Prediction of Glycine/NMDA Receptor Antagonist Inhibition from Molecular Structure

S. J. Patankar and P. C. Jurs*

Department of Chemistry, 152 Davey Laboratory, Penn State University, University Park, Pennsylvania 16802

Received November 6, 2001

The design and blood brain barrier crossing of glycine/NMDA receptor antagonists are of significant interest in pharmaceutical research. The use of these antagonists in stroke or seizure reduction have been considered. Measuring the inhibitory concentrations, however, can be time-consuming and costly. The use of quantitative structure−activity relationships to estimate IC_{50} values for these receptor antagonists is an attractive alternative compared to experimental measurement. A data set of 109 compounds with measured log(IC_{50}) values ranging from −0.57 to 4.5 is used. Structural information is encoded with numerical descriptors for topological, electronic, geometric, and polar surface properties. A genetic algorithm with a computational neural network fitness evaluator is used to select the best descriptor subsets. Multiple linear regression and computational neural network models are developed. Additionally, a quantitative radial basis function neural network (QRBFNN) was developed with the intent of introducing nonlinearity at a faster speed. A genetic algorithm using the radial basis function network as a fitness evaluator was also developed to search descriptor space for optimum subsets. All models are tested using an external prediction set. The nonlinear computational neural network model has root-mean-square errors of approximately half a log unit.

INTRODUCTION

A great deal of evidence supports the theory that overactivation of *N*-methyl-D- aspartate (NMDA) subtype of central excitatory amino acid receptor plays a role in a number of neurodegenerative disorders.[1] There are several sites on the receptor ion channel complex such as glutamate, glycine, polyamines, Mg^{2+}, Zn^{2+}, at which antagonists may act.[2] Considerable attention is focused on the glycine site on the NMDA receptor as glycine acts as a coagonist in the presence of glutamic acid.[3,4] Several factors affect central nervous system penetration and glycine/NMDA receptor antagonist activity of potential pharmacophores.[5−10] Structure activity relationship studies of several classes of glycine antagonists have been reported.[11−13] However these compounds lack in vivo activity after intravenous administration. Several variants of hydroxyquinolin exhibit greatly improved in vivo properties.[14]

In this paper we report the application of quantitative structure activity relationships (QSAR) to predict the inhibitory concentrations of a varied set of hydroxyquinolins. This data set is unusual as it contains large substituent variations, chiral centers, and significant changes in acidity as well as the basicity of the compounds.

Blood brain barrier penetration ability of these compounds can be significantly changed by any of these factors. Prediction of inhibitory concentrations was done on the basis of molecular structure alone and also using the reported log *P* values. A number of studies,[15−20] in different areas, have stressed the benefits of QSAR, such as cost saving, safety, lack of consumption of test sample, and shortened time, all of which are applicable here.

The QSAR methodology used in this study consists of three main parts: representation of molecular structure, feature selection, and mapping. Since the general assumption in QSAR modeling is that molecular structure causes the observed behavior of a compound, a series of chemical structures are linked to properties of interest. In this case the property of interest is the inhibitory concentrations for glycine/NMDA receptor binding. A necessary first step involves encoding the structures. This encoding is done by using calculated structural descriptors, which are mathematical representations of molecular structure. For example, the molecular volume can provide some information regarding the size of that structure. To best encode the structures, it is typically useful to calculate a multitude of descriptors, each helpful in describing the structure from a different prospective.

Once the structures have been encoded, the subset of descriptors that best encodes the property of interest must be found. Feature selection methods, employing the genetic algorithm (GA)[16,21] coupled with computational neural networks[22] are used for this purpose. The GA is effective in finding minima for complex problems without any knowledge of the form of the objective function. As a large number of descriptors are calculated in this particular QSAR approach, feature selection routines must be utilized.

In an effort to increase the speed of training, a GA routine that used radial basis function as a fitness evaluator was developed and used for descriptor space subsearching. Once a subset of descriptors is found, the descriptors are then mapped to the property of interest, using either a multiple linear regression equation or a nonlinear computational neural network. These mapping methods effectively provide a mechanism for linking a chemical structure with its corresponding IC_{50} values.

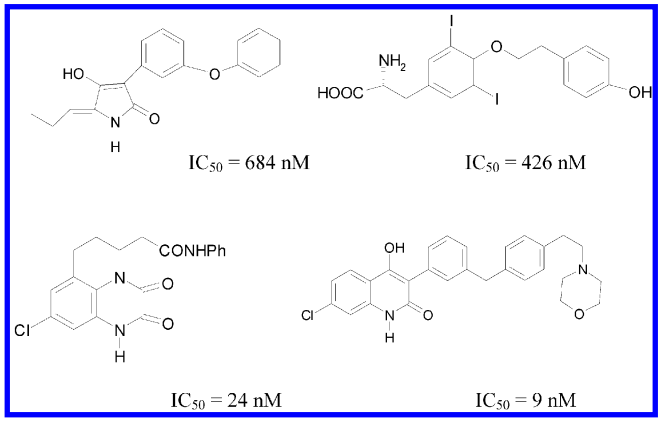
EXPERIMENTAL AND METHODOLOGY

This study was performed on a combined data set taken from three papers, two by Rowely et al.[23,14] and one by

* Corresponding author phone: (814)865-3739; fax: (814)865-3314; e-mail: pcj@psu.edu.

10.1021/ci010114+ CCC: $22.00
Published on Web 09/04/2002

Leeson et al.[24] The concentration of test compounds required to inhibit 50% of the specific binding (IC_{50} values) was evaluated by displacement of glycine site antagonist binding to rat cortex/hippocampus membranes and reported as nanomoles.[25] Only about 2% of the IC_{50} values ranged in the 10 thousandths range so this data set is heavily biased toward smaller values of inhibitory concentration given the tendency toward successful development of orally active drug compounds. A cursory examination of the compounds in this data set reveals a large amount of structural diversity. An idea of the diversity of data set can be seen from the following structural representations.



Seventeen of the reported compounds had to be excluded from the study as their IC_{50} values were reported as inequalities.[24]

The data set of 126 compounds containing N, O, S, and halogens had 111 compounds containing halogens, 17 compounds containing S, and all compounds containing at least one N and one O. The molecular formulas for this data set ranged from 13 non-hydrogen atoms to 34 non-hydrogen atoms. The total number of atoms in the compounds varied from 19 to 61. The molecular weight varied from 206 to 567 Da. The reported IC_{50} values were in the range from 0.27 to 30 000 nM. To compress the range of the data, log(IC_{50}) (-0.57 to 4.5) was used as the dependent variable.

However about two-thirds of the compounds had log(IC_{50}) values less than two, as the search is done for more potent inhibitors of glycine/NMDA binding site.

Out of this data set of 126 compounds, 17 compounds had their IC_{50} values reported as inequalities. These 17 compounds were set aside as an exclusion set. A subset of 11 compounds out of the remaining 109 compounds was chosen randomly as a prediction set (PSET) and was used for external validation of all the models that were developed in this study. The external prediction set was chosen in such a way as to cover the entire range of IC_{50} values of the data set. The compounds in the external prediction set were never used during the model development process but were reserved to validate potential models. For the development of a multiple linear regression model, the training set (TSET) included all the remaining 98 compounds. For the generation of nonlinear CNN models, the training set was further subdivided into a training set containing 89 compounds and a cross-validation set (CVSET) containing 9 compounds. Table 1 lists these compounds and their experimental as well as calculated log(IC_{50}) values.

