Journal of the Serbian Chemical Society
JSCS-info@shd.org.rs • www.shd.org.rs/JSCS
J. Serb. Chem. Soc. 75 (10) 1391–1404 (2010)
JSCS–4061
UDC 547.538:532.14:532.74:541.4:615.214
Original scientific paper

Artificial neural network prediction of the psychometric activities of phenylalkylamines using DFT-calculated molecular descriptors

MINA HAGHDADI[1]* and MOHAMMAD H. FATEMI[2]

[1]*Department of Chemistry, Islamic Azad University, Babol Branch and*
[2]*Department of Chemistry, Mazandaran University, Babolsar, Iran*

(Received 8 April, revised 11 May 2010)

Abstract: In the present work, a quantitative structure–activity relationship (QSAR) method was used to predict the psychometric activity values (as mescaline unit, log *MU*) of 48 phenylalkylamine derivatives from their density functional theory (DFT) calculated molecular descriptors and an artificial neural network (ANN). In the first step, the molecular descriptors were obtained by DFT calculation at the 6-311G* level of theory. Then the stepwise multiple linear regression method was employed to screen the descriptor spaces. In the next step, an artificial neural network and multiple linear regressions (MLR) models were developed to construct nonlinear and linear QSAR models, respectively. The standard errors in the prediction of log *MU* by the MLR model were 0.398, 0.443 and 0.427 for training, internal and external test sets, respectively, while these values for the ANN model were 0.132, 0.197 and 0.202, respectively. The obtained results show the applicability of QSAR approaches by using ANN techniques in prediction of log *MU* of phenylalkylamine derivatives from their DFT-calculated molecular descriptors.

Keywords: density functional theory; artificial neural network; multiple linear regression; quantitative structure–property relationship; phenylalkylamines.

INTRODUCTION

Phenylalkyamines form a class of hallucinogenic agents which are pharmacologically diverse and a heterogeneous group of agents.[1] Different properties of phenylalkylamine derivatives, (which were known to display hallucinogenic, central stimulant, empathogenic activity, or a combination of activities) have been studied for a long time using different approaches.[1,2] Phenylalkylamines are one of the few types of psychotomimetic compounds whose structure–activity relationships (SAR) have been investigated.[3] Snyder and Merril first reported a cor-

* Corresponding author. E-mail: mhaghdadi2@yahoo.co.uk
doi: 10.2298/JSC100408116H

relation of hallucinogenic activity with a quantum index that was calculated using the Hückel molecular orbital theory.[4] They found that high activity is associated with the highest occupied molecular orbital (HOMO) energy in a small number of phenylalkylamines. Moreover, some quantitative structure–activity relationship (QSAR) studies were reported recently on this class of compounds.[5–7] Thakur and coworkers reported QSAR studies on a series of psychotomimetric phenylalkylamines by combining the minimum topological difference method and topological descriptors.[6] They developed a penta-parametric linear model, which had the statistics: correlation coefficient R = 0.932, standard error SE = 0.272 and Fisher value F = 56.77. They evaluated the predictive power of their model by a cross-validation test, which provided the statistics: R^2_{CV} = 0.864 and SPRESS = = 0.262. They used two experimentally determined parameters in their model, which were the octanol–water partition coefficient and the molar refraction. In another work, the structure-hallucinogenic activity relationships for a series of phenylethylamine and phenylisopropylamine derivatives was investigated by Altum *et al*. within the framework of electron-conformational methods.[7] In addition, B. W. Clare developed a QSAR model of a series of psychotomimetric phenylalkylamines by using CNDO/2 calculation to derive molecular the volume and electronic structure of the compounds of interest.[8] These QSAR studies were limited to the use of physicochemical descriptors and electronic structure or quantum chemical descriptors, which were obtained based on semi-empirical molecular orbital (MO) calculations. The latest development in computer technology and software for electronic structure theory allows the calculation of quantum chemical descriptors at first principal levels, such as the density functional theory (DFT), with a higher accuracy including some effective consideration of electron correlation effects.[9] DFT Methods are, in general, capable of calculating a variety of isolated molecular properties, such as ionization energies, dipole moment, electron affinities, electronegativities, hardness and softness, electrostatic potential *etc*., quite accurately.[10–17] In the present work, a quantitative structure–activity relationship method was used to predict the psychometric activities value of some phenylalkylamine derivatives from their DFT-calculated molecular descriptors and an artificial neural network.

