Applications of Neural Networks in Quantitative Structure–Activity Relationships of Dihydrofolate Reductase Inhibitors

T. A. Andrea* and Hooshmand Kalayeh†

E. I. du Pont de Nemours, Stine-Haskell Research Center, Newark, Delaware 19711, and Engineering Physics Laboratory, E. I. du Pont de Nemours, Experimental Station, Wilmington, Delaware 19898. *Received December 20, 1990*

Back propagation neural networks is a new technology useful for modeling nonlinear functions of several variables. This paper explores their applications in the field of quantitative structure–activity relationships. In particular, their ability to fit biological activity surfaces, predict activity, and determine the "functional forms" of its dependence on physical properties is compared to well-established methods in the field. A dataset of 256 5-phenyl-3,4-diamino-6,6-dimethyldihydrotriazines that inhibit dihydrofolate reductase enzyme is used as a basis for comparison. It is found that neural networks lead to enhanced surface fits and predictions relative to standard regression methods. Moreover, they circumvent the need for ad hoc indicator variables, which account for a significant part of the variance in linear regression models. Additionally, they lead to the elucidation of nonlinear and "cross-products" effects that correspond to trade-offs between physical properties in their effect on biological activity. This is the first demonstration of the latter two findings. On the other hand, due to the complexity of the resulting models, an understanding of the local, but not the global, structure–activity relationships is possible. The latter must await further developments. Furthermore, the longer computational time required to train the networks is somewhat inconveniencing, although not restrictive.

Introductions

The field of quantitative structure–activity relationships (QSAR) was introduced in the early 1960s with the pioneering work of Hansch and his co-workers.[1,2] In a sequence of publications, these investigators convincingly demonstrated that biological activity of chemical compounds is a mathematical function of their physicochemical characteristics such as hydrophobicity, size, and electronic properties. Their methods have been widely adopted in the pharmaceutical and agrochemical industries.

The embodiment of these ideas into a concrete model is effected by fitting biological activity to linear or parabolic functions of physicochemical properties (X, Y, ...) of the form

$$A = C_0 + C_1X + C_2X^2 + C_3Y + C_4Y^2 + ... \qquad (1)$$

Multiple linear regression is used to determine the values of C_0, C_1, ..., which minimiize the variance between the data and the model. In these equations, third and higher order terms as well as cross-products terms corresponding to interactions between physicochemical properties are not used in practice. The most commonly used physicochemical properties are linear free energy (LFE) based parameters like Hammett's σ, Taft's Es, and Hansch's π hydrophobicity parameter, derived from in vitro reaction systems.

Functions of several variables represent surfaces or hypersurfaces over the space of independent variables. In practice, the parabolas of eq 1 have negative curvatures, i.e., they are convex upward. For the special case in which activity depends on only two physicochemical properties, eq 1 corresponds to one of three possible surface shapes: plane, parabolically curved plane ("barnroof"), or paraboloid of revolution ("eggshell"), depending on the terms that survive the statistical fitting procedure (Figure 1). These three shapes correspond to equations in which the function is linear in both variables, linear in one variable and parabolic in the other, or parabolic in both variables, respectively.

While eq 1 has an attractively simple form, its flexibility is somewhat limited. The only adjustable degrees of freedom are the statistically calculated coefficients which determine the heights of the surfaces, tilts of the planes, curvatures of the parabolas, and location of their maxima. The lack of third and higher order terms restricts the surface from further undulations. Furthermore, the absence of cross-product terms dictates that dependence of activity on a particular physicochemical property is invariant to the values of other properties. For example, if activity depends parabolically on hydrophobicity, the curvature of the parabola and the location of its maximum is invariant to the values of steric and/or electronic terms in the correlation equation. The consequence of this limited flexibility is the emergence of outliers whose biological activities cannot be adequately accounted for solely on the basis of their physicochemical properties. The occurrence of outliers is commonplace, especially in datasets having more than 50–70 data points. While such "stiff" surfaces may, in theory, be made more "pliable" by including higher order and cross-product terms in eq 1, the staggering diversity of such terms render them unfeasible in practice.

In order to alleviate this weakness, Hansch and his coworkers introduced indicator variables[3,4] and used them as adjuncts to the usual LFE parameters. Typically, such variables flag specific chemical structural features by assigning them a value of 1 for molecules having the feature and 0 otherwise. Geometrically, they represent two parallel surfaces corresponding to the values 0 and 1, which are separated by a "vertical" distance equal to the coefficient of the indicator variable in the regression equation. Usually, several indicator variables are required for a particular modeling exercise. Significantly enough, in many cases they account for a major part of the variance.[3,4] Since indicator variables are specifically designed to deal with outliers in a particular dataset, they are not useable for modeling other datasets. Furthermore, they are developed by several iterative cycles of modeling, identification of outliers and determination of their commonalities, assigning indicator variables, followed by remodeling. The laboriousness of this process increases with the size of the dataset. A modeling method that avoids indicator variables is therefore desirable.

Neural networks is a newly emerging field of information processing technology that has captured the interest of

[†] Engineering Physics Laboratory.

[1] Hansch, C. *Acc. Chem. Res.* **1969**, *2*, 232.
[2] Martin, Y. C. *Quantitative Drug Design, Medicinal Research Series*; Marcel Dekker: New York, 1978; Vol. 8.
[3] Silipo, C.; Hansch, C. *J. Am. Chem. Soc.* **1975**, *97*, 6849.
[4] Kim, K. H.; Hansch, C.; Fukunaga, J. Y.; Steller, E. E.; Jow, P. Y. C.; Craig, P. N.; Page, J. *J. Med. Chem.* **1979**, *22*, 366.

0022-2623/91/1834-2824$02.50/0

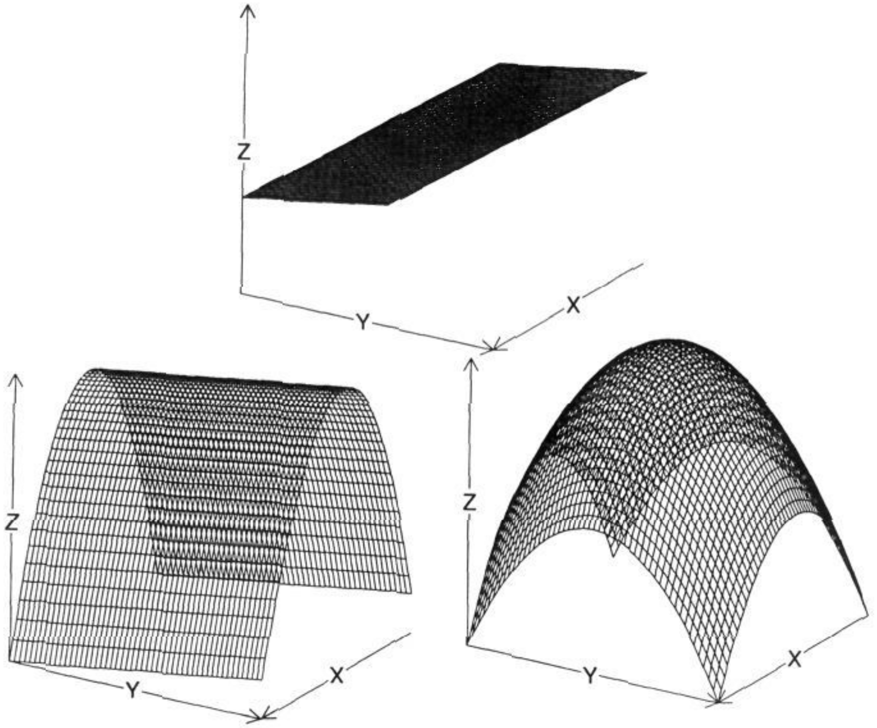


Figure 1. *Biological activity surfaces generated by standard regression using eq 1.*

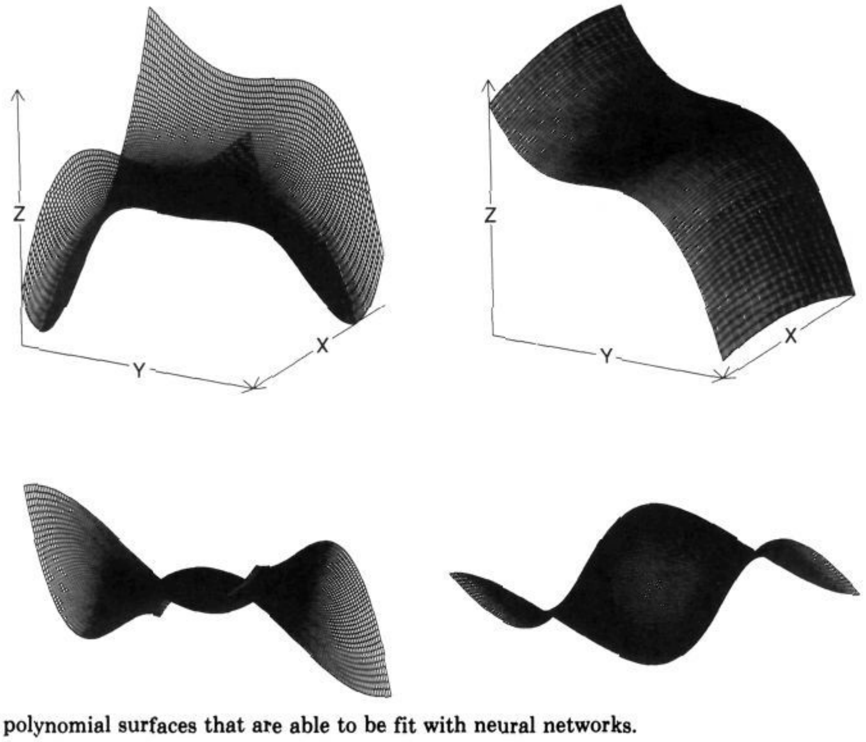


Figure 2. Complex polynomial surfaces that are able to be fit with neural networks.

scientists from diverse fields.[5-9] It evolved from attempts to understand and emulate the brain's information processing capability. The brain consists of multimodule neural networks that extract and recombine relevant information received from their environments and are capable of making decisions that satisfy the needs of the

[5] Hopfield, J. J. *Proc. Nat. Acad. Sci.* **1982**, *79*, 2554.