The computations for this work were performed at Penn State University on a DEC 3000 AXP Model 500 workstation running the OSF/1 V3.0B operating system. Those calculations involving HyperChem[26] were performed on a Pentium PC. The compounds were sketched as 2-D representations using HyperChem,[26] and optimized 3-D conformations were generated. These were further refined to their lowest energy states using MOPAC,[27] a semiempirical molecular modeling routine, using the PM3 Hamiltonian.[28] These optimum 3-D conformations were used for generation of descriptors, which were dependent on geometry. The ADAPT (**A**utomated **D**ata **A**nalysis and **P**attern Recognition **T**oolkit) software package[29,30] was used to calculate more than 200 molecular structure descriptors for each compound. These descriptors encode the geometric, topological, electronic, and polar surface area features of these compounds.

Geometric descriptors included solvent-accessible surface area,[31] molecular volume,[31] molecular polarizability,[32] and moments of inertia.[33] Accurate three-dimensional geometries of the molecules are necessary to calculate descriptors of this nature. Topological descriptors are derived from information about the 2-D structure of the molecule. Graph theory can be applied to the 2-D structures to generate a multitude of topological indices. Topological descriptors[34−37] included counts of atom types, bond types, numbers of basis rings, and functional groups as well as molecular connectivity indices to represent size and degree of branching. Electronic descriptors[38] stored the partial atomic charge descriptors, such as the most positive or most negative atoms, energy of the highest occupied molecular orbital, energy of the lowest unoccupied molecular orbital, and dipole moments. Polar surface area descriptors that combined geometric as well as electronic information such as hydrogen bonding and charged partial surface area (CPSA) were also included.[39,40] As the data set had the dissociable acidic hydrogens and heteroatoms, intramolecular as well as intermolecular hydrogen bonding was expected. Therefore hydrogen bonding descriptors were generated for pure and solvated states.

In addition a set of six descriptors was calculated using information from MOPAC runs using the MNDO Hamiltonian. These descriptors[41] were developed from semiempirical molecular orbital calculations and include a term represented by van der Waals volume, a volume-independent molecular polarizibility term, covalent hydrogen bonding acidity and basicity, and electrostatic hydrogen bonding acidity and basicity. In the current work five descriptors were calculated, and the steric term was represented by the volume descriptor generated by other ADAPT routines. As a measure of the reactivity of each atom a set of eight electro-topological state descriptors[42] were also calculated.

They encoded the information regarding intermolecular attractions. A new set of topological descriptors[43] was found to be useful in this study. The molecular distance-edge (MDE) vector consisted of the descriptor values for 10 distance-edge terms between four different types of carbon atoms.

The next step was to use objective feature selection (selection not using the dependent variable) to discard descriptors, which contained redundant or minimal information. Three methods of objective feature selection were employed. First, all descriptors that had greater than 90% identical values were removed since they were not encoding

Table 1. Structures, Experimental, and Predicted (IC_{50}) Values of NMDA Inhibitors

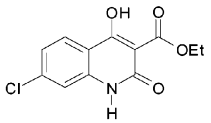
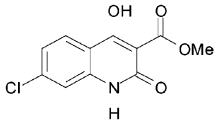
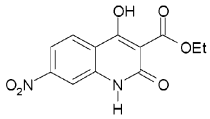
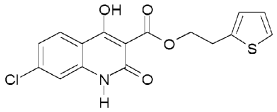
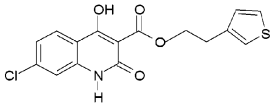
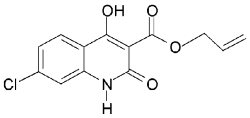
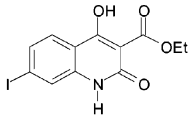
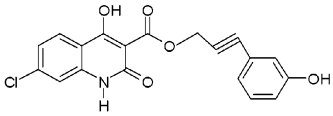
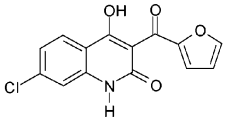
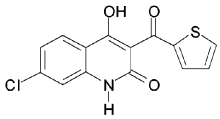
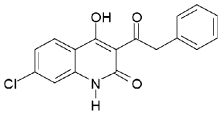
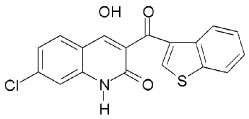
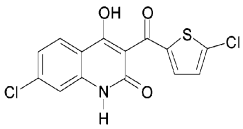
Comp. No.	Structure	IC_{50}	Experimental log IC_{50}	Calculated [a] log IC_{50}
1		16700	4.223	4.069
2		6450	3.810	3.622
3		25400	4.405	2.884
4		3000	3.477	3.346
5		3000	3.477	3.238
6		6420	3.808	3.641
7		30000	4.477	4.030
8		175	2.243	0.560
9[b]		2390	3.373	3.277
10		1090	3.037	3.277
11[c]		932	2.969	0.775
12[c]		2230	3.348	3.110
13[b]		1520	3.182	3.443

Table 1. (Continued)

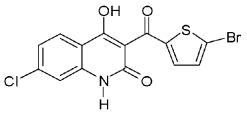
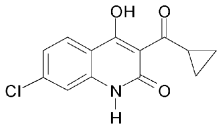
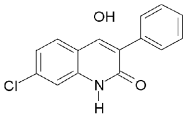
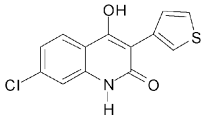
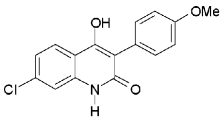
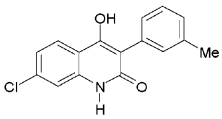
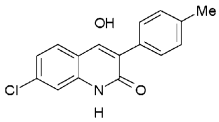
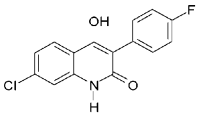
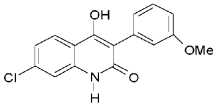
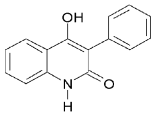
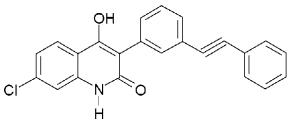
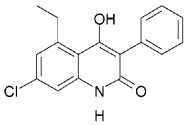
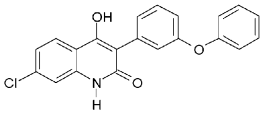
Comp. No.	Structure	IC_{50}	Experimental log IC_{50}	Calculated [a] log IC_{50}
14		954	2.980	2.806
15		419	2.622	3.902
16[c]		172	2.236	2.588
17		352	2.547	2.360
18		421	2.624	1.609
19		84	1.924	2.599
20		391	2.592	2.612
21		658	2.818	3.088
22[b]		204	2.310	1.599
23[c]		7500	3.875	4.568
24		34.3	1.535	0.922
25		6.9	0.839	1.066
26		2	0.301	0.844

Table 1. (Continued)

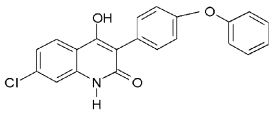
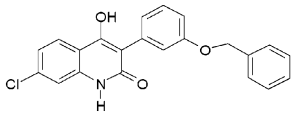
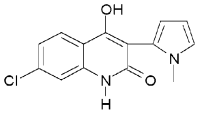
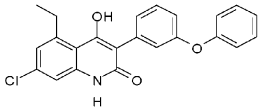
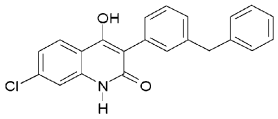
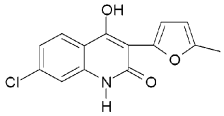
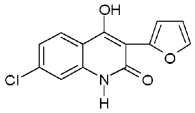
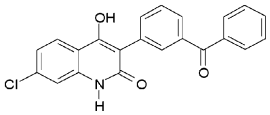
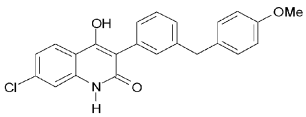
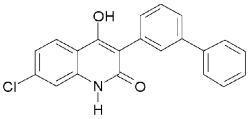
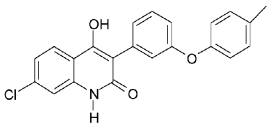
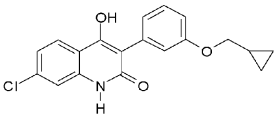
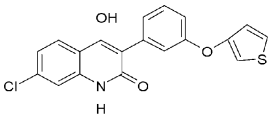
Comp. No.	Structure	IC_{50}	Experimental log IC_{50}	Calculated [a] log IC_{50}
27		12.1	1.083	0.892
28[c]		12.2	1.086	0.708
29		818	2.913	2.507
30[c]		1.95	0.290	0.533
31		4	0.602	0.996
32		481	2.682	2.794
33		747	2.873	3.037
34[b]		3.6	0.556	0.815
35		4.5	0.653	0.754
36		22.8	1.358	1.726
37[c]		7.8	0.892	0.863
38		37.8	1.577	0.740
39		1.4	0.146	0.684