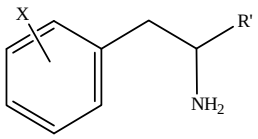
METHODOLOGY

Data set

The psychometric activity of drugs are generally expressed in mescaline units (*MU*), defined as the ratio of the effective dose of mescaline to the effective dose of the tested compound. The potency is usually expressed as log *MU*, where *MU* is taken as mole of mescaline/mole of the tested phenylalkylamine. The data set in this work consisted of the log *MU* values of 48 phenylalkylamine derivatives, which were taken from the literature.[6] The structural features of these compounds and their experimental log *MU* values are presented in Table I. The values of log *MU* ranged from –0.67 to 2.72 for compounds 1 and 38, res-

pectively. The data set was split by y-ranking methods into training; internal and external test sets, which consisted of 38, 5 and 5 members, respectively. In this method, the data were sorted according to their log *MU* values and the training, internal and external test sets were chosen from these lists with desired distances from each other.[18] The training set was used to adjust the ANN parameters, the internal test set was used to prevent overtraining during ANN development and the external test set was used to evaluate the obtained models.

TABLE I. Data set and corresponding observed and calculated values of log *MU* (*MU* is taken as the moles of mescaline/moles of the tested phenylalkylamine)



No.	X	R'	log *MU* Exp	log *MU* ANN	Residual	log *MU* MLR	Residual
1	3,5-OMe	H	–0.67	–0.52	0.15	–0.014	0.655
2	3,4-OMe	Me	–0.06	–0.13	–0.07	0.001	0.061
3	3,4,5-OMe	H	0.00	0.09	0.09	0.349	0.349
4	3-OEt,4,5-OMe	H	0.03	–0.15	–0.18	0.303	0.273
5	3,4-OEt,5-OMe	H	0.23	0.32	0.09	0.440	0.210
6	3,4,5-OMe	Me	0.33	0.43	0.10	0.525	0.195
7	3-Set,4-OEt,5-OMe	H	0.38	0.31	–0.07	0.697	0.317
8	3,4-OEt,5-OMe	H	0.38	0.34	–0.04	1.025	0.645
9	3-OCH_2O-4	Me	0.41	0.47	0.06	0.734	0.323
10	3-OMe,4-OCH_2O-5	Me	0.43	0.51	0.08	0.757	0.327
11	3,5-OMe,4-SBut	H	0.58	0.45	–0.13	1.482	0.901
12	4-Me	Me	0.59	0.67	0.08	–0.146	–0.736
13	3-OEt,4-SMe,5-OMe	H	0.66	0.57	–0.09	0.991	0.331
14	2,4-OMe	Me	0.67	0.61	–0.06	0.843	0.173
15	2.3-OMe,4-OCH_2O-5	Me	0.76	0.67	–0.09	0.829	0.069
16	3,4-OMe,5-SMe	H	0.81	0.85	0.04	1.072	0.262
17	3,5-OMe,4-OPr	H	0.83	0.93	0.10	1.073	0.243
18	3,5-OMe,5-SEt	H	0.84	0.89	0.05	0.457	–0.382
19	2,3,4,5-OMe	Me	0.86	0.94	0.08	1.216	0.356
20	3,5-OMe,4-OEt	H	0.87	0.75	–0.12	0.972	0.102
21	2-OMe,3-OCH_2O-4	Me	1.00	0.95	–0.05	0.575	–0.425
22	2-OMe,4-OCH_2O-5	Me	1.00	1.05	0.05	1.377	0.377
23	3,5-OMe,4-Phenyl	Me	1.10	1.24	0.14	1.261	0.161
24	3,5-OMe,4-SMe	H	1.11	1.21	0.10	1.311	0.201
25	2,5-OMe,3-CH_2O-4	Me	1.14	0.95	–0.19	0.631	–0.508
26	2,5-OMe,4-Et	H	1.25	1.35	0.10	1.354	0.104
27	3,5-OMe,4-SPr	H	1.29	1.12	–0.17	0.896	–0.394
28	2,4,5-OMe	Me	1.33	1.21	–0.12	1.188	–0.141
29	2,5-OMe,4-OEt	Me	1.36	1.25	–0.11	1.236	–0.123
30	3,5-OMe,4-SEt	H	1.36	1.44	0.08	1.368	0.008
31	2,5-OMe,4-SMe	Me	1.66	1.54	–0.12	1.317	–0.342
32	2,5-OMe,4-Bu	Me	1.68	1.76	0.08	1.391	–0.289