[6] Hopfield, J. J. *Proc. Nat. Acad. Sci.* **1984**, *81*, 3088.

[7] Kohonen, T. *Self-organization and Associative Memory*, 2nd ed.; Springer-Verlag: Berlin, 1987.

[8] Rumelhart, D. E.; McClelland, J. L. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*; MIT Press: Cambridge, 1986; Vols. I and II.

[9] Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. *Parallel Distributed Processing, Volume 1, Foundations*; Rumelhart, D. E., McClelland, J. L., Eds.; MIT Press: Cambridge, 1986. Widrow, B. *Generalization and Information Storage In Networks of ADELINE Neurons: Self Organizing Systems*; Yovitt, M., Ed.; Spartan Books: New York, 1962.

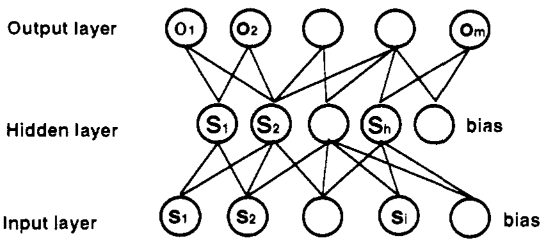


Figure 3. Three layer back propagation neural network topology.

organism. Such capabilities are featured by higher organisms as well as ones with only few neurons. The latter have been demonstrated to have robust and invariant feature extraction capabilities. These biological neural systems can be emulated with artificial neural networks, which can be "taught" complex nonlinear input–output transformations. They represent a unified and general purpose method for solving pattern recognition and functional mapping problems, providing satisfactory solutions in cases where there are no viable alternatives. In our experience, they have proved valuable for modeling complex polynomial surfaces such as those exemplified in Figure 2. Their nonlinear feature extraction capability suggests their potential usefulness in QSAR problems.

While there are several neural network topologies and a variety of training methods,[5-9] this report utilizes a back propagation network (BPN) trained by the algorithm of Owens and Filkin,[10] using a stiff differential equations solver.

The objective of this paper is to compare the performance of neural networks with regression methods with regard to their ability to fit biological activity surfaces, predict activity, and explore the nonlinear aspects of the dependence of activity on properties. A dataset of 256 diaminodihydrotriazines (I) that inhibit dihydrofolate reductase enzyme provides a basis for this comparison. This dataset has been extensively analyzed by QSAR experts using regression methods with and without indicator variables.[3]

Methods

Neural Networks. An artificial neural network (ANN) consists of layers of brainlike neurons with feedforward and feedback interconnections. During the past few years several ANN paradigms, such as Hopfield's[5,6] and the supervised back propagation network[8] (BPN), have been developed. Topologically, the latter consists of an input, hidden, and output layers of neurons or nodes connected by bonds as shown in Figure 3. Each input layer node corresponds to a single independent variable with the exception of the bias node. Similarly, each output layer node corresponds to a different dependent variable.

Associated with each node is an internal state designated by s_i, S_h, and o_m for the input, hidden, output layers, respectively. Each of the input and hidden layer has an additional unit, termed a bias unit, whose internal state is assigned a value of 1. The input layer's s_i values are related to the corresponding independent variables by the scaling equation

$$s_i = 0.8\frac{V_i - V_{i,\min}}{V_{i,\max} - V_{i,\min}} + 0.1 \tag{2}$$

where V_i is the value of the ith independent variable, $V_{i,\min}$ and $V_{i,\max}$ are its minimum and maximum values, respectively. The state S_h of each hidden unit is calculated by the squashing function

$$S_h(\varphi_h) = \frac{1}{1 + e^{-\varphi_h}} \tag{3a}$$

$$\varphi_h = \sum_i w_{hi}s_i + \theta_h \tag{3b}$$

where w_{hi} is the weight of the bond that connects hidden unit h with input unit i and θ_h is the weight of the bond connecting hidden unit h to the input layer bias unit. The state o_m of output unit m is calculated by

$$o_m(\varphi_m) = \frac{1}{1 + e^{-\varphi_m}} \tag{4a}$$

$$\varphi_m = \sum_h W_{mh}S_h + \theta_m \tag{4b}$$

where W_{mh} is the weight of the bond that connects output unit m to hidden unit h and θ_m is the weight of the bond that connects output unit m to the hidden layer bias unit. The network calculated o_m's have values in the range [0, 1].

Training of the neural network of Figure 3 is achieved by minimizing an error function E with respect to the bond weights $\{w_{hi}, W_{mh}\}$

$$E = \sum_p E_p = \tfrac{1}{2}\sum_p \sum_m (a_{pm} - o_{pm})^2 \tag{5}$$

where E_p is the error of the pth training pattern, defined as the set of independent and dependent variables corresponding to the pth data point, or chemical compound; a_{pm} corresponds to the experimentally measured value (A_{pm}) of the mth dependent variable, in this case biological activity, of the pth pattern, scaled by

$$a_{pm} = 0.8\frac{A_{pm} - A_{m,\min}}{A_{m,\max} - A_{m,\min}} + 0.1 \tag{6}$$

$A_{m,\min}$ and $A_{m,\max}$ are the minimum and maximum values of A_{pm} over the dataset.

E depends on the bond weights $\{w_{hi}, W_{mh}\}$ through o_{pm}. It is minimized by following its gradient with respect to the weights, given by

$$-\frac{\partial E}{\partial W_{mh}} = \sum_p -\frac{\partial E_p}{\partial W_{mh}} \tag{7a}$$

$$-\frac{\partial E}{\partial w_{hi}} = \sum_p -\frac{\partial E_p}{\partial w_{hi}} \tag{7b}$$

whereby, using the chain rule,

$$-\frac{\partial E_p}{\partial W_{mh}} = -\frac{\partial E_p}{\partial o_{pm}}\frac{\partial o_{pm}}{\partial W_{mh}} = (a_{pm} - o_{pm})o_{pm}(1 - o_{pm})S_{ph} \tag{8a}$$

$$-\frac{\partial E_p}{\partial w_{hi}} = \sum_m -\frac{\partial E_p}{\partial o_{pm}}\frac{\partial o_{pm}}{\partial w_{hi}} = \sum_m (a_{pm} - o_{pm})o_{pm}(1 - o_{pm})W_{mh}S_{ph}(1 - S_{ph})s_{pi} \tag{8b}$$

In the latter two equations, the p index in S_{ph} and S_{pi} refer to the pth training pattern. The derivative with respect to θ_h and θ_m are similarly calculated.

[10] Owens, A. J.; Filkin, D. L. *Joint IEEE/INNS International Joint Conference of Neural Networks*, Washington, D.C., June 11, 1989, p 381.

[11] Sokal, R. R.; Michener, C. D. *A Statistical Method for Evaluating Systematic Relationships*, University of Kansas Science Bulletin, 38, 1409.

[12] SAS Institute Inc. *SAS User's Guide: Statistics*, Version 5 Edition; SAS Institute Inc.: Cary, NC, 1985; Chapter 5.

[13] Wonnacott, T. H.; Wonnacott, R. J *Introductory Statistics*, 2nd ed.; Wiley: New York, 1972.

[14] Cramer, R. D. III; Patterson, D. E.; Bunce, J. D. *J. Am. Chem. Soc.* 1988, *110*, 5959.

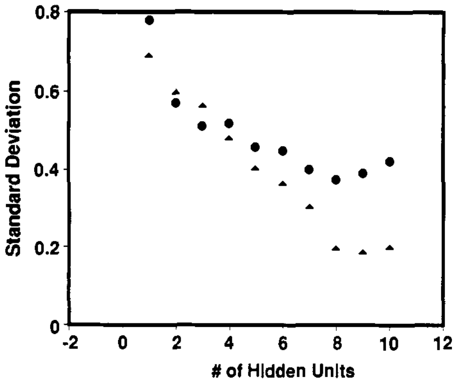


Figure 4. Standard deviation of training (▲) and test (●) sets as a function of the number of hidden units.

