

QSAR modeling of mono- and bis-quaternary ammonium salts that act as antagonists at neuronal nicotinic acetylcholine receptors mediating dopamine release

Fang Zheng,^a Ersin Bayram,^b Sangeetha P. Sumithran,^a Joshua T. Ayers,^a
Chang-Guo Zhan,^a Jeffrey D. Schmitt,^c Linda P. Dwoskin^a and Peter A. Crooks^{a,*}

^aDepartment of Pharmaceutical Sciences, College of Pharmacy, University of Kentucky, Lexington, KY 40536-0082, USA

^bDepartment of Biomedical Engineering, Wake Forest University Health Sciences, Medical Center Boulevard,
Winston-Salem, NC 27157-1022, USA

^cDepartment of Physiology and Pharmacology, Wake Forest University Health Sciences, Medical Center Boulevard,
Winston-Salem, NC 27157-1022, USA

Received 27 June 2005; revised 8 December 2005; accepted 12 December 2005

Available online 20 January 2006

Abstract—Back-propagation artificial neural networks (ANNs) were trained on a dataset of 42 molecules with quantitative IC_{50} values to model structure–activity relationships of mono- and bis-quaternary ammonium salts as antagonists at neuronal nicotinic acetylcholine receptors (nAChR) mediating nicotine-evoked dopamine release. The ANN QSAR models produced a reasonable level of correlation between experimental and calculated $\log(1/IC_{50})$ ($r^2 = 0.76$, $r_{cv}^2 = 0.64$). An external test for the models was performed on a dataset of 18 molecules with IC_{50} values $>1 \mu M$. Fourteen of these were correctly classified. Classification ability of various models, including self-organizing maps (SOM), for all 60 molecules was also evaluated. A detailed analysis of the modeling results revealed the following relative contributions of the used descriptors to the trained ANN QSAR model: $\sim 44.0\%$ from the length of the *N*-alkyl chain attached to the quaternary ammonium head group, $\sim 20.0\%$ from Moriguchi octanol–water partition coefficient of the molecule, $\sim 13.0\%$ from molecular surface area, $\sim 12.6\%$ from the first component shape directional WHIM index/unweighted, $\sim 7.8\%$ from Ghose–Crippen molar refractivity, and 2.6% from the lowest unoccupied molecular orbital energy. The ANN QSAR models were also evaluated using a set of 13 newly synthesized compounds (11 biologically active antagonists and two biologically inactive compounds) whose structures had not been previously utilized in the training set. Twelve among 13 compounds were predicted to be active which further supports the robustness of the trained models. Other insights from modeling include a structural modification in the bis-quinolinium series that involved replacing the 5 and/or 8 as well as the 5' and/or 8' carbon atoms with nitrogen atoms, predicting inactive compounds. Such data can be effectively used to reduce synthetic and in vitro screening activities by eliminating compounds of predicted low activity from the pool of candidate molecules for synthesis. The application of the ANN QSAR model has led to the successful discovery of six new compounds in this study with experimental IC_{50} values of less than $0.1 \mu M$ at nAChR subtypes responsible for mediating nicotine-evoked dopamine release, demonstrating that the ANN QSAR model is a valuable aid to drug discovery.

© 2006 Elsevier Ltd. All rights reserved.

1. Introduction

Tobacco smoking is a leading health problem accounting for more illnesses and health problems in the US than any other single factor.¹ Several drugs are market-

ed for smoking cessation, including nicotine (as a replacement therapy) and bupropion, an antidepressant agent with nicotinic receptor antagonist properties. Unfortunately, relapse rates are high with these agents, indicating that novel medications are still needed.²

Previous research^{3–10} in our laboratories has led to the discovery of a new class of nicotinic acetylcholine receptor (nAChR) antagonists resulting from *N*-*n*-alkylation of the pyridine moiety of either the *S*-(–)-nicotine molecule or of analogs of nicotine. These novel compounds exhibit potent and competitive inhibition of nAChR

Keywords: Neural network; Simulated annealing; QSAR; nAChR antagonist; Dopamine release; Nicotinic acetylcholine receptor; Self-organizing map.

* Corresponding author. Tel.: +1 859 257 1718; fax: +1 859 257 7585; e-mail: pcrooks@uky.edu

subtype(s) that mediate *S*(–)-nicotine-evoked dopamine release from dopaminergic nerve terminals in striatum.^{3,4} Further research has demonstrated that quaternization of the pyridine-nitrogen of the nicotine molecule with a lipophilic *N*-alkyl substituent to afford *N*-alkylnicotinium analogs and/or interconnecting various quaternary ammonium moieties with a lipophilic linker to afford *N,N'*-bis-analogs generates subtype-selective nAChR antagonists.^{3,4,7} These antagonists could have potential as novel smoking cessation agents, and are of considerable interest, due to their selective antagonist activity at specific nAChR subtypes. However, little is yet known about the SAR or the nature of the interaction site(s) on the target protein.^{3–10}

Due to the complexity of nAChR molecular recognition and the lack of knowledge about how these antagonists interact with the binding site(s), structural information on the target neuronal nAChR subtypes^{3–5} that these ligands bind to is not available, although the availability of the crystal structure of the acetylcholine binding protein (AChBP) has shed light on structure/function aspects of pentameric ligand-gated ion-channels which mediate and modulate chemical synaptic transmission, and currently, homolog modeling of nAChR structures with the AChBP crystal structure is possible.¹⁶ Therefore, the identification of initial leads, usually in the <1 μ M range, for this kind of drug discovery process is best achieved from ligand-based models, for example, pharmacophore identification and structure–activity relationships (SARs). Several structure-based or ligand-based approaches for identifying molecules that interact at nAChRs, and particularly at the $\alpha 4\beta 2$ nAChR subtype, have been exploited to gain information about specific receptor pharmacophores. For example, Glennon et al. performed quantitative structure–activity relationship (QSAR) studies on a series of agents structurally related to nicotine and epibatidine, and found their affinities to parabolically correlate with the respective N–N distances in these molecules (optimal distance 5.1–5.5 Å).^{11–17} Successful QSARs, however, were usually predicated on data sets with a uniform mode of action and on congeneric chemical frameworks of these types of ligands. In this respect, it is clear that models satisfactorily explaining activity of the various nAChR ligands with multi-binding modes are far from being identified, and little structure–activity work has been reported so far in this area.¹⁸

The development of nonlinear modeling approaches, such as artificial intelligence-based algorithms, opened up the field to the concurrent analysis of a wider variety of structures with potentially varying modes of action and noncongeneric chemicals.^{19–23} These artificial systems emulate the function of the brain, where a very high number of information-processing neurons are interconnected and are known for their ability to model a wide set of functions, including linear and nonlinear, without knowing the analytic forms in advance. Although there are a number of different neural network models, the most frequently used type of neural network in QSAR is the feed-forward back-propagation network. In the present paper, we used the artificial neural

network approach to build a QSAR model of mono- and bis-quaternary ammonium salts that can be used to predict the potency of these analogs in the inhibition of nicotine-evoked [³H]dopamine release mediated by nAChR subtypes. Based on currently available structure–activity data for the mono- and bis-quaternary ammonium salts generated as IC₅₀ values for the nAChR subtype involved in nicotine-evoked dopamine release, descriptors selected by stepwise regression from various molecular properties were used to train back-propagation artificial neural networks (ANNs). As a second approach, supervised self-organizing map²⁸ (sSOM)-based classification modeling was performed coupled with hill-climbing-based²⁹ descriptor selection.²⁸ sSOM is different from the usual SOM approach, in that information about class-identity is taken into account in the learning phase. Leave-one-out cross-validation was employed in both SOM classification and ANN QSAR modeling. The classification ability of the different models was evaluated and discussed. Additionally, new ligands were designed and their potencies for mediating nicotine-evoked dopamine release were predicted using the trained models. Some of these ligands have been synthesized and their activities have been evaluated. The results demonstrate that the performance of the generated neural networks is very good, and consistent with our expectations, based on the validation measurements.

