $R_o^2$ ,  $R_o'^2$ ,  $k$  and  $k'$  are based on regression of the observed activities against predicted activities and vice versa (regression of the predicted activities against observed activities). The definitions are presented clearly in Golbraikh et al. [18], but are not repeated here for brevity.

*Defining model applicability domain*

The domain of application [17, 19] of a QSAR model must be defined if the model is to be used for screening new compounds. Predictions for only those compounds that fall into this domain may be considered reliable. *Extent of Extrapolation* [17] is one simple approach to define the applicability of the domain. It is based on the calculation of the leverage  $h_i$  [20] for each chemical, where the QSAR model is used to predict its activity:

$$h_i = x_i^T (X^T X)^{-1} x_i \quad (10)$$

In Equation (10)  $x_i$  is the descriptor-row vector of the query compound and  $X$  is the  $k \times n$  matrix containing the  $k$  descriptor values 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 not be reliable.

**Results and discussion**

For the selection of the most important descriptors, the aforementioned stepwise multiple regression technique was used. The procedure was automated using a software developed in-house that realizes the ES-SWR algorithm. The seven most significant descriptors according to the ES-SWR algorithm are: the Modified Randic index (ChiMod) followed by Lipophilicity (CLogP), Randic Information 4 (ChiInfo4), Repulsion Energy (NRE), Randic Information 3 (ChiInfo3), Randic Information 1 (ChiInfo1) and finally LUMO Energy (LUMO). This selection resulted to the following full linear equation for the prediction of the inhibitory activity ( $1/IC_{50}$ ):

$$\begin{aligned} \log(1/IC_{50}) = & -0.332CLogP + 0.226LUMO \\ & + 0.021 \cdot 10^{-3}NRE + 4.22ChiInfl \\ & - 2.95ChiInf3 - 1.30ChiInf4 \\ & + 0.058ChiMod - 2.58 \\ F = 32.36 \quad R^2 = 0.837 \quad RMSE = 0.298 \\ R_{CV}^2 = 0.755 \quad S_{PRESS} = 0.366 \\ n = 52 \end{aligned}$$

(11)

Lipophilicity is known to be important for absorption, permeability, and *in vivo* distribution of organic compounds [21] and has been used as a physicochemical descriptor in QSARs with great success [22, 23]. Molecular orbital (MO) surfaces visually represent the various stable electron distributions of a molecule. According to the Frontier Orbital Theory, the shapes and symmetries of the highest-occupied and lowest-unoccupied molecular orbitals (HOMO and LUMO) are crucial in predicting the reactivity of a species and the stereochemical and regiochemical outcome of a chemical reaction. All the structures were fully optimized using the AM1 basis set before the calculation of the LUMO Energy (eV). The NRE (eV) contains the energy which is required to keep two electrons, each on separate  $\pi$  atoms, from moving apart and the energy which is required to keep two electrons, occupying the same orbital on the same  $\pi$  atom, from moving apart. The NRE is more positive as the atom becomes more electronegative. Modified Randic index is based on reciprocal distance of a molecular graph. Randic information topological descriptors (ChiInfo1, ChiInfo3, ChiInfo4) are combinations of topostructural and topochemical descriptors [12]. Topostructure indices encode information on the adjacency and distance of atoms in the molecular structure. Topochemical indices quantify information on topology but also specific chemical properties of atoms such their chemical identity and hybridization state. In our recent work Melagraki et al. [24] topological information descriptors were used with great success.

A correlation analysis on the seven selected descriptors (Table 2) was performed to test for internal correlations. All the values deviate from unity considerably so there is no significant corre-

lation between the seven independent variables. In order to investigate the possibility of having included outliers in our data set, the extent of the extrapolation method was applied to the 52 compounds that constitute the entire data set. The leverages for all 52 compounds were computed (Table 3) and found to be inside the domain of the model (warning leverage limit 0.461).

The predictive ability of the selected descriptors was further explored, by dividing the full data set consisting of 52 1-(3,3-diphenylpropyl)-piperidinyl phenylacetamides amides into a training set of 39 compounds, and a validation set of 13 compounds. The data set was partitioned in a way that we obtained a representative training set and at the same time a diverse test set in terms of molecular structure [25]. More specifically the selection of the derivatives in the training set was made according to the structure and the scale of the biological action, so that representatives of a wide range of structures (in terms of the different substituents, atoms and action) were included. The distribution of the activity values for the test set follows the distribution of the activity values for the training set. According to Golbraikh and Tropsha [26] this approach is correct since representative points of the test set must be close to those of training set and vice versa.

