

Novel semi-automated methodology for developing highly predictive QSAR models: application for development of QSAR models for insect repellent amides

Jayendra B. Bhonsle · Apurba K. Bhattacharjee · Raj K. Gupta

Received: 10 January 2006 / Accepted: 21 June 2006 / Published online: 20 September 2006
© Springer-Verlag 2006

Abstract Conventional 3D-QSAR models are built using global minimum conformations or quantum-mechanics based geometry-optimized conformations as bioactive conformers. QSAR models developed using the global minima as bioactive conformers, employing the GFA, PLS and G/PLS methodologies, gave good non-validated r^2 (0.898, 0.868 and 0.922) and performed well on an internal validation test with leave-one-out correlation q^2_{LOO} (0.902, 0.726 and 0.924), leave-10%-out correlation q^2_{L10O} (0.874, 0.728 and 0.883) and leave-20%-out q^2_{L20O} (0.811, 0.716 and 0.907). However, they showed poor predictive ability on an external data set with best predictive r^2 (Pred- r^2) of 0.349, 0.139 and 0.204 respectively. A novel methodology to mine bioactive conformers, from clusters of conformations with good 3D-spatial representation around pharmacophoric moiety, furnishes highly predictive 3D-QSAR models. The best QSAR model (model A) showed r^2 of 0.989, q^2_{LOO} of 0.989, q^2_{L10O} of 0.980, q^2_{L20O} of 0.963 and Pred- r^2 on eight test compounds of 0.845. The methodology is based on mimicking the multi-way Partial Least Squares (PLS) technique by performing several automated sequential PLS analyses. The poses/shapes of

the mined bioactive conformers provide valuable insight into the mechanism of action of the insect repellents. All of the repetitive tasks were automated using Tcl-based Cerius2 scripts.

Keywords 3D QSAR · Insect repellents · Bioactive conformer mining · Cerius2 scripts

Abbreviations

QSAR	quantitative structure–activity relationship
PLS	partial least squares
GFA	genetic function approximation
G/PLS	genetic partial least squares
RMS	root mean squares
Pred r^2	predictive r^2
LOO	leave-one-out
L10O	leave-10%-out
L20O	leave-20%-out
GPCR	G-protein coupled receptors
DSP	descriptor significance percentage
PPE	percentage prediction error
CtoBA	contribution to BioActivity
PT	protection time
Tcl	tool command language
DEET	N,N-diethyl-3-methyl benzamide
OBP	odorant binding protein
ODE	odorant degrading enzyme
JH	juvenile hormone
CoMFA	comparative molecular field analysis
MSA	molecular shape analysis
CoMSIA	comparative molecular similarity index analysis
COMPASS	condensed phase optimized molecular potentials for atomistic simulation studies

J. B. Bhonsle (✉) · A. K. Bhattacharjee
Department of Medicinal Chemistry,
Division of Experimental Therapeutics,
Walter Reed Army Institute of Research,
503 Robert Grant Avenue,
Silver Springs, MD 20910, USA
e-mail: Jayendra.Bhonsle@na.amedd.army.mil

A. K. Bhattacharjee
e-mail: Apurba.Bhattacharjee@na.amedd.army.mil

R. K. Gupta
Office of the Director, Research, Plans and Programs,
Ft. Detrick, MD 21702, USA

HASL	hypothetical active site lattice
MTD-ADJ	minimal topologic difference using adjusted biological activities
SD	sum of squared deviations
PRESS	Prediction Error Sum of Squares
S(y)	standard error for the y estimate
MTI	minimum threshold index

Introduction

Quantitative Structure Activity Relationships (QSAR) are among the most widely used techniques in rational drug design. Following the pioneering work of Hansch et al. [1] in 2D-QSAR, several sophisticated techniques like Comparative Molecular Field Analysis (CoMFA) [2], Molecular Shape Analysis (MSA) [3], Comparative Molecular Similarity Index Analysis (CoMSIA) [4], Condensed Phase Optimized Molecular Potentials for Atomistic Simulation Studies (COMPASS) [5], and Hypothetical Active Site Lattice (HASL) [6] have been developed for three-dimensional QSAR (3D-QSAR). Several novel two-dimensional QSAR (2D-QSAR) descriptors to quantify the topology and information-content of molecules have been reported recently. Among them are the Weiner [7], Zagreb [8], and Hosoya indices [9], the Kier and Hall molecular connectivity indices [10], the Kier and Hall subgraph count indices [10], Kier's shape indices [11], molecular flexibility indices [12], and the Balaban indices [13]. Some 2D-molecular-graph-based graph-theoretic descriptors recently reported are the information-content-info of atomic composition descriptors [14], information index based on adjacency matrix (A-matrix), distance matrix (D-matrix), Edge matrix (E-matrix) and edge-distance matrix (ED-matrix) [8], the sum of atomic polarizability [15], and the multi-graph information content indices [8]. Several novel 3D-descriptors to capture the conformational electronic and spatial information have also been reported. Among the recently reported 3D-descriptors are shadow indices [16] and Jurs indices [17]. All of the 2D and 3D-descriptors have been used widely in QSAR models. For example, in anti-tubercular agents [18], sulfamates have been used to distinguish sweet, sweet-bitter and bitter tasting molecules [19], and octopaminergic agonists to inhibit sex-pheromone production in insects [20].

Selection of the bioactive conformer is among the most important challenges in QSAR analysis [21]. Numerous sophisticated techniques have been reported to address this challenge, such as by Hopfinger et al. [22] using conformational averaging or conformational ensembles; by Hasagewa et al. [23] employing several conformers in multi-way data arrays; by Vedani et al. [24] using multi-

conformational ligand representation; by Appell et al. [25] invoking tensor decomposition; by Hasagewa et al. [21] employing three-way-PLS analysis; by Xiao et al. [26] propounding the Targacept Active Conformational Search algorithm; and by Sulea et al. [27] employing the multi-conformational minimal topologic difference (MTD-ADJ) using adjusted biological activities.

Previously, we have shown that employing several conformers of highly flexible cyclic pentapeptides in a CoMFA-based QSAR study coupled with several sequential partial least square analyses mimicking the multi-way-Partial Least Square analysis, we could develop highly predictive QSAR models [28]. In this paper we extend this method with a semi-automated heuristic using the Cerius2 software package [29] to develop highly predictive 3D-QSAR models for insect repellents.

Mosquitoes transmit a variety of parasites and pathogens including those that cause malaria, yellow fever, dengue fever, filariasis and viral encephalitis [30]. Keeping the mosquitoes away using insect repellents is, therefore, a significant preventive approach against these deadly diseases. The factors involved in attracting mosquitoes to their hosts are complex and not well understood [31]. Mosquitoes have chemo-receptors on their antenna that are involved in the host-sensing mechanism [32]. Davis et al. [33] have reported that mosquitoes' chemo-receptors may be inhibited by *N,N*-diethyl-3-methyl benzamide (DEET). The DEET molecules in the vapor state have access to the chemosensilla and membranes in the body *via* the pores in the cuticle and tracheal system. The interaction of DEET molecules with the dendrite membrane lipids is thought to perturb them so that the normal response of the mosquitoes to other attractants is altered [34]. The currently reported participating entities in the mechanism of action [35] of DEET and other odorants are the odorant-binding proteins (OBPs), the G-protein-coupled receptors (GPCRs) and the odorant-degrading enzymes (ODEs). The OBPs are believed to bind to odorants that are typically hydrophobic, and facilitate their movement through the hydrophilic hemolymph to the GPCRs on the cell (neurons) surfaces. It is believed that only the OBPs and odorant complex alone can bind with the GPCRs. The ODEs prevent continued stimulation of the olfactory receptors by degrading the molecules associated with the olfactory stimulus.

Amides, both aliphatic and aromatic, are well known mosquito repellents [36, 37]. Skinner et al. [38] have studied the relationship of repellency/potency of *N,N*-diethyl benzamides with their boiling points, polarizabilities and partition coefficients. Other physico-chemical characteristics that are correlated with repellency/potency are lipophilicity [34], molecular size [39], and molecular shape [40]. Suryanarayana et al. [41] have reported the synthesis and mosquito

repellency testing of forty aromatic and cyclohexyl carboxylic acid amides. Further, they have also shown the structure–activity relationship of lipophilicity, molecular length and molar refractivity to repellency/potency. Ma et al. [42] studied the electronic properties of several insect repellent benzamides and benzylamides. They have demonstrated that a specific range of the van der Waals surface electrostatic potential of the amide nitrogen and oxygen atoms, and the atomic charges and dipole moments is required for the compounds to exhibit potent repellency activity. Previously, we have reported a molecular similarity analysis study of several DEET analogs and the insect juvenile hormone (JH) [43]. We have also reported observing similarity of stereoelectronic attributes such as the electrostatic potentials of the amide and/or ester moieties in the benzamides, benzylamides, JH and JH-mimic compounds and the large distribution of hydrophobic regions in these molecules. These features play a crucial role in molecular recognition between repellent compounds and JH receptors. More recently, we have reported [44] a pharmacophore for insect repellent activity using a CATALYST-based QSAR study of eleven known insect repellents. Here we illustrate the semi-automated quasi-multiway PLS methodology by building the QSAR model for forty insect repellents based on different aliphatic and aromatic amides reported by Suryanarayana et al. [41] We have also compared our results with the GFA, G/PLS and PLS based QSAR models built using global minima.

Materials and methods

Cerius2 (C2) version 4.9 [29] running on a Silicon Graphics Octane workstation under the IRIX 6.5 operating system was used for all of the modeling work presented here. Gasteiger–Marsili [45] charges and the Dreiding 2.21 force field [46] were used for all of the computations in this study. Unless otherwise noted, default C2 settings were used.

Data set

We used a collection of forty compounds that included benzamides, benzyl amides and cyclohexyl amide derivatives for this QSAR study. Suryanarayana et al. [41] have reported the protection time (PT) of these compounds against the mosquito species *Anopheles aegypti*. The PT was determined as follows. The test compound was applied at a dosage of 1 mg cm⁻² to an alcohol cleaned human fist. The compound laced fist was then exposed to 200 female (5–7 days old) day-biting *Aedes aegypti* mosquitoes for five minutes. This was repeated every thirty minutes. PT is

defined as the period of protection offered until two consecutive bites are obtained in that half-hour interval. The data set was divided into a two parts: training set of thirty compounds and test set of ten compounds. The activity-ranking algorithm described by Golbraikh et al. [47] was used for training and test set selection. Table 1 summarizes the chemical structures, vapor pressure at 30 °C and biological activity data of all the compounds.

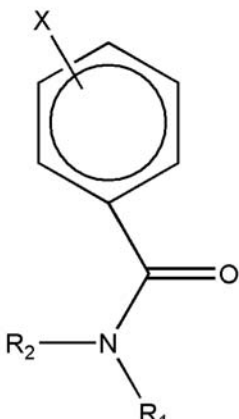
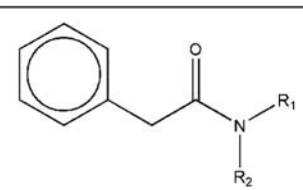
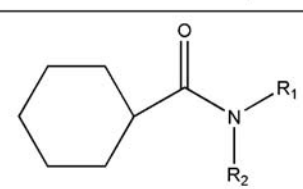
Molecular structure building, conformational search and cluster analyses

Each structure was built using the C2 3D-sketcher and minimized. Exhaustive conformational searches were performed using the Grid Scan method [48]. The grid scan step size was selected as follows: For compounds with three or fewer torsions 30° was used; for five or more torsions, 45° was used; for four torsions, 45° was used for bonds attached to the amidic N and 30° for the rest. The torsion bond is defined as a single bond connecting different groups, which on rotation would give rise to potential local-minimum conformers. Tcl-based Cerius2 scripts [49] were developed to automate the repetitive conformational searches.

We performed cluster analysis based on the RMS (root mean squares) differences of the torsion angles between the conformers. The steps in the algorithm [48] are as follows: All of the conformers are sorted by energy. The lowest energy conformer is assigned to the first cluster and it becomes the cluster nucleus. Next, all the conformers that have an RMS difference below the specified threshold value are placed in the first cluster. The lowest energy conformer of the remaining unclustered conformers is placed in the second cluster as its cluster nucleus. Again, all the conformers that have an RMS difference below the specified threshold value are placed in the second cluster. The above two steps are repeated until all the conformers are placed into clusters.

Preliminary cluster analysis was performed to generate 10–15, 15–20, 20–25, 25–30 and 30–35 conformers per cluster. The nuclei of each cluster were aligned using the amide group, the common core, as the template. The hydrophobic moiety on the carbonyl side of the amide and the aliphatic side chain on the nitrogen side of the amide, of each cluster set, were examined for 3D-spatial representation. The cluster nuclei for the 10–15 and 15–20 set showed no representative conformers in the region around the amide. Nuclei in the 25–30 and 30–35 conformer sets showed crowding in some region around the amide. The nuclei of the 20–25 conformer set showed good 3D-sampling with little or no vacant volume, with much less crowding or over-representation. Consequently, we chose 21–24 conformational clusters for our QSAR analysis.

Table 1 Compounds structure and bioactivity data

Compound Structure	Compd #	X	R1	R2	PT Hrs	Training or Test Set	VP
	1a	4-OCH ₃	Et	H	0.08	Test	0.0062
	1b		CH ₃	CH ₃	1.00	Training	0.0039
	1c		Et	Et	1.00	Training	0.0037
	1d		iPr	iPr	1.17	Training	0.0155
	1e		C ₅ H ₁₀		0.75	Training	0.1486
	2a	4-CH ₃	Et	H	0.08	Training	0.0063
	2b		CH ₃	CH ₃	4.00	Training	0.0110
	2c		Et	Et	2.83	Training	0.0244
	2d		iPr	iPr	0.50	Test	0.0159
	2e		C ₅ H ₁₀		1.00	Training	0.0313
	3a	H	Et	H	0.58	Training	0.0015
	3b		CH ₃	CH ₃	1.67	Test	0.0015
	3c		Et	Et	4.00	Training	0.1015
	3d		iPr	iPr	3.00	Training	0.0116
	3e		C ₅ H ₁₀		3.00	Test	0.0559
	4a	3-CH ₃	Et	H	0.67	Training	0.0013
	4b		CH ₃	CH ₃	3.00	Training	0.0055
	4c		Et	Et	5.00	Test	0.0260
	4d		iPr	iPr	2.67	Test	0.0151
	4e		C ₅ H ₁₀		1.42	Training	0.0001
	5a	2-Cl	Et	H	0.58	Training	0.0006
	5b		CH ₃	CH ₃	5.00	Training	0.0076
	5c		Et	Et	3.00	Training	0.0602
	5d		iPr	iPr	1.00	Test	0.7728
	5e		C ₅ H ₁₀		1.00	Training	0.0281
6a	2-OEt	Et	H	0.08	Training	0.0003	
6b		CH ₃	CH ₃	2.83	Training	0.0264	
6c		Et	Et	3.50	Training	0.0012	
6d		iPr	iPr	1.08	Test	0.0144	
6e		C ₅ H ₁₀		1.33	Training	0.0030	
	7a		Et	H	1.00	Training	0.0058
	7b		CH ₃	CH ₃	2.17	Test	0.0020
	7c		Et	Et	6.00	Training	0.1043
	7d		iPr	iPr	1.00	Test	0.0014
	7e		C ₅ H ₁₀		2.58	Training	0.1814
	8a		Et	H	0.50	Training	0.0168
	8b		CH ₃	CH ₃	3.00	Training	0.0136
	8c		Et	Et	4.00	Training	0.1638
	8d		iPr	iPr	2.00	Training	0.2843
	8e		C ₅ H ₁₀		2.00	Training	0.0315

PT=Protection Time; VP=Vapor Pressure @ 30 °C

Molecular/conformational alignment, descriptor computation and QSAR model building

All conformers were aligned using the amidic carbonyl carbon, carbonyl oxygen and amide nitrogen as the common-core substructure. C2 Align was used to align the conformers. A total of 127 descriptors were computed for all of the conformers and the correlation matrix computed. PLS was used to compute the QSAR models with the descriptor column auto scaled and means removed. The number of components to explore was set to six unless otherwise noted.

GFA, G/PLS, PLS and Quasi-multi-way PLS analyses

In the genetic function approximation (GFA) method [50], QSAR models are constructed from a randomly chosen proper subset of the independent variables (descriptors) and then these models are “evolved.” A *generation* is the set of models obtained from multiple linear regression analysis on each model. The best models selected become the next generation. Newer models are obtained by “cross-over” operations on the best models selected, which take some variables from each of the two models to produce an *offspring*. After the specified number of iterations, the best models are returned by the method. For this study, the GFA was used with C2 default settings, except for the following parameters: Initial equation length used was 15, constants were added to the equation and no fixed length was set for the final equation.

The partial least squares (PLS) method [51] is used when there are far more independent variables (descriptors) than observations and when there is co-linearity in the independent variables. We used the following PLS parameters: 6 components to explore, the column means removed and the column data auto-scaled. The internal ‘regression-only’ cross-validation was used during the model building process.