The most common procedure for minimizing E utilizes the delta rule,[8] whereby bond weights are iteratively changed from their initially assigned small random values by

$$W_{mh}{}^{n+1} = W_{mh}{}^{n} - \eta\frac{\partial E}{\partial W_{mh}} \tag{9a}$$

$$w_{hi}{}^{n+1} = w_{hi}{}^{n} - \eta\frac{\partial E}{\partial w_{hi}} \tag{9b}$$

The n and $n + 1$ superscripts designate consecutive iterations in the minimization sequence, and η is the learning rate with values typically much less than 1. Similar equations are used for the evolution of θ_h and θ_m.

In the current work, the minimization of the error function E was done by the algorithm of Owens and Filkin[10] in which eqs 9a and 9b are replaced by

$$\frac{\mathrm{d}W_{mh}}{\mathrm{d}t} = -\frac{\partial E}{\partial W_{mh}} \tag{10a}$$

$$\frac{\mathrm{d}w_{hi}}{\mathrm{d}t} = -\frac{\partial E}{\partial w_{hi}} \tag{10a}$$

where t corresponds to sequential training iterations. The set of coupled stiff differential eqs 10 are then solved by the algorithm of Gear.[15]

Determining the Number of Hidden Units. The number of hidden units determines the number of adjustable parameters of the neural network model. Few hidden units may be insufficient to extract all the pertinent features of the data, while too many units causes the network to "memorize" the dataset. The optimal number is that which minimizes the variance of a test set, not used in training the network. This is illustrated in Figure 4 in which the variance of the 100 data point training set and the corresponding 32 data point test set (Figure 5) are graphed as a function of the number of hidden units. This figure reflects a typical result whereby the training-set variance is a decreasing function of the number of hidden units while the test-set variance is an upward concave function with a minimum. Increasing the number of hidden units from 1 to 3 or 4 significantly reduces the test-set variance, which eventually minimizes at 8 hidden units. It should be noted however, that the reduction in the test-set variance between 3 and 8 hidden units is small and may or may not be significant.

Data Set. The dataset used for comparing neural networks (NN), multiple linear regression without indicator variables (MLR), and multiple linear regression with indicator variables (MLRI) consists of physicochemical properties and dihydrofolate reductase (DHFR) inhibitory activities of 256 2,4-diamino-6,6-dimethyl-5-phenyldihydrotriazines (I) that are variously mono- and disubstituted in the ortho, meta, and para positions of the phenyl ring (Table I). Of these, 11 had a non-hydrogen R_2 (Figure 5). The other 245 compounds are in two categories: 132 were tested on DHFR enzymes from Walker 256 leukemia tumors (Table I, $I_1 = 1$) and 113 were tested on DHFR from L1210 leukemia tumors (Table I, $I_1 = 0$).

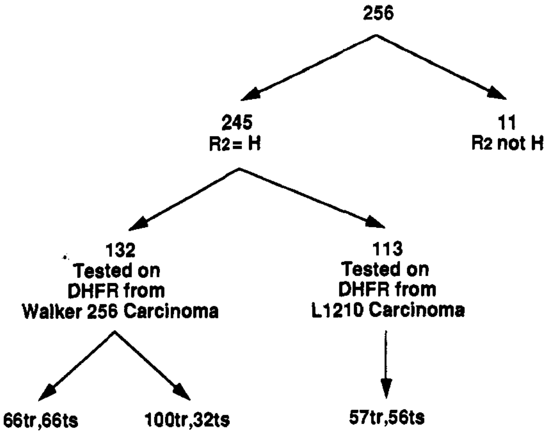
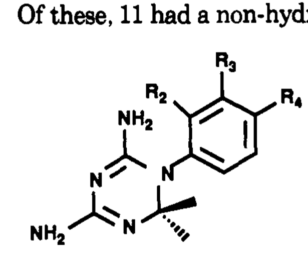


Figure 5. Subgroups of the dataset of Table I.



I

This dataset has been exhaustively analyzed by Hansch and Silipo using MLR.[3] Their best fit equation shows that DHFR inhibitory activity is a function of π_3, MR_4, and a set of six indicators variables, I_1–I_6. The first two variables correspond to the hydrophobicity of R_3 and size of R_4, respectively, I_1 has a value of 1 for compounds tested on DHFR from Walker 256 leukemia tumor and 0 for the enzyme from L1210 leukemia tumors. $I_2 = 1$ for compounds with a non-hydrogen substituent at R_2 and 0 otherwise. $I_3 = 1$ for compounds in which R_3 or R_4 = Ph, CHPh, CONHPh, or C=CHCONHPh. $I_4 = 1$ for analogues with the group $C_6H_4SO_2OC_6H_4X$ and 0 otherwise. I_5 takes a value of 1 for R_3 or R_4 = CH_2Ph, CH_2CH_2Ph, $(CH_2)_4Ph$, $(CH_2)_6Ph$, and $(CH_2)_4O$-Ph between an N-phenyl moiety and a second phenyl ring but is 0 otherwise. I_6 takes a value of 1 for bridges of the type $CH_2NHCONHC_6H_4X$, $CH_2CH_2C(=O)N(R)C_6H_4X$, and $CH_2CH_2CH_2C(=O)N(R)C_6H_4X$ (R = H or Me).

On the other hand, the NN models were exclusively based on π_2, π_3, π_4, MR_2, MR_3, MR_4, and $\Sigma\sigma_{3,4}$, corresponding to the π values of the 2, 3, and 4, substituents, their MR values, and the sum of σ values of the 3 and 4 substituents. Indicator variables were not used in the NN models.

Statistical Analyses. Program FIT MULTIPLE in RS1, BBN Software Products Corporation, MA, was used to perform MLR and MLRI. The best fit equations were selected on the basis of lowest standard deviation and highest correlation coefficients. Cluster analysis was performed using the average linkage method, due to Sokal and Michener,[11] of procedure cluster, in the SAS suite of statistical programs.[12]

[15] Gear, C. W. *Numerical Initial Value Problems in Ordinary Differential Equation*; Prentice Hall: Englewood Cliffs, NJ, 1971.

Table I. Structure, Experimentally Determined DHFR Inhibitory Activity, and Physicochemical Properties of Diaminodihydrotriazines I[a]