The compounds that constituted the training and validation sets are clearly presented in Tables 4 and 5, where the 52 compounds are separated in two groups (amides and phenylacetamides). The validation examples are marked with <sup>b</sup>. The rest of the study will be concentrated on the model which is constructed from the training set. Using the seven selected descriptors, we developed a new MLR equation based on only the 39 training examples:

Table 2. Correlation matrix of the seven selected descriptors.

	ClogP	Lumo	NRE	ChiInf1	ChiInf3	ChiInf4	ChiMod
ClogP	1						
Lumo	-0.09	1					
NRE	0.13	-0.29	1				
ChiInf1	-0.18	-0.04	-0.22	1			
ChiInf3	-0.17	-0.10	0.47	0.43	1		
ChiInf4	-0.28	-0.19	0.38	0.32	0.74	1	
ChiMod	0.11	0.002	-0.70	0.18	-0.53	-0.45	1

Table 3. Leverages for the entire data set.

Compound Id	Leverages
1	0.127
2	0.120
3	0.073
4	0.172
5	0.142
6	0.103
7	0.145
8	0.253
9	0.237
10	0.087
11	0.291
12	0.421
13	0.089
14	0.085
15	0.057
16	0.079
17	0.116
18	0.174
19	0.373
20	0.047
21	0.078
22	0.072
23	0.083
24	0.091
25	0.082
26	0.117
27	0.121
28	0.099
29	0.334
30	0.230
31	0.119
32	0.159
33	0.103
34	0.077
35	0.142
36	0.384
37	0.073
38	0.106
39	0.115
40	0.217
41	0.199
42	0.141
43	0.160
44	0.265
45	0.149
46	0.226
47	0.203
48	0.184
49	0.063

Table 3. Continued.

Compound Id	Leverages
50	0.155
51	0.110
52	0.149

$$\begin{aligned} \log(1/IC_{50}) = & -0.339CLogP + 0.262LUMO \\ & + 0.024 \cdot 10^{-3}NRE + 4.21ChiInfl \\ & - 2.66ChiInf3 - 1.60ChiInf4 \\ & + 0.06ChiMod - 2.81 \\ F = & 17.18 R^2 = 0.801 RMSE = 0.324 \\ R_{CV}^2 = & 0.657 S_{PRESS} = 0.425 \\ n = & 39 \end{aligned} \quad (12)$$

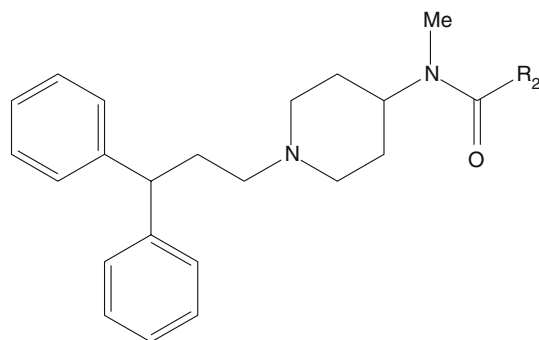
This equation was used to estimate the CCR5 binding affinities for the training and validation examples resulting in  $R^2$  statistics equal to 0.801 (as shown above) and 0.921 respectively. The outcomes of the model are presented in the last two columns of Tables 4 and 5. Graphically, observed vs. predicted inhibitory activities for the training and the validation data sets are shown in Figure 1. The results illustrated once more that the linear MLR technique combined with a successful variable selection procedure are adequate to generate an efficient QSAR model for predicting the CCR5 binding affinity of different compounds.