G/PLS is a hybrid of the best features of GFA and PLS. G/PLS is reported to give better QSAR models than GFA or PLS alone [52].

The definitions of the statistical terms used are as follows:

Equation (1) gives the conventional correlation coefficient or the non-validated correlation coefficient r^2 .

$$r^2 = 1 - \left(\frac{\sum (Y - Y_{\text{pred}})^2}{\sum (Y - Y_{\text{mean}})^2} \right) \quad (1)$$

where Y is the observed bioactivity, Y_{pred} is the predicted bioactivity and Y_{mean} is the mean bioactivity of all the training set compounds.

Equation (2) gives the cross-validation correlation coefficient q^2 .

$$q^2 = 1 - \left(\frac{\sum (Y - Y_{CV\text{pred}})^2}{\sum (Y - Y_{\text{mean}})^2} \right) \quad (2)$$

where $Y_{CV\text{pred}}$ is the cross-validated predicted bioactivity.

Equation (3) gives the predictive correlation coefficient r^2_{Pred} .

$$r^2_{\text{Pred}} = (\text{SD} - \text{PRES})/\text{SD} \quad (3)$$

where SD is the sum of squared deviations between the bioactivity of compounds in the test set and the mean bioactivity of the training set compounds, PRES is the sum of the squared deviation between the predicted and observed bioactivity for every test set compound.

Prediction Error Sum of Squares (PRESS) for the training set compounds is given by Eq. (4).

$$\text{PRESS} = \sum (Y - Y_{\text{pred}})^2 \quad (4)$$

The F -value or F -statistics is a variance-related parameter used to compare models developed using varying numbers of independent variables (descriptors). This is used to determine if a complex model (more descriptors) is significantly better than a less complex model. Equation (5) gives the F -value.

$$F = \frac{\left(\sum (Y_{\text{pred}} - Y_{\text{mean}})^2 / \nu_1 \right)}{\left(\sum (Y - Y_{\text{pred}})^2 / \nu_2 \right)} \quad (5)$$

where ν_1 and ν_2 are degrees of freedom associated with the regression sum of squares (numerator, variance explained) and the residual sum of squares (denominator, variance unexplained), respectively. A confidence level of 95% ($\alpha=0.05$) is generally considered significant.

The standard error for the y estimate $S(y)$, is also the unexplained variance, and is given by Eq. (6).

$$S(y) = \left(\sum (Y - Y_{\text{pred}})^2 / \nu_2 \right)^{1/2} \quad (6)$$

where ν_2 is the number of degrees of freedom associated with the residual sum of squares.

The multi-way PLS method, developed by Bro et al. [53] was used to develop the 3D-QSAR models of insecticidal neonicotinoid compounds [54]. Each dimension of the multi-way data corresponds to the compounds in training set, CoMFA field variables, conformations and alignments. The conformers and alignments that gave the best correla-

tion to observed bioactivities were determined from the multi-way PLS solution. We have mimicked the multi-way-PLS analyses by performing several sequential two-way PLS analyses on our data. We used a Tcl based Cerius2 script [55] to automate the repetitive task of several PLS analyses.

QSAR pharmacophore model using CATALYST

The previously reported pharmacophore model for insect repellents was generated from a training set of eleven structurally diverse arthropod repellents [44]. CATALYST [56] was used to develop the model by placing suitable constraints on the number of available chemical features, such as aromatic hydrophobic or aliphatic hydrophobic interactions, hydrogen bond donors, hydrogen bond acceptors, hydrogen bond acceptors (lipid) and ring aromatic sites, to describe the arthropod repellent activity of the compounds. Earlier reported [43] quantum chemical calculations and the stereoelectronic properties of these compounds provided guidance for the selection of these physico-chemical features. Molecules were initially mapped to the features with their predetermined conformations generated using the “fast fit” algorithm in CATALYST. A conformational energy range of 0 to 20 kcal mol⁻¹ was used for developing the set of three-dimensional conformers. The Fischer randomization test was used to rule out the possibility of chance correlation models.

Results and discussion

Data set

The collection of forty compounds was divided into two sets: thirty training-set compounds and ten test-set compounds using activity ranking [47]. All compounds were sorted based on activity (PT) into five categories as shown in Table 2.

QSAR models of the global minimum conformers based on contemporary (GFA, PLS, G/PLS) methods

Contemporary QSAR models based on the global minimum conformations of the training set were computed with 127 descriptors and the 30 descriptors selected. Descriptors were selected as relevant if they had correlation with bioactivity greater than 0.1 ($|r| > 0.1$) and if the cross correlation with other descriptors was not larger than 0.9 ($|r| < 0.9$). Table 3 summarizes the statistical details of the QSAR models.

All models (GFA, PLS and G/PLS) were apparently significant based on the internal cross-validation tests of

leave-one-out, leave-10%-out and leave-20%-out with $q^2 > 0.7$, and the randomization tests. The best mean random r was 0.873 ($r^2 = 0.76$) compared to the best non-random r of 0.976 ($r^2 = 0.95$). Despite these apparently good statistics, all models performed poorly on external validation test of ten compounds with the best predictive r^2 of 0.349 for the GFA models. The 30-descriptor GFA model showed modest predictive power with a predictive r^2 of 0.514.

Conformational search and cluster analysis

All compounds were subjected to Grid Scan method of conformational search. Compounds with three or fewer torsional bonds gave less than 2,000 conformers within 20 kcal mol⁻¹ energy range of their global minimum conformer. The global minimum conformer for each compound was obtained by exhaustive minimization of the lowest energy conformer. Compounds with four or more torsional bonds yielded numbers of conformers varying from 3,747 for **6e**, to 73,139 for **1c**. The C2 cluster analysis algorithm is limited to 2,000 conformers. Thus, all compounds with 2,000 or more conformers were reprocessed using appropriate (within 10 kcal mol⁻¹ of global minima) energy cutoff values. The overlay of the original conformations (i.e. before reprocessing) and the overlay of conformation obtained after reprocessing showed no significant loss of the 3D-spatial encompassment around the amide, the putative pharmacophoric moiety [42]. Table 4 summarizes the conformational search and cluster analysis data. We used Tcl-based Cerius2 scripts [49] to automate the repetitive task of conformational searches and cluster analyses.

Our novel methodology mines the 3D-encompassing conformations cluster nuclei to identify the conformer that most closely correlates with bioactivity. Further, the use of the gradual, stepwise refinement gives steady enrichment of bioactive conformers in each successive model. This allows us to identify plausible 3D-spatial requirements for bioactivity and suggests plausible mechanistic roles for various molecular moieties.

Molecular/conformer alignment

We used C2 Align for all alignments. We aligned 940 conformers with the amide template. This necessitated dividing the task into 40 segments because of limitations in C2 Align. For each of the forty compounds, we selected all of the twenty-three to twenty-five conformers, a common conformer viz. **1a_0**, and the amide template for effecting the alignment. All the aligned training set and test set compounds showed complete 3D-spatial representation

Table 2 Data set — training and test set classification

Compound activity class	Bioactivity range (PT-hours)	Total number of compound in range	Number of compounds in training set	Compound IDs in training set	Number of compounds in test set	Compound IDs in test set
A	0.0–0.79	9	7	1e, 2a, 3a, 4a, 5a, 6a, 8a	2	1a, 2d
B	0.8–1.10	8	5	1b, 1c, 2e, 5e, 7a	3	5d, 6d, 7d
C	1.11–2.6	8	6	1d, 4e, 6e, 7e, 8d, 8e	2	3b, 7b
D	2.61–3.0	8	6	2c, 3d, 4b, 5c, 6b, 8b	2	3e, 4d
E	3.01–6.0	7	6	2b, 3c, 5b, 6c, 7c, 8c	1	4c

around the hypothesized pharmacophore amide moiety in the overlaid models. Figure 1 shows the overlay of all 940 conformers, demonstrating the 3D-spatial encompassment around the amide, the putative pharmacophore moiety. We used Cerius2 scripts [57] to automate the repetitive alignment tasks.

3D-QSAR model development

Descriptor computation

A total of 127 different 2D- and 3D-descriptors were calculated for all the compounds. The ADME module

Table 3 Global minimum conformers GFA, PLS and G/PLS based QSAR model's statistical data

Global minimum Conformers GFA, PLS and G/PLS based QSAR model's statistical data							
Statistical method	Models built using 127 descriptors			Models built using selected 30 descriptors			
	GFA	PLS	G/PLS	GFA	PLS	G/PLS	
Non validated r^2	0.792	0.852	0.935	0.898	0.868	0.922	
Regression only CV-LOO q^2	-0.316	0.361	0.73	0.489	0.483	0.409	
PRESS	89.676	43.53	18.37	34.849	35.215	40.252	
Global minimum Conformers GFA, PLS, G/PLS, QSAR models validation results							
Tests		Model built using 127 descriptors			Model built using 30 selected descriptors		
	Model	GFA	PLS	G/PLS	GFA	PLS	G/PLS
Leave-one-out	q^2	0.896	0.900	0.951	0.902	0.726	0.924
	PRESS	7.101	6.831	3.368	6.712	18.673	5.174
Leave-10%-out	q^2	0.852	0.901	0.866	0.874	0.728	0.883
	PRESS	10.077	6.735	9.129	8.583	18.543	7.954
Leave-20%-out	q^2	0.834	0.754	0.826	0.811	0.716	0.907
	PRESS	11.290	16.772	11.884	12.902	19.343	6.306
Randomization tests							
99 trails at 99% confidence level; ((# Random r) > (non-Random r))=0							
		Models built using 127 descriptors			Models built using selected 30 descriptors		
Statistical method		GFA	PLS	G/PLS	GFA	PLS	G/PLS
r from non-random		0.963	0.923	0.976	0.945	0.932	0.968
Mean value of r from random trials		0.731	0.187	0.873	0.732	0.083	0.787
Std deviation of random trials		0.078	0.280	0.046	0.087	0.198	0.070
External test set validation results							
		Models built using 127 descriptors			Models built using selected 30 descriptors		
Statistical method		GFA	PLS	G/PLS	GFA	PLS	G/PLS
For 10 compounds	Predictive r^2	0.349	0.139	0.204	0.324	0.133	0.176
	$s(y)$	1.126	1.735	1.242	0.957	1.526	1.606
	F -value	4.296	1.286	2.053	3.840	1.231	1.699
For 8 compounds (w/o 1a and 7d)	Predictive r^2	0.164	0.0002	0.166	0.514	0.008	0.156
	$s(y)$	0.761	0.692	0.655	0.483	0.403	0.702
	F -value	1.177	0.001	1.196	6.355	0.048	1.106

Table 4 Conformational search and cluster analysis data

Compd #	Global min energy kcal mol ⁻¹	# Torsion bonds	# Confs within 20 kcal mol ⁻¹	Energy cutoff values for reprocessing conformer files kcal mol ⁻¹	# Confs within the energy cutoff range	RMS (torsion) cluster analysis			# Clusters obtained
						Min threshold index	Max threshold index	Chosen threshold index	
1a	17.6	4	6,191	25	676	0	159	59	22
1b	58.761	3	1,575	–	1,575	0	180	50	23
1c	58.603	5	73,139	68	992	0	180	76	23
1d	39.504	5	13,713	49	142	0	180	63	23
1e	46.226	3	1,369	–	1,369	0	180	59.7	24
2a	25.429	3	1,329	–	1,329	17	180	65.58	22
2b	38.243	2	144	–	144	0	180	46	24
2c	41.461	4	7,678	51	1,721	0	180	60	24
2d	49.357	4	5,528	60	466	0	180	55.37	23
2e	44.875	2	131	–	131	20	180	46	21
3a	23.067	3	1,329	–	1,329	17	180	66.86	22
3b	35.885	2	144	–	144	21	180	43	22
3c	38.805	4	7,754	48	1,433	0	180	66.3	21
3d	47.001	4	5,529	57	367	0	180	56.8	21
3e	44.201	2	130	–	130	0	180	43.69	21
4a	25.746	3	1,334	–	1,334	17	180	67	22
4b	38.415	2	144	–	144	0	180	45.85	22
4c	41.311	4	7,706	51	1,640	0	180	69.2	22
4d	49.44	4	5,520	59	322	0	180	46.15	24
4e	46.747	2	132	–	132	0	180	44.05	21
5a	26.516	3	1,293	–	1,293	17	180	61.87	24
5b	37.274	2	135	–	135	21	180	45	24
5c	39.433	4	5,938	49	1,106	0	180	60	23
5d	47.68	4	4,330	65	1,508	0	180	65	21
5e	50.8	2	135	–	135	20	180	45	22
6a	28.44	5	14,607	38	1,904	0	180	69	24
6b	39.852	4	5,898	49	1,073	0	180	66	21
6c	42.259	6	19,837	52	1,272	0	180	78	21
6d	50.351	6	8,674	60	207	0	180	55	22
6e	47.64	4	3,747	57	740	0	180	60	24
7a	18.964	4	6,490	27	1,910	0	180	60	24
7b	30.089	3	1,532	–	1,532	16	174	65	24
7c	32.245	5	2,0320	42	1,805	0	180	80	23
7d	40.337	5	13,258	50	199	0	180	55	22
7e	37.351	3	1,289	–	1,289	16	180	62	21
8a	18.667	3	1,273	–	1,273	16	180	62	21
8b	29.784	2	144	–	144	21	180	45	22
8c	32.599	4	5,872	43	702	0	180	60	23
8d	43.155	4	3,972	62.5	1,832	0	180	65	23
8e	40.03	2	144	–	144	20	180	45	23

provides the seven ADME descriptors, polar surface area (ADME_PSA_2D), intestinal absorption values as a multivariate distance (T2) from the center of the polar surface area (PSA) - ellipse surface (ADME_Absorption_T2_2D), log P values calculated based on the Ghosh and Crippen atom-types [58, 59] (ADME_AlogP68), blood–brain barrier ratios (ADME_BBB_2D), the

corresponding blood–brain barrier penetration level (ADME_BBB_level_2D), aqueous solubility at 25 °C (ADME_Solubility) and aqueous solubility ranking (ADME_Solubility_level). The Electrotopological State Descriptors (E-state) included in this study are S_sCH3, S_ssCH2, S_aaCH, S_sssCH, S_dssC, S_aasC, S_ssNH, S_sssN, S_dO, S_ssO and S_sCl. The meaning of the

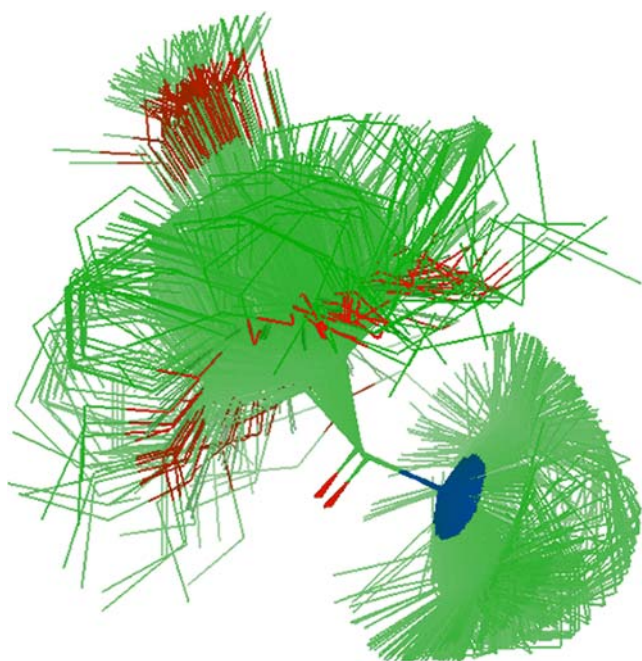


Fig. 1 Overlay of all 940 conformers showing the alignment

E-state symbol S_{xxx} is that it is the sum of type of xxx, where 'x' can be 's': single bond, 'd': double bond, 't': triple bond and 'a': aromatic bond. (For example, S_{aasC} stands for the sum descriptor for carbon with two aromatic bonds and one single bond). The values for atomic E-state indices are described by Kier et al. [60] The thermodynamic descriptors included are *n*-octanol/water partition coefficient (LogP), the desolvation free energy for water (Fh2o), the desolvation free energy for *n*-octanol (Foct), the partition coefficient computed on atom types reported by Ghosh et al. [58, 59] (AlogP and AlogP98), the molar refractivity (MR) computed based on refractive index, molecular weight, compound density and the molar refractivity (MolRef) computed based on the atom-types with additive contributions reported by Ghosh et al. [58, 59] The nine Ghosh and Crippen atom type descriptors, $Atype_C_1$, $Atype_C_2$, $Atype_C_5$, $Atype_C_6$, $Atype_C_24$, $Atype_C_25$, $Atype_H_46$, $Atype_H_47$ and $Atype_H_52$ are counts of atom types reported by Ghose et al. [58, 59] The descriptor JX is the Balaban index [13], which is evaluated by taking into account the bond orders, heteroatom electronegativities and covalent radii. The Kier's shape indices included in this study are $Kappa-1$, $Kappa-2$, $Kappa-3$, $Kappa-1-AM$, $Kappa-2-AM$ and $Kappa-3-AM$. These indices capture different aspects of molecular shape by comparing the molecular graph with minimal and maximal graphs [11]. The last three indices are refinements of the first three by taking into account the covalent radii and hybridization states [61].