	R	log 1/C	π_2	π_3	π_4	MR$_2$	MR$_3$	MR$_4$	$\Sigma\sigma_{3,4}$	I_1	I_2	I_3	I_4	I_5	I_6
1	2,5-Cl$_2$	3.43	0.71	0.00	0.00	0.60	0.10	0.10	0.37	1	1	0	0	0	0
2	2-OCH$_3$	3.68	−0.02	0.00	0.00	0.79	0.10	0.10	0.00	1	1	0	0	0	0
3	2,4-Cl$_2$	3.82	0.71	0.00	0.71	0.60	0.10	0.60	0.23	1	1	0	0	0	0
4	2-CH$_3$	4.00	0.56	0.00	0.00	0.57	0.10	0.10	0.00	1	1	0	0	0	0
5	2-Cl	4.15	0.71	0.00	0.00	0.60	0.10	0.10	0.00	1	1	0	0	0	0
6	2-Br	4.25	0.86	0.00	0.00	0.89	0.10	0.10	0.00	1	1	0	0	0	0
7	2,4,5-Cl$_3$	4.38	0.71	0.00	0.71	0.60	0.10	0.60	0.60	1	1	0	0	0	0
8	2-I	4.62	1.12	0.00	0.00	1.39	0.10	0.10	0.00	1	1	0	0	0	0
9	4-CONHC$_6$H$_4$-4′-SO$_2$F	4.68	0.00	0.00	1.50	0.10	0.10	4.23	0.36	1	0	1	0	0	0
10	4-CONHC$_6$H$_4$-3′-SO$_2$F	4.68	0.00	0.00	1.50	0.10	0.10	4.23	0.36	1	0	1	0	0	0
11	4-C$_6$H$_5$	4.70	0.00	0.00	1.96	0.10	0.10	2.54	−0.01	1	0	1	0	0	0
12	2-F	4.74	0.14	0.00	0.00	0.09	0.10	0.10	0.00	1	1	0	0	0	0
13	3-OCH$_2$CON(CH$_2$CH$_2$)$_2$O	4.85	0.00	−1.39	0.00	0.10	3.32	0.10	0.12	1	0	0	0	0	0
14	4-CN	5.14	0.00	0.00	−0.57	0.10	0.10	0.63	0.66	1	0	0	0	0	0
15	4-CH=CHCONHC$_6$H$_4$-4′-SO$_2$F	5.19	0.00	0.00	1.99	0.10	0.10	5.22	−0.01	1	0	1	0	0	0
16	3-OCH$_2$CONMe$_2$	5.44	0.00	−1.36	0.00	0.10	2.41	0.10	0.12	1	0	0	0	0	0
17	4-CH(Ph)CH$_2$CONHC$_6$H$_4$-4′-SO$_2$F	5.74	0.00	0.00	3.53	0.10	0.10	7.59	−0.09	1	0	1	0	0	0
18	4-Cl-3-(CH$_2$)$_2$C$_6$H$_4$-4′-SO$_2$F	5.82	0.00	2.71	0.71	0.10	4.39	0.60	0.16	0	0	0	0	1	0
19	4-CH=CHCONHC$_6$H$_4$-3′-SO$_2$F	5.89	0.00	0.00	1.99	0.10	0.10	5.22	−0.01	1	0	1	0	0	0
20	3-CONHC$_6$H$_4$-4′-SO$_2$F	5.96	0.00	1.50	0.00	0.10	4.33	0.10	0.35	1	0	1	0	0	0
21	3-NHCOCH$_2$Br-4·O(CH$_2$)$_3$C$_6$H$_5$	6.11	0.00	−0.37	2.66	0.10	2.11	4.15	−0.27	1	0	0	0	0	0
22	3-CH$_2$NHCONEt$_2$	6.11	0.00	−0.29	0.00	0.10	3.56	0.10	−0.07	0	0	0	0	0	0
23	3-OCH$_3$	6.17	0.00	−0.02	0.00	0.10	0.62	0.10	0.12	1	0	0	0	0	0
24	4-OCH$_2$CON(Me)C$_6$H$_5$	6.17	0.00	0.00	0.12	0.10	0.10	4.55	−0.27	1	0	0	0	0	0
25	4-CH$_2$CH(CH$_2$CH$_2$Ph)CONHC$_6$H$_4$-4′-SO$_2$F	6.20	0.00	0.00	4.23	0.10	0.10	8.52	−0.17	1	0	0	0	0	0
26	3-COCH$_2$Cl	6.21	0.00	−0.16	0.00	0.10	1.45	0.10	0.38	1	0	0	0	0	0
28	4-CH$_2$CH(α-C$_{10}$H$_7$)CONHC$_6$H$_4$-4′-SO$_2$F	6.24	0.00	0.00	5.02	0.10	0.10	9.13	−0.17	0	0	0	0	0	0
28	4-OCH$_2$CONMe$_2$	6.26	0.00	0.00	−1.36	0.10	0.10	2.58	−0.27	1	0	0	0	0	0
29	4-CH$_2$CH-(Ph-2″-OCH$_3$)CONHC$_6$H$_4$-4′-SO$_2$F	6.33	0.00	0.00	3.51	0.10	0.10	8.27	−0.17	0	0	0	0	0	0
30	3-Cl-4-OCH$_2$C$_6$H$_{10}$CH$_2$OC$_6$H$_4$-4′-SO$_2$F	6.37	0.00	0.71	5.16	0.10	0.49	7.25	0.10	0	0	0	0	0	0
31	3-CH(CH$_2$NHCOCH$_2$Br)(CH$_2$)$_3$C$_6$H$_5$	6.37	0.00	2.94	0.00	0.10	6.94	0.10	−0.07	1	0	0	0	0	0
32	3-CH$_2$NHCON(CH$_2$CH$_2$)$_2$O	6.43	0.00	−1.32	0.00	0.10	3.53	0.10	−0.07	0	0	0	0	0	0
33	4-COCH$_2$Cl	6.45	0.00	0.00	−0.16	0.10	0.10	1.62	0.50	1	0	0	0	0	0
34	4-CH$_2$CH(Ph-3″-OCH$_3$)CONHC$_6$H$_4$-4′-SO$_2$F	6.46	0.00	0.00	3.51	0.10	0.10	8.27	−0.17	0	0	0	0	0	0
35	4-CH(CH$_2$NHCOCH$_2$Br)(CH$_2$)$_3$C$_6$H$_5$	6.52	0.00	0.00	2.94	0.10	0.10	7.03	−0.17	1	0	0	0	0	0
36	2,3-Cl$_2$	6.52	0.71	0.71	0.00	0.60	0.49	0.10	0.37	1	1	0	0	0	0
37	2-Cl-4-(CH$_2$)$_4$C$_6$H$_5$	6.54	0.71	0.00	3.66	0.60	0.10	4.39	−0.17	1	1	0	0	1	0
38	3-Cl-4-O(CH$_2$)$_4$OC$_6$H$_4$-4′-SO$_3$C$_6$H$_4$-4″-Cl	6.55	0.00	0.71	4.92	0.10	0.49	8.90	0.10	0	0	0	0	0	0
39	3-CH$_2$NHCOCH$_2$Br	6.58	0.00	−0.52	0.00	0.10	2.57	0.10	−0.07	1	0	0	0	0	0
40	3-CONHC$_6$H$_4$-3′-SO$_2$F	6.60	0.00	1.50	0.00	0.10	4.33	0.10	0.35	1	0	1	0	0	0
41	4-CH$_2$CONMe$_2$	6.63	0.00	0.00	−1.70	0.10	0.10	2.37	−0.17	1	0	0	0	0	0
42	4-OCH$_2$CON(CH$_2$)$_4$	6.66	0.00	0.00	−0.72	0.10	0.10	3.31	−0.27	1	0	0	0	0	0
43	3·OCH$_2$CON(Me)C$_6$H$_5$	6.68	0.00	0.12	0.00	0.10	4.46	0.10	0.12	1	0	0	0	0	0
44	4-OCH$_2$CONEt$_2$	6.72	0.00	0.00	−0.36	0.10	0.10	3.51	−0.27	1	0	0	0	0	0
45	3-CH$_2$CH(CH$_2$NHCOCH$_2$Br)C$_6$H$_5$	6.72	0.00	1.94	0.00	0.10	6.01	0.10	−0.07	1	0	0	0	0	0
46	4-Cl-3-O(CH$_2$)$_5$OC$_6$H$_4$-4′-SO$_2$F	6.72	0.00	4.43	0.71	0.10	6.09	0.60	0.10	0	0	0	0	0	0
47	4-CH$_2$CONEt$_2$	6.77	0.00	0.00	−0.70	0.10	0.10	3.29	0.10	1	0	0	0	0	0
48	4-Cl-3-(CH$_2$)$_4$C$_6$H$_4$-4′·SO$_2$F	6.77	0.00	4.01	0.71	0.10	5.32	0.60	0.16	0	0	0	0	1	0
49	3-Cl-4-OCH$_2$C$_6$H$_4$-4′-CH$_2$OC$_6$H$_4$-4″-SO$_2$F	6.82	0.00	0.71	4.33	0.10	0.49	7.10	0.10	0	0	0	0	0	0
50	3-OCH$_2$CONHC$_6$H$_5$	6.85	0.00	0.60	0.00	0.10	4.00	0.10	0.12	1	0	0	0	0	0
51	3-C$_6$H$_5$	6.85	0.00	1.96	0.00	0.10	2.51	0.10	0.06	1	0	1	0	0	0
52	4-CH$_2$CH(Ph)CONHC$_6$H$_4$-3′-SO$_2$F	6.89	0.00	0.00	3.53	0.10	0.10	7.56	−0.17	0	0	0	0	0	0
53	3-Cl-4-OCH$_2$C$_6$H$_4$-3′-CONHC$_6$H$_4$-4″-SO$_2$F	6.92	0.00	0.71	3.16	0.10	0.49	7.34	0.10	0	0	0	0	0	0
54	3-Cl-4-OCH$_2$C$_6$H$_4$-4′-CONHC$_6$H$_4$-4″-SO$_2$F	6.92	0.00	0.71	3.16	0.10	0.49	7.34	0.10	0	0	0	0	0	0
55	3-OCH$_2$CONHC$_6$H$_4$-4′-SO$_2$F	6.92	0.00	1.61	0.00	0.10	4.95	0.10	0.12	0	0	0	0	0	0
56	4-CH$_2$CN	6.92	0.00	0.00	−0.57	0.10	0.10	1.01	0.01	1	0	0	0	0	0
57	H	6.92	0.00	0.00	0.00	0.10	0.10	0.10	0.00	1	0	0	0	0	0
58	3·OCH$_2$C$_6$H$_4$-3′-NHCOCH$_2$Br	6.92	0.00	1.29	0.00	0.10	5.24	0.10	0.12	1	0	0	0	0	0
59	4-CH$_2$CON(Me)C$_6$H$_5$	7.00	0.00	0.00	−0.19	0.10	0.10	4.34	−0.17	1	0	0	0	0	0
60	4-(CH$_2$)$_2$CONMe$_2$	7.05	0.00	0.00	−1.20	0.10	0.10	2.83	−0.17	1	0	0	0	0	0
61	3-Cl-4-(CH$_2$)$_4$C$_6$H$_3$-5′-Cl-2′-SO$_2$F	7.06	0.00	0.71	4.72	0.10	0.49	5.66	0.20	0	0	0	0	1	0
62	3-Cl-4-O(CH$_2$)$_3$OC$_6$H$_4$-4′-SO$_2$F	7.07	0.00	0.71	4.21	0.10	0.49	5.13	0.10	0	0	0	0	0	0
63	3-NO$_2$	7.07	0.00	−0.28	0.00	0.10	0.72	0.10	0.71	1	0	0	0	0	0
64	3-(CH$_2$)$_2$COCH$_2$Cl	7.10	0.00	0.20	0.00	0.10	2.38	0.10	−0.07	1	0	0	0	0	0
65	3-(CH$_2$)$_4$COCH$_2$Cl	7.10	0.00	1.20	0.00	0.10	3.31	0.10	−0.07	1	0	0	0	0	0
66	4-OCH$_2$CON(CH$_2$)$_5$	7.12	0.00	0.00	−0.32	0.10	0.10	3.78	−0.27	1	0	0	0	0	0
67	4-CH$_2$CON(CH$_2$CH$_2$)$_2$O	7.12	0.00	0.00	−1.70	0.10	0.10	3.27	−0.17	1	0	0	0	0	0
68	4-(CH$_2$)$_6$C$_6$H$_4$-4′-SO$_2$F	7.12	0.00	0.00	5.01	0.10	0.10	6.09	−0.17	0	0	0	0	1	0
69	3-Cl-4-OCH(CH$_3$)CONHC$_6$H$_4$-4′-SO$_2$F	7.13	0.00	0.71	1.91	0.10	0.49	5.37	0.10	0	0	0	0	0	0
70	4-CH$_2$CH(Ph)CONHC$_6$H$_4$-4′-SO$_2$F	7.13	0.00	0.00	3.53	0.10	0.10	7.59	−0.17	1	0	0	0	0	0
71	3-Cl-4-O(CH$_2$)$_2$O(CH$_2$)$_2$OC$_6$H$_4$-4′-SO$_2$F	7.14	0.00	0.71	3.38	0.10	0.49	5.80	0.10	0	0	0	0	0	0
72	3-Cl-4·O(CH$_2$)$_3$CONHC$_6$H$_4$-4′-SO$_2$F	7.15	0.00	0.71	2.38	0.10	0.49	5.84	0.10	0	0	0	0	0	0
73	3-Cl-4-OCH$_2$CONMe$_2$	7.16	0.00	0.71	−1.36	0.10	0.49	2.58	0.10	1	0	0	0	0	0
74	3·Cl-4-O(CH$_2$)$_3$CONHC$_6$H$_4$-3′-SO$_2$F	7.17	0.00	0.71	2.38	0.10	0.49	5.84	0.10	0	0	0	0	0	0
75	4-Cl-3-O(CH$_2$)$_4$OC$_6$H$_4$-4′-SO$_2$F	7.17	0.00	3.92	0.71	0.10	5.63	0.60	0.16	0	0	0	0	0	0