The proposed model (Equation 12) also passed the rest of the tests that we utilized for illustrating its predictive ability (Equations 6–9)

$$\begin{aligned} R_{CV,ext}^2 &= 0.915 > 0.5 \\ R^2 &= 0.921 > 0.6 \\ \frac{(R^2 - R_o^2)}{R^2} &= -0.1787 < 0.1 \\ \text{or } \frac{(R^2 - R_o'^2)}{R^2} &= -0.1595 < 0.1 \\ k &= 1.006 \text{ and } k' = 0.910 \end{aligned}$$

The model was further validated by applying the  $Y$ -randomization test. Several random shuffles of the  $Y$  vector were performed and the results are shown in Table 6. The low  $R^2$  and  $R_{CV}^2$  values show that the good results in our original model are not due to a chance correlation or structural dependency of the training set.

Table 4. Binding biological data of amides. training and test data.



	R <sub>2</sub>	IC <sub>50</sub> (μM) observed	Log(1/IC <sub>50</sub> ) observed	Training data log(1/IC <sub>50</sub> ) predicted	Test data log(1/IC <sub>50</sub> ) predicted
1 <sup>b</sup>	–	4.1	–0.61		–0.813
2	4-Pyridinyl	6.1	–0.78	–0.478	
3	4-F-C <sub>6</sub> H <sub>4</sub>	7.2	–0.86	–0.611	
4	3-NO <sub>2</sub> -phenyl	5.1	–0.71	–0.066	
5 <sup>b</sup>	2-Thienyl	8.7	–0.94		–0.832
6	2-Furanyl	7.9	–0.90	–0.608	
7 <sup>b</sup>	Cyclobutyl	7.4	–0.87		–0.727
8	Isobutyl	3.4	–0.53	–0.916	
9 <sup>b</sup>	Neopentyl	5.5	–0.74		–0.481
10	Benzyl	0.81	0.09	–0.253	
11		5.9	–0.77	–0.897	
12		6.8	–0.83	–0.962	

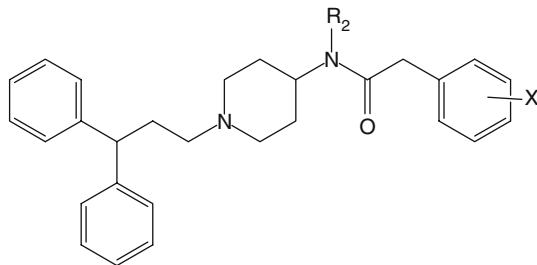
<sup>b</sup>Test data.

Finally, the extent of extrapolation method was applied to the compounds that constitute the test set. The leverages for all 13 compounds are presented in Table 7. None of the 13 compounds fell outside the domain of the model (warning leverage limit 0.615).

The proposed method, due to the high predictive ability [17, 27], can provide a useful aid to the costly and time consuming experiments for determining the CCR5 binding affinity. The method can also be used to screen existing databases or virtual libraries in order to identify new potentially active compounds. In this case, the applicability domain serves as a valuable tool to filter out “dissimilar” compounds.

Such a group of new derivatives, previously not tested for the specific biological action, was subjected to virtual screening using the produced model (Tables 8–11). 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. We have searched for optimized pharmacophore patterns by insertions, substitutions, and deletions of pharmacophoric substituents of the main building block scaffolds. The searching strategy was similar to the one followed in [28]. Finally, we identified the structural trends that lead to improved CCR5 binding affinity.

Table 5. Binding biological data of phenylacetamides. training and test data.