A total of fifty-five different 3D-descriptors were calculated for all the conformers. The thirty Jurs descriptors

based on partial charges mapped onto the surface area were reported by Stanton et al. [17] The descriptors included are as follows:

- a) The total molecular solvent accessible surface (Jurs-SASA).
- b) The sum of the solvent-accessible surface area of all partially positively charged atoms (Jurs-PPSA-1).
- c) The sum of the solvent-accessible surface area of all partially negatively charged atoms (Jurs-PNSA-1).
- d) The difference (Jurs-DPSA-1) between the partial positive solvent accessible area (PPSA-1) and partial negative solvent accessible surface area (PNSA-1).
- e) The partial positive solvent-accessible surface area times the total positive charge (Jurs-PPSA-2).
- f) The partial negative solvent-accessible surface area times the total negative charge (Jurs-PNSA-2).
- g) The difference (Jurs-DPSA-2) between Jurs-PPSA-2 and Jurs-PNSA-2.
- h) The sum of the products of solvent accessible surface area and partial charge for all positively charged atoms (Jurs-PPSA-3).
- i) The sum of the products of solvent accessible surface areas and partial charge for all negatively charged atoms (Jurs-PNSA-3).
- j) The difference (Jurs-DPSA-3) between Jurs-PPSA-3 and Jurs-PNSA-3.
- k–p) These six descriptors are the fractionally charged surface areas, Jurs-FPSA-1, Jurs-FNSA-1, Jurs-FPSA-2, Jurs-FNSA-2, Jurs-FPSA-3 and Jurs-FNSA-3, which are obtained by dividing each of the descriptor PPSA-1, PNSA-1, PPSA-2, PNA-2, PPSA-3 and PNSA-3 by total molecular solvent-accessible surface area (SASA) respectively.
- q–v) These six descriptors are the surface-weighted charged partial surface areas, Jurs-WPSA-1, Jurs-WNSA-1, Jurs-WPSA-2, Jurs-WNSA-2, Jurs-WPSA-3 and Jurs-WNSA-3, which are obtained by multiplying each of the descriptors PPSA-1, PNSA-1, PPSA-2, PNSA-2, PPSA-3 and PNSA-3 by SASA and dividing by 1,000 respectively.
- w) The Jurs-RPCG descriptor is the relative positive charge computed by dividing the charge of the most positive atom by the total positive charge.
- x) The Jurs-RNCG descriptor is the relative negative charge computed by dividing the charge of the most negative atom by the total negative charge.
- y) The Jurs-RPCS descriptor is the relative positive charge surface area, which is computed as the solvent-accessible surface area of the most positive atom divided by RPCG.
- z) The Jurs-RNCS descriptor is the relative negative charge surface area, which is obtained by dividing

Table 5 List of descriptors with correlation of less than 0.1 with Bioactivity (BA)

Descriptor	Abs(BA)
ADME_AlogP98	0.00722
ADME_Solubility	0.01544
AlogP98	0.00722
Area	0.02170
Atype_C_1	0.02218
Atype_C_2	0.00378
Atype_C_24	0.01618
Atype_C_25	0.00829
CHI-3_C	0.02531
CHI-3_P	0.03172
CHI-V-0	0.02719
CHI-V-1	0.03795
CHI-V-2	0.03881
CHI-V-3_C	0.02744
Dipole-mag	0.02350
IC	0.01588
Jurs-PPSA-1	0.00020
MolRef	0.03867
PHI	0.01783
Rotlbonds	0.01565
S_aaCH	0.00869
S_sCl	0.03094
S_ssCH2	0.02771
SC-3_C	0.01202
Shadow-XZ	0.01320
Shadow-XZfrac	0.00400
Shadow-Ylength	0.03485
SIC	0.02293
Vm	0.03713
Vap Press @ 30 °C	0.06720
AlogP	0.05078
Apol	0.07722
Atype_C_6	0.09263
Atype_H_46	0.06029
Atype_H_52	0.09255
BIC	0.04456
CHI-0	0.07652
CHI-V-3_P	0.04719
CIC	0.04263
HOMO	0.04317
HOMO_MOPAC	0.04508
IAC-Total	0.09304
Jurs-DPSA-1	0.05433
Jurs-RNCS	0.03958
Jurs-SASA	0.08612
Jurs-WPSA-1	0.04565
Kappa-1	0.06079
Kappa-1-AM	0.04699
Kappa-2	0.08430
Kappa-2-AM	0.06206
LogP	0.05800
MR	0.09360
S_dO	0.04796
S_sCH3	0.05810
S_sssCH	0.07320

Table 5 (continued)

Descriptor	Abs(BA)
SC-3_P	0.08261
Shadow-XYfrac	0.04764
Shadow-YZfrac	0.07132
Sr	0.04534

the solvent-accessible surface area of the most negative atom divided by RNCG.

- aa) The Jurs-TASA descriptors is the total hydrophobic surface area, which is computed as the sum of the solvent-accessible surface area of atoms with absolute partial charge less than 0.2.
- ab) The Jurs-TPSA descriptor is the total polar surface area, which is the sum of the solvent-accessible surface areas of atom with absolute partial charges greater than or equal 0.2.
- ac) The Jurs-RASA descriptor is the relative hydrophobic surface area, which is computed as the TASA divided by SASA.
- ad) The Jurs-RPSA descriptor is the relative polar surface area, which is obtained by dividing TPSA by SASA.

The ten shadow indices are based on the surface area of molecular projections on the XY, YZ and XZ planes, as reported by Rohrbaugh et al. [16] The descriptors shadow-XY, shadow-YZ and shadow-XZ are areas of the molecular shadow in the XY, YZ and XZ planes, respectively. The descriptors shadow-Xlength, shadow-Ylength and shadow-Zlength are the lengths of the molecule in X, Y and Z dimensions, respectively. The descriptors shadow-XYfrac, shadow-YZfrac and shadow-XZfrac are fractions of the area of molecular shadows in the XY, YZ and XZ planes on the areas of the enclosing rectangles, respectively. The areas of molecular shadows are the appropriate products of X-length, Y-length and Z-length of the molecular shadows. The descriptor shadow-nu is the ratio of the largest to the smallest dimension. The four quantum-mechanical descriptors included are HOMO_MOPAC, LUMO_MOPAC, DIPOLE_MOPAC and HF_MOPAC. These are the HOMO, LUMO, dipole moment and heat of formation calculated by semiempirical methods, which are generally known to provide more accurate values. The 3D-spatial descriptors are Density and PMI-mag. The descriptor Density is defined as the ratio of molecular weight to molecular volume. The descriptor PMI-Mag is the magnitude of the principal moments of inertia about the principal axes of the conformers as described by Hill [62]. The descriptor Hf is a thermodynamic descriptor that gives the enthalpy of formation of the conformer as described by Dewar et al. [63] The conformational descriptor 'Energy' gives the energy of the conformer.

Descriptor selection

The selection of descriptors is an important first step in a QSAR study. A good correlation between the selected variables and the bioactivity implies better bioactivity predictions [64]. Several techniques for descriptor selection, to reduce dimensionality, have been reported recently. L'Heureux et al. [65], have employed local linear embedding techniques, Olah et al. [66] demonstrated the use of automated PLS search for biologically relevant descriptors, Sutter et al. [67] used a generalized simulated annealing algorithm in a computational neural network for automated descriptor selection and Zheng et al. [68] have demonstrated the use of the k -nearest neighbor principle for descriptor selection. We adapted the descriptor selection strategy reported earlier by Yao et al. [69] First, all descriptors that had very low correlations with the bioactivity ($|r| < 0.1$) were discarded. Next, the highly collinear descriptors ($|\text{cross correlation coefficient}| > 0.9$) were identified. Those descriptors with more physical significance to offer mechanistic insight into the QSAR information were retained. For example, given a choice between Jurs-DPSA-2, CHI-1, E-DIST-mag, SC-0, Weiner and Zagreb, the Jurs-DPSA-2 was retained because it provides information about the difference between the positively and negatively charged solvent-accessible surface areas.

The cross correlation matrix was computed. The descriptors that showed very poor correlation with bioactivity ($r < 0.1$) were removed. Table 5 shows the 59 descriptors discarded and their correlation coefficients with bioactivity.

The cross correlation matrix showed that 38 of the remaining 68 descriptors exhibited very high cross correlation ($|r| > 0.9$). Table 6 summarizes the descriptor types, names and their cross-correlation coefficient values. These 38 descriptors were removed to leave the following 30 final descriptors, which are presented in nine descriptor categories:

- (1) ADME descriptors: ADME_Solubility_level, ADME_BBB_2D, ADME_BBB_level_2D and ADME_Absorption_T2_2D;
- (2) E-state descriptors: S_dssC, S_ssO, S_aasC and S_ssNH;
- (3) Graph theory based descriptors: Kappa-3-AM;
- (4) Jurs descriptors: Jurs-DPSA-2, Jurs-DPSA-3, Jurs-FPSA-1, Jurs-FPSA-3, Jurs-FNSA-2, Jurs-RPCS and Jurs-RASA;
- (5) Shadow index descriptors: Shadow-XY, Shadow-nu, Shadow-Xlength and Shadow-Zlength;
- (6) Quantum mechanical and Electronic descriptors: LUMO_MOPAC, Hf_MOPAC and DIPOLE_MOPAC;
- (7) Conformational descriptors: Energy;
- (8) 3D spatial descriptor: Density and PMI-mag; and

- (9) Miscellaneous descriptors: JX, Fh2o, Atype_C_5 and Atype_H_47.

Quasi-multi-way PLS analyses

Bhonsle et al. [28] have reported the use of automated quasi-multi-way PLS analyses for CoMFA-based 3D-QSAR of cyclic pentapeptides CXCR4 inhibitors. They have mimicked multi-way-PLS analyses by employing several automated two-way-PLS analyses using the SYBYL [70] software. We have used a similar approach here. The PLS analysis procedure in C2 provides for a quick cross-validation of QSAR models. In this cross-validation procedure, only the “regression” part of the model development is cross-validated. This “regression-only” cross-validation was computed for all QSAR models generated. The non-validated r^2 and the sum of squares of predicted residuals (PRESS) were used to guide successive generations of model development. The conventional method of selecting conformers (or conversely outliers) is based on absolute residual values. This method gives an unfair advantage to the low activity compounds vis-à-vis the high activity ones. Thus, to remove this bias, we have coined the term ‘Percentage Prediction Error (PPE)’. The PPE is computed as follows:

$$PPE = \text{absolute value} (\text{Bioactivity} - \text{Predicted_Bioactivity}) \\ * 100 / \text{Bioactivity}$$

The selection of conformers for all generations of QSAR models was based on the PPE values.

The first generation QSAR model was obtained by performing a PLS analysis on 706 conformers of the thirty training set compounds. The number of conformers for each training-set compound is the number of clusters plus the global minimum conformer (see Table 4). The computed QSAR model showed a non-validated r^2 of 0.883 and sum of squares of predicted residuals (PRESS) of 200.06. The second (IInd) generation model of 501 conformers was obtained as follows. The predicted residual values of several conformers of the same compound in the first generation model showed almost identical values. A closer examination of the descriptor values of all such conformers showed that they were also almost identical. Thus, all such ‘duplicate’ conformers were removed. The computed QSAR model showed a non-validated r^2 of 0.879 and PRESS value of 135.01. Figure 2 shows the plot of actual versus predicted bioactivity for IInd generation QSAR model.

Table 6 Highly correlated ($|r| > 0.9$) descriptors and their cross correlation coefficients

Descriptors and cross correlation coefficients					
CHI-1	CHI-2	E-ADJ-mag	E-DIST-mag	Jurs-DPSA-2	Jurs-PPSA-2
	0.92	0.958	0.986	0.901	0.849
	Jurs-WPSA-3	log Z	MW	PMI-mag	SC-0
	0.921	0.994	0.954	0.871	0.991
	V-DIST-mag	Wiener	Zagreb	Jurs-RNCG	Jurs-TASA
	0.984	0.984	0.977	-0.848	0.889
	SC-1	SC-2	V-ADJ-mag	Jurs-WPSA-2	
0.992	0.958	0.991	0.879		
CHI-2	CHI-1	E-ADJ-mag	E-DIST-mag	Jurs-DPSA-2	Jurs-TASA
	0.92	0.983	0.939	0.863	0.876
	SC-0	SC-1	SC-2	V-ADJ-mag	V-DIST-mag
	0.954	0.94	0.982	0.94	0.956
	Jurs-WPSA-2	log Z	MW	Wiener	Zagreb
0.847	0.903	0.925	0.941	0.97	
E-ADJ-mag	CHI-1	CHI-2	E-DIST-mag	Jurs-DPSA-2	Jurs-TASA
	0.958	0.983	0.976	0.855	0.898
	PMI-mag	SC-0	SC-1	SC-2	V-ADJ-mag
	0.85	0.973	0.978	1	0.978
	Jurs-WPSA-2	log Z	MW	V-DIST-mag	Wiener
	0.845	0.954	0.95	0.97	0.957
E-DIST-mag	Zagreb				
	0.996				
	CHI-1	CHI-2	E-ADJ-mag	Jurs-DPSA-2	Jurs-TASA
	0.986	0.939	0.976	0.872	0.884
	MW	PMI-mag	SC-0	SC-1	SC-2
	0.941	0.862	0.978	0.996	0.975
Zagreb	Jurs-WPSA-2	Jurs-WPSA-3	log Z	V-ADJ-mag	
0.989	0.857	0.888	0.99	0.997	
V-DIST-mag	Wiener				
0.977	0.976				
Jurs-DPSA-2	CHI-1	CHI-2	E-ADJ-mag	E-DIST-mag	Jurs-FPSA-2
	0.901	0.863	0.855	0.872	0.936
	Jurs-RPCG	Jurs-TASA	Jurs-WPSA-2	Jurs-WPSA-3	log Z
	-0.866		0.986	0.955	0.867
	SC-2	V-ADJ-mag	V-DIST-mag	Wiener	Zagreb
0.853	0.872	0.925	0.925	0.865	
Jurs-PPSA-2	Jurs-PPSA-3	Jurs-RNCG	SC-1	SC-0	
0.976	0.874	-0.937	0.872	0.918	
Jurs-PPSA-2	Jurs-WPSA-3	Jurs-DPSA-2	Jurs-FPSA-2	Jurs-RNCG	Jurs-RPCG
	0.91602	0.97551	0.98663	-0.924	-0.87832
	SC-0	V-DIST-mag	Wiener	Jurs-TASA	Jurs-WPSA-2
0.86272	0.86833	0.86111	0.86197	0.9952	
Jurs-RNCG	Jurs-WPSA-3	Jurs-DPSA-2	Jurs-FPSA-2	Jurs-PPSA-2	Jurs-PPSA-3
	-0.92241	-0.9367	-0.9045	-0.924	-0.87995
	SC-0	V-DIST-mag	Wiener	Jurs-RPCG	Jurs-WPSA-2
-0.85604	-0.85231	-0.84985	0.89619	-0.92139	
Jurs-RPCG	Jurs-DPSA-2	Jurs-FPSA-2	Jurs-PPSA-2	Jurs-PPSA-3	Jurs-RNCG
	-0.86588	-0.89565	-0.87832	-0.88167	0.89619
	Jurs-WPSA-2	Jurs-WPSA-3	Kappa-3-AM		
-0.85497	-0.86542	-0.87115			
Jurs-TASA	CHI-1	CHI-2	E-ADJ-mag	E-DIST-mag	Wiener
	0.88929	0.87557	0.89766	0.88447	0.86703
	log Z	MW	SC-0	SC-1	SC-2
	0.88591	0.8583	0.8954	0.89832	0.9
	Zagreb	V-DIST-mag	Jurs-WPSA-2	Jurs-WPSA-3	V-ADJ-mag
0.90406	0.88069	0.87719	0.84887	0.89641	

Table 6 (continued)