2. Methods

2.1. Generation of the molecular database

Molecular modeling was carried out with the aid of the Sybyl discovery software package.^{24a} This software was used to construct the initial molecular structures used in the geometry optimization (energy minimization) for all molecules involved in this study. The geometry optimization was first performed by using the molecular mechanics (MM) method with the Tripos force field and the default convergence criterion. In construction of the initial molecular structures, a formal charge of +1 and the Npl3 atom-type was assigned to the positively charged nitrogen atom in the structures of these compounds. To examine the accuracy of this atom-type assignment, we compared the MM-optimized molecular geometry of NPNI (see Table 1) with the corresponding geometry optimized by using a more sophisticated first-principles electronic structure method at the B3LYP/6-31+G* level with the Gaussian03 program.^{24b} The comparison indicated that the MM-optimized geometric parameters of NPNI are in good agreement with the corresponding geometric parameters optimized at the B3LYP/6-31+G* level. In consideration of various possible conformations of a molecule associated with local minima on the corresponding potential energy surface, we first carefully examined various possible conformations of a representative molecule for each series of compounds collected in Tables 1, 7, and 8. For a given representative molecule, we obtained a variety of conformations associated with local minima on the potential energy surface from an internal coordinate Monte Carlo conformational search

Table 1. The structures of the analogs synthesized, the observed IC₅₀ values and log(1/IC₅₀) values, and their error ranges

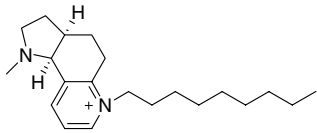
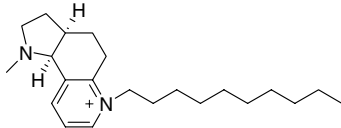
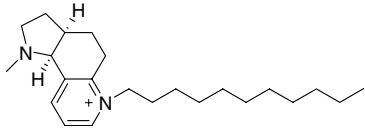
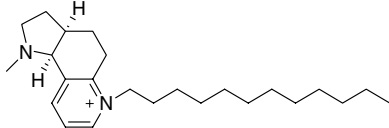
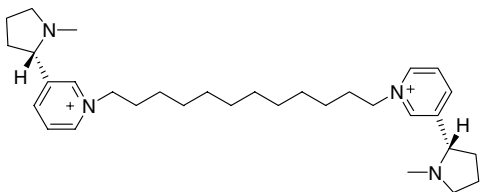

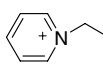
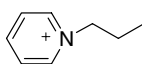
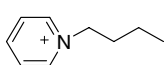
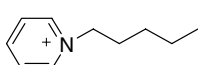
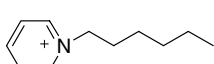
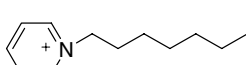
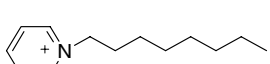
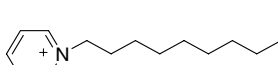
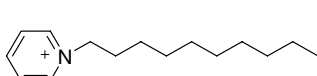
Compound	Structure	IC ₅₀ -obsd (μM) ^f	log (1/IC ₅₀) (obsd) ^f
<i>N</i> -Alkylnicotinium salts			
T1		24.6 ^a	−1.39
T2		37.4 ^a	−1.57
T3		9.07 ^a	−0.96
T4		3.45 ^a	−0.54
T5		0.80 ^a	0.1
T6		0.62 (0.20–1.9) ^a	0.2 (−0.28–0.70)
T7		0.21 ^a	0.68
P1		>1.0 ^a	
T8		0.009 (0.003–0.03) ^a	2.05 (1.52–2.52)
T9		25	−1.4
T10		20	−1.3
T11		0.61 ± 0.43 ^e	0.21 (−0.017–0.74)

(continued on next page)

Table 1 (continued)

Compound	Structure	IC _{50-obsd} (μM) ^f	log(1/IC ₅₀) (obsd) ^f
T12		0.27 ± 0.14 ^c	0.57 (0.39–0.89)
T13		2.44 ± 1.85 ^c	−0.39 (−0.63–0.23)
T14		0.15 ± 0.08 ^c	0.82 (0.64–1.15)
P2		>1.0	
T15		0.04 ± 0.03 ^c	1.40 (1.15–2)
T16		0.9	0.046
<i>Conformationally restricted N-alkylnicotinium salts (syn conformation)</i>			
T17		0.08 ± 0.04 ^b	1.10 (0.92–1.40)
T18		0.66 ± 0.03 ^b	0.18 (0.16–0.20)
T19		0.58 ± 0.55 ^b	0.24 (−0.053–1.52)
T20		0.04 ± 0.02 ^b	1.40 (1.22–1.70)
T21		0.22 ± 0.15 ^b	0.66 (0.43–1.15)
<i>Conformationally restricted N-alkylnicotinium salts (anti-conformation)</i>			
T22		0.04 ± 0.02 ^b	1.4 (1.22–1.70)

Table 1 (continued)

Compound	Structure	IC ₅₀ -obsd (μM) ^f	log(1/IC ₅₀) (obsd) ^f
T23		0.31 ± 0.15 ^b	0.51 (0.34–0.80)
T24		0.03 ± 0.01 ^b	1.52 (1.40–1.70)
T25		0.04 ± 0.03 ^b	1.4 (1.15–2)
P3		>1.0 ^b	
<i>N,N'-(1,12-Dodecanediyl) bis-[3-[(2S)-1-methyl-2-pyrrolidinyl]-pyridinium dibromide]</i>			
T26		0.17 ± 0.13 ^b	0.77 (0.52–1.40)
<i>N-Alkylpyridinium salts</i>			
P4		>1.0	
P5		>1.0	
P6		>1.0	
P7		>1.0	
P8		>1.0	
P9		>1.0	
P10		>1.0	
P11		>1.0	
P12		>1.0	
T27		0.13 (0.02–0.87) ^c	0.89 (0.06–1.70)

(continued on next page)

Table 1 (continued)

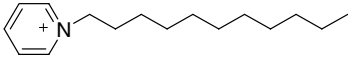
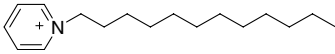
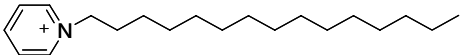
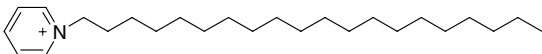
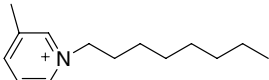
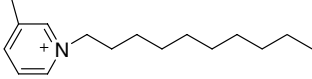
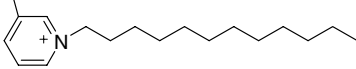
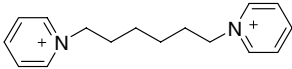
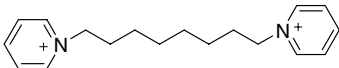
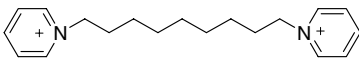
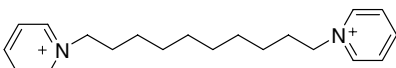
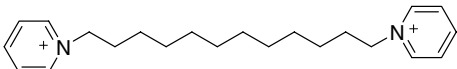
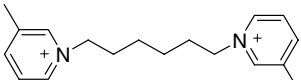
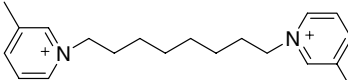
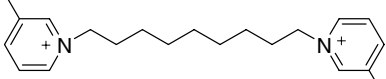
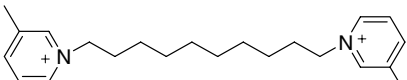
Compound	Structure	IC ₅₀ -obsd (μM) ^f	log(1/IC ₅₀) (obsd) ^f
T28		0.49 ^c	0.31
T29		0.26 (0.02–4.23) ^c	0.59 (−0.63–1.70)
T30		0.32 (0.11–0.87) ^c	0.49 (0.06–0.96)
T31		0.12 (0.01–2.38) ^c	0.92 (−0.38–2)
<i>N-Alkylpicolinium salts</i>			
T32		1.0 ± 0.09 ^d	0.00 (−0.037–0.041)
T33		0.3 ± 0.05 ^d	0.52 (0.46–0.60)
T34		0.03 ± 0.02 ^d	1.52 (1.30–2)
<i>N,N'-bis-Alkylpyridinium salts</i>			
P13		>1.0	
P14		>1.0	
P15		>1.0	
P16		>1.0	
T35		1.0 ± 0.38 ^b	0.00 (−0.14–0.21)
<i>N,N'-bis-Alkyl-(3-picolinium) salts</i>			
T36		1.66 ± 0.85 ^d	−0.22 (−0.40–0.091)
T37		0.01 ± 0.009 ^d	2.00 (1.72–3.00)
T38		1.52 ± 0.34 ^d	−0.18 (−0.27–0.072)
T39		0.03 ± 0.01 ^d	1.52 (1.40–1.70)

Table 1 (continued)

Compound	Structure	IC ₅₀ -obsd (μM) ^f	log(1/IC ₅₀) (obsd) ^f
T40		0.005 ± 0.003 ^d	2.30 (2.10–2.70)
<i>N,N'-(1,12-Dodecanediyl) bis-quinolinium dibromide</i>			
T41		0.021 ± 0.01 ^e	1.70 (1.51–1.96)
<i>N,N'-(1,12-Dodecanediyl) bis-isoquinolinium dibromide</i>			
T42		0.007 ± 0.003 ^e	2.15 (2.00–2.40)
<i>N,N,N',N',N',N'-Hexamethylaminium alkyldiiodides</i>			
P17		>1.0 ^e	
P18		>1.0 ^e	

^a From Wilkins, Jr. L. H. et al., *J. Pharmacol. Exp. Ther.*, **2002**, 301, 1088–1096.^b From Crooks, P. A. et al., *Bioorg. Med. Chem. Lett.*, **2004**, 14, 1869–1874.^c From Grinevich V. P. et al., *J. Pharmacol. Exp. Ther.*, **2003**, 306, 1011–1020.^d From Dwoskin, L. P. et al., *Bioorg. Med. Chem. Lett.*, **2004**.^e From the presentation given by Dr. Crooks at University of Michigan.^f For the error range of each observed IC₅₀ value, ± indicates standard error of the mean (SEM), ranges indicate confidence intervals; The log(1/IC₅₀) error ranges were simply calculated from the maximum and minimum values from either the SEM values or the confidence intervals.