Table 1. (Continued)

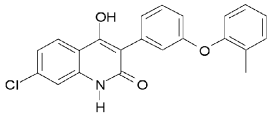
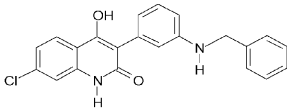
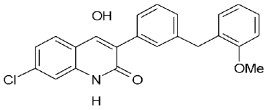
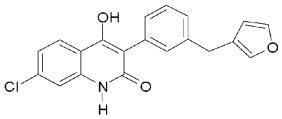
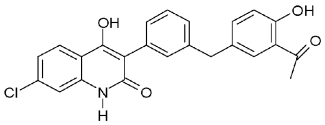
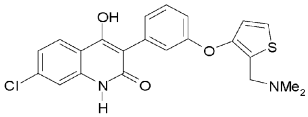
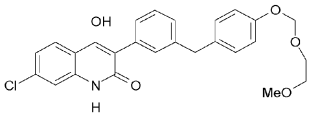
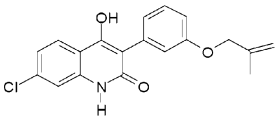
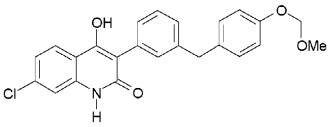
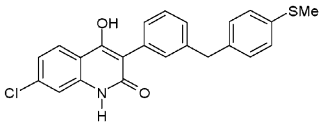
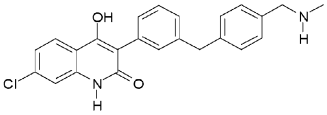
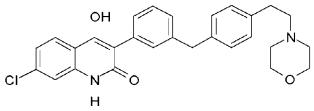
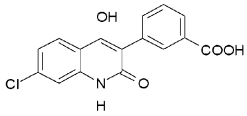
Comp. No.	Structure	IC_{50}	Experimental log IC_{50}	Calculated [a] log IC_{50}
40		1.3	0.114	0.813
41		9.6	0.982	0.881
42[b]		8	0.903	0.719
43[c]		8.9	0.949	1.414
44		42.7	1.630	1.538
45		32.6	1.513	0.704
46		3.6	0.556	0.491
47		10.9	1.037	0.949
48		2.2	0.342	0.534
49		2.4	0.380	0.660
50		124	2.093	1.203
51[b]		9	0.954	0.514
52		75.6	1.879	2.006

Table 1. (Continued)

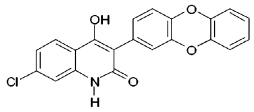
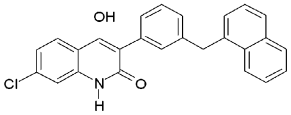
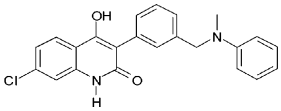
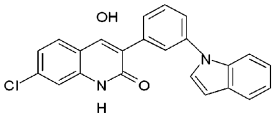
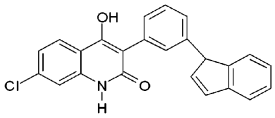
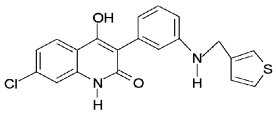
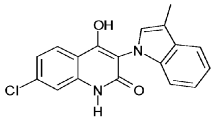
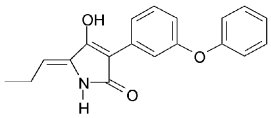
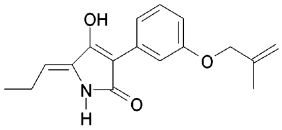
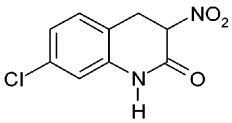
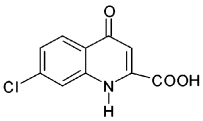
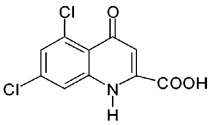
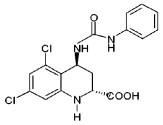
Comp. No.	Structure	IC_{50}	Experimental log IC_{50}	Calculated [a] log IC_{50}
53		3000	3.477	2.797
54		27.9	1.446	1.041
55		11.7	1.068	0.721
56		10.2	1.009	0.920
57		6	0.778	1.467
58		3.3	0.519	0.654
59		157	2.196	2.148
60		684	2.835	3.093
61		3030	3.481	3.440
62[c]		414	2.617	2.672
63		320	2.505	2.438
64		64	1.806	1.014
65		4	0.602	0.903

Table 1. (Continued)

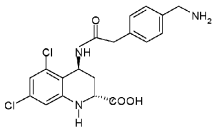
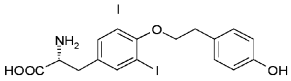
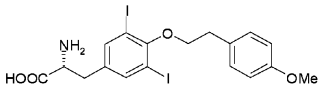
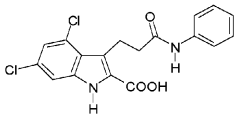
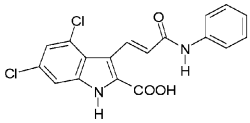
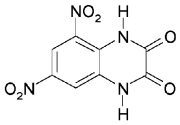
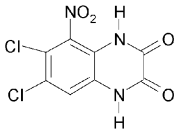
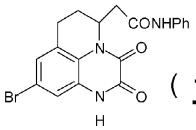
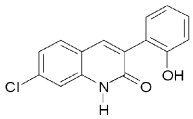
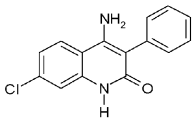
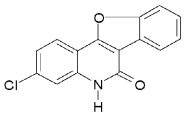
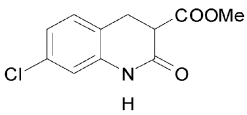
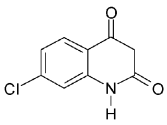
Comp. No.	Structure	IC_{50}	Experimental log IC_{50}	Calculated [a] log IC_{50}
66[c]		38.5	1.585	2.433
67[b]		426	2.629	2.925
68		545	2.736	2.654
69[b]		36.4	1.561	0.761
70		1	0.000	0.826
71		170	2.230	2.398
72		2.8	0.447	1.285
73	 (±)	24	1.380	1.331
74		563	2.751	3.273
75		6750	3.829	3.189
76		8340	3.921	3.988
77[d]		101	2.004	2.035
78[d]		101	2.004	2.714

Table 1. (Continued)