Descriptors and cross correlation coefficients

Jurs-WPSA-2	CHI-1	CHI-2	E-ADJ-mag	E-DIST-mag	Jurs-DPSA-2
	0.87864	0.84707	0.84519	0.85731	0.98641
	Jurs-RPCG	Jurs-TASA	Jurs-WPSA-3	log Z	SC-0
	-0.85497	0.87719	0.93093	0.84862	0.89487
	Wiener	Zagreb	Jurs-FPSA-2	Jurs-PPSA-2	Jurs-RNCG
	0.89622	0.85268	0.9663	0.9952	-0.92139
Jurs-WPSA-3	SC-1	V-ADJ-mag	V-DIST-mag		
	0.85563	0.85629	0.90211		
	CHI-1	E-DIST-mag	Jurs-DPSA-2	Jurs-FPSA-2	Jurs-PPSA-2
	0.92055	0.88814	0.95471	0.86693	0.91602
	Jurs-TASA	Jurs-WPSA-2	log Z	PMI-mag	SC-0
	0.84887	0.93093	0.9046	0.84588	0.90874
log Z	Wiener	Zagreb	Jurs-PPSA-3	Jurs-RNCG	Jurs-RPCG
	0.92032	0.8628	0.96366	-0.92241	-0.86542
	SC-1	V-ADJ-mag	V-DIST-mag		
	0.89301	0.89272	0.90388		
	CHI-1	CHI-2	E-ADJ-mag	E-DIST-mag	Jurs-DPSA-2
	0.99398	0.90294	0.95359	0.98971	0.86659
MW	MW	PMI-mag	SC-0	SC-1	SC-2
	0.93493	0.87162	0.97427	0.99493	0.95414
	Zagreb	Jurs-TASA	Jurs-WPSA-2	Jurs-WPSA-3	V-ADJ-mag
	0.97557	0.88591	0.84862	0.9046	0.9946
	V-DIST-mag	Wiener			
	0.96398	0.96688			
PMI-mag	CHI-1	CHI-2	E-ADJ-mag	E-DIST-mag	Jurs-DPSA-2
	0.95364	0.9255	0.95012	0.94069	0.84938
	SC-1	SC-2	V-ADJ-mag	V-DIST-mag	Wiener
	0.94614	0.95038	0.94553	0.96351	0.94636
	Jurs-TASA	log Z	SC-0	Zagreb	
	0.8583	0.93493	0.9675	0.95367	
SC-0	CHI-1	E-ADJ-mag	E-DIST-mag	Jurs-WPSA-3	log Z
	0.87081	0.85008	0.86244	0.84588	0.87162
	SC-2	V-ADJ-mag	V-DIST-mag	Wiener	Zagreb
	0.8503	0.87084	0.85115	0.87987	0.86331
	RadOfGyration	SC-0	SC-1		
	0.90049	0.86366	0.87152		
SC-1	CHI-1	CHI-2	E-ADJ-mag	E-DIST-mag	Jurs-DPSA-2
	0.99114	0.95426	0.9731	0.97755	0.91846
	Jurs-WPSA-2	Jurs-WPSA-3	log Z	MW	PMI-mag
	0.89487	0.90874	0.97427	0.9675	0.86366
	V-DIST-mag	Wiener	Zagreb	Jurs-PPSA-2	Jurs-RNCG
	0.99572	0.99084	0.98206	0.86272	-0.85604
SC-2	SC-1	SC-2	V-ADJ-mag	Jurs-TASA	
	0.98253	0.97317	0.98213	0.8954	
	CHI-1	CHI-2	E-ADJ-mag	E-DIST-mag	Jurs-DPSA-2
	0.99154	0.93997	0.978	0.99595	0.87168
	log Z	MW	PMI-mag	SC-0	SC-2
	0.99493	0.94614	0.87152	0.98253	0.97827
Jurs-DPSA-2	Zagreb	Jurs-TASA	Jurs-WPSA-2	Jurs-WPSA-3	V-ADJ-mag
	0.99216	0.89832	0.85563	0.89301	0.99988
	V-DIST-mag	Wiener			
	0.97441	0.9728			
	CHI-1	CHI-2	E-ADJ-mag	E-DIST-mag	Jurs-DPSA-2
	0.95827	0.98214	0.99985	0.97515	0.85291
Jurs-RNCG	PMI-mag	SC-0	SC-1	V-ADJ-mag	V-DIST-mag
	0.8503	0.97317	0.97827	0.97826	0.96855

Table 6 (continued)

Descriptors and cross correlation coefficients

	Jurs-TASA	log Z	MW	Wiener	Zagreb
	0.9	0.95414	0.95038	0.95559	0.99651
V-ADJ-mag	CHI-1	CHI-2	E-ADJ-mag	E-DIST-mag	Jurs-DPSA-2
	0.99112	0.94037	0.97823	0.99721	0.87212
	log Z	MW	PMI-mag	SC-0	SC-1
	0.9946	0.94553	0.87084	0.98213	0.99988
	Zagreb	Jurs-TASA	Jurs-WPSA-2	Jurs-WPSA-3	SC-2
	0.9921	0.89641	0.85629	0.89272	0.97826
	V-DIST-mag	Wiener			
	0.97532	0.97391			
V-DIST-mag	CHI-1	CHI-2	E-ADJ-mag	E-DIST-mag	Jurs-DPSA-2
	0.98351	0.95572	0.96989	0.97724	0.92468
	Jurs-WPSA-2	Jurs-WPSA-3	log Z	MW	PMI-mag
	0.90211	0.90388	0.96398	0.96351	0.85115
	V-ADJ-mag	Wiener	Zagreb	Jurs-PPSA-2	Jurs-RNCG
	0.97532	0.99485	0.976	0.86833	-0.85231
	Jurs-TASA	SC-0	SC-1	SC-2	
	0.88069	0.99572	0.97441	0.96855	
Wiener	CHI-1	CHI-2	E-ADJ-mag	E-DIST-mag	Jurs-DPSA-2
	0.98425	0.9412	0.95714	0.97623	0.92535
	Jurs-WPSA-2	Jurs-WPSA-3	log Z	MW	PMI-mag
	0.89622	0.92032	0.96688	0.94636	0.87987
	V-ADJ-mag	V-DIST-mag	Zagreb	Jurs-PPSA-2	Jurs-RNCG
	0.97391	0.99485	0.96754	0.86111	-0.84985
	Jurs-TASA	SC-0	SC-1	SC-2	
	0.86703	0.99084	0.9728	0.95559	
Zagreb	CHI-1	CHI-2	E-ADJ-mag	E-DIST-mag	Jurs-DPSA-2
	0.97669	0.97034	0.99632	0.98865	0.86495
	log Z	MW	PMI-mag	SC-0	SC-1
	0.97557	0.95367	0.86331	0.98206	0.99216
	Wiener	Jurs-TASA	Jurs-WPSA-2	Jurs-WPSA-3	SC-2
	0.96754	0.90406	0.85268	0.8628	0.99651
	V-ADJ-mag	V-DIST-mag			
	0.9921	0.976			
ADME_PSA_2D	Fh2o	Foct	Fh2o	ADME_PSA_2D	Foct
	-0.922	-0.885		-0.922	0.902
Jurs-WNSA-3	Jurs-DPSA-3	Jurs-PNSA-2	Jurs-PNSA-3	Jurs-WNSA-1	Jurs-WNSA-2
	-0.9275	0.91715	0.88193	-0.85335	0.89103
Jurs-RPSA	Jurs-RASA	Jurs-TPSA	Jurs-WNSA-2	Jurs-DPSA-3	Jurs-PNSA-2
	-1	0.94917		-0.84799	0.96571
Jurs-FPSA-2	Jurs-DPSA-2	Jurs-PPSA-2	Jurs-RNCG	Jurs-RPCG	Jurs-WPSA-2
	0.9361	0.98663	-0.9045	-0.89565	0.9663
Jurs-PNSA-2	Jurs-DPSA-2	Jurs-PPSA-2	Jurs-RNCG	Jurs-RPCG	Jurs-WPSA-2
	0.9361	0.98663	-0.9045	-0.89565	0.9663
Jurs-PNSA-3	Jurs-FNSA-2	Jurs-FNSA-3	Jurs-WNSA-3	LUMO	LUMO_MOPAC
	0.87575	0.87464	0.88193		0.94779
Jurs-DPSA-3	Jurs-PNSA-2	Jurs-WNSA-2	Jurs-WNSA-3	LUMO_MOPAC	LUMO
	-0.847	-0.848	-0.928		0.94779
Jurs-FNSA-1	Jurs-FNSA-3	Jurs-FPSA-1	Jurs-PNSA-1	S_sssN	S_ssNH
	-0.85712	-1	0.92642		-0.98072
Jurs-FNSA-3	Jurs-FNSA-1	Jurs-FPSA-1	Jurs-PNSA-3	S_ssNH	S_sssN
	-0.85712	0.85712	0.87464		-0.98072
Jurs-FPSA-1	Jurs-FNSA-1	Jurs-FNSA-3	Jurs-PNSA-1	HF_MOPAC	Hf
	-1	0.85712	-0.92642		0.9664
Shadow-YZ	Shadow-Zlength	Shadow-Zlength	Shadow-YZ	Hf	HF_MOPAC
	0.87557		0.87557		0.966

Table 6 (continued)

Descriptors and cross correlation coefficients					
Jurs-RASA	Jurs-RPSA	Jurs-TPSA	RadOf Gyration	PMI-mag	Shadow-Xlength
	–1	–0.94917		0.90049	0.88054
Jurs-TPSA	Jurs-RASA	Jurs-WPSA-3	Foet	ADME_PSA_2D	Fh2o
	–0.94917	0.86693		–0.885	0.902
Jurs-FNSA-2	Jurs-PNSA-1	Jurs-RPSA	Kappa-3-AM	Jurs-RPCG	Kappa-3
	–0.91545	0.94917		–0.87115	0.94634
Jurs-PNSA-1	Jurs-FNSA-1	Jurs-PNSA-2	Jurs-PNSA-3	Jurs-WNSA-1	
	0.92642	0.92624	0.87575	–0.94703	
Jurs-WNSA-1	Jurs-FNSA-2	Jurs-FNSA-2	Jurs-FPSA-1	Jurs-WNSA-1	
	–0.94703	–0.91545	–0.92642	0.9098	
Jurs-PPSA-3	Jurs-DPSA-2	Jurs-PNSA-1	Jurs-PNSA-2	Jurs-WNSA-3	
	0.8736	0.9098	–0.93312	–0.85335	
Kappa-3	Kappa-3-AM	Jurs-RNCG	Jurs-RPCG	Jurs-WPSA-3	
	0.94634	–0.87995	–0.88167	0.96366	

Selection of best, worst and moderate performing conformers approach

In the quasi-multi-way PLS analyses approach reported by Bhonsle et al. [28] the two least residual value conformers were selected for the intermediate QSAR model, followed by the selection of the least residual value conformer for the final QSAR model. Thus, our first attempt was to try selecting six conformers of each compound, such that two would have the least residual values, two with the worst residual values and two with the mean or median residual values. The idea behind this approach is that the PLS regression analysis will have the complete gamut of the descriptor values to be able to create a broadly predictive QSAR model. The third (IIIrd) generation model had 180 conformers with r^2 of 0.885 and PRESS of 57.52. The fourth (IVth) generation model obtained by selecting three conformers for each compound, with best residual value, worst residual value and the mean or median residual value. This model of 90 conformers showed r^2 of 0.881 and PRESS of 36.55. Two fifth (Vth) generation models of 60 conformers were obtained using the following two approaches. In the first approach, conformers showing the best and the worst residual values were selected to give a model with r^2 of 0.852 and PRESS of 46.50. In the second approach, conformers showing the best and the mean or median residual values were selected to give a model with r^2 of 0.934 and PRESS of 22.77. The final or sixth (VIth) generation models from the above two approaches were obtained by selecting one conformer for every compound with the best residual values. Both of these models, with 30 conformers, showed poor internal validation (leave-one-out regression-only cross-validation) correlation coefficient values. The model built *via* the first approach gave non-validated r^2 of 0.896 and q^2_{LOO} of 0.450. While, the model obtained from the second approach showed non-validated r^2 of 0.873 and q^2_{LOO} of 0.415.

Stepwise, gradual, worst residual value conformer elimination r^2 and PRESS-guided conformer selection approach

Since the steep approach of selecting the best, worst and moderate performing conformers failed to provide a good QSAR model, we tried a more gentle approach. We eliminated the worst residual value conformer in a stepwise and gradual fashion. We used the non-validated r^2 and PRESS as measures to guide the model improvement.

The IIIrd generation model of 300 conformers was obtained by selecting 10 least PPE values conformers and it showed a non-validated r^2 of 0.921 and PRESS value of 60.43. The IVth generation model of 150 conformers was constructed with the five least residual value conformers from the IIIrd generation model. This model displayed a non-validated r^2 of 0.965 and PRESS value of 15.12. The Vth generation model of 60 conformers was obtained with two least residual value conformers and it exhibited non-validated r^2 of 0.988 and PRESS value of 3.10. The VIth generation model was constructed by eliminating the worst residual value conformers of all compounds with PT less than 3.0. For the remaining nine compounds, **2b** (4.0), **3c** (4.0), **3d** (3.0), **5b** (5.0), **5c** (3.0), **6c** (3.50), **7c** (6.0), **8b** (3.0) and **8c** (4.0), two conformers each were retained in the model. This QSAR model had non-validated r^2 of 0.991 and PRESS value of 2.565. At this juncture, there were 18 conformers from which the best set of nine conformers could be chosen in 512 (2^9) ways. We used a Tcl-based Cerius2 script [55] to compute these 512 seventh (VIIth) generation models. We found six models with leave-one-out (regression-only) cross-validated r^2 of 0.67 or larger. The best VIIth generation QSAR model showed a non-validated r^2 of 0.989, leave-one-out (regression-only) cross-validated r^2 of 0.701 and PRESS value of 20.37. Figure 3 shows the best final QSAR model.

IIInd Generation 501 rows model

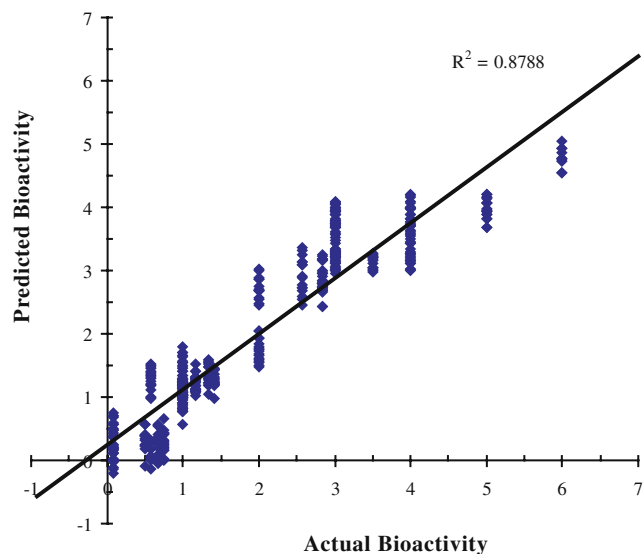


Fig. 2 Second-generation QSAR model with 501 conformers

The statistical data of the six QSAR models selected (Models A–F) with the selected conformer number of the nine compounds **2b**, **3c**, **3d**, **5b**, **5c**, **6c**, **7c**, **8b** and **8c** is shown in Table 7. The statistical data and selected conformer numbers for QSAR model G built using the original pool of 127 descriptors are also included in Table 7. It is noteworthy that in the six models selected all but four (**2b**, **5b**, **6c** and **8c**)

VIIth Generation QSAR Model A

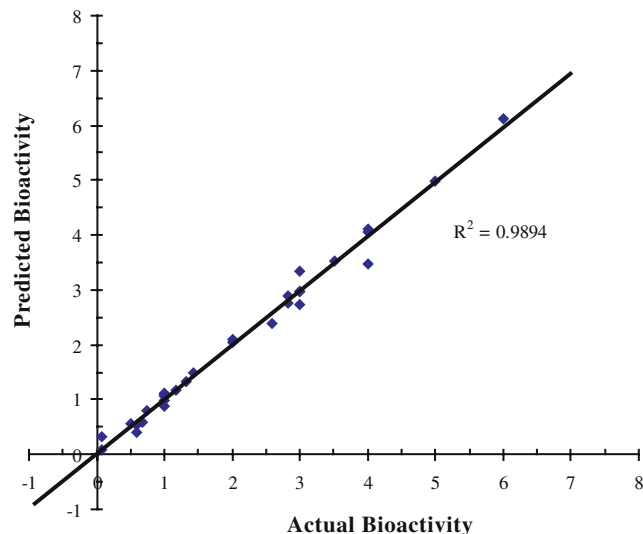


Fig. 3 Best QSAR Model A of thirty training set compounds

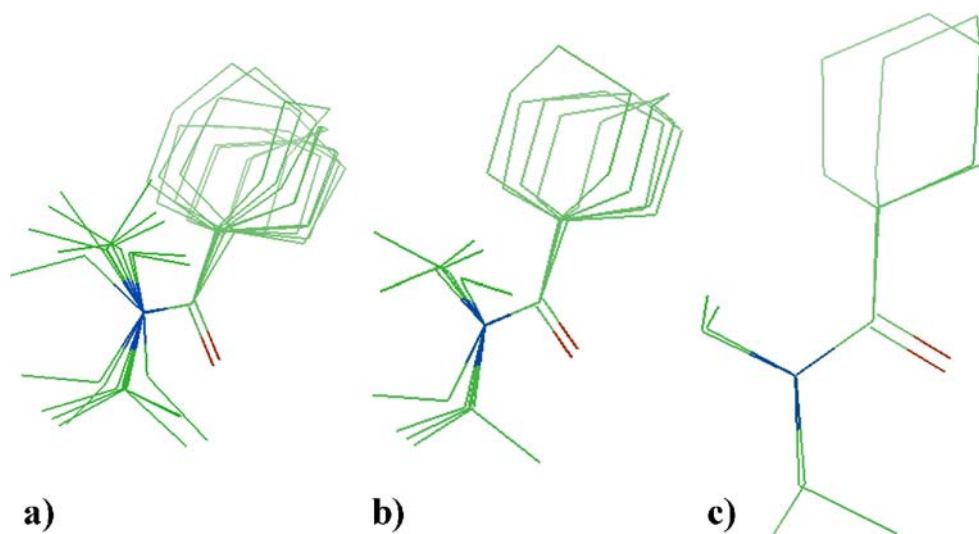
compounds have the same conformer numbers. The overlay of the conformers of compounds **2b**, **5b**, **6c** and **8c** show that the substitutions on the amidic N are spatially very close to each other. A representative overlay of conformer **8c_0** and **8c_5** is shown in Fig. 4c.