TABLE I. Continued

No.	X	R'	log *MU* Exp	log *MU* ANN	Residual		log *MU* MLR	Residual
33	2,5-OMe,4-Br	H	1.69	1.56	–0.13		2.275	0.585
34	2,5-OMe,4-Me	Me	1.90	1.81	–0.09		1.319	–0.581
35	3,5-OMe,4-Br	Me	1.91	2.24	0.33		1.484	–0.426
36	2,5-OMe,4-Pr	Me	1.95	1.67	–0.28		1.363	–0.586
37	2,5-OMe,4-Et	Me	2.02	2.24	0.22		2.078	0.058
38	2,5-OMe,4-Br	Me	2.72	2.56	–0.16		2.275	–0.444
			Internal test set					
39	2,3,5-OMe	H	–0.03	–0.12	–0.09		0.568	0.597
40	3,5-OMe,4-$CH_2C_6H_5$	Me	0.46	0.64	0.18		0.662	0.201
41	3-OMe,4-OEt,5-SMe	H	0.84	0.68	–0.16		1.176	0.336
42	2,5-OMe,4-Me	H	1.27	1.14	–0.13		0.920	–0.349
43	2,5-OMe,4-S-iPr	Me	1.71	1.92	0.21		1.191	–0.518
			External test set					
44	3,5-OMe,4-OBu	H	0.38	0.45	0.07		0.029	–0.350
45	3-OEt,4-SEt,5-OMe	H	0.68	0.88	0.20		1.040	0.360
46	3,5-OMe,4-OEt	Me	1.05	0.78	–0.27		0.409	–0.640
47	2,5-OMe,4-OPr	Me	1.38	1.22	–0.16		0.929	–0.450
48	2,5-OMe,4-SEt	Me	1.96	2.16	0.20		1.560	–0.399

Quantum chemical descriptors

The structures of the 48 substituted phenyalkylamine compounds were optimized using the Gaussian 2003W computational package.[19] The Becke three parameter exact exchange functional (B3)[20] combined with the gradient corrected correlation functional of Lee–Yang–Parr (LYP)[21] of the DFT methods implementing the 6-311G* basis sets were employed. The nature of optimized geometries at the B3LYP levels was checked with frequency calculations. The total energy and corrected zero point (*ZPE*) were calculated for all optimized structures. Then, the quantum chemical descriptors that were taken from the DFT calculations were used to analyze the variations in the biological activities of the studied compounds. The minimum energy conformations were used to calculate the electronic descriptors, such as the energy of the highest occupied molecular orbital, the lowest unoccupied molecular orbital (LUMO), hardness (η), softness (S), electronegativity (χ), electrophilicity (ω) and the dipole moment (*DP*). It was found that the stability of the molecules is related to their hardness and electronegativity, defined as the power of an atom in a molecule to attract electrons to it. The hardness and electronegativity are given as follow:

$$\eta = 1/2(\partial^2 E/\partial N^2)v(r) \quad (1)$$

$$\chi = -\mu = -(\partial E/\partial N)v(r) \quad (2)$$

where E and $v(r)$ are the electronic energy and the external potential of an N-electron system, respectively, and μ is the chemical potential. In addition, the operational definition of absolute hardness and electronegativity is given as:

$$\eta = 1/2(IP - EA) \quad (3)$$

$$\chi = -\mu = -1/2(IP + EA) \quad (4)$$

where *IP* and *EA* are the ionization potential and electron affinity, respectively. According to the DFT counterpart of the Koopman theorem, the *IP* is simply the eigenvalue of the HOMO

with a change of sign and *EA* is the eigenvalue of the LUMO with a change of sign;[22,23] hence Eqs. (3) and (4) may be written as:

$$\eta = \frac{1}{2}\left(\varepsilon\mathrm{LUMO} - \varepsilon\mathrm{HOMO}\right) \tag{5}$$

$$\chi = -\mu = \frac{1}{2}\left(\varepsilon\mathrm{LUMO} + \varepsilon\mathrm{HOMO}\right) \tag{6}$$

The other calculated molecular parameter is softness, which measures the extent of chemical reactivity and it is the reciprocal of hardness:

$$S = \frac{1}{\eta} \tag{7}$$

Parr *et al*. proposed an electrophilicity index (ω) as a measure of energy lowering due to maximal electron flow among donor and acceptor.[24] They defined the electrophilicity index as follow:

$$\omega = \frac{\mu^2}{2\eta} \tag{8}$$

These DFT calculated parameters can encode features of phenylalkylamines which could affect their psychometric activities.

Modeling

In the first step of feature screening, constant and near-constant variables were excluded from the pool of descriptors and the pair wise correlation cut off selection procedure was employed to exclude the correlated descriptors.[18] In the development of QSPRs, a major decision is when to stop adding descriptors to the model during the stepwise regression procedure. An excessive number of descriptors lead to overcorrelated equations that are difficult to interpret in terms of interaction mechanisms. A simple procedure used to control the model expansion is the so-called "break point procedure" in the improvement of the statistical quality of the model.[25] From analyses of the plot of the number of descriptors involved *vs*. the correlation coefficient (*R*), and also the standard error (*SE*) of model, it appears that the statistical improvement of the model is high until a point (the "break point") is reached after which the improvement is negligible. Consequently, the model corresponding to the break point is considered to be the best/optimum model (Fig. 1). This figure shows that no significant improvements in the *R* and *SE* values of the model appear after the addition of more than 5 parameters to the MLR model. Therefore, these five descriptors were considered as independent variable in the development of MLR and ANN models. These parameters are: hardness, chemical potential, lowest unoccupied molecular orbital energy, dipole moment and vibrational energy, the numerical values of which are given in Table II. The specifications of the obtained 5-parameter MLR model are presented in Table III.