(implemented as RANDOMSEARCH in Sybyl). All of the obtained conformations optimized at the MM level were examined further by performing semi-empirical molecular orbital (MO) energy calculations at the PM3 level.²⁴ MOPAC charges were loaded onto the structures to perform MM energy minimization again. The molecular conformation with the lowest energy at the PM3 level was considered to be the most stable conformation. For example, a total of 37 low-energy conformations of *N,N'*-bis-*n*-dodecyl-(3-picolinium) dibromide (bPiDDb) with PM3 calculated energy differences < 15 kcal/mol were found, and the total energies calculated at the PM3 level revealed that a straight alkyl chain attached to the two 3-picolinium sub-structures had the lowest energy. The conformation determined to be the most stable of this representative molecule was then used as a template to build the initial structures of other molecules in the same series, and these initial structures were used in the geometry optimizations using the same MM method.

The 60 molecules listed in Table 1 constituted a database for the structure-activity correlation analysis. A dataset of 42 molecules (**T1–T42**) was used for model training and leave-one-out validation. A dataset of 18 molecules (**P1–P18**) from various compound series was used for testing. Table 1 also lists the experimental IC₅₀ values,

a pharmacological measure of the antagonist activity of these compounds at the nAChR subtype(s) responsible for mediating nicotine-evoked dopamine release.^{3–10} We considered that a compound is active when its IC₅₀ < 1 μM or is inactive when its IC₅₀ ≥ 1 μM. For the set of 60 molecules utilized, 32 were active and 27 were inactive. The IC₅₀ values of 4 molecules were ≤ 0.01 μM, 29 molecules had IC₅₀ values within 0.01–1 μM, and 27 molecules had IC₅₀ values ≥ 1.

The newly designed molecules listed in Tables 7 and 8 were used for prediction purposes.

2.2. Supervised self-organizing map classification modeling

Five hundred and nineteen (519) molecular descriptors consisting of zero-dimensional (constitutional descriptors), one-dimensional (functional groups, empirical descriptors, and physical properties), two-dimensional (topological descriptors) as well as three-dimensional (WHIM descriptors) variables were created by the DRAGON program.²⁵ A reduced descriptor set of 79 was obtained after the constant and near constant descriptors, and the highly inter-correlated (>0.95) descriptors were discarded.

The supervised SOM-based modeling was implemented following the procedure of Bayram et al.²⁸ In this study, however, a standard iterative hill-climbing²⁹ algorithm with a steepest ascent approach was implemented for variable selection instead of the genetic algorithm, to reduce the computation time. To avoid the model getting too complex, a preset maximum limit of 10 was set on the number of selected descriptors. For the first iteration, a random set of descriptors was used, with at most two descriptors from the selected set being randomly changed after each successive iteration. Newly selected sets of descriptors were put through a leave-one-out (LOO)-based SOM modeling and the new solution was moved forward to the next iteration, if the correct classification rate of the sSOM model on the left-out compounds improved compared to the best solution found thus far. The process continued until a pre-selected number of iterations (400 in our case) were complete.

Supervised SOM was coupled with hill-climbing and classification models were then constructed. The SOM, also known as Kohonen networks, converts high dimensional, nonlinear relationships into simple geometric relationships.³⁴ The reduced representation is constructed in such a way as to best preserve the input data's original topology and density.

It is worthwhile to discuss how training is performed on a supervised self-organizing map. The difference between SOM and supervised SOM is in the learning or map formation phase and is relatively simple. The supervised SOM forces the early learning phase to be influenced by actual class distribution of the input samples information as part of the descriptor vector space, as demonstrated in Figure 1. Comparing Figure 1a and b, one can see that the only difference is the added class information vector. For example, for a three-way classification, the vector [100] would be appended to the descriptor vector for compounds that belong to group 1. By adding this class information, one can influence the map formation process and improve the classification accuracy. More specifically, the appended class information vectors will have the same values for the compounds that belong to the same group. Therefore, the same group compounds will be more likely to be placed together, as input data distribution more closely resembles the class information around compared to the unsupervised case. In other words, the supervised self-organizing nature of the system dictates training to be a guided self-learning process. During training, class information of each compound is appended to its descriptor vector in binary format. This combined descriptor vector is fed into the SOM as input to guide the map organization, allowing class information to influence the topological ordering of the map. During the prediction phase, the map that was created during the learning phase is used to relate the descriptors of the compounds to the unknown class information. Another important point that needs to be emphasized is that SOM is strictly a classification system. Therefore, percent correct classification was used to drive training in this study.

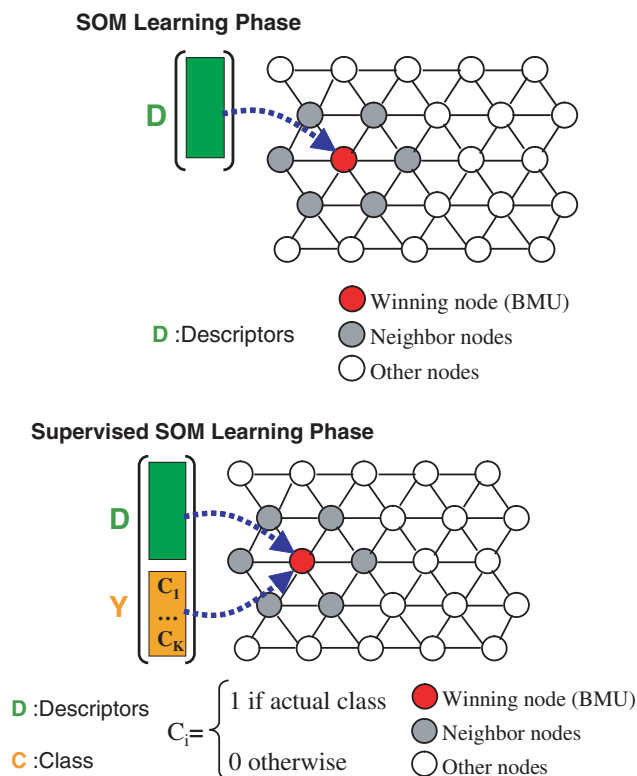


Figure 1. SOM versus sSOM learning.

2.3. Stepwise descriptor selection by multiple linear regressions

The constitutional, topological, empirical, functional group, molar refractivity, octanol–water partition coefficient, and WHIM descriptors were generated as described in Section 2.2. In addition, a variety of other molecular properties, for example, area of molecular surface, polar surface area, were produced by use of the Tripos software package.^{24a} Other steric descriptors were measured from the optimized 3D molecular structures, including the intra-molecular distance between the pyridine ring nitrogen and the last (most distant) heavy atom of the alkyl chain attached to the quaternary ammonium nitrogen atom. The PM3 method was used to determine the LUMO, HOMO energies, dipole moment, and atomic charges, etc., of each molecule.

The multiple linear regression (MLR) analysis was performed by use of an in-house Fortran 77 program. Starting from the entire set of descriptors, variable selection by a forward and reverse stepwise regression procedure was performed, in which forward selection was followed by backward elimination of variables, resulting in an equation in which only variables that significantly increased the predictability of the dependent variable were included.

2.4. Target properties

Experimental IC₅₀ values of the synthesized quaternary ammonium analogs of nicotine were measured according to the procedure described by Dwoskin and co-

workers.^{4,5} The $\log(1/\text{IC}_{50})$ (with IC_{50} value in μM) was used as the target property to derive the QSARs.

2.5. ANN QSAR modeling

Feed-forward, back-propagation-of-error networks were developed using a neural network C program.²⁶ Network weights ($W_{ji}(s)$) for a neuron ' j ' receiving output from neuron ' i ' in the layer ' s ' were initially assigned random values between -0.5 and $+0.5$. The sigmoidal function was chosen as the transfer function that generates the output of a neuron from the weighted sum of inputs from the preceding layer of units. Consecutive layers were fully interconnected; there were no connections within a layer or between the input and the output. A bias unit with a constant activation of unity was connected to each unit in the hidden and output layers.

The input vector was the set of descriptors for each molecule in the series, as generated by the previous steps. All descriptors and targets were normalized to the $[0, 1]$ interval using the following formula:

$$X'_{ij} = \frac{X_{ij} - X_{j,\min}}{X_{j,\max} - X_{j,\min}}, \quad (1)$$

where X_{ij} and X'_{ij} represent the original value and the normalized value of the j th ($j = 1, \dots, k$) descriptor for compound i ($i = 1, \dots, n$). $X_{j,\min}$ and $X_{j,\max}$ represent the minimum and maximum values for the j th descriptor. The network was configured with one or two hidden layers. During the ANN learning process, each compound in the training set was iteratively presented to the network. That is, the input vector of the chosen descriptors in normalized form for each compound was fed to the input units, and the network's output was compared with the experimental 'target' value. During one 'epoch,' all compounds in the training set were presented, and weights in the network were then adjusted on the basis of the discrepancy between network outputs and observed $\log(1/\text{IC}_{50})$ values by back-propagation using the generalized delta rule.

2.6. Cross-validation and testing

Models were cross-validated using the 'leave-one-out (LOO)' approach.²⁷

Generated models were then tested using a subset of 18 compounds (**P1–P18** in Table 1) for which only their qualitative activity properties in the experimental screening array have been obtained, that is, active ($\text{IC}_{50} < 1 \mu\text{M}$) or inactive ($\text{IC}_{50} \geq 1 \mu\text{M}$).