The statistical data of all the seven generation models obtained during the QSAR model development phases are shown in Table 8.

Table 7 Statistical data of selected 30 descriptors six (A–F) models and 127 descriptors (G) Model

Model #	Conformer # of 21 compounds common to all six(A–F) Models	Conformer # selected for the 9 active/very active compounds	NV r^2	Leave-one-out (regression only) cross-validated	
				q^2	PRESS
A	1b_21; 1c_6; 1d_18; 1e_24; 2a_11; 2c_3; 2e_16; 3a_0; 4a_0; 4b_2; 4e_12; 5a_16; 5e_4; 6a_1; 6b_7; 6e_4; 7a_22; 7e_5; 8a_5; 8d_5; 8e_23	2b_13; 3c_15; 3d_6; 5b_1; 5c_5; 6c_4; 7c_15; 8b_14; 8c_5	0.989	0.701	20.37
B		2b_13; 3c_15; 3d_6; 5b_22; 5c_5; 6c_16; 7c_15; 8b_14; 8c_5	0.988	0.674	22.2
C		2b_13; 3c_15; 3d_6; 5b_22; 5c_5; 6c_16; 7c_15; 8b_14; 8c_0	0.991	0.673	22.28
D		2b_13; 3c_15; 3d_6; 5b_1; 5c_5; 6c_16; 7c_15; 8b_14; 8c_5	0.989	0.674	22.21
E		2b_13; 3c_15; 3d_6; 5b_1; 5c_5; 6c_16; 7c_15; 8b_14; 8c_0	0.991	0.673	22.29
F		2b_7; 3c_15; 3d_6; 5b_22; 5c_5; 6c_16; 7c_15; 8b_14; 8c_5	0.988	0.674	22.24
G	1b_17; 1c_7; 1d_9; 1e_5; 2a_9; 2b_3; 2c_4; 2e_5; 3a_9; 3c_13; 3d_15; 4a_14; 4b_9; 4e_17; 5a_4; 5b_2; 5c_4; 5e_2; 6a_2; 6b_8; 6c_13; 6e_5; 7a_10; 7c_2; 7e_8; 8a_12; 8b_7; 8c_3; 8d_1; 8e_12		0.984	0.719	19.12

Fig. 4 a) IIIrd Generation **8c** Conformers. b) IVth Generation **8c** Conformers. c) Vth Generation **8c** Conformers



The final best QSAR model A with the selected conformers, predicted bioactivities and percent prediction errors is shown in Table 9.

It is noteworthy that QSAR Model A predicts all (thirteen) of the potent (PT>2.5 h) insect repellents within a PPE of 14%. Of these thirteen compounds, eleven are within a PPE of 10% and nine, quite accurately, within a PPE of 3%. The algorithm used for discovering the bioactive conformers, leading to QSAR models A–F is summarized in Fig. 5.

The QSAR models/equations

Bioactive conformer mining and insights into the mechanism of action

The gradual and stepwise refinement of successive generation QSAR models by selecting the least PPE (residual) value conformers affords the conformers that correlate best with the observed bioactivity. Thus, these selected conformers are the bioactive conformations of the respective compounds. On examination of the ten, five and two selected conformers in the IIIrd, IVth and Vth generation models respectively, we find that our novel methodology selects the cluster of

Table 8 Statistical data of all generation QSAR models

QSAR model generation number	Number of Conformers in model	Non-validated r^2	Leave-one-out cross-validated (regression only) q^2	PRESS
1	706	0.883	0.877	200.06
2	501	0.879	0.869	135.01
3	300	0.921	0.911	60.43
4	150	0.965	0.956	15.12
5	60	0.988	0.977	3.10
6	39	0.991	0.974	2.565
7	30	0.989	0.701	20.37

Table 9 Best QSAR Model A selected conformers predicted bioactivities and percent prediction errors

Compound_Conformer #	Actual bioactivity	Model # A predicted bioactivity	Percent error in model A predicted bioactivity=abs(BA–PredBA)*100/BA
1b_21	1	0.888	11.17
1c_6	1	0.975	2.53
1d_18	1.17	1.158	1.02
1e_24	0.75	0.803	7.10
2a_11	0.08	0.310	288.01
2b_13	4	4.108	2.69
2c_3	2.83	2.755	2.65
2e_16	1	1.096	9.60
3a_0	0.58	0.395	31.96
3c_15	4	4.065	1.62
3d_6	3	2.979	0.69
4a_0	0.67	0.588	12.18
4b_2	3	2.720	9.34
4e_12	1.42	1.499	5.56
5a_16	0.58	0.547	5.64
5b_1	5	4.988	0.24
5c_5	3	2.974	0.86
5e_4	1	1.121	12.10
6a_1	0.08	0.087	9.14
6b_7	2.83	2.892	2.18
6c_4	3.5	3.531	0.87
6e_4	1.33	1.320	0.78
7a_22	1	1.068	6.85
7c_15	6	6.114	1.90
7e_5	2.58	2.400	6.99
8a_5	0.5	0.560	12.08
8b_14	3	3.352	11.73
8c_5	4	3.465	13.36
8d_5	2	2.089	4.46
8e_23	2	2.054	2.71

bioactive conformers and further refines the cluster over successive generations, finally to yield the best bioactivity correlating conformer. This phenomenon is illustrated in Figs. 4, 6, 7 and 8. Figure 4 shows the IIIrd, IVth and Vth generation selected conformers of compound **8c** (PT=4) of activity class E. The same observations are also shown for compound **3d** (PT=3) of activity class D in Fig. 6 and for compound **8d** (PT=2) of activity class C in Fig. 7.

Figure 8a–c show the overlay of selected conformers for the VIIth generation model A of all compounds of the high activity classes C, D and E, respectively.

As is evident from Figs. 4, 6, 7 and 8 that the α -to-amide C–N bonds ‘a’ ‘b’ and ‘c’ ‘d’ (see Fig. 9) form a cluster within a 60 degree range (e.g. Figs. 4a, 6a and 7a) and the

cluster is enriched with the bioactive conformers with narrower angle ranges for bonds ‘a’ ‘b’ and ‘c’ ‘d’ (e.g. see Figs. 4b, 6b, 7b and 4c, 6c, 7c) over the successive QSAR model generations.

The shapes of the selected conformers over the successive generations also allude to the roles of various moieties around the putative amide pharmacophore in the mechanism of action and the structure activity relationship. Closer study of the cluster of conformers in the successive generation models, their shapes around the amide group and their PPE values gave the following observations.

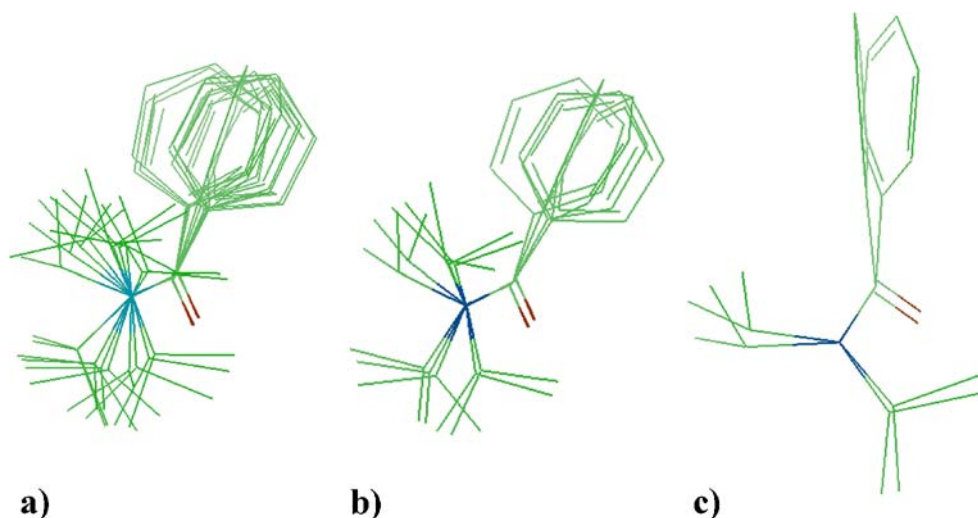
- i) The 3D-spatial location of the phenyl, benzyl or cyclohexyl moiety attached to the carbonyl C does not have any significant effect on the bioactivity.

Fig. 5 Flow chart of algorithm for QSAR model development

Flow Chart of Algorithm for QSAR model development

1. Build & minimize Structures.
2. Grid Scan Conformational Search.
3. Cluster Conformations on RMS of torsions to get 22 - 24 clusters. Select the cluster Nuclei as representative conformer.
4. Compute 2D & 3D descriptors.
5. Perform Cross Correlation Analysis of the descriptors.
6. Remove descriptors with less than 0.1 correlations to bioactivity. Remove descriptors having greater than 0.9 cross correlation.
7. Perform PLS analysis and Compute PPE to get 1st Generation Model.
8. Examine conformers with identical PPE values. If descriptor values are nearly identical, remove these duplicate conformers.
9. (i) Perform PLS & compute PPE to get 2nd Generation Model.
(ii) Sort Conformers for each compound based on PPE.
(iii) Select 10 conformers for each compound with the least PPE value.
10. (i) Perform PLS & compute PPE to get 3rd Generation Model.
(ii) Sort Conformers for each compound based on PPE.
(iii) Select 5 conformers for each compound with the least PPE value.
Remove the other 5 conformers.
11. (i) Perform PLS & compute PPE to get 4th Generation Model.
(ii) Sort Conformers for each compound based on PPE.
(iii) Select 2 conformers for each compound with the least PPE value.
Remove the other 3 conformers.
12. (i) Perform PLS & compute PPE to get 5th Generation Model.
(ii) Sort Conformers for each compound based on PPE.
(iii) For 21 compounds with PT < 3.00, select the least PPE conformer.
For 9 compounds with PT > 3.00, keep both the conformers.
13. (i) Perform automated PLS analysis on 2⁹ = 512 models obtained by selecting each combination of the conformer of the 9 compounds.
(ii) Select those QSAR Models which have $q^2_{LOO} > 0.67$
14. (i) Validate the QSAR Models with external test set.
(ii) Validate the QSAR Models with Fischer Randomization Test.

Fig. 6 a) IIIrd Generation **3d** Conformers. b) IVth Generation **3d** Conformers. c) Vth Generation **3d** Conformers



- ii) The bioactive conformers of all compounds show preferential positioning of the methyl, ethyl, isopropyl etc. moieties on the amidic N within a range of 60 to 70 degrees. (See Figs. 4c, 6c, 7c and 8c) Thus, the shape on the amidic N side of the molecule is critical for bioactivity.
- iii) The narrower the PPE of the conformers, the closer are the α -to-amide C–N bonds ‘a’, ‘b’ and ‘c’, ‘d’ (see Fig. 9) to each other. For example, for compound **8c** (PT=4) (PPE 0.65%–7.4%) see Fig. 4a–c. Thus, suggesting that particular 3D-spatial location of alkyl and cycloalkyl groups on the nitrogen are important for bioactivity.

The elementary requirement for the odorant to possess insect repellency is that it has a high enough vapor pressure for a significant number of molecules to reach the arthropods. Suryanarayana et al. [41] and Johnson et al. [71] reported that odorants with too low or too high vapor pressures than DEET (**4c**) exhibit poor PT. The mechanism of action of odorants on arthropods is reported to comprise three steps [35], which are detailed as follows. The first step is that of the odorant (Od) binding to the lipophilic odorant binding

protein (OBP). The complex (Od–OBP) then binds with the G-coupled protein receptor (GPCR) on the neuron cells in the second step, giving rise to the repellency action. In the third step, the Od–OBP degrading protein (ODP) degrades the Od–OBP complex to prevent continued stimulus to the neurons. Our SAR observations indicate that the second and third steps could compete. Thus if the Od–OBP complex is not strong enough or if the Od–OBP complex cannot bind/dock effectively with the GPCR, then ODP would degrade the Odorant faster than the Od–OBP complex binding to the GPCR, thus resulting in poor PT.

The OBPs are usually about 20 kDa in size and comprise a single peptide with six characteristic highly conserved cysteine residues and hydrophobic domain residues between residues numbers 40 and 60 [72, 74]. The hydrophobic region of the compounds is primarily on the carbonyl side of the amide. We believe that the phenyl, benzyl and cyclohexyl moieties attached to the carbonyl group dock with the OBPs between residues 40 and 60. This step being the first step in the mechanism of action is important in the

Fig. 7 a) IIIrd Generation **8d** Conformers. b) IVth Generation **8d** Conformers. c) Vth Generation **8d** Conformers

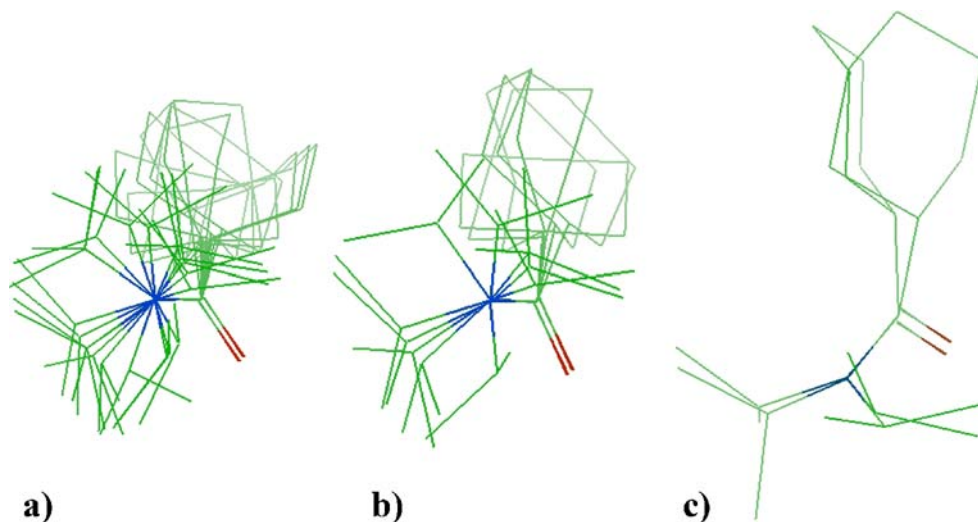
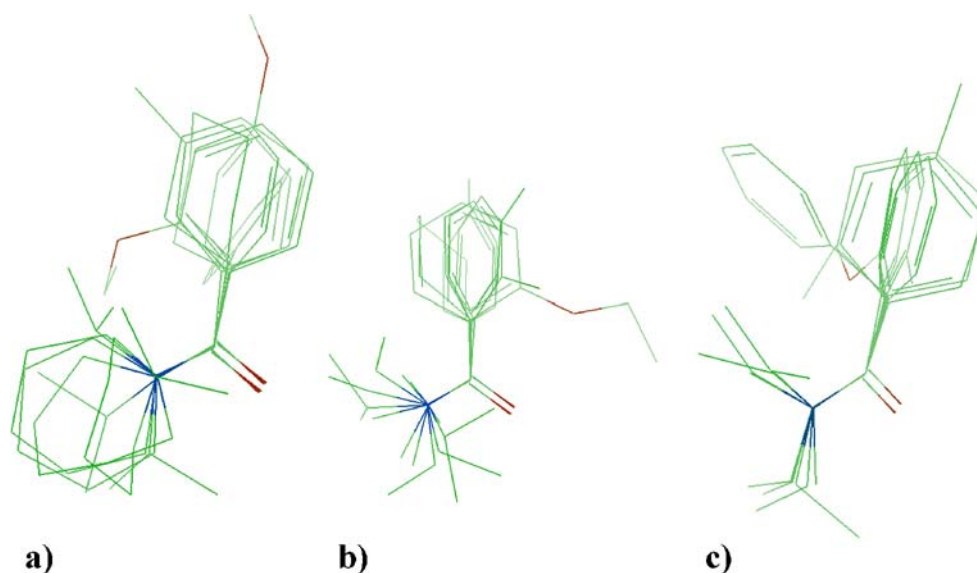


Fig. 8 a) QSAR Model A Activity class C Conformers (1.11<PT<2.6). b) QSAR Model A Activity class D Conformers (2.61<PT<3.0). c) QSAR Model A Activity class E Conformers (3.01<PT<6.0)



overall SAR. Thus, compounds that do not have sufficient hydrophobic groups on the carbonyl side would show poor activity. The SAR observations supporting this theory are as follows. All five *para*-methoxy phenyl compounds, four of the five *ortho*-ethoxy phenyl compounds and three of the five *ortho*-chloro phenyl compounds exhibit PTs less than 3.0 h. The fact that **5b** and **5c** show PT of 3.0 h or larger indicates that smaller sized *ortho* polar substituents are tolerable for the hydrophobic pocket of the OBP.