TABLE II. The DFT- calculated values of the molecular descriptors

No.	LUMO eV	Hardness eV	Chemical potential eV	Vibrational energy Hartree	Dipole moment D
1	–0.346	2.953	–3.299	–595.42	1.252
2	–0.352	2.955	–3.307	–634.475	1.186
3	–0.198	2.875	–3.073	–709.969	2.366

TABLE II. Continued

No.	LUMO eV	Hardness eV	Chemical potential eV	Vibrational energy Hartree	Dipole moment D
4	–0.059	2.898	–2.957	–749.297	3.488
5	–0.165	2.879	–3.044	–788.63	2.282
6	–0.117	2.857	–2.974	–749.297	2.524
7	–0.606	2.653	–3.259	–1111.61	2.579
8	–0.239	2.703	–2.942	–1111.62	2.817
9	–0.353	2.723	–3.077	–594.224	1.526
10	–0.038	2.849	–2.887	–708.772	1.886
11	–0.487	2.512	–2.999	–1150.93	2.032
12	–0.324	2.977	–3.301	–444.976	1.258
13	–0.604	2.579	–3.184	–1072.29	2.273
14	–0.226	2.669	–2.896	–749.309	3.562
15	–0.214	2.724	–2.957	–823.316	2.869
16	–0.332	2.729	–3.062	–1032.96	0.965
17	–0.575	2.567	–3.142	–1111.6	2.411
18	–0.679	2.675	–3.354	–1072.28	2.768
19	–0.239	2.674	–2.913	–863.838	1.524
20	–0.606	2.574	–3.179	–1072.28	2.465
21	–0.019	2.902	–2.921	–708.772	2.036
22	–0.188	2.571	–2.579	–708.775	2.48
23	–0.231	2.612	–2.843	–788.627	2.359
24	–0.518	2.523	–3.041	–1032.96	2.172
25	–0.051	2.82	–2.872	–823.318	3.458
26	–0.201	2.609	–2.811	–674.076	1.774
27	–0.498	2.519	–3.016	–111.609	1.998
28	–0.26	2.61	–2.87	–749.298	2.455
29	–0.247	2.609	–2.856	–788.629	2.428
30	–0.511	2.52	–3.031	–1072.29	2.08
31	–0.686	2.447	–3.133	–1072.29	2.357
32	–0.294	2.585	–2.879	–792.052	1.623
33	–0.682	2.536	–3.218	–3208.29	1.586
34	–0.248	2.601	–2.849	–674.08	1.746
35	–0.444	2.789	–3.233	–3168.96	2.978
36	–0.296	2.586	–2.882	–752.728	1.624
37	–0.216	2.606	2.822	–713.402	1.821
38	–0.682	2.536	–3.218	–3208.29	1.585
39	–0.089	2.847	–2.936	–709.966	2.6
40	–0.204	2.848	–3.051	–980.405	1.876
41	–0.219	2.741	–3.046	–1072.29	1.266
42	–0.486	2.604	–2.838	–634.754	1.672
43	–0.783	2.445	–3.228	–1150.94	2.533
44	–0.437	2.901	–3.107	–827.947	1.856
45	–0.586	2.576	–3.161	–1111.61	2.322
46	–0.093	2.861	–2.954	–788.627	3.451
47	–0.435	2.614	–2.842	–788.627	2.359
48	–0.45	2.454	–2.803	–1111.61	2.876

TABLE III. Specification of the multiple linear regression model

Descriptor	Notation	Coefficient	S. E.
Lowest unoccupied molecular orbital	LUMO	1.588	0.530
Hardness	η	–3.605	0.619
Chemical potential	μ	1.019	0.079
Vibrational energy	*Ev*	–0.00053	0.000
Dipole moment	*DP*	–0.171	0.114
Constant		11.388	1.720

To examine the existence of any nonlinear relations among the selected molecular descriptors and the log *MU* values, an artificial neural network was used. Artificial neural networks are mathematical systems that simulate biological neural networks. Detail descriptions of the theory of ANNs are presented in the literature.[26–28] In the first step in the development of an ANN, a model of a three-layer network with a sigmoid transfer function was designed both for the hidden and output nodes. This ANN program was coded in MATLAB 7 for windows.[29] In this network, the number of nodes in the input layer was the number of descriptors that were selected by the stepwise multiple linear variable selection method. The adjustable weights among neurons have a random distribution between –0.3 and 0.3. The ANN parameters, including the number of nodes in the hidden layer, the weights of the learning rate, the biases of the learning rate and the momentum should be optimized to find the most accurate results. The procedure of optimizing these parameters is given in previous papers.[30,31] The optimized values of these terms and the ANN topology are given in Table IV. The network was then trained using the training set by the back propagation strategy for optimization of the weights and bias values. The goal in the training of the network is to change the weights between the layers to minimize output errors. The root-mean-square error (*RMSE*) value was calculated to measure how good the outputs were in comparison with the target values. To prevent over fitting, the training of the network had to be stopped when the *RMSE* in the prediction of log *MU* of the internal test set commenced to increase. Then the trained network was used to predict the log *MU* values of the external test set.