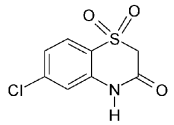
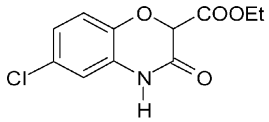
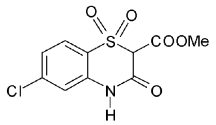
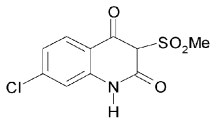
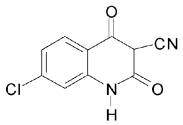
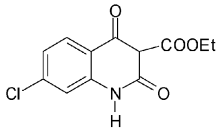
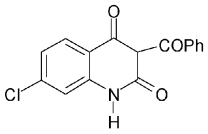
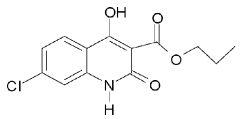
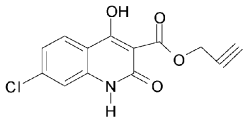
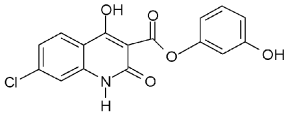
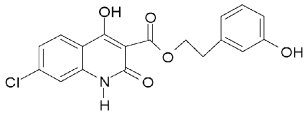
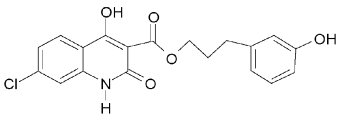
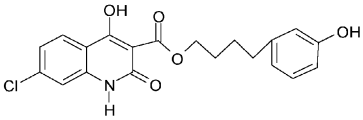
Comp. No.	Structure	IC_{50}	Experimental log IC_{50}	Calculated [a] log IC_{50}
79[d]		101	2.004	1.508
80[d]		101	2.004	1.715
81		13.6	1.134	1.397
82		15.4	1.188	0.764
83		3.9	0.591	0.907
84		16.7	1.223	0.776
85		3.16	0.500	0.636
86[d]		101	2.004	2.549
87		9	0.954	1.037
88		6.13	0.787	0.592
89		1.82	0.260	0.524
90		2.64	0.422	0.510
91		2.62	0.418	0.496

Table 1. (Continued)

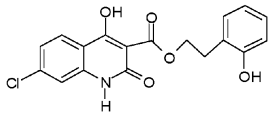
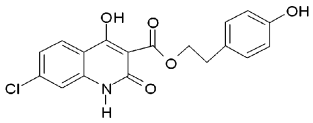
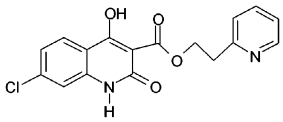
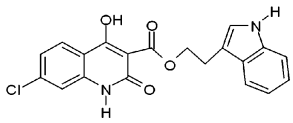
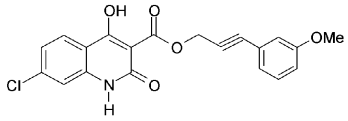
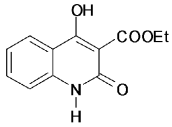
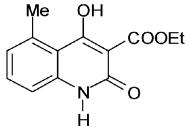
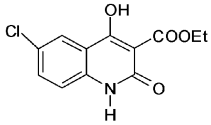
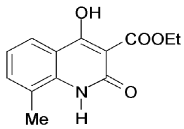
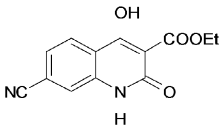
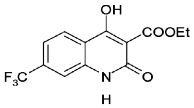
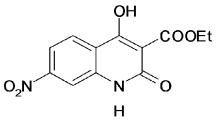
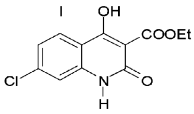
Comp. No.	Structure	IC_{50}	Experimental log IC_{50}	Calculated [a] log IC_{50}
92		7.12	0.852	0.520
93		2.93	0.467	0.540
94		27.5	1.439	0.536
95		0.43	-0.367	0.612
96		0.27	-0.569	0.509
97[d]		101	2.004	1.714
98[d]		1001	3.000	2.932
99[d]		101	2.004	1.535
100[d]		101	2.004	1.696
101		26	1.415	2.471
102		12.3	1.090	0.485
103		25.4	1.405	1.405
104		1.61	0.207	0.644

Table 1. (Continued)

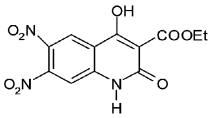
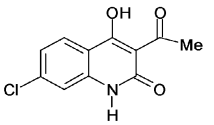
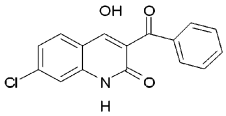
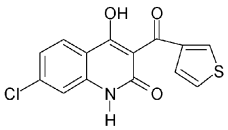
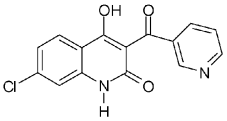
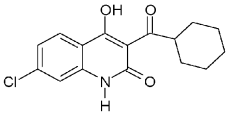
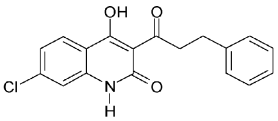
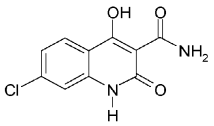
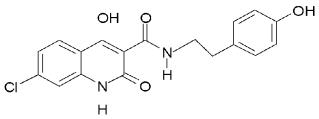
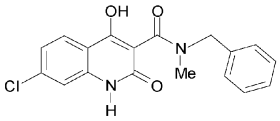
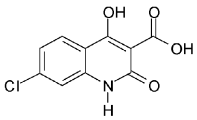
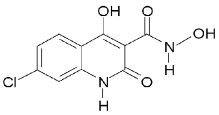
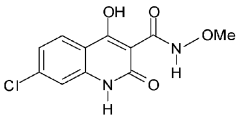
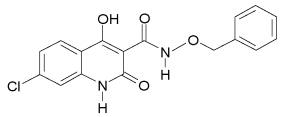
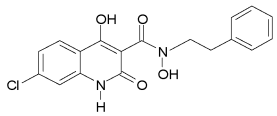
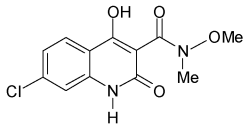
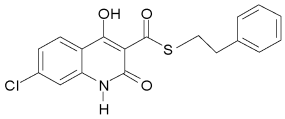
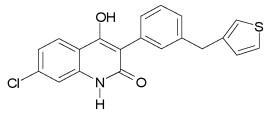
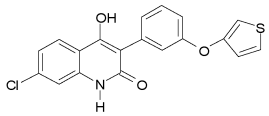
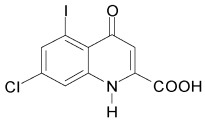
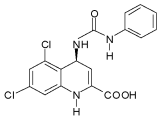
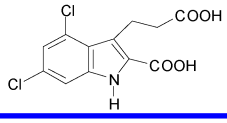
Comp. No.	Structure	IC_{50}	Experimental log IC_{50}	Calculated [a] log IC_{50}
105		10.5	1.021	1.282
106[d]		101	2.004	1.923
107		3.16	0.500	0.854
108[c]		1.97	0.294	0.729
109		6.85	0.836	0.935
110[d]		101	2.004	1.882
111[d]		101	2.004	1.758
112[d]		101	2.004	1.550
113[d]		101	2.004	1.903
114[d]		101	2.004	2.202
115		5.82	0.765	0.710
116		6.32	0.801	0.995
117		3.09	0.490	0.999

Table 1. (Continued)

Comp. No.	Structure	IC_{50}	Experimental log IC_{50}	Calculated [a] log IC_{50}
118[d]		101	2.004	1.721
119		9.3	0.968	0.600
120		47.1	1.673	1.825
121[d]		11	1.041	1.522
122		2.3	0.362	0.850
123		2.4	0.146	0.684
124		14	1.146	1.399
125		4	0.602	0.685
126		69	1.839	1.639

[a] All values calculated on th basis of Type 3 CNN model. [b] Cross-validation set compounds. [c] Prediction set compounds. [d] Exclusion set compounds.

the structural differences between the compounds. Second, pairs of descriptors were examined for redundancy. If two descriptors were pairwise correlated with an $r \geq 0.93$, one of them was removed from the pool. This value is used as an empirical cutoff as it worked well from past experience. These two steps reduced the 240 descriptor pool to 98 descriptors. This reduced pool of descriptors was further subjected to a vector space descriptor analysis routine. This routine treats the descriptors as multidimensional vectors and uses a stepwise orthogonalization procedure to find a subset of mutually orthogonal descriptors, which further lowers the likelihood of chance correlation. However, this reduced pool of descriptors still contains most of the variance in the data. At this stage the 98 descriptor pool was further reduced to 42 descriptors. This was deemed acceptable as the ratio of descriptors to TSET observations was well below 0.6 for all cases investigated.