The second step in the mechanism of action is the binding of the Od–OBP complex to the GPCR on the neuron cells. We suggest that the alkyl and cycloalkyl moieties on the amidic N probably interact with the GPCR and that this, being the crucial step for repellency, would be the ‘rate determining step’ in the mechanism of action. Thus, any compound that does not have favorable substitution on the amidic nitrogen in terms of 3D-spatial and physico-chemical requirements would exhibit poor PT. For example, disubstitution at the amidic N is crucial for bioactivity, thus all Xa (monosubstituted ethylamine) compounds consistently exhibit poor PTs of 0.05 to 1.0 h. Further, all diisopropyl and cyclohexyl substituted amides (except **3c** and **3d**) exhibit poor PT (PT<3.0). All diethyl and

dimethyl substituted amides (except **1b**, **1c**, **3b** and **7b**) exhibit PT higher than 2.83 with six out of twelve exhibiting PT larger than 4.0 h. The poor PT of **1b** and **1c** have already been explained as they, probably, bind poorly with OBP in the first step, while the poor PT of **3b** and **7b** could be attributed to poor vapor pressures at 30 °C of 0.0015 and 0.0020, respectively (DEET VP @ 30 °C=0.026).

QSAR equation analysis

The best QSAR model A is described by the following equation:

$$\begin{aligned}
 \text{PLS Predicted Bioactivity} = & 0.53848 * \\
 & \text{ADME_Absorption_T2_2D} - 0.682301 * \text{ADME_BBB_} \\
 & \text{2D} - 0.689156 * \text{ADME_BBB_Level_2D} - 1.20977 * \\
 & \text{ADME_Solubility_Level} - 0.04213 * \text{Energy} - 0.531331 * \\
 & S_dssC - 0.192429 * S_aasC - 0.367471 * S_ssNH + \\
 & 0.054425 * S_ssO + 0.00082393 * \text{LUMO_MOPAC} + \\
 & 0.432683 * \text{DIPOLE_MOPAC} + 0.0041452 * \\
 & \text{HF_MOPAC} - 0.00016215 * \text{Jurs_DPSA-2-} \\
 & 0.014325 * \text{Jurs_DPSA-3} + 1.28755 * \text{Jurs_FPSA-} \\
 & 1 + 0.530519 * \text{Jurs_FNSA-2} + 66.4923 * \text{Jurs-} \\
 & \text{FPSA-3} + 0.536246 * \text{Jurs_RPCS} + 12.5082 * \text{Jurs-} \\
 & \text{RASA} - 0.008205 * \text{Shadow_XY} - 0.530922 * \text{Shadow-} \\
 & \text{nu} - 0.285412 * \text{Shadow_Xlength} - 0.05695 * \text{Shadow-} \\
 & \text{Zlength} + 0.311959 * \text{Density} - 0.0013471 * \text{PMI-} \\
 & \text{mag} - 0.074066 * \text{Atype_C_5} + 0.195987 * \text{Atype_H-} \\
 & 47 + 0.097114 * \text{Fh2o} + 1.5147 * \text{JX} + 1.29936 * \\
 & \text{Kappa-3-AM} - 12.4913
 \end{aligned}$$

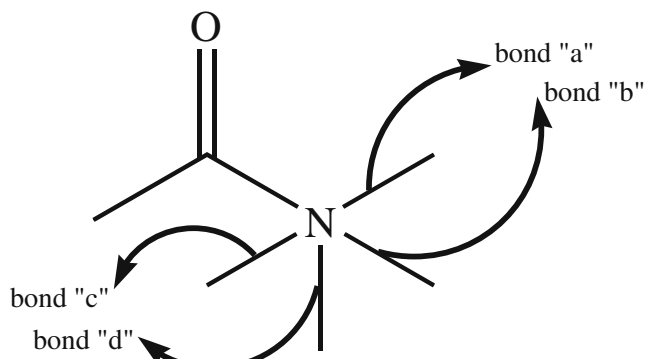


Fig. 9 α -to-amide C–N bonds

The values and signs of the QSAR equation coefficients provide a qualitative insight into the correlation of the physicochemical properties with biological activity. The quantitative contribution of any physicochemical property

to the bioactivity of the compound is judged from both its QSAR equation coefficient and the value of the descriptor quantifying the property. The product of the QSAR coefficients and the descriptor mean value

$$\left(\text{Descriptor_Mean} = \frac{\sum \text{Descriptor_values_all_training_set_compounds}}{30} \right)$$

would provide the contribution value of that descriptor to the overall bioactivity (Contribution_to_BioActivity – CtoBA).

$$\text{CtoBA} = \text{QSAR_Coefficient} * \text{Descriptor_mean_value}$$

The significance of CtoBA of any descriptor vis-à-vis the CtoBA of all the other descriptors can be computed by dividing the individual CtoBA by the sum total of all the

CtoBA of all descriptors. The percentage value of this quotient is termed as ‘Descriptor Significance Percentage – DSP’.

$$\text{DSP} = (\text{CtoBA} * 100) / \sum \text{abs}(\text{CtoBA})$$

The DSP values would provide a better insight in the quantitative contributions of the descriptors to the bioactivities of the compounds. The QSAR coefficients for the

Table 10 Computation of Descriptor Significance Percentage (DSP) for Model A

Descriptor	QEC	MVD	CtoBA	DSP
Jurs-RASA	12.508	0.882	11.028	25.038
Jurs-FPSA-3	66.492	0.074	4.916	11.161
JX	1.515	2.527	3.828	8.690
ADME_Solubility_Level	-1.210	3.100	-3.750	-8.514
Shadow-Xlength	-0.285	11.888	-3.393	-7.703
Kappa-3-AM	1.299	2.568	3.337	7.577
Energy	-0.042	51.357	-2.164	-4.912
Atype_H_47	0.196	7.933	1.555	3.530
DIPOLE_MOPAC	0.433	3.553	1.537	3.490
ADME_Absorption_T2_2D	0.538	2.736	1.473	3.344
Shadow-nu	-0.531	1.909	-1.014	-2.301
Jurs-FPSA-1	1.288	0.770	0.992	2.252
ADME_BBB_Level_2D	-0.689	1.300	-0.896	-2.034
Jurs-DPSA-3	-0.014	50.052	-0.717	-1.628
Shadow-XY	-0.008	59.513	-0.488	-1.109
Fh2o	0.097	-4.866	-0.473	-1.073
PMI-mag	-0.001	324.800	-0.438	-0.993
Shadow-Zlength	-0.057	6.306	-0.359	-0.815
Density	0.312	1.004	0.313	0.711
Jurs-RPCS	0.536	0.488	0.262	0.595
S_ssNH	-0.367	0.642	-0.236	-0.536
Jurs-FNSA-2	0.531	-0.438	-0.232	-0.528
S_aasC	-0.192	1.148	-0.221	-0.501
Jurs-DPSA-2	0.000	808.217	-0.131	-0.298
S_ssO	0.054	1.394	0.076	0.172
ADME_BBB_2D	-0.682	0.103	-0.071	-0.160
S_dssC	-0.531	0.114	-0.061	-0.138
HF_MOPAC	0.004	-11.390	-0.047	-0.107
Atype_C5	-0.074	0.533	-0.040	-0.090
LUMO_MOPAC	0.001	0.094	0.000	0.000

QEC: QSAR Model A Equation Coefficient values

MVD: Mean value of descriptors of all training compounds = $(\sum \text{descriptor_value}/30)$

CtoBA: Contribution of the descriptor to bioactivity = $(\text{QEC} * \text{MVD})$

DSP: Descriptor Significance Percentage = $(\text{CtoBA} * 100 / \sum \text{abs}(\text{CtoBA}))$

QSAR model A, CtoBA and DSP values are shown in Table 10.

The largest contribution to the bioactivity is from the descriptor Jurs-RASA (Jurs-Relative Hydrophobic Surface Area) with a value of 25% and with a positive effect. Jurs-RASA is defined as the ratio between the total hydrophobic surface area (Jurs-TASA) and the total solvent accessible surface area (Jurs-SASA). Thus, molecules with larger hydrophobic surface area and smaller total solvent-accessible area would exhibit high potency. This observation is consistent with the first step in the mechanism of action of the molecules binding to the OBP, where hydrophobicity/lipophilicity plays a key role. The other QSAR model A coefficients that support the role of hydrophobicity in the mechanism of action are ADME_Solubility_Level with a negative contribution of 8.5%, Atype_H_47 with a positive contribution of 3.5% and Fh2o with a negative contribution of 1.1%. These observations agree with the observations of McIver et al. [34] and Suryanarayana et al. [41] that lipophilicity is directly related to repellency. The CATALYST-based pharmacophore model [44] is also consistent with this observation because potent insect repellents require three hydrophobic sites in the molecules for activity.

The next largest contribution to the bioactivity is from the descriptor Jurs-FPSA-3 (Jurs-Fractional Positive Surface Area 3) with a value of 11% and a positive effect. Jurs-FPSA-3 is the quotient of the Jurs PPSA-3 and Jurs-SASA. Jurs-PPSA-3 (Jurs-Partial Positive Surface Area 3) is the summation of the products of solvent-accessible surface area and partial charge of all positively charged atoms. Thus, QSAR model A indicates that molecules with larger partial positive surface areas and larger partial positive charges along with smaller total solvent accessible surface areas would show high activity. This probably refers to the second step in the mechanism of action where the odorant–OBP complex binds to the neuron GPCR; where the partial positively charged amidic N of the odorant molecules is involved in hydrogen bonding with the GPCR peptide residues. The diffuse or soft positively charged moiety's contribution towards bioactivity is also supported by the following QSAR model A coefficients. Jurs-FPSA-1 (Jurs-Fractional Positive Surface Areas-1), defined as the sum of the solvent-accessible area of all partial positively charged atoms, has a positive contribution of 2.3%. The Balaban index JX, which is inversely proportional to the electronegativities and covalent radii of the atoms in the repellent molecules, shows a positive contribution of 8.7%. Jurs-RPCS (Jurs-Relative Positively Charged Surface Area) and Jurs-FNSA-2 (Jurs-Fractional Negatively Charged Surface Areas-2) exhibit positive 0.6% and negative 0.5% contributions, respectively. The importance of soft positively charged moieties, optimal atomic charges and dipole moments has been reported by Ma et al. [42] in their

electrostatic potential studies of DEET analogs. The contribution of appropriate partial positive and negative charge separation in the QSAR model A is evident in the coefficients DIPOLE_MOPAC with a positive 3.5% contribution and Jurs-DPSA-3 (Jurs-Differential Partial Positive Solvent-accessible Surface Area-3) with a negative 1.6% contribution. The optimal values of dipole moment for the compounds would be relevant in both the first step of odorant–OBP complex formation and in the second odorant–OBP complex binding to the neuron GPCR.

The fifth largest DSP contribution to the QSAR model A is from Shadow-Xlength, which is the projection measure of the repellent compound on the *x*-axis, with a negative 7.7% contribution. The contributions of other shadow indices to the QSAR model A are as follows: Shadow-Zlength (projection measure on the *z*-axis) negative 0.8%, Shadow-XY (the area of the shadow of the molecule in the XY plane) negative 1.1% and Shadow-nu (ratio of the largest to the smallest shadow measures) negative 2.3%. This combination of shadow indices indicates that molecules with an elongated rectangular box (parallelepiped)-like structure would be more potent than other shapes. This also alludes to the shape of the binding pocket of the OBP involved in the first step of the mechanism of action. The other significant QSAR model A coefficients that contribute to this observation are as follows. Kappa-3-AM, which is directly proportional to the number of vertices and inversely proportional to the number of edges in the molecular graph, has larger values for larger but denser molecules. Kappa-3-AM has a positive 7.6% contribution. Density (molecular density) has a positive contribution of 0.7%, while PMI-mag (Principal moment of inertia magnitude) has a negative 0.99% contribution. These observations also agree with those of Suryanarayana et al. [41] and Wright et al. [40] that molecular size and shape have a large effect on repellency activity. Further, the pharmacophore reported by Bhattacharjee et al. [44] also has a parallelepiped (rectangular) shape with the aliphatic hydrophobic moiety at one end, the aromatic hydrophobic moiety at the other end and the hydrogen bond acceptor moiety around the central portion of the parallelepiped. A closer study of the shapes of the conformers selected for the QSAR model A, which correlate best with the observed bioactivities, gave the following observations. The molecular shape on the amidic nitrogen side of the molecule is more critical to the bioactivity than the carbonyl side of the amidic moiety. Thus, compounds with a piperidine moiety like **1e**, **2e**, **3e** etc. consistently exhibit poor (PT=0.08) to moderate (PT=2.58) bioactivity. In case of the 1 series compounds, viz. **1a**, **1b**, **1c** etc. the methoxy group on the aromatic ring falls outside the favorable area of the parallelepiped, thus showing poor (max PT=1.17) bioactivity. For compounds of the Xa or Xd series, the

terminal methyl group on the ethyl or isopropyl moieties falls in the disfavored region on the critical amidic nitrogen side, thus consistently exhibiting poor (PT=0.08) to moderate (PT=2.0) with the exceptions of **4d** (2.67) and **3d** (3.0).

The rest of the compounds that have relatively favorable 3D-spatial disposition with regard to the parallelepiped show bioactivities of 3.00 or larger with the exceptions of **2c** (PT=2.83), **6b** (PT=2.83) and **7b** (PT=2.17).

The other significant (DSP>1%) QSAR model A coefficients are Energy (conformational energy) with a negative 4.9% contribution, ADME_Absorption_T2_2D with a positive 3.3% contribution and ADME_BBB_Level_2D with a negative 2% contribution. Our QSAR study showed poor correlation between molar refractivity (MR and MolRef) and repellency, unlike that reported by Suryanarayana et al. [41].

QSAR Models A–G Validation

Internal validation tests

Internal validation (cross-validation) tests of selected QSAR models (see Table 11) were performed at three levels. All models showed $q^2_{LOO}>0.983$ for the leave-one-out cross-validation tests. For the leave-10%-out (leave-three-out) cross-validation tests, four models viz. A, C, D

and E showed $q^2_{L10O}>0.98$, whereas models B, F and G showed q^2_{L10O} values of 0.978, 0.976 and 0.955, respectively. Five models viz. A–E showed $q^2_{L20O}>0.96$ for the leave-20%-out (leave-six-out) cross-validation tests, while models F and G showed q^2_{L20O} values of 0.705 and 0.884 respectively.

QSAR model validation by randomization tests

It is known that even with large number of observations and fewer terms, QSAR models can be poorly predictive. Thus, with fewer observations (in this study thirty compounds) and many more terms (in this study one hundred and twenty seven descriptors down selected to thirty), QSAR models are prone to chance correlation. In the randomization test, the dependent variables (bioactivity values) are randomly reassigned to different compounds and new regression models are recomputed. This process is repeated several times. If the statistical data of these randomized models is comparable to the QSAR model developed, then the QSAR model developed is not predictive and the number of observations is insufficient. We performed two sets of randomization tests of ninety-nine trials each at 99% confidence level for all QSAR models A–G. The results of the randomization tests are shown in Table 11. The best mean random r value obtained for models A–F is 0.133 ($r^2=0.018$) and model G is 0.240 ($r^2=0.058$). The best

Table 11 Validation tests results six (A–F) models built using selected 30 descriptors and (G) Model built using 127 descriptors

Internal validation tests								
Validation tests	Model #	A	B	C	D	E	F	G
Leave-one-out	q^2	0.989	0.988	0.991	0.989	0.991	0.988	0.983
	PRESS	0.759	0.801	0.631	0.762	0.593	0.795	1.127
Leave-10%-out	q^2	0.980	0.978	0.991	0.988	0.984	0.976	0.955
	PRESS	1.329	1.469	0.586	0.805	1.089	1.644	3.051
Leave-20%-out	q^2	0.963	0.977	0.978	0.978	0.981	0.705	0.884
	PRESS	2.552	1.570	1.472	1.450	1.280	20.110	7.933
Randomization tests								
99 trails at 99% confidence level								
((# Random r) > (non-Random r))=0								
r from non-Random								
Test 1	Mean value of r from random trial	0.071	0.111	0.100	0.113	0.112	0.133	0.240
	Std deviation of random trial	0.192	0.248	0.233	0.244	0.226	0.253	0.311
Test 2	Mean value of r from random trial	0.087	0.103	0.110	0.118	0.105	0.131	0.225
	Std deviation of random trial	0.219	0.222	0.248	0.255	0.233	0.257	0.312
External validation tests								
All 10 test set compounds	Predictive r^2	0.335	0.333	0.334	0.319	0.333	0.334	0.228
	$s(y)$	1.321	1.321	1.330	1.326	1.333	1.319	1.251
	F -value	4.028	4.002	4.012	3.743	3.996	4.007	2.361
8 compounds (w/o compounds 1a and 7d)	Predictive r^2	0.845	0.663	0.669	0.651	0.676	0.666	0.219
	$s(y)$	0.242	0.418	0.416	0.410	0.410	0.415	0.608
	F -value	32.764	11.789	12.127	11.208	12.520	11.960	1.684

random r value possible (based on the standard deviation) is about 0.4 ($r^2=0.16$) for models A–F and about 0.55 ($r^2=0.303$) for model G. These correlation-coefficient values are far lower than the non-Random r values of 0.995 ($r^2=0.99$), thus indicating that the QSAR models A–G are not obtained by chance.