2.7. Evaluation of the QSAR models

QSAR models were assessed by the Pearson correlation coefficient r^2 , root mean square deviation (RMSD), maximum and minimum deviations (max dev and min dev), and predictive r^2_{cv} , which is defined as

$$r^2_{\text{cv}} = \frac{\text{SD} - \text{PRESS}}{\text{SD}}, \quad (2)$$

where SD is the sum of squared deviations of each measured $\log(1/\text{IC}_{50})$ value from its mean and PRESS is the predictive sum of squared differences (the sum of squared differences between actual and predicted values).

3. Results and discussion

3.1. Descriptors chosen for use in this study

Except for building mathematical QSAR models, selection of descriptors used to train ANN also focused on choosing some descriptors that could best represent physical characteristics of the current set of compounds by considering the property that each descriptor represents, with the intent to find the factors that govern the functional behavior of these antagonists.

The need for inclusion of electronic, lipophilic, and geometrical terms in the QSAR equation was clearly indicated in many of the QSAR analyses. It could be particularly true when a set of molecules included chiral centers and a different number of positive charges. During the descriptor selection, descriptors that were important and/or with apparent physical meaning from MLR stepwise analysis to the reduced 79 descriptor dataset are listed in Table 2a. Among them, the lipophilicity parameter MLOGP represents the extent of hydrophilic/hydrophobic interactions. In addition, descriptor results from quantum chemical calculations are included to describe electronic effects. Geometric effects are represented by molecular surface area, topological indices, and WHIM indices, etc., as defined in Table 2a.

Variable selection from Table 2a by a stepwise MLR procedure based on the forward-selection and backward-elimination methods located six descriptors, that is, Area, LUMO, DISNC, MR, MLOGP, and P1u, which were used as input to build ANNs. The variation range (VR) of a descriptor can be calculated by using the maximum (V_{\max}) and minimum (V_{\min}) values of the descriptor for the examined molecules as $\text{VR} = (V_{\max} - V_{\min})/V_{\min}$. The VR values calculated for the six descriptors are 46% (LUMO), 94% (P1u), 161% (Area), 192% (MR), 257% (MLOGP), and 1638% (DISNC). The Pearson correlation coefficients r for the linear correlation between these descriptors and the experimental pIC_{50} values are 0.23 (LUMO), 0.76 (P1u), 0.70 (Area), 0.66 (MR), 0.75 (MLOGP), and 0.68 (DISNC). Mathematically, in a multiple linear regression (MLR) analysis, when the correlation coefficients of two used descriptors are equal to 1, this will cause matrix singularity and the regular matrix inverse cannot be calculated. Therefore, there is a paradox between choosing the most correlated descriptors and eliminating collinearity. Nevertheless, unlike MLR, neural network training/analyses do not need to calculate the matrix inverse, and thus do not share the problems of multicollinearity, as clearly discussed in the literature.³⁵

Table 2a. Brief description of the descriptors used in the stepwise regression analysis

No.	Descriptor	Definition
1	PV	Polar volume (includes O, N, S atoms and covalently bonded Hs)
2	AREA	Total surface area
3	PSA	Polar surface area (includes all O, N, S atoms and covalently bonded Hs)
4	HOMO	Highest occupied molecular orbital energy
5	LUMO	Lowest unoccupied molecular orbital energy
6	DIPOLEM	Dipole moment
7	DISNC	Length of alkyl chain connected with the pyridine nitrogen
8	DISNN	Distance between nitrogen atom in a quaternary ammonium group and the closet positively charged nitrogen atom
9	MR	Ghose–Crippen molar refractivity
10	MLOGP	Moriguchi octanol–water partition coefficient ($\log P$)
11	UI	Unsaturation index
12	Hy	Hydrophilic index
13	ARR	Aromatic index
14	W	Wiener W index
15	WA	Mean Wiener index
16	RDSUM	Reciprocal distance Wiener-type index
17	Qpos	Total positive charge
18	nR05	Number of 5-element rings
19	nR06	Number of 6-element rings
20	nR09	Number of 9-element rings
21	nR10	Number of 10-element rings
22	PyrrineNC	Charge of 5-element ring nitrogen
23	PrroNC	Charge of pyridium ring nitrogen
24	CIC1	Complementary information content (neighborhood symmetry of 1-order)
25	Plu	1st component shape directional WHIM index/unweighted

Table 2b. Brief description of the descriptors selected by hill-climbing-SOM

No.	Descriptor	Definition
1	nR05	Number of five-membered rings
2	X0Av	Average valence connectivity index χ -0
3	SEigv	Eigenvalue sum from van der Waals weighted distance matrix
4	D/Dr05	Distance/detour ring index of order 5
5	T (N...N)	Sum of topological distances between N...N
6	G2s	2nd component symmetry directional WHIM index/weighted by atomic electrotopological states
7	nCp	Number of total primary C(sp ³)
8	Hy	Hydrophilic factor
9	MLOGP	Moriguchi octanol–water partition coefficient ($\log P$)

Nine descriptors selected by supervised SOM when coupled with hill climbing are listed in Table 2b. Selected descriptors are from many different groups, including two-dimensional topological indices, molecular properties, Eigen-value-based indices, functional group counts, constitutional descriptors, and WHIM descriptors. It is very plausible that the self-organizing map selected the key descriptors from such a diverse group to form the final classification model.

3.2. Neural network configuration

Different ANN configurations were used to train, cross-validate, and test the network, in order to find a configuration that would give the best prediction. Taking the descriptors selected from Table 2a as neural network input, a set of 12 configurations were first examined after obtaining the optimal learning rate (0.01), momentum rate (0.05), and total training cycles (10^6) (i.e., the mean squared errors of the neural networks were converged to 10^{-4} for at least the last 100 consecutive epochs). The network parameters used for these 12 configurations

were the same, except for the number of input neurons, hidden neurons, and the number of hidden layers. For all the cases, input and output data were normalized between 0 and 1. Models were evaluated on the basis of correlation and root-mean-square deviation (RMSD)

Table 2c. Neural network configuration analysis

Neural network	r^2	RMSD	r_{cv}^2	LOORMSD
NN501	0.72	0.52	0.61	0.61
NN511	0.76	0.49	0.64	0.58
NN521	0.83	0.41	0.58	0.64
NN531	0.85	0.38	0.44	0.74
NN541	0.85	0.37	0.28	0.83
NN551	0.86	0.37	0.30	0.82
NN5111	0.76	0.49	0.64	0.59
NN5211	0.81	0.41	0.56	0.65
NN5221	0.85	0.38	0.37	0.77
NN5311	0.83	0.39	0.48	0.71
NN5321	0.85	0.38	0.12	0.92
NN611	0.76	0.49	0.64	0.59

between the calculated outputs and target values, comparing the training to leave-one-out validation results.

Table 2c shows that when the input vector is composed of five elements (eliminating LUMO from the descriptors chosen from Table 2a), the correlation coefficient (r^2) of the whole training set increases with increasing number of hidden neurons from 0 to 5, and RMSD decreases. However, on increasing the number of hidden neurons from 2 to 5, the leave-one-out validation correlation coefficient (r_{cv}^2) decreases and RMSD increases, indicating signs of over-fitting. So, the optimal number of hidden neurons was determined to be 1, corresponding to model NN511, which indicates a three-layer neural network, that is, 1 input layer with 5 input neurons, 1 hidden layer with 1 neuron, and 1 output layer with one output neuron.

The number of hidden layers was then increased to 2. Various number of hidden neurons was chosen for calculation. The data listed in the third and fourth rows of Table 2c show that models NN611 and NN5111 afford the best prediction. To avoid underdetermining, we also tested the configuration with an input vector including 7 (adding CIC1) and 8 (adding CIC1 and PV) elements to train NN711 and NN811 networks, the correlation coefficients (r^2) and RMSDs of the two configurations being similar to those of NN611, suggesting that a 5 or 6-element input vector was suitable for the best prediction of the ANNs with the descriptors selected from Table 2a.

Then, the total training cycles were gradually decreased for each configuration during training. The LOORMSD is compared with the lowest LOORMSD 0.58 (0.59) of the three models (NN611, NN511, and NN5111) at training cycle 10^6 in Table 2c. While for the configurations with one hidden node, LOORMSD tended to decrease as training cycles increased until a point was reached after which it did not change anymore, as shown in Figure 3; for other configurations, a minimum LOORMSD exists at some point within 10^6 training cycles. For example, with 5 input neurons and a single hidden layer, when the total training cycles used was 10^5 , the training RMSD value for NN521, NN531, NN541

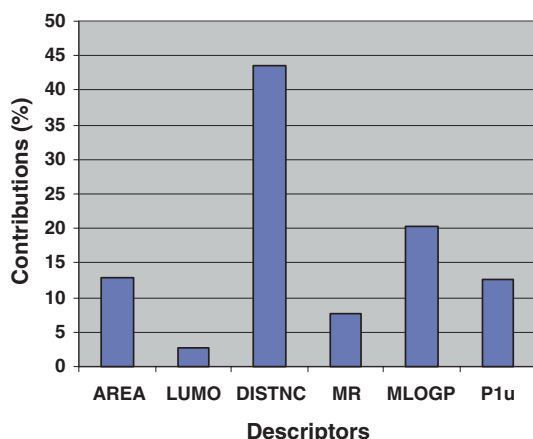


Figure 2. Contributions of the six descriptors to the structure-activity relationship.