TABLE IV. Architecture and specifications of optimized ANN model

Parameter	Value
Number of nodes in the input layer	5
Number of nodes in the hidden layer	4
Number of nodes in output layer	1
Weights learning rate	0.6
Biases learning rate	0.8
Momentum	0.4

RESULTS AND DISCUSSION

Molecular diversity analysis

In this study, diversity analysis was performed on the data set to ensure that the structures of the training or test sets can represent those of the whole ones.[32] A data base of n compounds generated from m highly correlated chemical descriptors $\{x_j\}_{j=1}^{m}$. Each compound X_i is represented as a vector of:

$$X_i = (x_{i1}, x_{i2}, x_{i3},..., x_{im}) \text{ for } i = 1, 2,..., n \quad (9)$$

where x_{ij} denotes the value of descriptor j of compound X_i. The collective database, $X = \{X_i\}_{i=1}^{N}$ is represented by the $n \times m$ matrix **X** as follow:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{nm} \end{bmatrix} \quad (10)$$

A distance score for two different compounds X_i and X_j can be measured from their Euclidean distance norm, d_{ij}, based on the compound descriptors:

$$d_{ij} = \left\| X_I - X_J \right\| = \sqrt{\sum_{k=1}^{m} \left(x_{ik} - x_{jk} \right)^2} \quad (11)$$

The mean distances $\overline{d_i}$ of one sample to the remaining ones were computed as follows:

$$\overline{d_i} = \frac{\sum_{j=1}^{n} d_{ij}}{n-1}, \; i = 1, 2,....., n \quad (12)$$

Then the mean distances were normalized within the interval (0,1). The closer to one the distance is, the more diverse from each other the compounds are. The mean distances of samples *vs*. experimental log *MU* of data set are plotted in Fig. 2. Inspection to this figure illustrates the good distribution of test set throughout the whole of data set. The training set with a broad representation of the chemistry space was adequate to ensure the stability of the models and the diversity of test set can prove the predictive capability of the model.

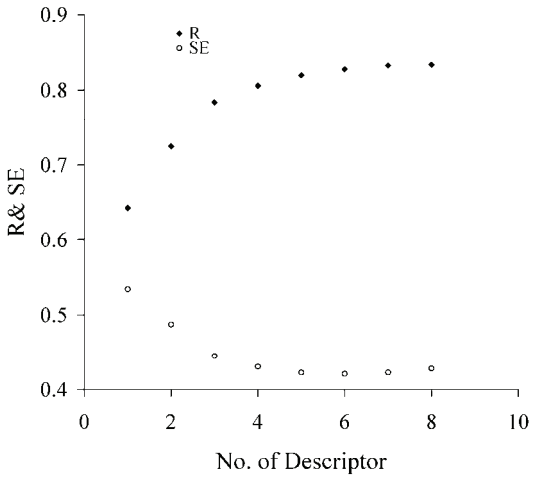


Fig. 1. The results of the break point procedure.

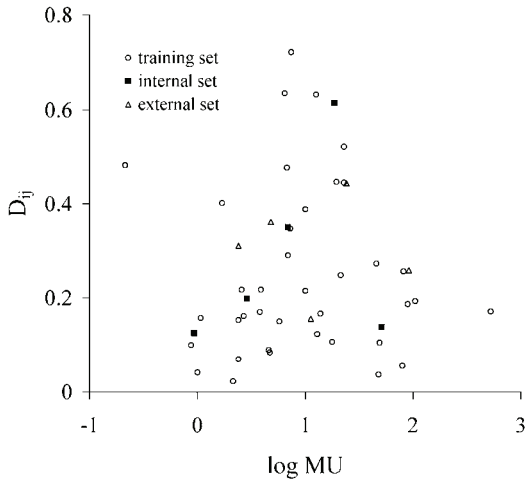


Fig. 2. The results of diversity analysis (the unit of *MU* is mmol of mescaline/mmol of the tested phenylalkylamine with the same effect).

Model development

The descriptors which were selected by the stepwise variable subset selection were chemical potential, hardness, lowest unoccupied molecular orbital energy, dipole moment and vibrational energy. The appearances of these descriptors in the model can represent different aspects of a molecule which affected the absorption, distribution, metabolism, and excretion (ADEME) properties of the molecule.