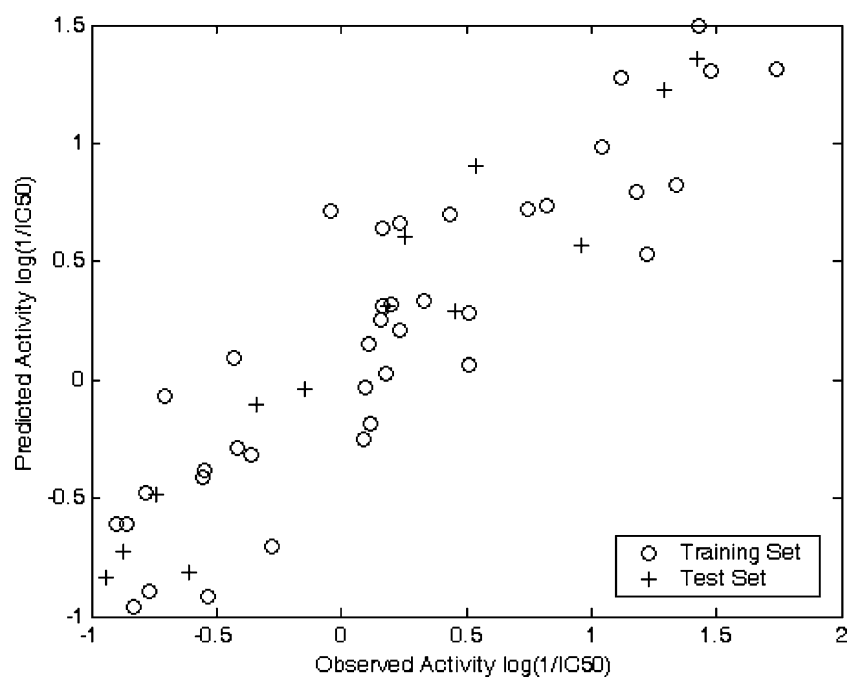


	R <sub>2</sub>	X	IC <sub>50</sub> (μM) observed	Log(1/IC <sub>50</sub> ) observed	Training data log(1/IC <sub>50</sub> ) predicted	Test data log(1/IC <sub>50</sub> ) predicted
13	Me	H	0.77	0.11	-0.182	
14	Me	2-Cl	3.60	-0.56	-0.411	
15 <sup>b</sup>	Me	3-Cl	2.20	-0.34		-0.108
16	Me	4-Cl	0.80	0.10	-0.031	
17	Me	3,4-di-Cl	0.78	0.11	0.152	
18	Me	2,4-di-Cl	2.60	-0.41	-0.289	
19	Me	2-F	1.90	-0.28	-0.703	
20 <sup>b</sup>	Me	3-F	1.40	-0.15		-0.042
21	Me	4-F	0.66	0.18	0.031	
22	Me	3,4-di-F	0.69	0.16	0.256	
23	Me	3-OMe	0.68	0.17	0.642	
24	Me	4-OMe	0.58	0.24	0.666	
25 <sup>b</sup>	Me	3,4-di-OMe	0.65	0.19		0.313
26	Me	3,5-di-OMe	2.70	-0.43	0.096	
27	Me	2,4,5-tri-OMe	1.10	-0.04	0.715	
28	Me	4-Br	0.58	0.24	0.207	
29	Me	4-Benzyloxy	3.50	-0.54	-0.384	
30	Me	4-Phenyl	2.30	-0.36	-0.319	
31	Me	4-CF <sub>3</sub>	0.37	0.43	0.703	
32 <sup>b</sup>	Me	4-OCF <sub>3</sub>	0.29	0.54		0.901
33	Me	4-NHCOMe	0.68	0.17	0.309	
34	Me	4-CN	0.06	1.22	0.528	
35	Me	4-SO <sub>2</sub> NH <sub>2</sub>	0.091	1.04	0.985	
36	Me	4-SO <sub>2</sub> N(Me) <sub>2</sub>	0.046	1.34	0.823	
37 <sup>b</sup>	Me	4-SMe	0.56	0.25		0.603
38	Me	4-CO <sub>2</sub> Me	0.63	0.20	0.322	
39	Me	4-OH	0.47	0.33	0.335	
40	Me	4-NO <sub>2</sub>	0.15	0.82	0.734	
41	Et	4-OCF <sub>3</sub>	0.31	0.51	0.066	
42	Et	4-CN	0.066	1.18	0.796	
43 <sup>b</sup>	Et	4-SO <sub>2</sub> NH <sub>2</sub>	0.038	1.42		1.359
44	Et	4-SO <sub>2</sub> N(Me) <sub>2</sub>	0.018	1.74	1.315	
45	Et	4-SO <sub>2</sub> Me	0.076	1.12	1.277	
46 <sup>b</sup>	Et	4-NO <sub>2</sub>	0.11	0.96		0.568
47	Cyclopropyl	4-SO <sub>2</sub> NH <sub>2</sub>	0.033	1.48	1.305	