Table I (Continued)

	R	log 1/*C*	π_2	π_3	π_4	MR_2	MR_3	MR_4	$\Sigma\sigma_{3,4}$	I_1	I_2	I_3	I_4	I_5	I_6
76	4-CH_2CH(Ph-3″-Me)CONHC_6H_4-4′-SO_2F	7.17	0.00	0.00	4.09	0.10	0.10	8.05	−0.17	1	0	0	0	0	0
77	3-$(CH_2)_2$CONHC_6H_4-4′-SO_2F	7.19	0.00	1.77	0.00	0.10	5.26	0.10	−0.07	1	0	0	0	0	1
78	4-CH_2CH(Ph-4″-Me)CONHC_6H_4-4′-SO_2F	7.24	0.00	0.00	4.09	0.10	0.10	8.05	−0.17	1	0	0	0	0	0
79	4-CH_2CH(Ph-2″-CH_3)CONHC_6H_4-4′-SO_2F	7.24	0.00	0.00	4.09	0.10	0.10	8.05	−0.17	1	0	0	0	0	0
80	3-Cl-4-O$CH_2C_6H_4$-3′-CONHC_6H_4-3″-SO_2F	7.24	0.00	0.71	3.16	0.10	0.49	7.34	0.10	0	0	0	0	0	0
81	3-Cl-4-O$CH_2C_6H_4$-2′-CONHC_6H_4-4″-SO_2F	7.24	0.00	0.71	3.16	0.10	0.49	7.34	0.10	0	0	0	0	0	0
82	3-Cl-4-O$(CH_2)_4$CONHC_6H_4-4′-SO_2F	7.24	0.00	0.71	2.88	0.10	0.49	6.30	0.10	0	0	0	0	0	0
83	3-Cl-4-O$CH_2C_6H_3$-5′-Cl-2′-SO_2F	7.27	0.00	0.71	2.42	0.10	0.49	4.48	0.10	0	0	0	0	0	0
84	4-Cl-3-O$(CH_2)_2$OC_6H_4-4′-SO_2F	7.27	0.00	3.00	0.71	0.10	4.70	0.60	0.16	0	0	0	0	0	0
85	3-SO_2F	7.27	0.00	0.05	0.00	0.10	0.95	0.10	0.80	1	0	0	0	0	0
86	3-Cl-4-O$(CH_2)_3$NHCONHC_6H_4-3′-SO_2F	7.28	0.00	0.71	2.72	0.10	0.49	6.18	0.10	0	0	0	0	0	0
87	4-$(CH_2)_2$CONEt_2	7.28	0.00	0.00	−0.21	0.10	0.10	3.76	−0.17	1	0	0	0	0	0
88	3-Cl-4-OCH_2CON$(CH_2)_4$	7.29	0.00	0.71	−0.72	0.10	0.49	3.31	0.10	1	0	0	0	0	0
89	4-OCH_2CON$(CH_2CH_2)_2$O	7.29	0.00	0.00	−1.39	0.10	0.10	3.49	−0.27	1	0	0	0	0	0
90	4-CH$(CH_3)CH_2$CONHC_6H_4-4′-SO_2F	7.29	0.00	0.00	2.07	0.10	0.10	5.62	−0.17	1	0	0	0	0	0
91	4-CH_2CON(Me)$CH_2C_6H_5$	7.30	0.00	0.00	0.43	0.10	0.10	4.80	−0.17	1	0	0	0	0	0
92	4-$(CH_2)_2$CON(Me)$CH_2C_6H_5$	7.31	0.00	0.00	0.93	0.10	0.10	5.27	−0.17	1	0	0	0	0	0
93	4-$(CH_2)_2$CON$(CH_2CH_2)_2$O	7.32	0.00	0.00	−1.20	0.10	0.10	3.74	−0.17	1	0	0	0	0	0
94	4-O$(CH_2)_3$NHCONHC_6H_4-3′-SO_2F	7.32	0.00	0.00	2.72	0.10	0.10	6.18	−0.27	0	0	0	0	0	0
95	3-Cl-4-O$(CH_2)_3$NHCOC_6H_4-4′-SO_2F	7.34	0.00	0.71	1.42	0.10	0.49	5.84	0.10	0	0	0	0	0	0
96	3-CH_2CONHC_6H_4-4′-SO_2F	7.34	0.00	1.31	0.00	0.10	4.79	0.10	−0.07	1	0	0	0	0	0
97	4-CH_2NHCONHC_6H_4-4′-SO_2F	7.35	0.00	0.00	1.84	0.10	0.10	5.08	−0.17	0	0	0	0	0	1
98	4-$(CH_2)_2$CON$(C_3H_7)_2$	7.35	0.00	0.00	0.80	0.10	0.10	4.69	−0.17	1	0	0	0	0	0
99	3-Cl-4-O$CH_2C_6H_3$-6′-Cl-3′-SO_2F	7.38	0.00	0.71	2.42	0.10	0.49	4.48	0.10	0	0	0	0	0	0
100	3-Cl-4-O$CH_2C_6H_3$-2′-CH_3-4′-SO_2F	7.38	0.00	0.71	2.27	0.10	0.49	4.44	0.10	0	0	0	0	0	0
101	3-Cl-4-S$(CH_2)_2$CONHC_6H_4-4′-SO_2F	7.39	0.00	0.71	2.74	0.10	0.49	5.97	0.37	0	0	0	0	0	0
102	4-$(CH_2)_2C_6H_4$-4′-SO_2F	7.41	0.00	0.00	2.71	0.10	0.10	4.23	−0.17	0	0	0	0	1	0
103	3-Cl-4-O$CH_2C_6H_4$-4′-CONHC_6H_4-3″-SO_2F	7.41	0.00	0.71	3.16	0.10	0.49	7.34	0.10	0	0	0	0	0	0
104	4-$(CH_2)_2$NH$SO_2C_6H_4$-4′-SO_2F	7.41	0.00	0.00	1.01	0.10	0.10	5.48	−0.17	1	0	0	0	0	0
105	3-Cl-4-SCH_2CONHC_6H_4-4′-SO_2F	7.42	0.00	0.71	2.24	0.10	0.49	5.51	0.37	0	0	0	0	0	0
106	3-Cl-4-O$CH_2C_6H_3$-3′-Cl-2′-SO_2F	7.42	0.00	0.71	2.42	0.10	0.49	4.48	0.10	0	0	0	0	0	0
107	3-Cl-4-OCH_2CONHC_6H_4-4′-SO_2F	7.43	0.00	0.71	1.61	0.10	0.49	4.91	0.10	0	0	0	0	0	0
108	3-Cl-4-O$CH_2C_8H_4$-2′-SO_2F	7.43	0.00	0.71	1.71	0.10	0.49	3.98	0.10	0	0	0	0	0	0
109	3-Cl-4-O$CH_2C_6H_3$-3′-Cl-4′-SO_2F	7.43	0.00	0.71	2.42	0.10	0.49	4.48	0.10	0	0	0	0	0	0
110	3-Cl-4-O$CH_2C_6H_3$-2′-Cl-4′-SO_2F	7.44	0.00	0.71	2.42	0.10	0.49	4.48	0.10	0	0	0	0	0	0
111	3-Cl-4-O$(CH_2)_2$OC_6H_4-4′-SO_2F	7.44	0.00	0.71	3.00	0.10	0.49	4.66	0.10	0	0	0	0	0	0
112	3-$(CH_2)_4C_6H_3$-2′,4′-Cl_2	7.45	0.00	5.08	0.00	0.10	5.35	0.10	−0.07	1	0	0	0	1	0
113	3-Cl-4-O$(CH_2)_6$OC_6H_4-4′-SO_2F	7.46	0.00	0.71	5.00	0.10	0.49	6.52	0.10	0	0	0	0	0	0
114	4-$(CH_2)_2$CONHC_6H_3-3′-OMe-4′-SO_2F	7.46	0.00	0.00	1.75	0.10	0.10	5.84	−0.17	1	0	0	0	0	1
115	3-Cl-4-OCH_2CON$(CH_3)C_6H_4$-4′-SO_2F	7.47	0.00	0.71	1.13	0.10	0.49	5.37	0.10	0	0	0	0	0	0
116	3-Cl-4-OCH_2CON$(CH_2)_5$	7.47	0.00	0.71	−0.32	0.10	0.49	3.78	0.10	1	0	0	0	0	0
117	3-Cl-4-O$CH_2C_6H_4$-4′-SO_2NMe_2	7.48	0.00	0.71	0.88	0.10	0.49	5.27	0.10	1	0	0	0	0	0
118	3-Cl-4-O$CH_2C_6H_3$-2′-Cl-3′-SO_2F	7.49	0.00	0.71	2.42	0.10	0.49	4.48	0.10	0	0	0	0	0	0
119	3-O$(CH_2)_4$OC_6H_4-4′-SO_2F	7.49	0.00	4.00	0.00	0.10	5.62	0.10	0.12	0	0	0	0	0	0
120	4-Cl-3-O$(CH_2)_3$OC_6H_4-4′-SO_2F	7.51	0.00	3.50	0.71	0.10	5.16	0.60	0.16	0	0	0	0	0	0
121	3-Cl-4-O$CH_2C_6H_4$-3′-CN	7.51	0.00	0.71	1.09	0.10	0.49	3.75	0.10	1	0	0	0	0	0
122	3-Cl-4-O$CH_2C_6H_5$	7.52	0.00	0.71	1.66	0.10	0.49	3.22	0.10	1	0	0	0	0	0
123	4-SCH_2CONHC_6H_4-4′-SO_2F	7.52	0.00	0.00	2.24	0.10	0.10	5.51	0.00	0	0	0	0	0	0
124	3-Cl-4-O$CH_2C_6H_3$-4′-Cl-2′-SO_2F	7.52	0.00	0.71	2.42	0.10	0.49	4.48	0.10	0	0	0	0	0	0
125	3-CH_2NHCONHC_6H_5	7.52	0.00	0.83	0.00	0.10	4.21	0.10	−0.07		0	0	0	0	1
126	4-CH_2CH(Me)CONHC_6H_4-4′-SO_2F	7.55	0.00	0.00	2.07	0.10	0.10	5.62	−0.17	1	0	0	0	0	0
127	3-O$(CH_2)_3$OC_6H_4-4′-NHCOCH_2Br	7.55	0.00	1.77	0.00	0.10	6.32	0.10	0.12	1	0	0	0	0	0
128	4-$(CH_2)_2$CON(Me)C_6H_5	7.56	0.00	0.00	0.31	0.10	0.10	4.80	−0.17	1	0	0	0	0	1
129	3-Cl-4-O$(CH_2)_4$OC_6H_4-4′-SO_2F	7.57	0.00	0.71	4.00	0.10	0.49	5.59	0.10	0	0	0	0	0	0
130	3-Cl-4-O$(CH_2)_5$OC_6H_4-4′-SO_2F	7.57	0.00	0.71	4.50	0.10	0.49	6.05	0.10	0	0	0	0	0	0
131	3-Cl-4-O$CH_2C_6H_4$-4′-SO_2F	7.58	0.00	0.71	1.71	0.10	0.49	3.98	0.10	0	0	0	0	0	0
132	4-$(CH_2)_2$CONHC_6H_4-4′-SO_2F	7.60	0.00	0.00	1.77	0.10	0.10	5.16	−0.17	0	0	0	0	0	1
133	3-Cl-4-$(CH_2)_2$CONHC_6H_4-4′-SO_2F	7.62	0.00	0.71	1.77	0.10	0.49	5.16	0.20	0	0	0	0	0	1
134	3-CH_2NHCONHC_6H_4-3′-SO_2F	7.62	0.00	1.84	0.00	0.10	5.16	0.10	−0.07	1	0	0	0	0	1
135	4-$(CH_2)_2$NH$SO_2C_6H_4$-3′-SO_2F	7.64	0.00	0.00	1.01	0.10	0.10	5.48	−0.17	1	0	0	0	0	0
136	3-Cl-4-OCH_2CONEt_2	7.64	0.00	0.71	−0.36	0.10	0.49	3.51	0.10	1	0	0	0	0	0
137	3-O$(CH_2)_3$OC_6H_4-3′-NHCOCH_2Br	7.64	0.00	1.77	0.00	0.10	6.32	0.10	0.12	1	0	0	0	0	0
138	3-O$(CH_2)_2$OC_6H_4-2′-NHCOCH_2Br	7.66	0.00	1.27	0.00	0.10	5.85	0.10	0.12	1	0	0	0	0	0
139	3-O$(CH_2)_2$OC_6H_4-3′-NHCOCH_2Br	7.66	0.00	1.27	0.00	0.10	5.85	0.10	0.12	1	0	0	0	0	0
140	3-Cl-4-SCH_2CONHC_6H_4-3′-SO_2F	7.66	0.00	0.71	2.24	0.10	0.49	5.51	0.37	0	0	0	0	0	0
141	3-Cl-4-O$(CH_2)_4$NHCOC_6H_4-4′-SO_2F	7.66	0.00	0.71	1.92	0.10	0.49	6.21	0.10	0	0	0	0	0	0
142	4-$(CH_2)_3$CONHC_6H_4-4′-SO_2F	7.66	0.00	0.00	2.27	0.10	0.10	5.62	−0.17	1	0	0	0	0	1
143	3-Cl-4-O$(CH_2)_3$NHCONHC_6H34-4′-SO_2F	7.68	0.00	0.71	2.27	0.10	0.49	6.18	0.10	0	0	0	0	0	0
144	3-Cl-4-O$(CH_2)_4$NHCONHC_6H_4-3′-SO_2F	7.70	0.00	0.71	3.22	0.10	0.49	6.65	0.10	0	0	0	0	0	0
145	3-$(CH_2)_4C_6H_3$-3′-Cl-4′-SO_2F	7.70	0.00	4.42	0.00	0.10	5.81	0.10	−0.07	0	0	0	0	1	0
146	3-Cl-4-$(CH_2)_4C_6H_3$-3′-Cl-4′-SO_2F	7.70	0.00	0.71	4.42	0.10	0.49	5.66	0.20	0	0	0	0	1	0
147	4-$(CH_2)_4C_6H_4$-4′-SO_2F	7.70	0.00	0.00	3.71	0.10	0.10	5.16	−0.17	0	0	0	0	1	0
148	4-CH_2CONHC_6H_4-4′-SO_2F	7.70	0.00	0.00	1.31	0.10	0.10	4.69	−0.17	1	0	0	0	0	0
149	4-O$(CH_2)_2$OC_6H_4-4′-NHCOCH_2Br	7.70	0.00	0.00	1.27	0.10	0.10	6.09	−0.27	1	0	0	0	0	0
150	3-Cl-4-O$CH_2C_6H_4$-3′-CONMe_2	7.72	0.00	0.71	0.15	0.10	0.49	5.02	0.10	1	0	0	0	0	0
151	3-Cl-4-O$CH_2C_6H_4$-4′-$SO_3C_6H_4$-3″-Cl	7.72	0.00	0.71	3.92	0.10	0.49	7.29	0.10	1	0	0	1	0	0