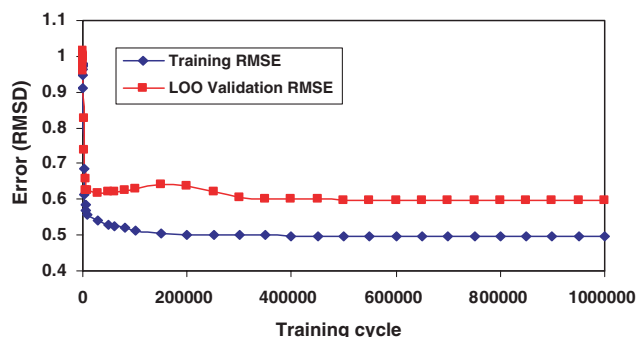


Figure 3. The training RMSD and LOORMSD as functions of the number of training cycles of NN611.

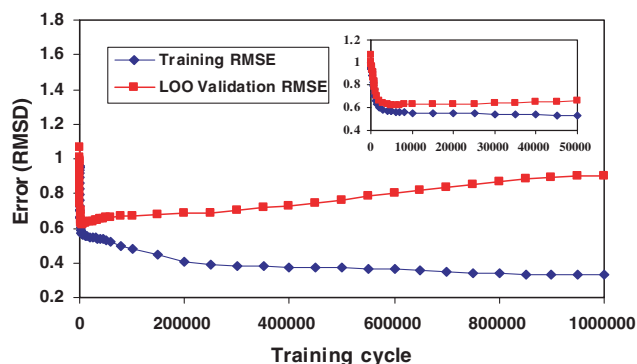


Figure 4. The training RMSD and LOORMSD as functions of the number of NN651.

or NN551 was within 0.48–0.50, and the LOORMSD value was within 0.67–0.70. For a configuration with two or more hidden neurons, by consecutively decreasing the number N of total training cycles, a model similar/closer to the best LOORMSD in Table 2c can be found. For example, when $N = 5010$, the trained NN651 has an RMSD of 0.57, and a LOORMSD of 0.62, as shown in Figure 4. By comparing the IC_{50} values predicted by leave-one-out validation of NN651 trained at $N = 5010$ with those predicted by NN611 trained at $N = 10^6$, the predictions of NN651 for boundary (extreme) values of the activity data of the 42 molecules are close to those predicted by NN611, for example, the IC_{50} values of bPiDDb and NPNI predicted by leave-one-out validation of NN651 are 0.07 and 10.59. Additionally, while testing the network with the 18 inactive molecules (P1–P18 in Table 1), one can conclude that the accuracy of classification of the networks with these models decreases as the number of hidden neurons increases. Only 11 of the 18 inactive molecules (P1–P18 in Table 1) were correctly classified by NN651, compared to 14 of 18 by NN611. NN621 has an RMSD of 0.46 and a LOORMSD of 0.63 at training cycles 300,000, and the model has the same accuracy when classifying the set of molecules. However, comparing the leave-one-out results with the observed results demonstrated that the IC_{50} values for biQDDb, bPiDDb and bQDDb predicted by leave-one-out cross-validation are 0.41, 0.05, and 0.005, respectively; these predictions are far from the observed

results. These data indicate that NN611 can generalize better than the networks with more hidden neurons for this small dataset. So the configuration with fewer parameters was retained.

It is worth mentioning that leave-one-out cross-validation is the special case of k -fold cross-validation, where k equals the sample size. Cross-validation is a method for estimating generalization error of the neural network based on ‘resampling’.^{27,31} The resulting estimates of generalization error are often used for choosing among various models, such as different network architectures. Cross-validation is also quite different from the ‘split-sample’ or ‘hold-out’ method that is commonly used for early stopping in neural networks. In the split-sample method, only a single subset (the validation set) is used to estimate the generalization error, instead of k different subsets; that is, there is no crossing. If the data in the validation set are included in the finally used model, a further test is needed by using the data in an external test set.

The distinction between cross-validation and split-sample validation is extremely important, because cross-validation is markedly superior for small datasets.³² While various investigators have suggested that cross-validation be applied to early stopping, the correct way of doing so is not obvious. However, Figures 3 and 4 shown in this study demonstrate that this can be done in a manner identical to ‘early stopping,’ by choosing stop at the cycle for a configuration that has the smallest estimated generalization error during the whole range of the chosen training cycles. For example, the stop cycle is at any point after $N = 500,000$ for NN611 and at $N = 5010$ for NN651.

3.3. Computational results

Table 3 lists the results of the overall $\log(1/IC_{50})$ values calculated by the three best ANN models (NN511, NN5111, and NN611), showing the experimental $\log(1/IC_{50})$ values, the NN611 leave-one-out predicted $\log(1/IC_{50})$ values, and the results calculated by MLR with the same six descriptors used in the NN611 model for comparison. Plotted out in Figure 5 are the relationships of the trained and leave-one-out predicted pIC_{50} versus experimental pIC_{50} values for the NN611 model. As can be seen, most of the leave-one-out predicted pIC_{50} values are very close to the calculated pIC_{50} values of the trained model, indicating that the model is not over-fitted.

As seen from the results listed in Table 3, the three models generate consistent results, showing that either adding descriptor LUMO or increasing an additional hidden layer with one neuron to the NN511 results in little improvement of the calculated results.

Statistical results for the calculated values of the three models, as well as MLR, are found in Table 4 where r^2 , r_{cv}^2 , and RMSD have the same definition as indicated in Table 2c. The maximum deviations of the calculated results by the four models come from those having the smallest IC_{50} values, such as that for NDDNI, as these

are few extreme points in the whole dataset. These extreme data values, however, were not in error, and should not be regarded as outliers, since the IC_{50} values were repeated many times during experimental measurement. A model with more hidden neurons did fit the activity of these compounds better in the training, but gave worse leave-one-out validation r_{cv}^2 and LOO-RMSD, as shown in Table 2c. This phenomenon should disappear as more potent antagonists with nanomolar IC_{50} values are added to the analysis database. Table 4 indicates that neural network models performed better than MLR, which yielded a correlation value $r^2 = 0.72$ and a $r_{cv}^2 = 0.59$.

Considering that a r_{cv}^2 value greater than 0.5 has been regarded as standard proof of the high predictive ability of a model³⁰ and the error (uncertainty) in the dependent target measurement of the training set, the current study has obtained models with reasonable predictive accuracy by a correlation coefficient of $r^2 = 0.76$ and cross-validation of $r_{cv}^2 = 0.64$. In the case of the currently investigated training set, in general, there is a considerable variability (about 20% measurement error) in the experimental IC_{50} values, as shown in Table 1. The ‘noise’ in the measurement makes the modeling difficult. This is also reflected by the fact that only ~80% training pattern recognition networks give better generalization error. For a noise-free quantitative target variable, twice as many training cases as weights may be more than enough to avoid overfitting. For a very noisy target variable, many more training cases than weights are needed to avoid overfitting.

Generally speaking, an IC_{50} value predicted by using a trained ANN is dependent on all of the used descriptors through some nonlinear relationship. The dependence of the predicted IC_{50} value on each individual descriptor is generally nonlinear and is usually complicated. However, one can estimate their relative contribution to the establishment of a QSAR model by excluding a descriptor together with its corresponding weights one by one from the model.³³ Considering NN611, descriptor i ($i = 1, 6$) together with its corresponding weights were eliminated from the 6-1-1 ANN and the resulting 5-1-1 was retrained as usual. The mean of the absolute values of the deviation ΔD_i between the experimental antagonist activity and the calculated activity for all compounds was evaluated. This procedure was repeated for each of all descriptors. The percent contribution (C_i) of the i th descriptor was given by

$$C_i = \frac{100 \cdot \Delta D_i}{\sum_{i=1}^6 \Delta D_i}. \quad (3)$$

Eq. 3 indicates that the contribution C_i of the i th descriptor to the trained neural network refers to the improvement of the overall predictions when the i th descriptor is added to the training.

Figure 2 shows the relative contribution of each descriptor. According to the plot in Figure 2, the dependence of the predicted IC_{50} value on two of the used descriptors looks apparent. One of them is the length of the alkyl

Table 3. $\log(1/IC_{50})$ values calculated by MLR, neural network models, and NN611 leave-one-out cross-validation for the 60 compounds listed in Table 1