The inter-correlations among these descriptors are shown in Table V, from which it can be seen that there are no high correlations between the descriptors. These descriptors were used as independent variables and log *MU* was considered as a dependent variable to develop the MLR model. The obtained MLR model was used to calculate the log *MU* values for the test and training sets. The MLR calculated values of log *MU* are given in Table I. The average error (*AE*), average absolute error (*AAE*) and standard error (*SE*) of this calculation were *AE* = –0.051, *AAE* = 0.334 and *SE* = 0.396 for training set; *AE* = –0.054, *AAE* = = 0.408 and *SE* = 0.475 for the internal test set and *AE* = 0.296, *AAE* = 0.440 and *SE* = 0.506 for the external test set. Other statistical parameters of the MLR model are given in Table VI.

TABLE V. Correlation matrix between selected descriptors

	LUMO	η	μ	*Ev*	*DP*
LUMO	1.000	0.62	0.21	0.45	0.143
η	–	1.000	–0.17	0.237	–0.32
μ	–	–	1.000	0.136	–0.44
Ev	–	–	–	1.000	–0.15
DP	–	–	–	–	1.000

TABLE VI. Statistical parameters obtained using the ANN and MLR models

Set	MLR							ANN						
	R	*F*	*SE*	*AE*	*AAE*	Q^2	*SPRESS*	*R*	*F*	*SE*	*AE*	*AAE*	Q^2	*SPRESS*
Training	0.82	74	0.398	–0.05	0.334	0.62	0.372	0.98	968	0.132	0.0065	0.113	0.91	0.147
External	0.82	6	0.443	–0.05	0.408	–	–	0.97	54	0.197	–0.002	0.154	–	–
Internal	0.80	5	0.427	0.296	0.44	–	–	0.96	33	0.202	0.072	0.180	–	–

Since the MLR model was unable to accurately predict the log *MU* values of the data set, an ANN model as a non-linear feature mapping technique was developed. After training this 5:4:1 network, its prediction power was examined on an independent data set which was named the external test set. The ANN predicted values of log *MU* for this set as well as for the training and internal test sets are given in Table I. The statistics of this calculation are: *AE* = 0.006, *AAE* = = 0.113 and *SE* = 0.130 for the training set; *AE* = –0.002, *AAE* = 0.154 and *SE* = = 0.178 for the internal test set and *AE* = 0.072, *AAE* = 0.180 and *SE* = 0.210 for the external test set. Other statistical parameters of the developed ANN model are given in Table VI. In the next step, the sensitivity analysis was carried out on the ANN model to determine the relative importance of each descriptor in the model. The procedure of this approach is based on the sequential removal of variables by zeroing the specific connection weights for that specific input variable in the first layer of the ANN.[33] For each sequentially zeroed input variable, the root mean square error of the prediction set (*RMSEP*), as the prediction error of this network, was calculated. Generally, the *RMSEP* value increases in this way. Then, differences between the *RMSEP* and the root mean square error (*RMSE*) of the established ANN was calculated and shown as *DRMSE*. Each variable which causes a greater value of *DRMSE* is important. The results of this test showed that the order of the importance of the selected molecular descriptors is $Ev > \mu >$ $> dp > \eta >$ LUMO.

Model validation

Validation techniques constitute a fundamental tool for the assessment of the validity of models. They are used to check the prediction power of the models, that is, to give a measure of their capability to perform reliable predictions of the modeled response for new cases for which the response is unknown. Cross-validation is a popular technique used to explore the predictive ability of statistical models.[34] Presuming that a data set consisting of *n* compounds is available, then several modified data sets are created by deleting in each case one or a small group (leave-some-out) of the objects.[23] For each data set, an input-output model is developed, based on the used modeling technique. The model is evaluated by measuring its accuracy in predicting the responses of the remaining data (the ones that have not been used in the development of the model). In particular, the leave-five-out (L5O) procedures were utilized in this study, which produce seve-

ral models, by deleting five objects, respectively from the data set. The maximum numbers of models produced by the L5O procedure is equal to $n!/5!(n-5)!$, where n is the number of molecules in the data set. The squared differences among the true response and the predicted response for each object left out are added to calculate the standardized prediction error sum of squares (*SPRESS*). In addition, the value of Q^2 was calculated from the following equation:

$$Q^2_{\mathrm{Imo}} = 1 - \frac{\Sigma\left(y_0 - y_i\right)^2}{\Sigma\left(y_i - y_m\right)^2} \tag{13}$$