Table 5. Continued

	R <sub>2</sub>	X	IC <sub>50</sub> (μM) observed	Log(1/IC <sub>50</sub> ) observed	Training data log(1/IC <sub>50</sub> ) predicted	Test data log(1/IC <sub>50</sub> ) predicted
48 <sup>b</sup>	Cyclopropyl	4-SO <sub>2</sub> Me	0.051	1.29		1.223
49	Cyclopropyl	4-NO <sub>2</sub>	0.31	0.51	0.283	
50 <sup>b</sup>	Allyl	4-OCF <sub>3</sub>	0.35	0.46		0.291
51	Allyl	4-SO <sub>2</sub> Me	0.037	1.43	1.494	
52	Allyl	4-NO <sub>2</sub>	0.18	0.74	0.721	

<sup>b</sup>Test data.Figure 1. Observed vs. predicted activity log(1/IC<sub>50</sub>) for the training and test set.Table 6. R<sup>2</sup> and R<sub>CV</sub><sup>2</sup> values for several Y-randomization tests.

Iteration	R <sup>2</sup>	R <sub>CV</sub> <sup>2</sup>
1	0.11	0.00
2	0.16	0.00
3	0.05	0.00
4	0.25	0.00
5	0.06	0.00
6	0.24	0.00
7	0.08	0.00
8	0.18	0.00
9	0.07	0.00
10	0.19	0.00

Initially alternative functionalities were considered for the acetamide core of the starting *N*[1-(3,3-diphenylpropyl)piperidin-4-yl]-*N*-methyl-2-[4-(methylsulfonyl)-phenyl]acetamide [log(1/IC<sub>50</sub>) 0.904] (Table 8). Introduction of a 2-hydroxy guanidine core [id 4n, log(1/IC<sub>50</sub>) 1.698] showed significant improvement in the activity and remained within the domain of applicability. This compound id 4n was therefore chosen for further manipulation. In Table 9 the 4-(methylsulfonyl)phenyl end group of compound id 4n is replaced with heteroaromatic analogues. With the exception of pyrrole id 19n all the 5-membered heteroaromatic systems investigated were outside

Table 7. Leverages for the test set.

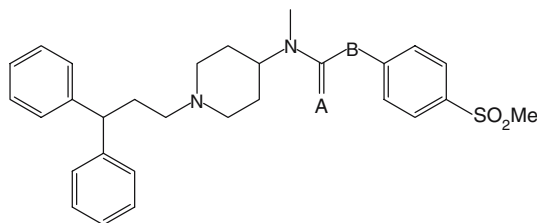
Compound Id	Leverages
1	0.183
5	0.195
7	0.245
9	0.425
15	0.079
20	0.066
25	0.125
32	0.247
37	0.105
43	0.243
46	0.385
48	0.283
50	0.219

of the domain of applicability and as such 5-membered heteroaromatic systems were not considered further. The 6-membered aza heterocycles, pyridine, pyridazine, pyrimidine and pyrazine were investigated. With these systems the methylsulfonyl group showed a preference to be *para* to the guanidine group and the diaza heterocycles pyrimidin-5-yl [id 14n,  $\log(1/IC_{50})$  2.095] and pyrazin-2-yl [id 17n,  $\log(1/IC_{50})$  2.095] gave the best improvements in activity within the domain of applicability. Since the model was unable to

differentiate between these two systems we chose only one of them [the pyrazin-2-yl (id 17n)] to proceed with our studies and varied the alkyl group  $R^2$  on the guanidine core. Increasing the length of the alkyl substituent  $R^2$  gave improved activity up to n-Pr [id 25n,  $\log(1/IC_{50})$  2.389], the n-Bu derivative gave reduced activity as did branching on the i-Pr analogue. An investigation of the branched butyl derivatives indicated that the iso-Bu analogue [id 30n,  $\log(1/IC_{50})$  2.524] had superior activity.

Next the piperidine core was replaced by 5-membered heterocycles pyrrolidine and pyrrole but in both cases the predicted activity was reduced (Table 10). The alkyl linker between the piperidine core and the diphenylmethane end group was also investigated (Table 10). Extending the alkyl chain by one carbon i.e. propyl ( $n = 2$ ) led to a peak in predicted activity [id 35n,  $\log(1/IC_{50})$  2.533] but this was improved even further when the end group diphenylmethane was replaced by diphenylamine [id 38n,  $\log(1/IC_{50})$  2.605]. Further modification to the carbazol led to a reduction in activity (Table 11). Many of the above structures show an increase in activity and fall well within the domain of applicability as such they are worthy of further study. Clearly the model tolerates a wide variety of structural modification demonstrating its potential for virtual screening studies.