Table I (Continued)

	R	log 1/*C*	π_2	π_3	π_4	MR$_2$	MR$_3$	MR$_4$	$\Sigma\sigma_{3,4}$	I_1	I_2	I_3	I_4	I_5	I_6
152	4-OCH$_2$CONHC$_6$H$_4$-4′-SO$_2$F	7.72	0.00	0.00	1.61	0.10	0.10	4.91	0.27	1	0	0	0	0	0
153	3-Cl-4-OCH$_2$CONHC$_6$H$_4$-3′-SO$_2$F	7.72	0.00	0.71	1.61	0.10	0.49	4.91	0.10	0	0	0	0	0	0
154	3-Cl-4·OCH$_2$C$_6$H$_4$-3′-SO$_2$F	7.72	0.00	0.71	1.71	0.10	0.49	3.98	0.10	0	0	0	0	0	0
155	3-Cl-4-OCH$_2$C$_6$H$_3$-6′-Cl-2′-SO$_2$F	7.72	0.00	0.71	2.42	0.10	0.49	4.48	0.10	0	0	0	0	0	0
156	4-CH$_2$NHCONHC$_6$H$_3$-3′-Me-4′-SO$_2$F	7.72	0.00	0.00	2.40	0.10	0.10	5.54	−0.17	0	0	0	0	0	1
157	4-(CH$_2$)$_2$CONHC$_6$H$_4$-3′-SO$_2$F	7.74	0.00	0.00	1.77	0.10	0.10	5.16	−0.17	0	0	0	0	0	1
158	3,5-Cl$_2$-4-OCH$_2$CONHC$_6$H$_4$-4′-SO$_2$F	7.74	0.00	0.71	1.62	0.10	0.49	4.91	0.47	0	0	0	0	0	0
159	3-Cl	7.76	0.00	0.71	0.00	0.10	0.49	0.10	0.37	1	0	0	0	0	0
160	3-CF$_3$	7.76	0.00	0.88	0.00	0.10	0.51	0.10	0.43	1	0	0	0	0	0
161	3-Cl-4-OCH$_2$C$_6$H$_4$-4′-SO$_3$C$_6$H$_4$-4″-Cl	7.77	0.00	0.71	3.92	0.10	0.49	7.29	0.10	1	0	0	1	0	0
162	3-CH$_2$NHCONH$_6$H$_4$-3′-CON(Me)$_2$	7.77	0.00	−0.68	0.00	0.10	6.01	0.10	−0.07	0	0	0	0	0	1
163	3-Cl-4-(CH$_2$)$_4$C$_6$H$_4$-2′-SO$_2$F	7.77	0.00	0.71	3.71	0.10	0.49	5.16	0.20	0	0	0	0	1	0
164	3-Cl-4-(CH$_2$)$_4$C$_6$H$_3$-2′-Cl-4′-SO$_2$F	7.77	0.00	0.71	4.42	0.10	0.49	5.66	0.20	0	0	0	0	1	0
165	3-Cl-4-CH$_2$NHCONHC$_6$H$_3$-3′-Me-4′-SO$_2$F	7.80	0.00	0.71	2.40	0.10	0.49	5.54	0.20	0	0	0	0	0	1
166	4-O(CH$_2$)$_2$OC$_6$H$_4$-4′-SO$_2$F	7.80	0.00	0.00	3.00	0.10	0.10	4.67	−0.27	1	0	0	0	0	0
167	4-(CH$_2$)$_3$CONHC$_6$H$_4$-2′-SO$_2$F	7.80	0.00	0.00	2.27	0.10	0.10	5.62	−0.17	1	0	0	0	0	1
168	3-Cl-4-O(CH$_2$)$_2$NHCONHC$_6$H$_3$-3′-Me-4′-SO$_2$F	7.82	0.00	0.71	2.78	0.10	0.49	6.18	0.10	0	0	0	0	0	0
169	3-O(CH$_2$)$_2$OC$_6$H$_4$-4′-SO$_2$F	7.82	0.00	3.00	0.00	0.10	4.70	0.10	0.12	0	0	0	0	0	0
170	3-Cl-4-(CH$_2$)$_4$C$_6$H$_3$-4′-Cl-2′-SO$_2$F	7.82	0.00	0.71	4.42	0.10	0.49	5.66	0.20	0	0	0	0	1	0
171	3-Cl-4-(CH$_2$)$_2$C$_6$H$_4$-4′-SO$_2$F	7.85	0.00	0.71	2.71	0.10	0.49	4.23	0.20	0	0	0	0	1	0
172	3-Cl-4-(CH$_2$)$_2$C$_6$H$_3$-5′-Cl-2′-SO$_2$F	7.85	0.00	0.71	3.42	0.10	0.49	4.73	0.20	0	0	0	0	1	0
173	3-Cl-4-(CH$_2$)$_2$C$_6$H$_3$-3′-Cl-4′-SO$_2$F	7.85	0.00	0.71	3.42	0.10	0.49	4.73	0.20	0	0	0	0	1	0
174	3-Cl-4-OCH$_2$CON(CH$_2$CH$_2$)$_2$O	7.85	0.00	0.71	−1.39	0.10	0.49	3.49	0.10	1	0	0	0	0	0
175	3-Cl-4-OCH$_2$C$_6$H$_4$-3′-CON(CH$_2$CH$_2$)$_2$O	7.85	0.00	0.71	0.13	0.10	0.49	5.93	0.10	1	0	0	0	0	0
176	3-Cl-4-OCH$_2$C$_6$H$_4$-3′-CON(CH$_2$)$_4$	7.85	0.00	0.71	0.80	0.10	0.49	5.75	0.10	1	0	0	0	0	0
177	3-Cl-4-OCH$_2$CON(Me)C$_6$H$_5$	7.89	0.00	0.71	0.12	0.10	0.49	4.55	0.10	1	0	0	0	0	0
178	4-OCH$_2$CONHC$_6$H$_5$	7.89	0.00	0.00	0.60	0.10	0.10	4.09	−0.27	1	0	0	0	0	0
179	4-(CH$_2$)$_2$C^{6}H$_5$	7.89	0.00	0.00	2.66	0.10	0.10	3.47	−0.17	0	0	0	0	1	0
180	4-(CH$_2$)$_2$CONHC$_6$H$_3$-3′-Me-4′-SO$_2$F	7.89	0.00	0.00	2.33	0.10	0.10	5.62	−0.17	1	0	0	0	0	1
181	3-Cl-4-CH$_2$NHCONHC$_6$H$_4$-4′-SO$_2$F	7.92	0.00	0.71	1.84	0.10	0.49	5.08	0.20	0	0	0	0	0	1
182	3-Cl-4-O(CH$_2$)$_2$NHCONHC$_6$H$_4$-4′·SO$_2$F	7.92	0.00	0.71	2.22	0.10	0.49	5.77	0.10	1	0	0	0	0	0
183	4-(CH$_2$)$_3$CONHC$_6$H$_4$-3′-SO$_2$F	7.92	0.00	0.00	2.27	0.10	0.10	5.62	−0.17	1	0	0	0	0	1
184	4-(CH$_2$)$_2$COCH$_2$Cl	7.92	0.00	0.00	0.20	0.10	0.10	2.47	−0.17	1	0	0	0	0	0
185	3-OC$_6$H$_4$-4′-NHCOCH$_2$Br	7.92	0.00	1.71	0.00	0.10	4.77	0.10	0.25	1	0	0	0	0	0
186	3-Cl-4-(CH$_2$)$_4$C$_6$H$_5$	7.92	0.00	0.71	3.66	0.10	0.49	4.39	0.20	0	0	0	0	1	0
187	4-(CH$_2$)$_4$C$_6$H$_3$-2′,4′-Cl$_2$	7.92	0.00	0.00	5.08	0.10	0.10	5.39	−0.17	0	0	0	0	1	0
188	3-Cl·4-(CH$_2$)$_4$C$_6$H$_5$	7.96	0.00	0.71	4.13	0.10	0.49	4.39	−0.17	1	0	0	0	1	0
189	3-O(CH$_2$)$_3$OC$_6$H$_4$-4′-SO$_2$F	7.96	0.00	3.50	0.00	0.10	5.16	0.10	0.12	0	0	0	0	0	0