External validation tests

The selection of bioactive conformation of the test compounds for activity prediction is challenging. Bhonsle et al. [28] have reported the use of a predictive r^2 approach to demonstrate that there always is a test compound conformer within the energy range of 5 kcal mol⁻¹ of the global minima that accurately predicts the bioactivity. We used the same predictive r^2 approach to discover the best conformer that predicted the bioactivity most accurately.

On the test set of ten compounds, all the QSAR models A–F showed predictive r^2 values of 0.33 and variance $s(y)$ values of 1.3, while model G showed predictive r^2 and $s(y)$ values of 0.23 and 1.25, respectively. Two low-activity test compounds viz. **1a** (PT=0.08) and **7d** (PT=1.0) were consistently found to be the outliers. The poor activity of **1a** and **7d** could be ascribed to their low vapor pressures of 0.0062 and 0.0014 at 30 °C as compared to that of DEET (0.026). Thus, justifiably, removing them from the test set models B–F gave good predictive r^2 values of 0.65 to 0.67 and $s(y)$ values of about 0.41, while model A gave the best predictive r^2 value of 0.845 with the smallest variance $s(y)$ value of 0.242. Model G performed poorly on the eight test compounds with predictive r^2 and $s(y)$ values of 0.219 and 0.608, respectively. The F -values for the ten test compounds regression for all models A–F are between 3.7 and 4.0. The reported [75] F -values for $\alpha=0.10$ (90% confidence level) for ten observations is 3.46. Thus, the predictions of all the models A–F on the ten test compounds are statistically significant with less than 10% probability that the null hypothesis is true. The reported [75] F -values for $\alpha=0.025$ (97.5% confidence level) for eight observations is 8.81. The predictions of all the models A–F on the eight test compounds are statistically significant with confidence levels larger than 98% for F -values ranging from 11.2 to 32.7. The prediction correlation of model A is the strongest with F -values of 32.8, which is within the 99.5% confidence level based on the reported F -value for $\alpha=0.005$ (99.5% confidence level) of 18.63. The F -values for model G of 2.4 and 1.7 for the ten and eight set compounds, respectively, indicate that the prediction regression is also less than 90% statistically significant. The best conformer numbers, predicted bioactivities and the PPE for QSAR Model A on the test set compounds are shown in Table 12.

Table 12 Best QSAR Model A external test set validation results

Compd #	Actual bioactivity	Best available conformer number	Best conformer predicted bioactivity	Percent error in best conformer predicted bioactivity=(BA–PredBA) *100/BA
1a	0.08	19	–1.781	2325.64
2d	0.50	1	1.782	256.48
3b	1.67	1	1.964	17.62
3e	3.00	0	2.810	6.35
4c	5.00	12	3.585	28.30
4d	2.67	17	2.663	0.28
5d	1.00	19	2.267	126.67
6d	1.08	11	2.362	118.72
7b	2.17	5	2.154	0.72
7d	1.00	18	3.749	274.89

GFA and G/PLS models based on mined bioactive conformers of QSAR model A

The GFA- and G/PLS-based QSAR models built using the global minimum conformers and the selected pool of thirty descriptors showed poor predictive performance vis-à-vis models A–G (see Tables 3 and 11). In order to investigate if the selected conformers in QSAR model A are indeed the bioactive conformers, we built GFA- and G/PLS-QSAR models based on these conformers using the selected pool of thirty descriptors. The statistical data of the GFA and G/PLS models is shown in Table 13. The GFA- and G/PLS-models showed non-validated r^2 of 0.989 and 0.991, respectively, and excellent (regression-only) cross-validated q^2_{LOO} of 0.949 and 0.981, respectively. Both models showed superior cross-validated q^2 on all the internal validation tests for leave-one-out, leave-10%-out and leave-20%-out of 0.924 or larger.

The GFA model showed good q^2_{L200} of 0.723. The randomization tests showed the best random mean r values for the GFA and G/PLS models as 0.503 ($r^2=0.253$) and 0.814 ($r^2=0.663$), with the possible approximate random r values (based on the standard deviation of random trials) of 0.665 ($r^2=0.442$) and 0.885 ($r^2=0.783$), respectively. In the external validation tests on the test set of ten compounds, both the GFA- and G/PLS-models exhibited poor predictive r^2 and variance $s(y)$ values of 0.375 and 1.335 for the GFA-model and 0.363 and 1.302 for the G/PLS-model. The F -value of 4.803 and 4.551 compared to 3.46 (at $\alpha=0.10$ or 90% confidence level) indicates that the prediction correlation is statistically significant within the 90% confidence level. However, eliminating the usual outlier compounds, **1a** and **7d**, from the test set furnished extraordinary predictive r^2 and $s(y)$ values of 0.973 and 0.126, respectively, for the GFA-model, with a very strong F -value of 214.7 (reported [75] F -critical value at $\alpha=0.001$ for the 99.9% confidence

Table 13 Statistical data of GFA and G/PLS Models using Bioactive Conformer from QSAR Model A

Model #	Bioactive Conformer # from QSAR Model A	NV r^2	Leave-one-out (regression only) cross-validated	
			q^2	PRESS
GFA	1b_21; 1c_6; 1d_18; 1e_24; 2a_11; 2b_13; 2c_3; 2e_16; 3a_0; 3c_15; 3d_6; 4a_0;	0.989	0.949	3.45
G/PLS	4b_2; 4e_12; 5a_16; 5b_1; 5c_5; 5e_4; 6a_1; 6b_7; 6c_4; 6e_4; 7a_22; 7c_15; 7e_5; 8a_5; 8b_14; 8c_5; 8d_5; 8e_23	0.991	0.981	1.295
Validation results				
Internal validation				
Tests	Model	GFA	G/PLS	
Leave-one-out	q^2	0.933	0.991	
	PRESS	4.585	0.591	
Leave-10%-out	q^2	0.924	0.992	
	PRESS	5.189	0.560	
Leave-20%-out	q^2	0.723	0.969	
	PRESS	18.891	2.086	
Randomization tests				
99 trails at 99% confidence level				
((# Random r) >= (non-Random r))=0				
	Test set 1		Test set 2	
Model	GFA	G/PLS	GFA	G/PLS
r from non-random	0.971	0.996	0.973	0.996
Mean value of r from random trials	0.503	0.814	0.470	0.808
Std deviation of random trial	0.162	0.071	0.152	0.066
External test set validation tests				
	For 10 compounds		For 8 compounds (w/o 1a and 7d)	
Model	GFA	G/PLS	GFA	G/PLS
Predictive r^2	0.375	0.363	0.973	0.687
$s(y)$	1.335	1.302	0.126	0.432
F -value	4.803	4.551	214.733	13.189

level is 35.51) indicating a highly statistically significant correlation within the 99.9% confidence level. The G/PLS-model shows fair predictive r^2 and $s(y)$ values on the eight-compound test set of 0.687 and 0.432, respectively, with a good F -value of 13.189 (reported [75] F -critical value at $\alpha=0.025$ for 97.5% confidence level is 8.81).

The plots of actual vs. predicted bioactivity for model A, the GFA-model and the G/PLS-model are shown in Fig. 10.

The best activity-predicting conformer numbers for all the models are shown in Table 14. It is noteworthy that among the ten test-set compounds, six compounds have the same conformer number as the best activity predictors. Of the remaining four compounds, three compounds have only two conformers as best predictors. Only compound **4d** has three different best predicting conformers for the three models. An overlay of the best predicting conformers of **7b** show near perfect overlap, the **3b** conformers show within 30° angle separation and **4d** conformers show 40° angle separation of the alkyl group on the amidic N (see Fig. 9).

The descriptor significance percentage (DSP) computation for the GFA and G/PLS is shown in Table 15. The GFA- and G/PLS-models are in 94% agreement with each other, except on descriptor *Atype_H_47*, where the GFA-model shows a negative 4% contribution and the G/PLS-model shows a positive 3.6% contribution. The GFA-model is in agreement with model A, having 80% concurrence in positive and negative DSP effects, while the G/PLS-model concurs with the model A on 95% of the DSP values. The G/PLS-model shows disagreement with model A on the descriptor *Jurs-FPSA-1* with a negative 5% contribution, while model A has a positive 2.3% contribution. The GFA-model showed disagreement with model A on three descriptors, viz. *Atype_H_47*, *ADME_Absorption_T2_2D* and *Jurs-FPSA-1* with negative 4%, negative 2.9% and negative 14.2% contributions, while model A exhibited positive 3.5%, positive 3.3% and positive 2.3%, respectively. The first two descriptors, *Atype_H_47* and *ADME_Absorption_T2_2D* relate to hydrophobicity/lipophilicity, while the latter two *ADME_Absorption_T2_2D* and *Jurs-FPSA-1* relate to polar surfaces in the repellent

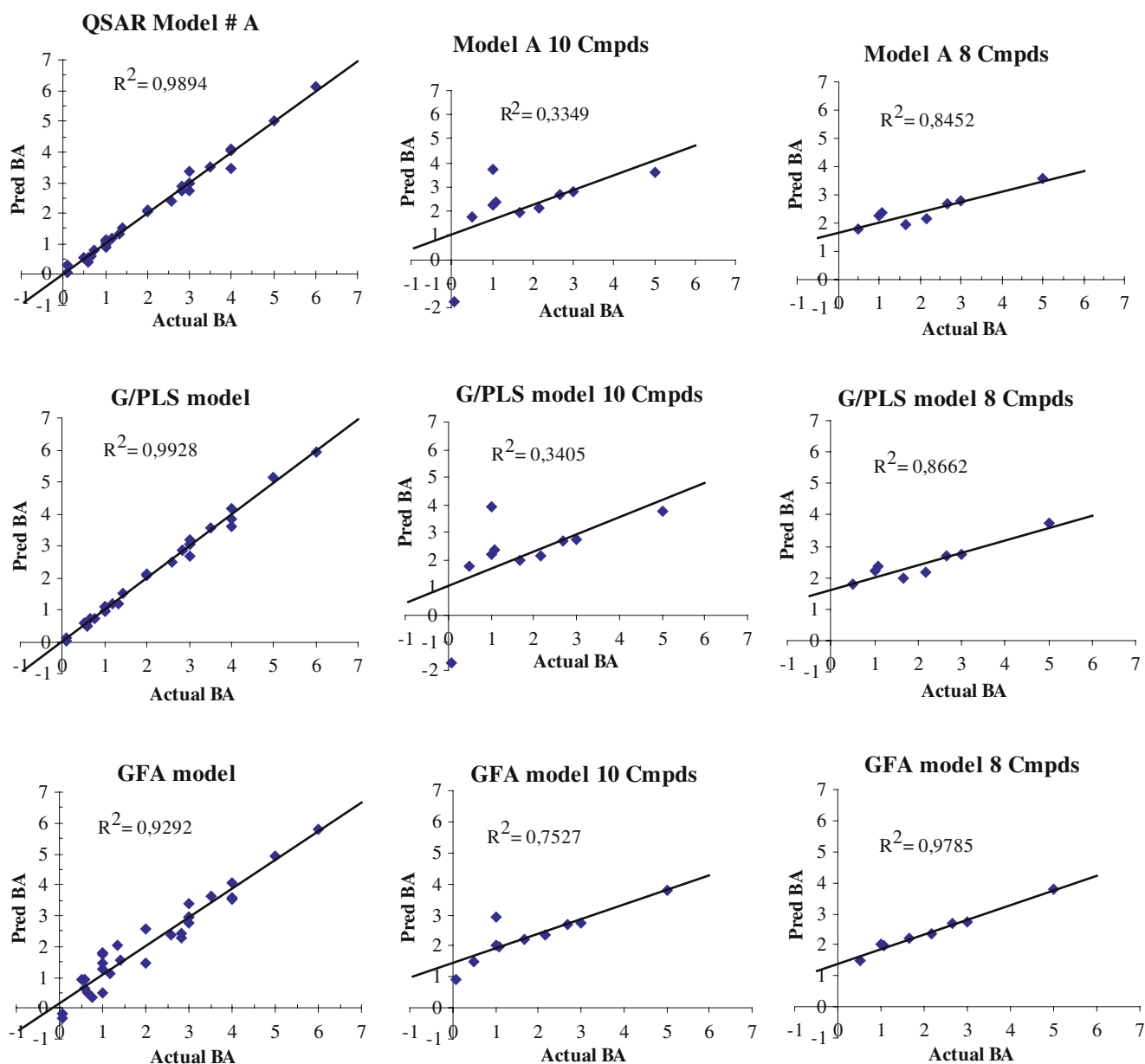


Fig. 10 Plots of actual bioactivity vs. predicted bioactivity for model A, GFA model and G/PLS

Table 14 Best activity predicting conformer numbers

Test compound	QSAR Model # A	GFA model	G/PLS model
1a	19	13	19
2d	1	1	1
3b	1	17	17
3e	0	0	0
4c	12	12	12
4d	17	14	9
5d	19	19	19
6d	11	11	11
7b	5	21	5
7d	18	18	18

molecules. As discussed in the previous section, the four aspects involved in the first two steps of the mechanism of action of the repellents are hydrophobicity/lipophilicity, positively charged surface area, positively and negatively charge separation or dipole and lastly molecular shape. Both the G/PLS and model A indicate that the soft or diffused positively charged surface on the repellent molecules contribute about 17 to 22% to the bioactivity, whereas the GFA-model suggests that this effect does not contribute positively to the bioactivity. The GFA-model disagrees with the other two models on the contribution towards the hydrophobicity/lipophilicity aspect from the descriptors Jurs-RASA and Atype_H_47. The disagreements among the three models could be ascribed to the significance each model associates to each of the steps in the mechanism of

Table 15 DSP computation for GFA and G/PLS Models built with conformers of model A

Descriptor	MVD	GFA			G/PLS			PLS
		QEC	CtoBA	DSP	QEC	CtoBA	DSP	DSP
Jurs-RASA	0.882	25.501	22.484	34.887	13.777	12.147	26.360	25.038
Jurs-FPSA-3	0.074	42.501	3.142	4.876	80.961	5.986	12.989	11.161
JX	2.527	1.663	4.202	6.520	1.980	5.004	10.859	8.690
ADME_Solubility_Level	3.100	-1.081	-3.350	-5.198	-0.781	-2.420	-5.252	-8.514
Shadow-Xlength	11.888	-0.589	-7.006	-10.871	-0.451	-5.365	-11.641	-7.703
Kappa-3-AM	2.568	1.025	2.632	4.084	0.810	2.081	4.517	7.577
Energy	51.357	-0.066	-3.410	-5.291	-0.067	-3.433	-7.451	-4.912
Atype_H_47	7.933	-0.331	-2.627	-4.077	0.209	1.655	3.591	3.530
DIPOLE_MOPAC	3.553	0.665	2.364	3.667	0.408	1.448	3.142	3.490
ADME_Absorption_T2_2D	2.736	-0.671	-1.835	-2.848	-	-	-	3.344
Shadow-nu	1.909	-	-	-	-	-	-	-2.301
Jurs-FPSA-1	0.770	-11.853	-9.130	-14.167	-3.248	-2.502	-5.429	2.252
ADME_BBB_Level_2D	1.300	-	-	-	-0.762	-0.991	-2.150	-2.034
Jurs-DPSA-3	50.052	-	-	-	-0.049	-2.472	-5.364	-1.628
Shadow-XY	59.513	-	-	-	-	-	-	-1.109
Fh2o	-4.866	-	-	-	-	-	-	-1.073
PMI-mag	324.800	-	-	-	-	-	-	-0.993
Shadow-Zlength	6.306	-	-	-	-	-	-	-0.815
Density	1.004	-	-	-	-	-	-	0.711
Jurs-RPCS	0.488	-	-	-	-	-	-	0.595
S_ssNH	0.642	-1.766	-1.134	-1.760	-0.598	-0.384	-0.833	-0.536
Jurs-FNSA-2	-0.438	-	-	-	-	-	-	-0.528
S_aasC	1.148	-0.294	-0.337	-0.524	-0.169	-0.194	-0.421	-0.501
Jurs-DPSA-2	808.217	-	-	-	-	-	-	-0.298
S_ssO	1.394	-	-	-	-	-	-	0.172
ADME_BBB_2D	0.103	-5.604	-0.580	-0.900	-	-	-	-0.160
S_dssC	0.114	-	-	-	-	-	-	-0.138
HF_MOPAC	-11.390	0.019	-0.213	-0.331	-	-	-	-0.107
Atype_C_5	0.533	-	-	-	-	-	-	-0.090
LUMO_MOPAC	0.094	-	-	-	-	-	-	0.000