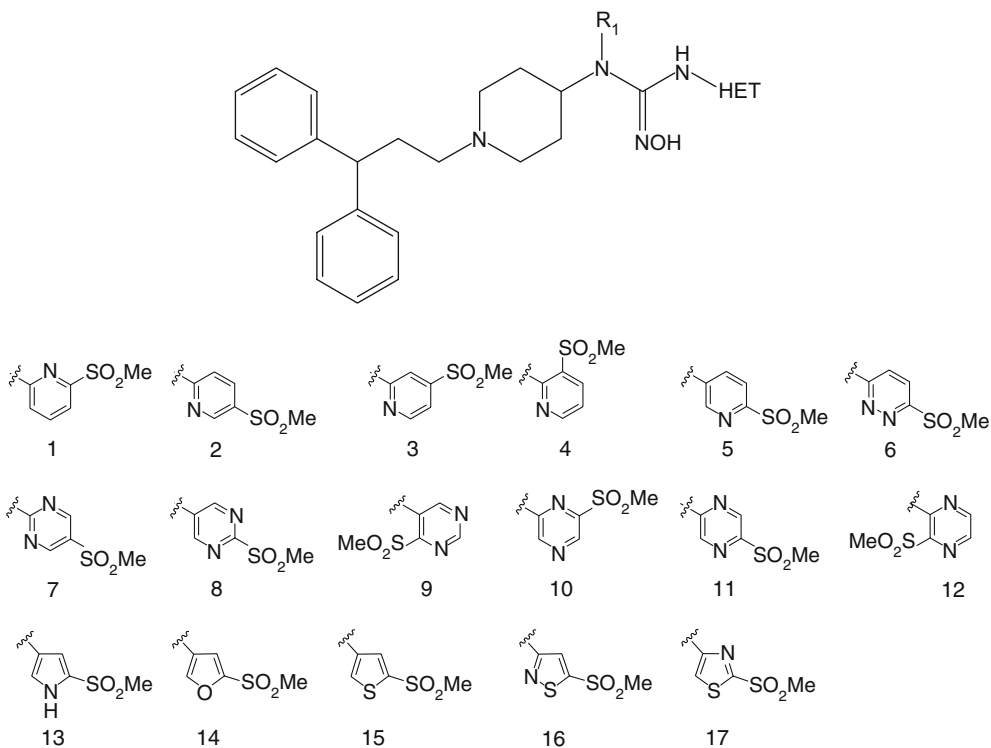
Table 8. Virtual screening results. Modification of *N*-[1-(3,3-diphenylpropyl)-piperidin-4-yl]-*N*-methyl-2-[4-(methylsulfonyl)-phenyl]acetamide.



Id	A	B	Log(1/IC <sub>50</sub> ) predicted	Leverages
1n	O	CH <sub>2</sub>	0.904	0.169
2n	S	CH <sub>2</sub>	0.923	0.155
3n	NOH	CH <sub>2</sub>	1.042	0.105
4n	NOH	NH	1.698	0.288
5n	NOH	O	1.505	0.201
6n	NOH	S	1.400	0.167

Leverage limit = 0.615.

Table 9. Virtual screening results. Modification of 1-[1-(3,3-diphenylpropyl)-piperidin-4-yl]-2-hydroxy-1-alkyl-3-[(methylsulfonyl)heteroaryl]guanidine.