Compound	Obsd $\log(1/IC_{50})$	NN (different configuration)			LOONN 611	Predicted by MLR
		611	511	5111		
T1	−1.39	−1.48	−1.49	−1.47	−1.49	−1.50
T2	−1.57	−1.35	−1.35	−1.34	−1.29	−1.16
T3	−0.96	−1.15	−1.14	−1.15	−1.21	−0.88
T4	−0.54	−0.48	−0.49	−0.52	−0.46	−0.32
T5	0.10	−0.12	−0.17	−0.18	−0.15	−0.09
T6	0.21	0.19	0.13	0.13	0.19	0.12
T7	0.68	0.42	0.35	0.36	0.41	0.29
T8	2.05	0.89	0.81	0.82	0.80	0.70
T9	−1.40	−1.33	−1.25	−1.24	−1.27	−1.41
T10	−1.30	−1.42	−1.41	−1.40	−1.44	−1.37
T11	0.21	0.48	0.47	0.50	0.55	0.29
T12	0.57	0.34	0.26	0.27	0.26	0.27
T13	−0.39	−0.20	−0.25	−0.27	−0.16	−0.12
T14	0.82	0.42	0.36	0.37	0.36	0.31
T15	1.40	0.92	0.88	0.90	0.80	0.64
T16	0.05	0.62	0.54	0.55	0.68	0.47
T17	1.10	0.96	0.94	0.96	0.82	0.74
T18	0.18	1.05	1.03	1.04	1.10	0.86
T19	0.24	1.15	1.16	1.17	1.22	1.07
T20	1.40	1.22	1.21	1.21	1.22	1.16
T21	0.66	1.30	1.33	1.32	1.40	1.46
T22	1.40	1.22	1.16	1.17	1.15	1.10
T23	0.51	1.22	1.24	1.23	1.30	1.23
T24	1.52	1.30	1.33	1.31	1.30	1.46
T25	1.40	1.40	1.40	1.38	1.40	1.68
T26	0.77	1.00	0.87	0.85	1.22	1.07
T27	0.89	0.28	0.24	0.25	0.12	0.29
T28	0.31	0.28	0.21	0.21	0.23	0.29
T29	0.59	0.70	0.65	0.67	0.70	0.57
T30	0.49	1.05	1.02	1.03	1.10	0.92
T31	0.92	1.30	1.26	1.25	1.30	1.27
T32	0.00	0.32	0.30	0.31	0.39	0.27
T33	0.52	0.77	0.75	0.78	0.80	0.61
T34	1.52	1.05	0.99	1.00	0.96	0.85
T35	0.00	0.59	0.54	0.53	0.89	0.66
T36	−0.22	−0.16	0.02	−0.01	−0.06	0.17
T37	2.00	1.00	1.03	1.03	0.70	1.04
T38	−0.18	0.19	0.27	0.26	0.34	0.32
T39	1.52	1.10	1.10	1.09	1.05	1.13
T40	2.30	1.22	1.22	1.21	1.10	1.36
T41	1.70	1.40	1.48	1.45	1.40	1.74
T42	2.15	1.05	1.10	1.10	0.66	1.03
P1	<0	0.57	0.49	0.51		0.37
P2	<0	1.10	1.13	1.14		0.90
P3	<0	1.30	1.33	1.32		1.16
P4	<0	−1.56	−1.57	−1.54		−2.22
P5	<0	−1.53	−1.54	−1.52		−1.51
P6	<0	−1.40	−1.44	−1.43		−0.81
P7	<0	−1.21	−1.27	−1.29		−0.51
P8	<0	−1.47	−1.47	−1.45		−1.49
P9	<0	−1.44	−1.43	−1.41		−1.47
P10	<0	−0.70	−0.72	−0.76		−0.31
P11	<0	−0.44	−0.44	−0.47		−0.18
P12	<0	−0.05	−0.08	−0.08		0.06
P13	<0	−1.39	−1.35	−1.36		−0.70
P14	<0	0.01	0.05	0.01		0.32
P15	<0	−0.24	−0.21	−0.26		0.17
P16	<0	−0.18	−0.17	−0.22		0.15
P17	<0	−1.57	−1.57	−1.55		−9.39
P18	<0	−1.57	−1.57	−1.55		−10.42

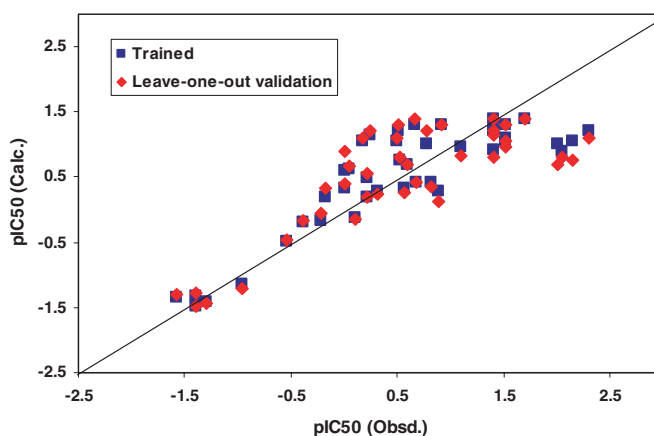


Figure 5. Trained and leave-one-out predicted pIC_{50} values versus experimental pIC_{50} values for NN611 model.

Table 4. Statistical results for the calculated $\log(1/IC_{50})$ values of 42 compounds in training set by MLR and neural network models

Method	r^2	RMSD	Max dev	Min dev
MLR	0.72	0.52	1.34	0.01
NN611	0.76	0.49	1.16	0.00
NN511	0.76	0.49	1.24	0.00
NN5111	0.76	0.49	1.22	0.02
Method	r_{cv}^2	RMSD	Max dev	Min dev
LOOMLR	0.59	0.62	1.42	0.02
LOONN611	0.64	0.59	1.49	0.00
LOONN511	0.64	0.58	1.33	0.00
LOONN5111	0.64	0.59	1.33	0.02

chain attached to a molecular head group, which explains 44% of the contributions. The other important descriptor is MLOGP, which explains 20% contribution of the total. Interestingly, MLOGP is among the key descriptors selected by hill-climbing-based SOM modeling as well. The remaining 36% is from AREA (13.0%), P1u (12.6%), MR (7.8%) and LUMO (2.6%). In most cases, it is shown that the IC_{50} value decreases by increasing the length of the alkyl chain. To obtain an antagonist activity $<0.1 \mu M$, the chain length needs to be between 10 and 23 Å, the area of molecules between 600 and 900 Å, P1u between 0.8–1 and MLOGP 3–5. A ligand with a MLOGP value less than 3 was experimentally measured as inactive for the nAChR subtypes responsible for mediating nicotine-evoked DA release, comparing the experimental IC_{50} values of the 42 molecules in the training set with the individual descriptor values in the selected variable set.

The compounds used in the ANN training set for determining the optimum ANN configurations did not include the 18 compounds P1–P18 in Table 1, since the experimental measurements for these molecules did not give quantitative IC_{50} values. However, these ligands from different compound series are good examples for testing the trained models. These samples were introduced to the trained ANNs, forcing the ANNs to predict their activity based on previous experience. The ANNs accurately classified 14 of them, while MLR accurately classified only 11.

Table 5. Classification of the activity of the 60 antagonists in Table 1

Model	R^2	R_{cv}^2	Training % correct	LOO % correct	Test % correct
MLR	0.77	0.74	88 (37/42)	88 (37/42)	61 (11/18)
NN511	0.86	0.77	90 (38/42)	88 (37/42)	78 (14/18)
NN5111	0.86	0.77	90 (38/42)	90 (38/42)	78 (14/18)
NN611	0.86	0.81	90 (38/42)	90 (38/42)	78 (14/18)
Hill-SOM	N/A	N/A	95 (40/42)	93 (39/42)	72 (13/18)

Table 5 summarizes the results using various models (MLR, NN511, NN5111, NN611, and Hill-SOM) as classifiers, by considering the leave-one-out validation and testing results together. The correlation coefficients (R^2 and R_{cv}^2) based on IC_{50} values are also listed where applicable to reflect the relationship between experimental and calculated IC_{50} values. In ANN approaches, NN611 and NN5111 have the best classification results, that is, 90% of leave-one-out validation, which is the same as that of their training. The supervised SOM model has the best classification overall in the training and LOO validation, that is, 95% and 93%, respectively. On the inactive test set compounds, supervised SOM has one less correct classification for P1–P18 than NN511, NN611, or NN5111. Not only does MLR have the worst classification ability compared to other models, but it also has lower predictive R_{cv}^2 values than ANN models that use variables chosen from Table 2a. Both statistics indicate that better predictive models are associated with the use of the variety of variables to include as much information of the target features as possible when modeling the current small dataset of compounds.

Table 6 provides more detailed information about the classifier's performance in the form of confusion matrices by considering the leave-one-out validation and test results together. It is shown that NN611 and NN5111 correctly classify 22 of 27 molecules as inactive compounds and 32 of 33 molecules as active compounds from the pool of 60 molecules (diagonal elements). The two models classify five inactive compounds as active and 1 active compound as inactive (nondiagonal elements), resulting in 3.0% false negatives (active compounds predicted to be inactive) and 18.5% false positives (inactive compounds predicted to be active).

Table 6. Confusion matrices based on the leave-one-out and testing results of the 60 compounds calculated by MLR, ANNs and SOM

		IC ₅₀ (Calc)									
		MLR		NN511		NN5111		NN611		SOM	
IC ₅₀ (Obsd)	<i>I</i>	<i>I</i>	<i>A</i>	<i>I</i>	<i>A</i>	<i>I</i>	<i>A</i>	<i>I</i>	<i>A</i>	<i>I</i>	<i>A</i>
	<i>A</i>	<i>A</i>									
		16	11	21	6	22	5	22	5	20	7
		1	32	1	32	1	32	1	32	1	32
% correct ^a	<i>I</i>	59.3	40.7	77.8	22.2	81.5	18.5	81.5	18.5	74.1	25.9
	<i>A</i>	3.0	97.0	3.0	97.0	3.0	97.0	3.0	97.0	3.0	97.0

I, the number of compounds with IC₅₀ value ≥ 1 μ M.

A, the number of compounds which IC₅₀ value < 1 μ M.