190	3-(CH$_2$)$_4$C$_6$H$_3$-5′-Cl-2′-SO$_2$F	7.96	0.00	4.42	0.00	0.10	5.81	0.10	−0.07	0	0	0	0	1	0
191	4-(CH$_2$)$_4$C$_6$H$_3$-2′·Cl-4′-SO$_2$F	7.96	0.00	0.00	4.42	0.10	0.10	5.66	−0.17	0	0	0	0	1	0
192	3-Cl-4-OCH$_2$C$_6$H$_3$-4′-Cl-3′-SO$_2$F	8.00	0.00	0.71	2.42	0.10	0.49	4.48	0.10	0	0	0	0	0	0
193	3-(CH$_2$)$_4$C$_6$H$_3$-2′-Cl-4′-SO$_2$F	8.00	0.00	4.42	0.00	0.10	5.81	0.10	−0.07	0	0	0	0	1	0
194	4-OCH$_2$CONHC$_6$H$_4$-3′-SO$_2$F	8.00	0.00	0.00	1.61	0.10	0.10	4.91	−0.27	1	0	0	0	0	0
195	3-Cl-4-OCH$_2$C$_6$H$_4$-3′-CONHC$_6$H$_5$	8.00	0.00	0.71	2.15	0.10	0.49	6.53	0.10	1	0	0	0	0	0
196	3·CH$_2$C$_6$H$_5$	8.00	0.00	2.01	0.00	0.10	2.97	0.10	−0.08	1	0	0	0	1	0
197	4-(CH$_2$)$_4$C$_6$H$_5$	8.00	0.00	0.00	3.66	0.10	0.10	4.39	−0.17	0	0	0	0	1	0
198	3-Cl-4-OCH$_2$C$_6$H$_4$-3′-CON(CH$_2$)$_5$	8.02	0.00	0.71	1.20	0.10	0.49	6.21	0.10	1	0	0	0	0	0
199	3-CH$_2$NHCONHC$_6$H$_4$-3′-OCH$_3$	8.02	0.00	0.81	0.00	0.10	4.83	0.10	−0.07	0	0	0	0	0	1
200	4-(CH$_2$)$_2$CONHC$_6$H$_3$-4′-Me-3′-SO$_2$F	8.02	0.00	0.00	2.33	0.10	0.10	5.62	−0.17	1	0	0	0	0	1
201	3-Cl-4-(CH$_2$)$_4$C$_6$H$_4$-3′-SO$_2$F	8.03	0.00	0.71	3.71	0.10	0.49	5.16	0.20	0	0	0	0	1	0
202	3-(CH$_2$)$_4$C$_6$H$_3$-2′,4′-Cl$_2$	8.03	0.00	5.08	0.00	0.10	5.35	0.10	−0.07	0	0	0	0	1	0
203	4-CH$_2$NHCONHC$_6$H$_4$-3′-SO$_2$F	8.04	0.00	0.00	1.84	0.10	0.10	5.08	−0.17	1	0	0	0	0	1
204	4-(CH$_2$)$_2$CON(Me)-C$_6$H$_4$-4′-SO$_2$F	8.04	0.00	0.00	1.28	0.10	0.10	5.62	−0.17	1	0	0	0	0	1
205	3-Cl-4-(CH$_2$)$_2$C$_6$H$_3$-4′-Cl-2′-SO$_2$F	8.05	0.00	0.71	3.42	0.10	0.49	4.73	0.20	0	0	0	0	1	0
206	4-CH$_2$C$_6$H$_5$	8.05	0.00	0.00	2.01	0.10	0.10	3.00	−0.09	1	0	0	0	1	0
207	3-CH$_2$NHCONHC$_6$H$_4$-3′-Cl	8.05	0.00	1.54	0.00	0.10	4.70	0.10	−0.07	0	0	0	0	0	1
208	3·Cl-4-O(CH$_2$)$_3$NHCONHC$_6$H$_3$-3′-Me-3′-SO$_2$F	8.06	0.00	0.71	3.28	0.10	0.49	6.64	0.10	0	0	0	0	0	0
209	4-CH$_2$CONHC$_6$H$_4$-3′-SO$_2$F	8.06	0.00	0.00	1.31	0.10	0.10	4.69	−0.17	1	0	0	0	0	0
210	4-(CH$_2$)$_2$CONHC$_6$H$_3$-6′-OMe-3′·SO$_2$F	8.08	0.00	0.00	1.75	0.10	0.10	5.84	−0.17	1	0	0	0	0	1
211	3-Cl-4-OCH$_2$C$_6$H$_4$-4′-SO$_3$C$_6$H$_4$-3″-CF$_3$	8.09	0.00	0.71	4.09	0.10	0.49	7.19	0.10	1	0	0	1	0	0
212	3-CH$_2$NHCONHC$_6$H$_4$-3′-NO$_2$	8.10	0.00	0.55	0.00	0.10	4.94	0.10	−0.07	0	0	0	0	0	1
213	3-(CH$_2$)$_4$C$_6$H$_4$-4′-SO$_2$F	8.10	0.00	3.71	0.00	0.10	5.32	0.10	−0.07	0	0	0	0	1	0
214	3-(CH$_2$)$_4$C$_6$H$_4$-3′-SO$_2$F	8.10	0.00	3.71	0.00	0.10	5.32	0.10	−0.07	0	0	0	0	1	0
215	3-(CH$_2$)$_2$C$_6$H$_4$-4′-SO$_2$F	8.10	0.00	2.71	0.00	0.10	4.39	0.10	−0.07	0	0	0	0	1	0
216	4-(CH$_2$)$_2$NHCOC$_6$H$_4$-4′-SO$_2$F	8.11	0.00	0.00	1.11	0.10	0.10	5.16	−0.17	1	0	0	0	0	1
217	3-Cl-4-(CH$_2$)$_4$C$_6$H$_3$-4′-Cl-3′-SO$_2$F	8.11	0.00	0.71	4.42	0.10	0.49	5.66	0.20	0	0	0	0	1	0
218	3-Cl-4-OCH$_2$C$_6$H$_4$-3′-CON(Me)C$_6$H$_5$	8.12	0.00	0.71	2.15	0.10	0.49	6.99	0.10	1	0	0	0	0	0
219	3-O(CH$_2$)$_2$OC$_6$H$_4$-4′-NHCOCH$_2$Br	8.13	0.00	1.27	0.00	0.10	5.85	0.10	0.12	1	0	0	0	0	0
220	3-Cl-4-OCH$_2$C$_6$H$_4$-3′-CONEt$_2$	8.14	0.00	0.71	1.15	0.10	0.49	5.95	0.10	1	0	0	0	0	0
221	3-Cl-4·(CH$_2$)$_4$C$_6$H$_4$-4′-SO$_2$F	8.14	0.00	0.71	3.71	0.10	0.49	5.16	0.20	0	0	0	0	1	0
222	3-Br·4-OCH$_2$CONHC$_6$H$_4$-4′-SO$_2$F	8.14	0.00	0.86	1.61	0.10	0.78	4.91	0.12	0	0	0	0	0	0
223	4-(CH$_2$)$_4$OC$_6$H$_4$-4′-SO$_2$F	8.14	0.00	0.00	4.62	0.10	0.10	5.37	−0.17	1	0	0	0	1	0
224	3-(CH$_2$)$_2$C$_6$H$_5$	8.19	0.00	2.66	0.00	0.10	5.64	0.10	−0.07	0	0	0	0	1	0
225	3-CH$_2$NHCONHC$_6$H$_4$-3′-CN	8.19	0.00	0.26	0.00	0.10	4.69	0.10	−0.07	0	0	0	0	0	1
226	3-Cl-4-OCH$_2$C$_6$H$_4$-4′-SO$_2$OC$_6$H$_5$	8.20	0.00	0.71	3.21	0.10	0.49	6.79	0.10	1	0	0	1	0	0
227	3-Cl-4·(CH$_2$)$_4$C$_6$H$_3$-3′-Cl-2′-SO$_2$F	8.20	0.00	0.71	4.42	0.10	0.49	5.66	0.20	0	0	0	0	1	0