Multiple linear regression models (Type 1 models) that link the property of interest to the structures can be developed using subsets of descriptors selected from the reduced descriptor pool. A simulated annealing[44,45] feature selection routine and a geneticalgorithm[46] feature selection routine was used to find good descriptor subsets. Model sizes ranging from 3 to 15 descriptors were investigated. For each model size the subset of descriptors that produced lowest rms errors were further investigated. Models of various sizes were compared to find the model providing the lowest rms error with the fewest descriptors. Once selected, the best model was then used to predict estimated log(IC_{50}) values for the PSET.

The set of descriptors chosen from the best linear model was subsequently submitted to computational neural networks (CNN)[47,48] to develop a nonlinear (Type 2) model. The CNNs used here are fully connected, three-layer, feed-forward neural networks. The set of descriptors chosen for the best linear model was used as input neurons. The hidden layer neurons were varied keeping only one output neuron to find the best nonlinear model. The best model was selected on the basis of fewer neurons and lower rms error.

Often the descriptors that best encode the linear relationship do not yield the best nonlinear relationship. Hence a feature selection routine which combined the genetic algorithm with a neural network fitness evaluator was also used in this study. These fully nonlinear CNN models are called Type 3 models. The models were then developed by training CNN to minimize rms error values for CVSET. Beyond this point the network was considered over trained as it learned the idiosyncrasies of training set. The quality of the model was based on the fitness function. The fitness function was defined as quality equal to rms error of TSET plus 0.4 times the difference in the rms errors of TSET and CVSET. The reduced pool of 42 descriptors was submitted to GA. Starting from 12:2:1, different architectures such as 11:3:1, 9:4:1, were investigated. The ratio of TSET observations to number of adjustable parameters was always kept above 2 so as to prevent over training. These models were then built using the same methods as nonlinear models.

All the models that were developed in this study were validated with the same external prediction set.

At this stage the development of quantitative radial basis function neural networks was explored. Radial basis functions are capable of modeling nonlinear data at a faster speed. However development of quantitative RBFNN is a complex task.[49] Several different radial basis functions (eqs 1−5) were evaluated for the quantitative RBFNN development. The cubic functions were not explored, as they are known to give rise to a nonsingular linear system.

$$h(r) = e^{-r2/2\sigma 2} \quad (1)$$

$$h(r) = (r^2 + {}^{\sigma 2})^{-1/2} \quad (2)$$

$$h(r) = r^2 * \ln r \quad (3)$$

$$h(r) = [r^2 + \sigma^2]^{-1/2} \quad (4)$$

$$h(r) = (r^2 + \sigma^2)^{1/2} \quad (5)$$

As multiparameter input by the user becomes tedious, the neural network was developed so that the input needed was minimal. This RBFNN has one hidden layer. The network has two phase training in the sense that centers and scaling parameters are determined first and subsequently the output layer is changed.

GA routines using the radial basis function NN as a fitness evaluator were used to search for the best subsets of descriptors from the 42-descriptor reduced pool. Model sizes from 3 to 7 descriptors were investigated. For each model size the subset of descriptors that produced the lowest rms errors were further investigated. Models of various sizes were compared to find the model providing the lowest rms error with the fewest descriptors.

Table 2. Descriptors Used in Linear Type 1 Model and Nonlinear Type 2 Model for NMDA Inhibitors

descriptor	coeff	error est[a]	range	explanation[b]
FPSA3	0.392	0.126	0.057−0.17	fractional positive SA
WNSA1	−0.476	0.106	54.6−211	weighted negative SA
CHDH1	−0.278	0.091	0.24−1.05	sum of charge on donatable H
CHDH2	−0.318	0.116	0.19−0.29	charge per donatable H
SCAA1	−0.443	0.170	−85.6−6.21	sum SA of acceptor atoms
SCAA2	−0.301	0.0913	12.0−37.2	SA per acceptor atom
MDE44	0.573	0.165	7.01−48.5	dist edge Q−Q 'C′
EMIN1	0.683	0.187	−2.17 − −0.28	min E-state value
V5PC13	−0.435	0.141	0.93−3.91	5th order valance clusters
PND6	0.423	0.166	0.00−35.6	halogens
S5C10	0.555	0.131	0.16−0.69	5th order mole connectivity
S6PC14	−0.838	0.183	4.97−14.0	6th order path cluster

[a] Values representing linear model. [b] FPSA3, fractional charged positive surface area;[39] WNSA1, weighted negative surface area;[39] CHDH1, charge on donatable hydrogen atoms;[40] CHDH2, charge per donatable hydrogen;[40] SCAA1, surface area of hydrogen bond acceptor atoms/number of hydrogen bond acceptor atoms;[40] SCAA2, surface area x charge of hydrogen bond acceptor atoms/number of hydrogen bond acceptor atoms;[40] MDE44, molecular distance edge between all quaternary quaternary carbons;[43] EMIN1, minimum atomic E-state value;[42] V5PC13, 5th order valence cluster; [35]PND6, only the pendent halogen vertices;[49] S5C10, 5th order molecular connectivity;[35] S6PC14, 6th order path cluster.[35]

RESULTS AND DISCUSSION

Many potential linear models were investigated. The multiple linear regression model containing smallest subset of descriptors with most favorable *F* values, multiple correlation coefficient, and rms error was chosen for further investigation. The best Type 1 linear model found consisted of twelve descriptors. Table 2 shows these descriptors with their coefficients and errors. Half of the descriptors in this model encoded information about hydrogen bonding and charge, which is not surprising as the potent inhibitors of glycine/NMDA need to penetrate the blood brain barrier. The balance of hydrophilicity and hydrophobicity would plays a key role in this inhibition process. The polar surface area descriptors such as FPSA and WNSA incorporated the effect of charge and size in the above model. The two CHDH and two SCAA descriptors influenced the contribution of donor nitrogen and acceptor hydrogen atoms. The other half of the descriptors encoded the topology of the inhibitors. The MDE descriptor accounted for connection information between quaternary carbons, while PND accounted for all the halogens. The EMIN describes reactivity of each atom and information regarding intermolecular attractions. The three connectivity indices V5PC, S5C, and S6PC encoded the degree of branching, such as the 5th order valence cluster term, 5th order molecular connectivity term, and 6th order path cluster term, respectively. The WNSA and S6PC descriptors had the maximum influence on the linear model.

A plot of experimental log(IC_{50}) values vs predicted log(IC_{50}) values from the Type 1 linear model is shown in Figure 1. The TSET had rms error of 0.862 and multiple correlation coefficient of 0.732. After diagnostic testing, compound 104 was flagged as an outlier. After removal of that single outlier, the multiple correlation coefficient increased from 0.732 to 0.759. Therefore compound 104 was not removed from the

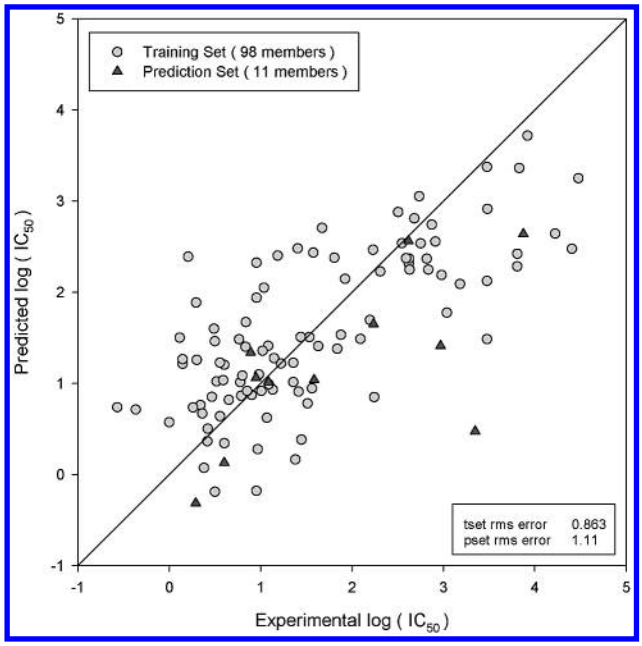


Figure 1. Predicted vs experimental log(IC_{50}) values for Type 1 linear model.