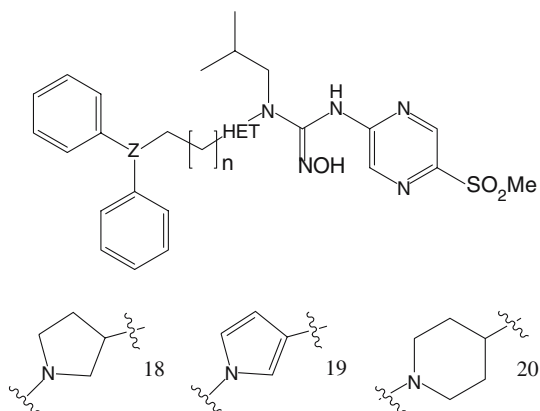
Id	HET	R <sub>1</sub>	Log(1/IC <sub>50</sub> ) predicted	Leverages
7n	1	Me	1.560	0.317
8n	2	Me	2.024	0.473
9n	3	Me	1.834	0.476
10n	4	Me	1.673	0.466
11n	5	Me	1.753	0.309
12n	6	Me	2.395	0.760
13n	7	Me	2.366	0.734
14n	8	Me	2.095	0.512
15n	9	Me	1.747	0.506
16n	10	Me	1.905	0.515
17n	11	Me	2.095	0.512
18n	12	Me	1.744	0.505
19n	13	Me	1.722	0.407
20n	14	Me	2.513	0.926
21n	15	Me	2.571	0.841
22n	16	Me	2.682	1.179
23n	17	Me	2.607	0.908
24n	11	Et	2.154	0.519
25n	11	Pr	2.389	0.504
26n	11	i-Pr	2.281	0.425
27n	11	c-Pr	2.406	0.770

Table 9. Continued.

Id	HET	R <sub>1</sub>	Log(1/IC <sub>50</sub> ) predicted	Leverages
28n	11	n-Bu	2.327	0.495
29n	11	sec-Bu	2.119	0.453
30n	11	i-Bu	2.524	0.424
31n	11	Tert-Bu	2.026	0.664

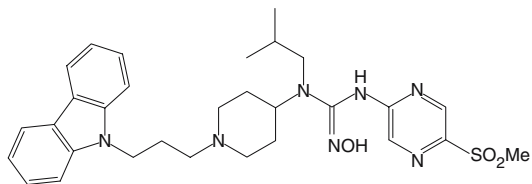
Leverage limit = 0.615.

Table 10. Virtual screening results. Modification of 1-[1-(3,3-diphenylpropyl)-piperidin-4-yl]-2-hydroxy-1-isobutyl-3-[5-(methylsulfonyl)pyrazin-2-yl]guanidine.



Id	HET	<i>N</i>	Z	Log(1/IC <sub>50</sub> ) predicted	Leverages
32n	18	1	CH	2.021	0.271
33n	19	1	CH	2.049	0.278
34n	20	0	CH	2.412	0.397
35n	20	2	CH	2.533	0.435
36n	20	3	CH	2.419	0.453
37n	20	4	CH	2.264	0.539
38n	20	2	N	2.605	0.466

Leverage limit = 0.615.

Table 11. Virtual screening results. Investigation of 1-[1-[3-(9*H*-carbazol-9-yl)-propyl]piperidin-4-yl]-2-hydroxy-1-isobutyl-3-[5-(methylsulfonyl)pyrazin-2-yl]-guanidine.

Id	Log(1/IC <sub>50</sub> ) predicted	Leverages
39n	2.3024	0.567

Leverage limit = 0.615.

## Conclusion

The successful results of this study led to the conclusion that CCR5 binding affinity can be successfully modeled with physicochemical constants and structural descriptors. The validation procedures utilized in this work (separation of data into independent training and validation sets, *Y*-randomization) illustrated the accuracy and robustness of the produced QSAR model not only by calculating its fitness on sets of training data, but also by testing the predictive ability of the model. The proposed method, due to the high predictive ability, offers a useful alternative to the costly and time consuming experiments for determining CCR5 binding affinity. Furthermore, the produced QSAR model can be used to screen existing databases or virtual libraries in order to identify novel potent compounds. An attempt in this direction was carried out. Synthesis of the molecules proposed by the aforementioned virtual screening procedure and experimental evaluation of their biological activity will show if the method can be used as a general rational drug discovery tool.

## Acknowledgements

A.A. wishes to thank Cyprus Research Promotion Foundation (Grant No. PENEK/ENISX/0603/05) for its financial support. A.A and G.M. wish to thank Leventis Foundation for its financial support.