QEC: QSAR Equation Coefficient values

MVD: Mean value of descriptors of all the training set compounds ($\sum \text{descriptor_value}/30$)

CtoBA: Contribution of the descriptor to bioactivity (QEC*MVD)

DSP: Descriptor Significance Percentage ($\text{CtoBA} * 100 / \sum \text{abs}(\text{CtoBA})$)

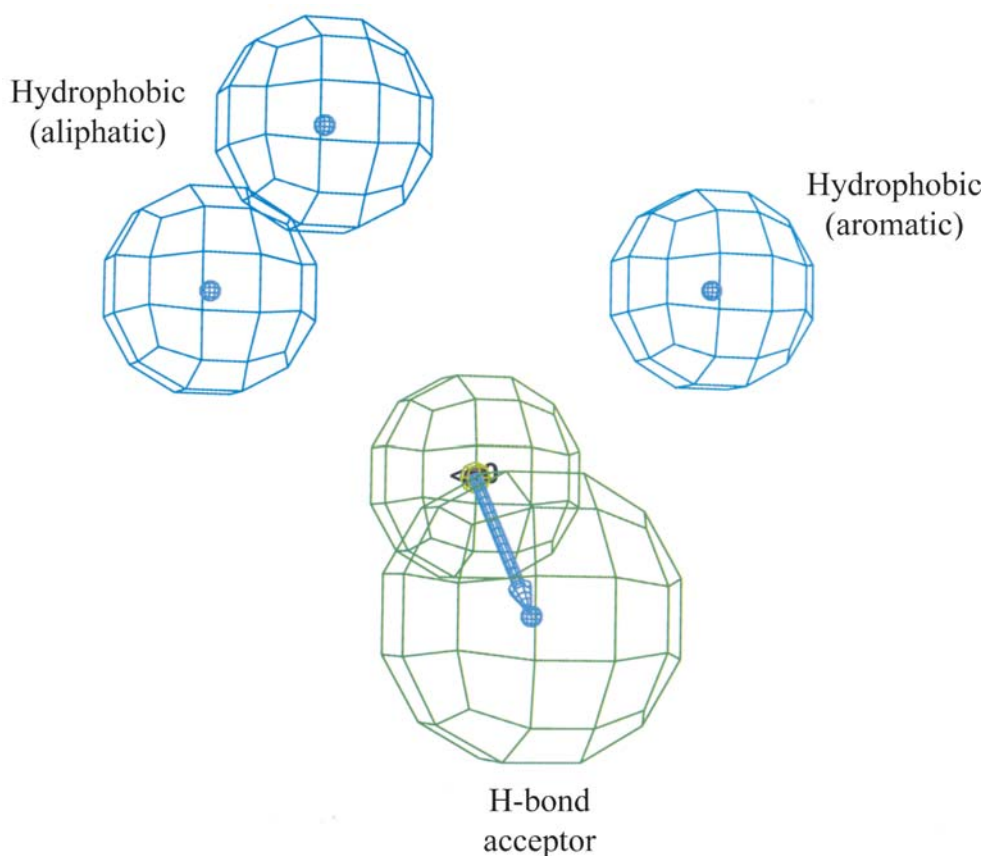
action towards the overall bioactivity. The GFA-model suggests that both hydrophobicity/lipophilicity and positively charged surface area play lesser roles in the overall bioactivity, while both the G/PLS-model and model A suggest otherwise. For the positively charged surface area property, the G/PLS-model suggests that, although this property does play a role in the overall bioactivity, the descriptor Jurs-FPSA-1 does not contribute positively, unlike Jurs-FPSA-3 and JX. Model A suggests that all Jurs-FPSA-3, JX and jurs-FPSA-1 contribute positively towards this property.

QSAR model comparison with earlier reported model developed using CATALYST

The QSAR model A was found to be qualitatively consistent with the earlier reported pharmacophore model. The CATA-

LYST based protocol reported earlier [44] resulted in the generation of ten pharmacophores. The correlation coefficients were found to range from 0.91 to 0.87 for six of the ten models. The total costs of the pharmacophores varied over a narrow range (45 to 51) and the difference between the fixed cost and the null cost was 71 bits, satisfying the acceptable range as recommended in the cost analysis of the CATALYST procedure [56]. A difference of 71 bits between the fixed and the null cost clearly indicates the robustness of the correlation. Moreover, as the cost difference between the first to the tenth hypothesis and the null hypothesis was found to be between 66 and 60 bits, it could be expected that for all these hypotheses there is a 80–92% chance of representing a true correlation in the data. The Fischer randomization test as implemented in the CatScramble module of CATALYST gave nineteen random spreadsheets from the training set. Sixteen of the randomized models

Fig. 11 CATALYST based pharmacophore model for insect repellency



generated required a total cost value lower than the model under investigation, indicating an approximately 85% confidence level of our pharmacophore model.

Significantly, the best pharmacophore characterized by two hydrophobic aliphatic functions, one aromatic ring function and one hydrogen-bond acceptor function (Fig. 11) is also statistically the most relevant pharmacophore.

A total of fifteen compounds were selected for the mapping experiments with the reported pharmacophore. The five most active ($PT > 4.0$) compounds selected were **2b**, **4c**, **5b**, **7c** and **8c**. The five moderately active ($4.0 > PT > 2.8$) compounds selected were **3d**, **3e**, **5c**, **6b** and **8b**. Lastly the five inactive ($PT < 0.5$) compounds selected were **1a**, **2a**, **2d**, **6a** and **8a**. The conformers selected for mapping were the same ones as those selected for the QSAR model A. The conformers of the most active compounds mapped all the functional features of the best hypothesis with high scores, whereas the less active compounds mapped fewer of the features.

Conclusion

A highly predictive QSAR model has been built for benzamides and benzylamides employing a semi-automated quasi multi-way PLS approach. The QSAR model concurs

with the reported physicochemical properties like lipophilicity, molecular shape and size and correlation to repellency bioactivity. The novel methodology of gradual and stepwise refinement of successive generation QSAR models results in selection of bioactive conformers. The selected bioactive conformers generate far superior GFA and G/PLS QSAR models than those obtained from the global minimum conformers. The poses/shapes of the selected bioactive conformers provide valuable insight into the mechanism of action of the insect repellents. The phenyl, benzyl or cyclohexyl moieties on the carbonyl carbon are proposed to bind to the odorant binding proteins, whereas the alkyl and cycloalkyl moieties on the amidic nitrogen are suggested to interact with the GPCRs on the insect neuron cells. Since the identity of the target for arthropod repellent activity remains unknown, these QSAR models and the related analysis should aid in the design of well-tolerated, target-specific arthropod-repellent agents. Effective and efficient use of use of Tcl-based Cerius2 scripts is demonstrated in developing highly predictive QSAR models.

Acknowledgements This research was performed while one of the authors (JBB) held a National Research Council Research Associate-ship Award at Walter Reed Army Institute of Research. The authors gratefully acknowledge funding from the U.S. Department of Defense Health Program. This work was first presented as a poster presentation

at the Fourth Indo-US Workshop on Mathematical Chemistry, Pune, India; January 2005. We are grateful to the referees for their helpful comments and suggestions. We gratefully acknowledge the technical help (Drs. Mehl, Shah and Salaniwal), customer support (Mr. Fankhanel and Dr. Robinson) of Accelrys, Inc. and help with statistics from Mr. Charles White, Senior Biostatistician, WRAIR. JBB is thankful to Dr. Donald Huddler for help in manuscript preparation.

References

- Hansch C (1969) *Acc Chem Res* 2:232–239
- Cramer RD III, Patterson DE, Bunce JD (1988) *J Am Chem Soc* 110:5959–5967
- Hopfinger AJ, Burke BJ (1990) In: Johnson MA, Maggiora GM (eds) *Molecular shape analysis: a formalism to quantitatively establish spatial molecular similarity*. Wiley, New York, pp 173–209
- Klebe G (1998) *Perspectives in drug discovery and design*. 12/13/14:87–104
- Jain AN, Dietterich TG, Lathrop RH, Chapman D, Critchlow RE Jr, Bauer BE, Webster TA, Lozano-Perez T (1994) *J Comput-Aid Mol Des* 8:635–652
- Doweyko AM (1991) *J Math Chem* 7:273–285
- Wiener H (1947) *J Am Chem Soc* 69:17–20
- Bonchev D (1983) *Chemometrics series*, vol 5. Information theoretic indices for characterization of chemical structures, p 249
- Hosoya H (1971) *Bull Chem Soc Jpn* 44:2332–2339
- Kier LB, Hall LH (1976) *Medicinal chemistry*, vol 14. Molecular connectivity in chemistry and drug research. Academic, New York, p 257
- Kier LB (1985) *Quant Struct–Act Relat* 4:109–116
- Kier LB, Hall LH, Frazer JW (1991) *J Math Chem* 7:229–241
- Balaban AT (1982) *Chem Phys Lett* 89:399–404
- Bonchev D, Mekenyan O, Trinajstić N (1981) *J Comput Chem* 2:127–148
- Marsili M, Gasteiger J (1981) *Croat Chem Acta* 53:601–614
- Rohrbaugh RH, Jurs PC (1987) *Anal Chim Acta* 199:99–109
- Stanton DT, Jurs PC (1990) *Anal Chem* 62:2323–2329
- Kharkar PS, Desai B, Gaveria H, Varu B, Loria R, Naliapara Y, Shah A, Kulkarni VM (2002) *J Med Chem* 45:4858–4867
- Drew MGB, Wilden GRH, Spillane WJ, Walsh RM, Ryder CA, Simmie JM (1998) *J Agricult Food Chem* 46:3016–3026
- Hirashima A, Rafaeli A, Gileadi C, Kuwano E (1999) *Bioorg Med Chem* 7:2621–2628
- Hasegawa K, Arakawa M, Funatsu K (2000) *Chemom Intell Lab Syst* 50:253–261
- Hopfinger AJ, Wang S, Tokarski JS, Jin B, Albuquerque M, Madhav PJ, Duraiswami C (1997) *J Am Chem Soc* 119:10509–10524
- Hasegawa K, Arakawa M, Funatsu K (2003) *Comput Biol Chem* 27:211–216
- Vedani A, McMasters DR, Dobler M (2000) *Quant Struct–Act Relat* 19:149–161
- Appell M, Dunn WJ, Reith MEA, Miller L, Flippen-Anderson JL (2002) *Bioorg Med Chem* 10:1197–1206
- Xiao Y-D, Hammond PS, Harris R, Schmitt JD, Klucik J (2004) *J Med Chem* 47:6831–6839
- Sulea T, Kurunczi L, Oprea TI, Simon Z (1998) *J Comput-Aid Mol Des* 12:133–146
- Bhonsle JB, Wang Z-X, Tamamura H, Fujii N, Peiper SC, Trent JO (2005) *QSAR & Combinat Sci* 24:620–630
- Accelrys Software Inc (2005) *Cerius2 Modeling Environment*, Release 4.9. San Diego, CA, USA
- Brewster D (2001) *BioMed J* 323:289
- Bock GR, Cardew G (1996) Symposium on Olfaction in Mosquito–Host Interactions, held in Collaboration with the World Health Organization at the Ciba Foundation, London, 31 October–2 November 1995. In: *Ciba Found Symp* 200. p 331
- Davis EE, Bowen MF (1994) *J Am Mosquito Control* 10:316–325
- Davis EE (1985) *J Med Entomol* 22:237–243
- McIver SB (1981) *J Med Entomol* 18:357–361
- Justice RW, Biessmann H, Walter MF, Dimitratos SD, Woods DF (2003) *BioEssays* 25:1011–1020
- Rao SS, Rao KM, Ramachandran PK (1988) *J Sci & Indust Res* 47:722–735
- McGovern TP, Schreck CE, Jackson J (1984) *Mosquito News* 44:11–16
- Skinner WA, Johnson HL (1980) *Med Chem* 11:277–305
- Sugawara R, Tominaga Y, Suzuki T (1977) *Insect Biochem* 7:483–485
- Wright RH (1975) *Sci Am* 233:104–111
- Suryanarayana MVS, Pandey KS, Prakash S, Raghuvveeran CD, Dangi RS, Swami RV, Rao KM (1991) *J Pharm Sci* 80:1055–1057
- Ma D, Bhattacharjee AK, Gupta RK, Karle JM (1999) *Am J Trop Med Hyg* 60:1–6
- Bhattacharjee AK, Gupta RK, Ma D, Karle JM (2000) *J Mol Recognit* 13:213–220
- Bhattacharjee AK, Dheranetra W, Nichols DA, Gupta RK (2005) *QSAR & Combinat Sci* 24:593–602
- Marsili M, Gasteiger J (1981) *Stud Phys Theor Chem* 16:56–67
- Mayo SL, Olafson BD, Goddard WAI (1990) *J Phys Chem* 94:8897–8909
- Golbraikh A, Tropsha A (2002) *Molecular Diversity* 5:231–243
- Accelrys Software Inc (2005) Performing a conformational analysis. In: *Cerius2*, version 4.10L. *Conformational Search & Analysis Manual*, vol San Diego, CA, USA, p 49
- Bhonsle JB (2005) *Cerius2_ConfSrchAndAnalysis*
- Accelrys Software Inc (2005) Genetic Function Approximation. In: *Cerius2*, version 4.10. *QSAR Manual*, vol San Diego, CA, p 237
- Accelrys Software Inc (2005) In: *Cerius2*, version 4.10. *QSAR Manual*, vol San Diego, CA, p 34
- Accelrys Software Inc (2005) *Theory: Statistical Methods*. In: *Cerius2*, version 4.10. *QSAR Manual*, vol San Diego, CA, p 33
- Bro R (1996) *J Chemomet* 10:47–61
- Hasegawa K, Arakawa M, Funatsu K (1999) *Chemom Intell Lab Syst* 47:33–40
- Bhonsle JB (2005) *Cerius2_QSARModel_computationScript*
- Accelrys Software Inc (2005) *Catalyst*, version 4.10. San Diego, California, USA
- Bhonsle JB (2005) *Cerius2_confAlignmentScript*
- Ghose AK, Crippen GM (1986) *J Comput Chem* 7:565–577
- Ghose AK, Crippen GM (1987) *J Chem Inf Comput Sci* 27:21–35
- Kier LB, Hall LH (1992) The electrotopological state index: an atom-centered index for QSAR. Academic, London, UK, pp 205
- Hall LH, Kier LB (1991) *Rev Comput Chem* 2:367–422
- Hill TL (1960) *An introduction to statistical thermodynamics*. Addison-Wesley Publishing Co, Reading, Massachusetts, p 508
- Dewar MJS, Thiel W (1977) *J Am Chem Soc* 99:4907–4917
- Richon AB, Young SS (1997) *An introduction to QSAR methodology*. Network Science Corporation, Saluda, NC
- L'Heureux PJ, Carreau J, Bengio Y, Delalleau O, Yue SY (2004) *J Comput-Aid Mol Des* 18:475–482
- Olah M, Bologa C, Oprea TI (2004) *J Comput-Aid Mol Des* 18:437–449
- Sutter JM, Dixon SL, Jurs PC (1995) *J Chem Inf Comput Sci* 35:77–84

68. Zheng W, Tropsha A (2000) *J Chem Inf Comput Sci* 40:185–194
69. Yao SW, Lopes VHC, Fernandez F, Garcia-Mera X, Morales M, Rodríguez-Borges JE, Cordeiro MNDS (2003) *Bioorganic & Medicinal Chemistry* 11:4999–5006
70. Tripos Inc (2003) Sybyl Molecular Modeling System, version 6.9.1. St. Louis, MO, USA
71. Johnson HL, Tsakotellis P, Skinner WA, Skidmore D, Maibach HI (1971) *J Pharm Sci* 60:84–89
72. Sandler BH, Nikonova L, Leal WS, Clardy J (2000) *Chem Biol* 7:143–151
73. McKenna MP, Hekmat-Scafe DS, Gaines P, Carlson JR (1994) *J Biol Chem* 269:16340–16347
74. Pikielny CW, Hasan G, Rouyer F, Rosbash M (1994) *Neuron* 12:35–49
75. Beyer WH (1985) *CRC Handbook of Tables for Probability and Statistics*, 2nd edn. CRC Press Inc, Boca Raton, FL, pp 305–310