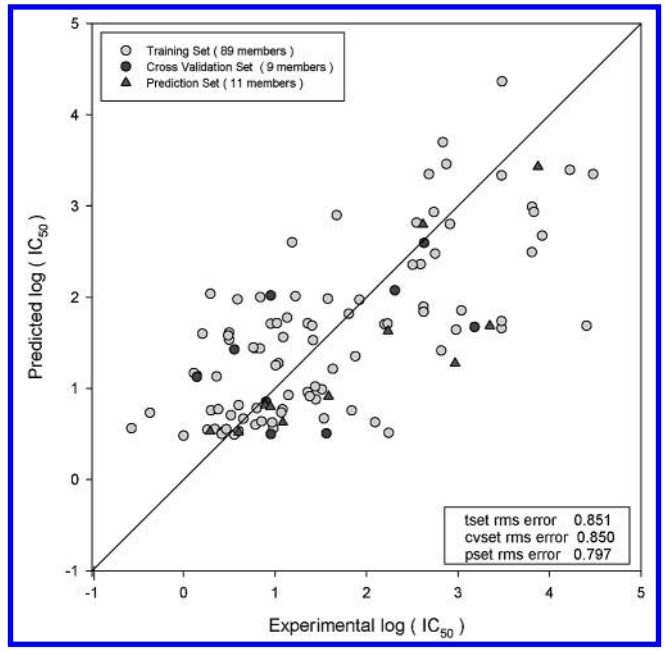


Figure 2. Predicted vs experimental log(IC_{50}) values for Type 2 nonlinear model.

TSET. It was also included in the Figure 1. In validation of this model, the external prediction set had an rms error of 1.11. This linear model shows a substantial amount of scatter. It is evident that the relationship between structure and log(IC_{50}) is not completely linear. This led to the exploration of the nonlinear relationships between the structure and the property.

The descriptors in the best linear model were submitted as inputs to CNN. The best architecture found after varying the number of hidden layer neurons was 12−2−1. This architecture had a ratio of adjustable parameters to number of observations above 2.0. The TSET, CVSET, and PSET rms errors were 0.851, 0.850, and 0.797, respectively. The PSET showed a significant improvement over the linear model PSET error by about 39%. Figure 2 shows the plot of experimental vs predicted log(IC_{50}) values by the Type 2 nonlinear model. This Type 2 model still shows a substantial amount of scatter.

The third type of model was developed by using genetic algorithm (GA) for thefeature selection process and by using computational neural network for model development. This process results in a fully nonlinear Type 3 CNN model. The descriptors chosen for model development are listed in Table 3. The nonlinear feature selection routine chose twelve descriptors, four of which are identical to the ones found in the linear model. The influence of PNSA, two CHDH, and EMIN descriptors on models have already been discussed earlier. The QSUM sums all atomic charges, RPCS encodes the area of the most positive atom and its relative charge. The ELOW and EDIFF calculate the influence of E-state indices. The topological descriptors MOLC, WTPT, and PND bring contributions of degree of branching, number of atoms, and connectivity to the model. Even though eight out of twelve descriptors are different in the two sets, most of the new descriptors chosen are from the same programs and encode features of the same type. CPSA descriptors and hydrogen bonding descriptors remain very significant in this model.

Table 3. Descriptors Used in the Nonlinear Neural Network Type 3 Model

descriptor	range	explanation[a]
QSUM	3.30−9.66	sum abs. values of atomic charges
PNSA1	132−336	diff. in positively charged partial SA
WNSA1	58.2−230	weighted negative SA
RPCS1	0.0056−8.89	SA most pos atom/rel. negative charge
CHDH1	0.087−1.04	sum of charge on donatable H
CHDH2	0.087−0.247	charge per donatable H
MOLC7	0.313−1.25	3rd order path clusters
WTPT2	1.94−2.06	sum of path weights/no. of atoms
EMIN1	−4.54 - −0.28	min E-state value
EDIFF1	12.1−15.5	max. atom. E-state − min. atom. E-state
ELOW1	1.26−11.5	through space dist. bet. max. min. E-state
PND1	19.0−231	all pendant vertices

[a] QSUM, sum of absolute values of atomic charge;[50] PNSA1, difference in negatively charged partial surface area;[39] WNSA1, weighted negative surface area;[39] RPCS1, surface area of the most positively charged atom/relative positive charge;[39] CHDH1, charge on donatable hydrogen atoms;[40] CHDH2, charge per donatable atom;[40] MOLC7, third-order path clusters;[51] WTPT2, sum of path weights/number of atoms;[52] EMIN1, minimum atomic E-state value;[42] EDIFF1, Difference between maximum E-state and minimum E-state;[42] ELOW1, through space distance between maximum and minimum E-staate;[42] PND1, all pendent vertices.[49]

Figure 3 is a plot of experimental vs predicted log(IC_{50}) for the fully nonlinear Type 3 model. The model shows considerable improvement from 0.851 to 0.511 in the TSET rms error, a 40% decrease. A similar improvement can also be seen in the CVSET error. However, the PSET error was only slightly improved from 0.797 to 0.776, which is about 3%. Less generality in prediction may be due to the fact that the small TSET might have been overtraining the network. Hence, in this study the nonlinear model was as good as the CNN model from the point of view of the external prediction set error. Overall, however, the plot for the type 3 model shows substantially less scatter than the previous two models.

One of the problems encountered while developing QSAR models is the possibility of models being found due to

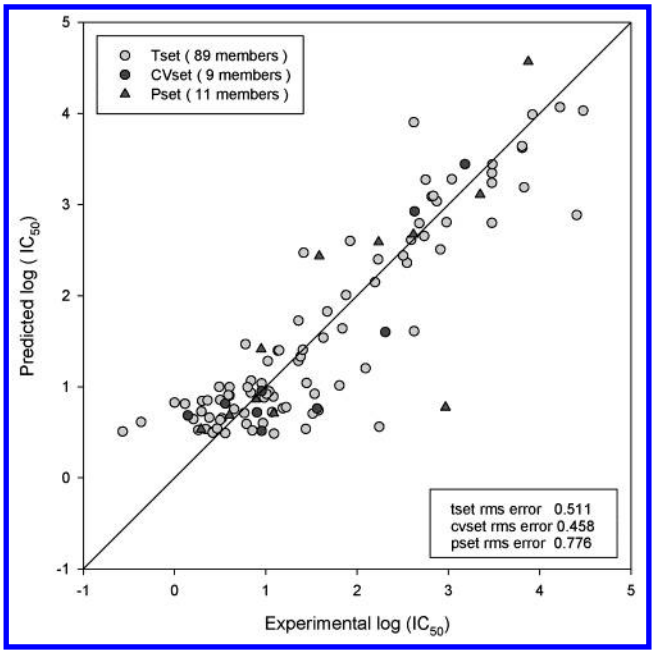


Figure 3. Predicted vs experimental log(IC_{50}) values for Type 3 CNN model.

chance. To ensure that the chance effects did not influence the current study, Monte Carlo experiments were studied. The dependent variables of all the compounds was scrambled, and the genetic algorithm feature selection routine was run again. No suitable models were found. Most of the runs found no models at all. A few of the runs found models exhibited errors of double the magnitude and the plots that looked more or less random.