where y_i, y_0 and y_m are the experimental, predicted and average experimental value of log *MU*, respectively. The statistical results of this test on the examined data set was Q^2 = 0.628 and *SPRESS* = 0.472 for the MLR model, while Q^2 = = 0.916 and *SPRESS* = 0.147 for the ANN model, which revealed the reliability of ANN model. In addition, a *Y*-randomization test was applied to examine the chance correlation among log *MU* and selected molecular descriptors.[35] In this method, a dependent variable vector is randomly shuffled and a new QSAR model is developed using the original independent variable. The result of 30 times randomization of log *MU* vectors gave $\overline{R}^2 = 0.111$, which revealed that there was no chance correlation in the data set. Comparison of the statistical results in the prediction of log *MU* using the ANN and the MLR models in Table VI reveals the superiority of the ANN over the MLR model. These observations show that there are some non-linear relation between the selected molecular descriptors and log *MU*. The ANN predicted values of log *MU* are plotted against their experimental values for the training, internal and external test sets in Fig. 3, which shows good correlation among these values. In addition, the residuals of this calculation are plotted against their experimental log *MU* value in Fig. 4. The random propagation of the residuals over the zero line shows that no systematic error existed in the developed ANN model.

As mentioned earlier, Thakur and coworkers proposed a QSAR model on the same data set.[6] The leave-one-out cross-validation test on their model provided the statistics Q^2 =0.7154 and *SPRESS* =0.87, while these statistics for the presented model are Q^2 =0.841 and *SPRESS* =0.34, which revealed the superiority of the present work. In addition, they used two experimentally determined parameters (log P_{ow} and *MR*) as independent parameter in their proposed model while it is preferred to use theoretically derived molecular descriptors in QSAR models. In contrast to their work, the present model uses only theoretically derived molecular descriptors.

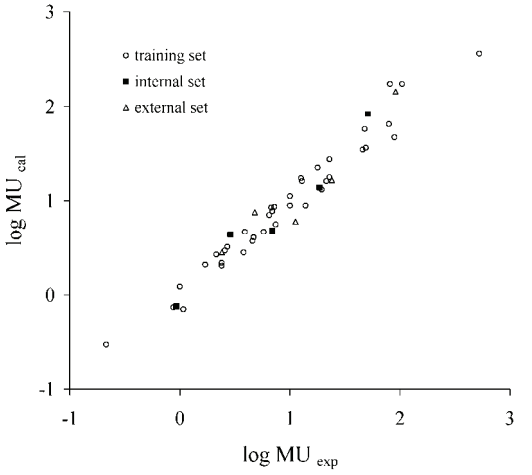


Fig. 3. The plot of ANN calculated *vs.* experimental value of log *MU*.

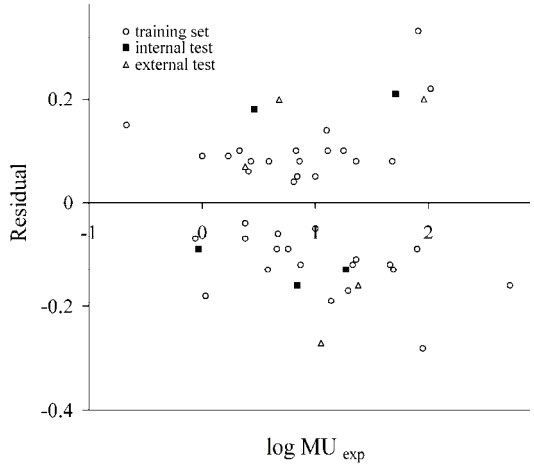


Fig. 4. The plot of ANN calculated residuals *vs.* experimental value of log *MU*.

CONCLUSIONS

The results of this study show that there are some non-linear relationships between the DFT calculated molecular descriptors and the psychometric activities of phenylalkylamines. The high correlation coefficients and low prediction errors of the developed ANN model confirm the good predictive ability of this QSAR model. Finally, the descriptors used in this work can encode some electronic and topological aspect of phenylalkylamines, which could affect on their steric and lipophilic interaction with target cell.

Acknowledgments. The authors wish to acknowledge Anna Ranson, Casa Latinoamericana, London, UK, for reading the manuscript and providing valuable suggestions. This work was supported by the Islamic Azad University, Babol Branch.

ИЗВОД

ПРЕДВИЂАЊЕ ПСИХОМЕТРИЈСКЕ АКТИВНОСТИ ФЕНИЛАЛКИЛАМИНА ПОМОЋУ НЕУРОНСКИХ МРЕЖА ПРИМЕНОМ МОЛЕКУЛСКИХ ДЕСКРИПТОРА РАЧУНАТИХ DFT МЕТОДОМ

MINA HAGHDADI[1] и MOHAMMAD H. FATEMI[2]

[1]*Department of Chemistry, Islamic Azad University, Babol Branch и* [2]*Department of Chemistry, Mazandaran University, Babolsar, Iran*

У раду је примењена метода QSAR за предвиђање психометријске активности (у мескалинским јединицама, log *MU*) за 48 фенилалкиламинских деривата, применом њихових молекулских дескриптора рачунатих DFT методом и помоћу вештачких неуронских мрежа (ANN). У првом кораку молекулски дескриптори су одређени DFT израчунавањима на нивоу теорије 6-311G*. Затим је примењена метода вишеструке линеарне регресије (MLR). У следећем кораку разрађена је метода базирана на ANN, одн. MLR, да би се добили нелинеарни одн. линеарни QSAR модели. Стандардне грешке за log *MU* у MLR-моделу су 0,398, 0,443 и 0,427 за тренинг, унутрашњи и спољашњи сет, док одговарајуће вредности за ANN модел износе 0,132, 0,197 и 0,202. Добијени резултати показују да се ANN-технике могу применити за QSAR предвиђања log *MU* вредности деривата фенилалкиламина на основу њихових молекулских дескриптора израчунатих DFT методом.