^a percent accuracy of a model for classifying active or inactive molecules.

during the classification. Since NN511 calculated the IC₅₀ value of T36 as 0.97 μ M compared to 1.41 μ M and 1.02 μ M calculated by NN611 and NN511, respectively, the percent accuracy of NN511 for classifying inactive compounds is 77.8% (i.e., 21/27). The ability of classifying inactive compounds by SOM (20/27 = 74.3%) is worse than that using the three ANN models, but better than that using the MLR (16/27 = 59.3%). It is noted that the accuracy of all the models in classifying active molecules is higher than that the accuracy in classifying inactive molecules. This is particularly true for the MLR model; a reasonable explanation for this is that there are more active molecules included in the training set.

With the trained models (NN511, NN5111, and NN611), we quantitatively calculated the IC₅₀ values of an additional 204 new compounds designed by modifying the structures of the compounds in the ANN training set. The values of the normalized descriptors in input vector for all of these compounds are between 0 and 1 using the same (maximum and minimum values in Eq. 1) scaling in the training set, which insured that the predicted activity values were the interpolation results of the QSAR models. Some of these previously predicted compounds have been synthesized and their IC₅₀ values have been experimentally determined. Table 7 lists the predicted and observed IC₅₀ values of 13 compounds. Among the 13 compounds, BQDI was evaluated as inactive for unknown reasons. The situation is similar to that observed for compound P1 (NDNI) in the initial 60 molecule set. Within the group of 12 new compounds, whose experimental IC₅₀ values are listed in Table 7, the RMSD of the predicted pIC₅₀ values from the corresponding experimental pIC₅₀ values is 0.47 for the NN611 model, 0.50 for the NN511 model, and 0.48 for the NN5111 model. The RMSD values for these new compounds are lower than the corresponding LOORMSD values of 0.58, 0.59, and 0.59, suggesting that the ANN models are robust. In addition, the Pearson correlation coefficients (r^2) for the pIC₅₀ values of these 12 compounds are 0.74, 0.73, and 0.74 for the NN611, NN511, and NN5111 models, respectively. The r^2_{cv} values calculated based on Eq. 2 for the pIC₅₀ of these 12 compounds are 0.56, 0.51, and 0.54 for the NN611, NN511, and NN5111 models, respectively.

Further examples of the predicted compounds are listed in Table 8. By cross-referencing the results of the three models shown in Table 8, the following conclusions are derived:

- (1) Modifying the structures of bis-quinolinium compounds by replacing the 5 and/or 8, as well as the 5' and/or 8' carbon atoms with nitrogen atoms is predicted to generate inactive antagonists at the nAChR subtypes responsible for mediating nicotine-evoked dopamine release (see 14–23 in Table 8).
- (2) Modifying the *N*-alkylnicotinium salts by attaching the terminus of the *N*-alkyl chain to a benzyl group is predicted to generate compounds with IC₅₀ values smaller than those compounds with two *S*-(–)-nicotinium head groups linked via a similar alkyl chain. (see T26 in Table 1 and 24–32 in Table 8).
- (3) *N*-alkylnicotinamide salts are predicted to be inactive (see 33–36 in Table 8).
- (4) Unsaturated *N*-alkylated pyridinium salts are generally predicted to be inactive (37–47 in Table 8).

The SAR insights gained from this current study are expected to be useful in the future design of new compounds that have potentially higher antagonist activity at the nAChR subtypes responsible for mediating nicotine-evoked dopamine release. These data will also aid in significantly reducing the need for synthesizing and screening large number of compounds by eliminating compounds with in silico-predicted low activities from the pool of candidate compounds. In practice, we set high priority to synthesize the compounds whose IC₅₀ values are predicted to be less than 0.1 μ M and use the most potent compounds to perform the subsequent pharmacology and animal behavior tests. However, whether such a compound will eventually become a clinically effective drug or not is dependent on many other 'drugability' factors, such as solubility, chemical stability, membrane transport, bioavailability, pharmacokinetics, and toxicity, etc., in addition to the receptor antagonist activity of the compound. The ANN QSAR model has already guided us to successfully discover six new compounds with experimental IC₅₀ values less than 0.1 μ M, as seen in Table 7, demonstrating that the

Table 7. Experimental verification for the predicted IC₅₀ values of some synthesized compounds^a

Compound	Structure	NN611 IC ₅₀ (μM)	NN511 IC ₅₀ (μM)	NN5111 IC ₅₀ (μM)	Obsd IC ₅₀ ^b (μM)
<i>4-Methyl-unsaturated N-alkylpyridinium salts</i>					
1		0.04	0.04	0.05	0.04 (0.009–0.22)
2		0.05	0.05	0.05	0.03 (0.006–0.20)
3		0.06	0.06	0.06	0.14 (0.03–0.59)
<i>Substituted-N,N'-bis-alkylpyridinium salts</i>					
4		0.05	0.05	0.05	0.07 (0.02–0.29)
5		0.04	0.03	0.04	0.19 (0.05–0.80)
6		0.04	0.04	0.04	0.08 (0.01–0.44)
<i>N,N'-bis-Alkylquinolinium salts</i>					
7		0.07	0.05	0.05	0.05 (0.01–0.10)
8		0.04	0.04	0.04	0.49 (0.20–0.78)
9		0.04	0.04	0.04	0.24 (0.07–0.41)
10		0.04	0.03	0.04	>1
11		0.04	0.03	0.03	0.27 (0.04–0.50)
<i>1,1'-Dimethyl-5,5'-alkylpyridinium salts</i>					
12		0.02	0.02	0.02	0.05 (0.005–0.49)
<i>1,1'-Dimethyl-2,3'-dipyrrolidine</i>					
13		37.26	37.32	35.81	21.9

^a Unpublished results. The detail will be reported by animal experimental scientists in the future.^b Numbers in parenthesis are confidence intervals.

Table 8. Designed salts of the halide ions, their structures, and predicted IC₅₀ values for the nAChR subtypes responsible for mediating nicotine-evoked dopamine release

Compound	Structure	NN611 IC ₅₀ (μM)	NN511 IC ₅₀ (μM)	NN5111 IC ₅₀ (μM)
<i>N,N'-bis-Alkyl-(1,4-dihydro-1,5-naphthyridinium) salts</i>				
14		33.3	34.52	32.37
15		32.81	34.23	32.12
16		22.94	27.24	26.65
17		28.69	31.62	30.02
18		32.41	34.26	32.17
<i>N,N'-bis-Alkyl-(1,4-dihydro-pyrido[2,3-b]pyrazinium) salts</i>				
19		36.89	37.16	35.36
20		36.78	37.11	35.28
21		34.83	36.39	34.30
22		36.47	37.00	35.09
23		35.21	36.56	34.49
<i>N-(Alkylphenyl)-nicotinium salts</i>				
24		0.06	0.06	0.06
25		0.08	0.08	0.08

(continued on next page)

Table 8 (continued)

Compound	Structure	NN611 IC ₅₀ (μM)	NN511 IC ₅₀ (μM)	NN5111 IC ₅₀ (μM)
26		0.07	0.07	0.07
27		0.03	0.03	0.03
28		0.03	0.03	0.03
<i>N,N'</i> -bis-Alkylpyridinium salts				
29		1.63	1.72	2
30		1.3	1.46	1.69
31		0.39	0.51	0.56
32		0.72	0.91	1.04
<i>N</i> -Alkylpyridiniumamide salts				
33		31.12	32.93	30.89
34		27.88	30.43	28.85
35		22.11	25.65	25.04
36		17.98	21.94	22.02

Table 8 (continued)

Compound	Structure	NN611 IC ₅₀ (μM)	NN511 IC ₅₀ (μM)	NN5111 IC ₅₀ (μM)
<i>Unsaturated N-alkylpyridinium salts</i>				
37		35.19	35.38	32.98
38		4.59	4.27	4.48
39		13.54	10.24	10.25
40		1.54	1.52	1.55
41		2.14	1.86	1.89
42		0.43	0.42	0.4
43		1.85	1.68	1.7
44		4.44	2.98	2.95
45		0.23	0.23	0.22
46		0.23	0.22	0.21
47		0.30	0.27	0.26

ANN QSAR model is a valuable aid to drug discovery.

4. Conclusion

With the neural network pattern recognition technique, we present here a QSAR modeling approach to predict the IC₅₀ values of a set of mono- and bis-quaternary ammonium salts that were developed in our laboratory as antagonists for the nicotine-evoked dopamine releasing nAChR subtype. The results have shown that ANN can be used to predict the activity of drugs from calculable information derived from structures and available physicochemical descriptors. A dataset of 60 mono- and bis-quaternary ammonium analogs was analyzed. A SOM classification model and 3 QSAR ANN models with consistent results ($r^2 = 0.76$, $r_{cv}^2 = 0.64$) were developed. The modeling study also shows that two factors are important for predicting the antagonist activity of these compounds of the generated models, that is, length of the alkyl chain attached to a molecular head group,

and the Moriguchi octanol–water partition coefficient. Compared with results from the MLR model, the neural network models are better than the corresponding multiple linear regressions in predicting the binding activities from a set of comparable molecular descriptors. Furthermore, classification-based results (Tables 5 and 6) demonstrate that MLR affords the worst classification performance when compared to those of ANN and SOM. Both of these results imply that nonlinear neural network models are preferred, while the linear MLR model is also predictive when trained with the experimental pattern utilizing the set of comparative descriptors.