As the plot of fully nonlinear Type 3 model shows substantially less scatter overall that model was considered as the best model for this study. Once the best model was found, log(IC_{50}) values were predicted for the prediction set on the basis of Type 3 model. The results are reported in Table 1.

At this point the previously set aside exclusion set compounds were examined. These 17 compounds had experimental IC_{50} values reported as inequalities. Their dependent variables were entered as the log values of the next higher integer. All the compounds with the exception of two had the reported values of >100. One compound had the reported value of >10, and the other had a reported value of >1000. Again the Type 3 model was used to predict their log(IC_{50}) values. The results are reported in Table 1.

Predicted log(IC_{50}) values for five of the exclusion set compounds were greater than the ones reported. Overall predicted values for more than half of these compounds are within the margin of error. About two-thirds of the compounds in this study have IC_{50} values below 100, comprising of the more potent inhibitors of the Glycine site of the NMDA receptors. So the training set is more skewed toward the lower values of the more potent inhibitors. This can lead to less representation and therefore less predictive ability. In addition all exclusion set compounds have high ratios of positive charge per hydrogens. So the compounds may have problem crossing blood brain barrier due to increase in the pH. In the case of compound number 80 the ability to lose the protons is beyond the range of values seen in the rest of the compounds in TSET, CVSET, and PSET, which could lead to poorer prediction. In case of compound number 113 the overall positive charge is quite high, and the surface area is quite small which can also lead to poor inhibitor properties and poor prediction. In addition this compound also exhibits very high level of branching so it may be spatially unapproachable, and as there are no examples of similar compounds it is being poorly predicted.

A fourth type of model was developed using quantitative RBFNN. The best model found was a three-descriptor model, with rms errors comparable to Type I model discussed above. Even though the rms errors are not close to nonlinear models (12:2:1 architecture) as the number of descriptors needed to build the model was reduced from 12 to 3, this seems to be a promising start. This quantitative RBFNN should be investigated further.

SUMMARY AND CONCLUSIONS

A series of models have been developed using QSAR methodology. The models clearly demonstrate connection between structure and inhibitory concentrations of glycine/NMDA antagonists. These are the first models that predict IC_{50} values for glycine/ NMDA antagonist activity of several variants of hydroxyquinolones. The structural information of glycine/NMDA antagonists is numerically encoded as molecular descriptors. Objective feature selection and vector space analysis led to development of linear models. These were further refined to develop nonlinear neural network models. A nonlinear feature selection routine, which combined the genetic algorithm with a neural network fitness evaluator was used to develop CNN models. All models were validated using an external prediction set.

This study confirms that IC_{50} values can be predicted on the basis of molecular structure alone, without the inclusion of any experimentally derived data such as partition coefficients. The models that have been derived from linear regression and computational neural network techniques can be applied to prediction of inhibitory activities that are not present in the data set used in this study as long as there are structural similarities. The predictive power of these models is useful in cases where biological assays are necessary to measure the activities of different pharmacophores. This type of prediction could save cost, time, and difficulty in measurements due to variations.

The quantitative RBFNN needs further investigation. The reduction in the number of descriptors needed to encode the structural information is significantly low, leading to the conclusion that this is a good start in the process of complex neural network development.

REFERENCES AND NOTES

(1) *Excitatory Amino Acid Antagonists*; Meldrum, B., Ed.; Blackwell Scientific Publications: Oxford, 1991.

(2) *The NMDA Receptor*; Watkins, J. C., Collingridge, G. C.; Eds; Oxford University Press: London, 1989.

(3) Kemp, J. A.; Leeson, P. D. The Glycine site of the NMDA Receptor − Five Years On. *Trends Pharmacol. Sci.* **1993**, *14*, 20−25.

(4) Lesson, P. D. Glycine-site *N*-Methyl-D-Aspartate Receptor Antagonist. In *Drug Design for Neurosciences*; Kozikowaski, A. P., Ed.; Raven Press: New York, 1993; pp 339−381.

(5) Hansch, C.; Bjorkroth, J. P.; Leo, A. J. Hydrophobicity and Central Nervous System Agents: On the Principle of Minimal Hydrophobicity in Drug Design, *J. Pharm. Sci.* **1987**, *76*, 663−687.