(Примљено 8. априла, ревидирано 11. маја 2010)

REFERENCES

1. R. A. Glennon, *Pharmacol. Biochem. Behav.* **64** (1999) 251
2. O. E. Taurian, R. H. Contreras, *J. Mol. Struct. (Theochem)* **504** (2000) 119
3. K. A. Nieforth, *J. Pharm. Sci.* **60** (1971) 655
4. S. H. Synder, C. R. Merril, *Natl. Acad. Sci. U.S.A.* **54** (1965) 258
5. M. Mracec, M. Mracec, L. Kurunczi, T. Nussar, Z. Simon, G. Naray-Szabo, *J. Mol. Str. (Theochem)* **367** (1996) 139
6. M. Thakur, A. Thakur, P. V. Khadikar, *Bioorg. Med. Chem.* **12** (2004) 825
7. A. Altun, K. Golcuk, M. Kumru, A. F. Jalbout, *Bioorg. Med. Chem.* **11** (2003) 3861
8. B. W. Clare, *J. Med. Chem.* **33** (1990) 687
9. R. R. Ruffolo, *Drug Design Discov.* **9** (1993) 351
10. P. K. Chattaraj, A. Cedillo, R. G. Parr, *J. Chem. Phys.* **103** (1995) 7645
11. P. W. Ayers, R. G. Parr, *J. Am. Chem. Soc.* **122** (2000) 2010
12. F. De Proft, J. M. L. Martin, P. Geerlings, *Chem. Phys. Lett.* **250** (1996) 393
13. P. Geerlings, F. De Proft, J. M. L Martin, *Recent Developments in Density Functional Theory*, Elseveir, Amsterdam, 1996, p. 773
14. P. Geerlings, F. De Proft, W. Langenaeker, *Adv. Quantum Chem.* **33** (1999) 303
15. R. G. Parr, R. A. Donnelly, M. Levy, W. E. Palke, *J. Chem. Phys.* **68** (1978) 3801
16. F. De Proft, J. M. L. Martin, P. Geerlings, *Chem. Phys. Lett.* **256** (1996) 400
17. F. De Proft, P. Geerlings, *J. Chem. Phys.* **106** (1997) 3270
18. R. Todeschini, V. Consonni, *Molecular Descriptors for Chemoinformatics*. Wiley-VCH, Weinheim, 2009
19. *Gaussian 98*, Revision A. 6, Gaussian. Inc., Pittsburgh, PA, 2003
20. C. Lee, W. Yang, R. G. Parr, *Phys. Rev. B* **37** (1988) 785
21. A. D. Becke, *J. Chem. Phys.* **98** (1993) 1372

22. R. G. Parr, R. A. Donnelly, M. Levy, W. E. Palke, *J. Chem. Phys.* **68** (1978) 3801
23. P. Politzer, A. Fakher, *Theor. Chim. Acta* **99** (1998) 83
24. A. R. Katritzky, D. C. Fara, H. Yang, M. Karelson, T. Suzuki, V. P. Solov'ev, A. Varnek, *J. Chem. Inf. Comput. Sci.* **44** (2004) 529
25. J. Zupan, J. Gasteiger, *Anal. Chim. Acta* **248** (1991) 1
26. N. K. Bose, P. Liang, *Neural Network-Fundamentals*, McGraw-Hill, New York, 1996
27. B. J. Wytho, *Chemom. Intell. Lab. Syst.* **18** (1993) 115
28. MATLAB version 7.0. The MathWorks, Inc., 2004 (www.mathworks.com)
29. M. Jalali-Heravi, M. H. Fatemi, *Anal. Chim. Acta* **415** (2000) 95
30. M. Jalali-Heravi, M. H. Fatemi, *J. Chromatogr. A* **897** (2000) 227
31. A. G. Maldonado, J. P. Doucet, M. Petitjean, B. T. Fan, *Mol. Divers.* **10** (2006) 39
32. L. I. Nord, S. P. Jacobsson, *Chemometr. Intell. Lab. Syst.* **44** (1998) 153
33. D. W. Osten, *J. Chemom.* **2** (1998) 39
34. B. Walczak, D. L. Massart, *Chemom. Intell. Lab. Sys.* **27** (1995) 2741.