## References

- Burrows, J.N., Cumming, J.G., Fillery, S.M., Hamlin, G.A., Hudson, J.A., Jackson, R.J., McLaughlin, S. and Shaw, J.S., *Bioorg. Med. Chem. Lett.*, 15 (2005) 25.
- Kazmierski, W., Bifoulco, N., Yang, H., Boone, L., DeAnda, F., Watson, C. and Kenakin, T., *Bioorg. Med. Chem.*, 11 (2003) 2663.
- Pipitone, N. and Pitzalis, C., *Curr. Opin. Anti-inflammat. Immunomodulat. Invset. Drugs*, 2 (2000) 9.
- Sellebjerg, F., Madsen, H.O., Jensen, C.V., Jensen, J. and Garred, P.J., *J. Neuroimmunol.*, 102 (2000) 98.
- Fischereder, M., Luckow, B., Wuthrich, R.P., Rothenpieler, U., Schneeberger, H., Panzer, U., Stahl, R.A.K., Hauser, I.A., Budde, K., Neumayer, H.-H., Kramer, B.K., Land, W. and Schlondorff, D., *Lancet*, 387 (2001) 1758.
- Andres, P.G., Beck, P.L., Mizoguchi, E., Mizoguchi, A., Bhan, A.K., Dawson, T., Kuziel, W.A., Maeda, N., MacDermott, N., Podolsky, R.P. and Reinecker, D.K., *J. Immunol.*, 164 (2000) 6303.
- Debnath, A.K., *J. Med. Chem.*, 46 (2003) 4501.
- Xu, Y., Liu, H., Niu, C., Luo, C., Shen, J., Chen, K. and Jiang, H., *Bioorg. Med. Chem.*, 12 (2004) 6193.
- Song, M., Breneman, C.M. and Sukumar, N., *Bioorg. Med. Chem.*, 12 (2004) 489.
- Leonard, J.T. and Roy, K., *QSAR Comb. Sci.*, 23 (2004) 387.
- Roy, K. and Leonard, J.T., *J. Chem. Inf. Model.*, 45 (2005) 1352.
- Todeschini, R., Consonni, V., Mannhold, R. (Series Editor), Kubinyi, H. (Series Editor) and Timmerman, H. (Series Editor), *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, 2000.
- Efroymsen, M.A., In Ralston, A. and Wilf, H.S. (Eds.), *Mathematical Methods for Digital Computers*, Wiley, NY, 1960.
- Efron, B., *J. Am. Stat. Assoc.*, 78 (1983) 316.
- Osten, D.W., *J. Chemom.*, 2 (1998) 39.
- Wold, S. and Eriksson, L., In Van de Waterbeemd, H. (Ed.), *Chemometrics Methods In Molecular Design*, VCH Weinheim, Germany, 1995.
- Tropsha, A., Gramatica, P. and Gombar, V.K., *QSAR Comb. Sci.*, 22 (2003) 1.
- Golbraikh, A. and Tropsha, A., *J. Mol. Graph. Mod.*, 20 (2002) 269.
- Shen, M., Beguin, C., Golbraikh, A., Stables, J., Kohn, H. and Tropsha, A., *J. Med. Chem.*, 47 (2004) 2356.
- Atkinson, A. *Plots, Transformations and Regression*. Clarendon Press, Oxford (UK), 1985.
- Walters, W.P.A. and Murcko, M.A., *Curr. Opin. Chem. Biol.*, 3 (1999) 384.
- Devillers, J. (Ed.), *Comparative QSAR*. Taylor and Francis, Washington, DC, 1998.
- Hansch, C. and Leo, A. *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology*. ACS, Washington, DC, 1995.
- Melagraki, G., Afantitis, A., Sarimveis, H., Igglessi-Markopoulou, O. and Supuran, C.T., *Bioorg. Med. Chem.*, 14 (2006) 1108.
- Melagraki, G., Afantitis, A., Sarimveis, H., Igglessi-Markopoulou, O. and Alexandridis, A., *Mol. Div.* (2006) In Press ID AP\_11030\_2005\_9008.
- Golbraikh, A. and Tropsha, A., *Mol. Div.*, 5 (2000) 231.
- Aptula, A.O., Jeliakova, N.G., Schultz, T.W. and Cronin, M.T.D., *QSAR Comb. Sci.*, 24 (2005) 385.
- Afantitis, A., Melagraki, G., Sarimveis, H., Koutentis, P.A., Markopoulos, J. and Igglessi-Markopoulou, O., *Mol. Div.* (2006) In Press DOI MODI28R2.