(6) Hansch, C.; Steward, A. R.; Anderson, S. M.; Bently, D. The Parabolic Dependence of Drug Action upon Lipophilic Character as Revealed by a Study of Hypnotics. *J. Med. Chem.* **1968**, *11*, 1−11.
(7) Gupta, S. P. QSAR Studies on Drugs Acting at the Central Nervous System. *Chem. Rev.* (Washington DC) **1989**, *89*, 1765−1800.
(8) Curry, S. H. Binding of Psychotropic Drugs to Plasma Protein and Its Influence on Drug Distribution. *Clin. Pharmacol. Psychiatr*. **1981**, 213−223.
(9) du Souich, P.; Verges, J.; Erill, S. Plasma Protein Binding and Pharmacological Response. *Clin. Pharmacokinet.* **1993**, *24*, 435−440.
(10) Rolan, P. E. Plasma Protein Binding Displacement Interactions-Why Are They Still Regarded as Clinically Important? *Br. J. Clin. Pharmacol.* **1994**, *37*, 125−128.
(11) Leeson, P. D.; Baker, R.; Carling, R. W.; Curtis, N. R.; Moor, K. W.; Williams, B. J.; Foster, A. C.; Donald, A. E.; Kemp, A. J.; Kemp, J. A.; Marshall, J. R. Kynurenic Acid Derivatives. Structure Activity Relationship For Excitatory Amino Acid antagonism and Identification of Potent and Selective Antagonists at the Glycine Site of the *N*-Methyl-D-Aspartate Receptor. *J. Med. Chem.* **1991**, *34*, 1243−1252.
(12) Leeson, P. D.; Carling, R. W.; Moore, K. W.; Mosely, A. M.; Smith, J. D.; Stevenson, G.; Chan, T.; Baker, R.; Foster, A. C.; Grimwood, S.; Kemp, J. A.; Marshall, G. R.; Hoogsteen, K. 4-Amido-2-carboxytetrahydroquinolines Structure−Activity Relationships for Antagonism at the Glycine Site of the NMDA Receptor. *J. Med. Chem.* **1992**, *35*, 1954−1968.
(13) Salituro, F. G.; Harrison, B. L.; Baron, B. M.; Nyce, P. L.; Stewart, K. T.; Kehne, J. H.; White, H. S.; McDonald, I. A. 3-(2-Carboxyindole3-yl) propionic Acid-Based Antagonists of the *N*-Methyl-D-Aspartic Acid Receptor Associated Glycine Binding Site. *J. Med. Chem.* **1992**, *35*, 1791−1799.
(14) Rowley, M.; Leeson, P. D.; Stevenson, G. I.; Moseley, A. M.; Stanfield, I.; Sanderson, I.; Robinson, L.; Baker, R.; Kemp, J. A.; Marshall, G. R.; Foster, C.; Grimwood, S.; Trickleblank, M. D.; Saywell, K. L. 3-Acyl-4 Hydroxyquinolin-2(1H)-ones. Symmetrically Active Anticonvulsants Acting by Antagonism at the Glycine Site of the *N*-Methyl-D-Aspartate Receptor. *J. Med. Chem.* **1993**, *36*, 3386−3396.
(15) Wessel, M. D.; Jurs, P. C.; Tolan, J. W.; Muskal, S. M. Prediction of Human Intestinal absorption of Drug Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 726−735.
(16) Johnson, S. R.; Jurs, P. C. Prediction of Acute Mammalian Toxicity from Molecular Structure for a diverse Set of Substituted Anilines Using Regression Analysis and Computationl Neural Networks. In *Computer-Assisted Lead Finding and Optimization*; van de Waterbeemd, H., Testa B., Folkers, G., Eds.; Verlag Helvetica Chimica Acta: Basel, 1997; pp 29−48.
(17) Mitchell, B. E.; Jurs, P. C. Development of QSAR Models to Predict the Infinite Dilution Activity Coefficients of Organic Compounds in Aqueous Solutions From Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 639−645.
(18) Patankar, S. J.; Jurs, P. C. Prediction of IC50 Values for ACAT Inhibitors from Molecular Structure. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 716−723.
(19) Bakken, G. A.; Jurs, P. C. Prediction of Hydroxyl Radical Rate Constants From Molecular Structure. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 1064−1075.
(20) Johnson, S. R.; Jurs, P. C. Prediction of the Clearing Temperatures of a Series of Liquid Crystals from Molecular Structure. *Chem. Mater*. **1999**, *11*, 1007−1023.
(21) *Chemometric Methods in Molecular Design*; van de Waterbeemed, H., Ed.; VCH: New York, 1995; Vol. 2.
(22) Dressman, J. B.; Amidon, G. L.; Fleisher, D. Absorption Potential: Estimating the fraction Absorbed for Orally Administered Compounds. *J. Pharm. Sci.* **1985**, *74*, 588−589.
(23) Rowley, M.; Kulagowski, J. J.; Watt, A. P.; Rathbone, D.; Stevenson, G. I.; Carling, R. W.; Baker, R.; Marshall, G. R.; Kemp, J. A.; Foster, A. C.; Grimwood, S.; Hargreaves, R.; Hurley, C.; Saywell, K. L.; Tricklebank, M. D.; Leeson, P. D. Effect of Plasma Protein Binding on in Vivo Activity and Brain Penetration of Glycine/NMDA receptor Antagonists. *J. Med. Chem.* **1997**, *40*, 4053−4068.
(24) Kulagowski, J. J.; Baker, R.; Curtis, N. R.; Leeson, P. D.; Mawer, I. M.; Mosely, A. M.; Ridgill, M. P.; Mawer, I. M.; Rowley, M.; Stansfield, I.; Foster, A. C.; Grimwood, S.; Hill, R. G.; Kemp, J. A.; Marshall, G. R.; Saywell, K. L.; Tricklebank, M. D. 3′-(Arylmethyl)- and 3′-(Aryloxy)-3- phenyl-hydroxyquinolin-2(1H)-ones: Orally Active Antagonists of the Glycine Site on the NMDA Receptor. *J. Med. Chem.* **1994**, *37*, 1402−1405.
(25) Grimwood, S.; Moseley, A. M.; Carling, R. W.; Lesson, P. D.; Foster, A. C. Characterisation of the binding of [^{3}H]-L-689, 560, an antagonist for the glycine site of the *N*-methyl-D-aspartate receptor, to rat brain membranes. *Mole. Pharmacol.* **1992**, *41*, 923−930.
(26) Hypercube Inc. Waterloo, ON.
(27) Stewart, J. P. P. MOPAC 6.0; Quantum Chemistry Program Exchange, Indiana University, Bloomsburg, IN, Program 455.
(28) Stewart, J. P. P. MOPAC− A Semiempirical Molecular-Orbital Package. *J. Comput.-Aided Mol. Des*. **1990**, *4*, 1−45.
(29) Stuper, A. J.; Brugger, w. E.; Jurs, P. C. *Computer-Assisted Studies of* chemical *Structure and Biological Function*; Wiley-Interscience: New York, 1979.
(30) Jurs, P. C.; Chou, T. J.; Yuan, M. *In Computer − Assisted Drug Design*; Olsen, E. C., Christoffersen, R. E., Eds.; American Chemical Society: Washington, DC, 1979; pp 103−129.
(31) Pearlman, R. S. *In Physical Chemical Properties of Drugs*; Yallowsky, S. H., Sinkula, A. A., Valvani, S. C., Eds.; Marcel Dekker: New York, 1980.
(32) Miller, K. J.; Savchik, J. A. A New Empirical Method to Calculate Average Molecular Polarizibilities. *J. Am. Chem. Soc.* **1979**, *101*, 7206.
(33) Goldstein, H. *Classical Mechanics*; Addison-Wesley: Reading, MA, 1950; pp 144−156.
(34) Kier, L. B. A Shape Index of Orders One and Three from Molecular Graphs. *Quant. Struct.-Act. Relat. Pharmacol. Chem. Biol*. **1986**, *5*, 1−7.
(35) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*; John Wiley & Sons Inc.: New York, 1986.
(36) Balaban, A. T.; Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett*. **1982**, *89*, 399.
(37) Randic, M. On Molecular Identification Numbers. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 164−175.
(38) Dixon, S. L.; Jurs, P. C. Atomic Charge Calculations for Quantitative structure- Property Relationships. *J. Comput. Chem*. **1992**, *13*, 492.
(39) Stanton, D. T.; Jurs, P. C. Development and Use of Charged Partial Surface Area Structural Descriptors in Computer-Assisted Quantitative Structure Property Relation Studies. *Anal. Chem*. **1990**, *62*, 2323−2329.
(40) Vinogradov, S. N.; Linnell, R. H. *Hydrogen Bonding*; Van Nostrand Reinhold: New York, 1971.
(41) Lowrey, A. H.; Cramer, C. J.; Urban, J. J.; Famini, G. R. Quantum Chemical Descriptors for Linear Solvation Energy Relationships. *Comput. Chem.* **1995**, *19*, 209−215.
(42) Kier, L. B.; Hall, L. H. The E-State as an Extended Free Valence. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 548−552.
(43) Cao, C. Distance-Edge Topological Index-Research on Structure−Property Relationships of Alkanes. *Huaxue Tongbao* **1996**, *22*, 1238−1244.
(44) Sutter, J. M.; Jurs, P. C. Selection of molecular structure descriptors for quantitative structure−activity relationships. In *Adaption of Simulated Annealing to Chemical Problems*; Kalivas, J. H., Ed.; Elsevier Science Publishers B. V.: Amsterdam, 1995.
(45) Sutter, J. M.; Dixon, S. L.; Jurs, P. C. Automated descriptor selection for quantitative structure−activity relationship using generalized simulated annealing. *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 77−84.
(46) Wessel, M. D. Computer-Assisted Development of Quantitative Structure−Property Relationships and Design of Feature Selection Routines. Ph.D. Dissertation, Pennsylvania State University, University Park, PA, 1996.
(47) Xu, L.; Ball, J.; Dixon, S. L.; Jurs, P. C. Quantitative Structure−Activity Relationships for Toxicity of Phenols. Using Regression Analysis and Computational Neural Networks. *Environ. Toxicol. Chem.* **1994**, *13*, 841−851.
(48) Cupid, B. C.; Beddel, C. R.; Lindon, J. C.; Wilson, I. D.; Nicholson, J. K. Quantitative Structure-Metabolism Relationships for Substituted Benzoic Acids in Rabbit: Prediction of Urinary Excretion of Glycine and Glucuronide Conjugates. *Xenobiotica* **1996**, *26*, 157−176.
(49) Bakken, G. A. Ph. D. Thesis, Penn State University, PA.
(50) Madan, A. K.; Gupta, S.; Singh, M. Superpendentic Index: A Novel Highly Discriminating Topological Descriptor for Predicting Biological Activity. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 272−277.
(51) Dixon, S. L. Ph.D. Thesis, The Pennsylvania State University, 1994; Chapter 4.
(52) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.
(53) Randic, M. On Molecular Identification Numbers. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 164−175.

CI010114+