Utilizing the neural network technique to determine structure–activity relationships for predicting binding activity of ligands at the nAChR subtypes responsible for mediating nicotine-evoked dopamine release, low-energy conformations of the ligands (generally with a straight alkyl chain) constructed in a consistent manner, have been used to generate the three-dimensional molecular descriptors. The reliability of the models developed in this study has been scored by the validation and test-

ing results of the models. Recently obtained activity data of some newly designed and synthesized compounds further support the robustness of the trained ANN models. Like many QSAR studies, our approach is also limited by the small size of the training set. However, the predictive models developed in this study appear to be sufficiently robust to interpolatively characterize a larger chemical library of new molecules, based on the modification of the structures of the compounds in the training set for the identification and optimization of lead compounds in the context of combinatorial chemistry. Thus, these models may therefore reduce the need for the synthesis and biological assay of large number of analogs by eliminating compounds of predicted low activity from the pool of candidate compounds. As more candidate antagonist molecules with predicted high activities ($IC_{50} < 0.1 \mu M$) are assayed, the resulting IC_{50} data can be used to challenge and refine the current QSAR models. The output of the recursive process from integration of the experimental and modeling efforts will afford a model with an improved capability for predicting IC_{50} values of ligands as antagonists at nAChR subtypes mediating nicotine-evoked dopamine release.

Acknowledgment

This work was supported by NIH Grant No. 1U19DA017548.

References and notes

1. Ericson, N. U.S. Department of Justice, 2001 May #17.
2. (a) Hurt, R. D.; Sachs, D. P. L.; Glover, E. D.; Offord, K. P.; Johnston, J. A.; Dale, L. C.; Khayrallah, M. A.; Schroeder, D. R.; Glover, P. N.; Sullivan, C. R.; Croghan, I. T.; Sullivan, P. M. *N. Eng. J. Med.* **1997**, *337*, 1195; (b) Jorenby, D. E.; Leischow, S. J.; Nides, M. A.; Rennard, S. I.; Johnston, J. A.; Hughes, A. R.; Smith, S. S.; Muramoto, J. L.; Daughton, D. M.; Doan, K.; Fiore, M. C.; Baker, T. B. A. *N. Eng. J. Med.* **1999**, *340*, 685; (c) Shiffman, S.; Johnston, J. A.; Khayrallah, M.; Elash, C. A.; Gwaltney, C. J.; Paty, J. A.; Gnys, M.; Evoniuk, G.; DeVeauh-Geiss *Psychopharmacology (Berl.)* **2000**, *148*, 33; (d) Rose, J. E.; Behm, F. M.; Westman, E. C.; Levin, E. D.; Stein, R. M.; Ripka, G. V. *Clin. Pharmacol. Ther.* **1994**, *56*, 86; (e) Rose, J. E.; Westman, E. C.; Behm, F. M.; Johnson, M. P.; Goldberg, J. S. *Pharmacol. Biochem. Behav.* **1999**, *62*, 165–172.
3. Crooks, P. A.; Ayers, J. T.; Xu, R.; Sumithran, S. P.; Grinevich, V. P.; Wilkins, L. H., Jr.; Deaciuc, A. G.; Allen, D. D.; Dwoskin, L. P. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 1869.
4. Dwoskin, L. P.; Sumithran, S. P.; Zhu, J.; Deaciuc, A. G.; Ayers, J. T.; Crooks, P. A. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 1863.
5. Xu, R.; Dwoskin, L. P.; Grinevich, V.; Sumithran, S. P.; Crooks, P. A. *Drug Dev. Res.* **2002**, *55*, 173–186.
6. Wilkins, L. H., Jr.; Grinevich, V. P.; Ayers, J. T.; Crooks, P. A.; Dwoskin, L. P. *J. Pharmacol. Exp. Ther.* **2003**, *304*, 400.
7. Dwoskin, L. P.; Wilkins, L. H., Jr.; Pauly, J. R.; Crooks, P. A. *Ann. N.Y. Acad. Sci.* **1999**, *868*, 617.
8. Wilkins, L. H., Jr.; Haubner, A.; Ayers, J. T.; Crooks, P. A.; Dwoskin, L. P. *J. Pharmacol. Exp. Ther.* **2002**, *301*, 1088.
9. Crooks, P. A.; Ravard, A.; Wilkins, L. H., Jr.; Teng, L. H.; Buxton, S. T.; Dwoskin, L. P. *Drug. Dev. Res.* **1995**, *36*, 91.
10. Dwoskin, L. P.; Crooks, P. A. *J. Pharmacol. Exp. Ther.* **2001**, *298*, 395.
11. (a) Nicolotti, O.; Pellegrini-Calace, M.; Altomare, C.; Carrieri, A.; Carotti, A.; Sanz, F. *Curr. Med. Chem.* **2002**, *9*, 1; (b) Gohlke, H.; Schwarz, S.; Gundisch, D.; Tilotta, M. C.; Weber, A.; Wegge, T.; Seitz, G. *J. Med. Chem.* **2003**, *46*, 2031; (c) Glennon, R. A.; Herndon, J. L.; Dukat, M. *Med. Chem. Res.* **1994**, *4*, 461.
12. Tonder, J. E.; Olesen, P. H. *Curr. Med. Chem.* **2001**, *8*, 651.
13. Gao, F.; Bren, N.; Little, A.; Wang, H.-L.; Hansen, S. B.; Talley, T. T.; Taylor, P.; Sine, S. M. *J. Biol. Chem.* **2003**, *278*, 23020.
14. Wang, H.-L.; Gao, F.; Bren, N.; Sine, S. M. *J. Biol. Chem.* **2003**, *278*, 32284.
15. Schapira, M.; Abagyan, R.; Totrov, M. *BMC Struct. Biol.* **2002**, *2*, 1.
16. Henchman, R. H.; Wang, H.-L.; Sine, S. M.; Taylor, P. *Biophys. J.* **2003**, *85*, 3007; Sixma, T. K.; Smit, A. B. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 311.
17. Bikadi, Z.; Simonyi, M. *Curr. Med. Chem.* **2003**, *10*, 1241.
18. Schmitt, J. D. *Curr. Med. Chem.* **2000**, *7*, 749.
19. (a) Ochoa, C.; Chana, A.; Stud, M. *Curr. Med. Chem—Central Nervous System Agents* **2001**, *1*, 247; (b) Kaiser, K. L. E. *Quant. Struct.-Act. Relat.* **2003**, *22*, 1.
20. Aoyama, T.; Suzuki, Y.; Ichikawa, H. *J. Med. Chem.* **1990**, *33*, 2583.
21. Zhu, J.; Chen, D.; Wu, P. *Comput. Chem.* **1999**, *23*, 97.
22. Yan, A.; Jiao, G.; Hu, Z.; Fan, B. T. *Comput. Chem.* **2000**, *24*, 171–179.
23. Bleckmann, A.; Meiler, J. E. *QSAR Comb. Sci.* **2003**, *22*, 722.
24. (a) Tripos discovery software package with SYBYL 6.8.1 Tripos, 1699 South Hanley Road, St. Louis, Missouri 63144, USA <http://www.tripos.com/sciTech/inSilicoDisc/dataAnalysis/lithium.html>; (b) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, Revision A.1*; Gaussian: Pittsburgh, PA, 2003.
25. DRAGON software version 3.0 2003 developed by Milano Chemometrics and QSAR Research Group (<http://www.disat.nimib.it/chm/Dragon.htm>). For Ghose–Crippen molar refractivity, see: Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. *J. Chem.*

- Inf. Comput. Sci.* **1989**, 29, 163–172; For Moriguchi octanol–water partition coefficient see: Moriguchi, I.; Hirono, S.; Liu, Q.; Nakagome, I.; Matsushita, Y. *Chem. Pharm. Bull.* **1992**, 40, 127–130.
26. Revised based on a C program originally from the neural network Senior Research Scientist Lars J. Kangas at Pacific Northwest National Laboratory.
27. David, M. *Skapura Building Neural networks*; Addison-Wesley: Reading, MA, 1996, P87.
28. Bayram, E.; Santago, P.; Harris, R.; Xiao, Y.; Clauset, A. J.; Schmitt, J. D. *J. Comput.-Aid. Mol. Des.* **2004**, 18, 483–493.
29. Rich, E.; Knight, K. *Artificial Intelligence*, 2nd ed.; McGraw-Hill: New York, 1991.
30. Schneider, G., *Modeling Structure–Activity Relationships*, Schneider, G.; Sung-San So, Landes Bioscience, 2004.
31. Shao, J.; Tu, D. *The Jackknife and Bootstrap*; Springer: New York, 1995.
32. Gotte, C. *Neural Comput.* **1997**, 9, 1211–1215; Zhu, H.; Rohwer, R. *Neural Comput.* **1996**, 8, 1421–1426.
33. (a) Cherqaoui, D.; Villemin, D. *J. Chem. Soc., Faraday Trans.* **1994**, 90, 97; (c) Rajarshi Guha; Peter, C. Jurs. *J. Chem. Inf. Model.* **2005**, 45, 800.
34. Kohonen, T. *Self-Organizing Maps*; Springer Series in Information Sciences; Vol. 3, extended edition, 2001.
35. De Veaus, R. D.; Ungar, L. H. Multicollinearity: A Tale of Two Nonparametric Regressions. In *Selecting Models from Data: AI and Statistics IV*; Cheeseman, P., Oldford, R. W., Eds.; Springer: New York, 1994; pp 293–302.