Table I (Continued)

	R	log $1/C$	π_2	π_3	π_4	MR_2	MR_3	MR_4	$\Sigma\sigma_{3,4}$	I_1	I_2	I_3	I_4	I_5	I_6
228	4-$(CH_2)_2CONHC_6H_3$-2′-Me-4′-SO_2F	8.24	0.00	0.00	2.33	0.10	0.10	5.62	–0.17	1	0	0	0	0	1
229	4-$(CH_2)_2CONHC_6H_3$-4′-OMe-3′-SO_2F	8.24	0.00	0.00	1.75	0.10	0.10	5.84	–0.17	1	0	0	0	0	1
230	3-Cl-4-$OCH_2C_6H_4$-4′-$SO_3C_6H_4$-3″-CN	8.24	0.00	0.71	2.64	0.10	0.49	7.32	0.10	1	0	0	1	0	0
231	4-$(CH_2)_4OC_6H_5$	8.24	0.00	0.00	3.61	0.10	0.10	4.61	–0.17	0	0	0	0	1	0
232	3-Cl-4-$OCH_2C_6H_4$-4′-$SO_3C_6H_3$-3″,4″-Cl_2	8.25	0.00	0.71	4.63	0.10	0.49	7.79	0.10	1	0	0	1	0	0
233	3-$(CH_2)_2C_6H_4$-4′-$NHCOCH_2Br$	8.26	0.00	2.29	0.00	0.10	5.55	0.10	–0.07	1	0	0	0	1	0
234	3-Cl-4-$(CH_2)_2C_6H_3$-4′-Cl-3′-SO_2F	8.27	0.00	0.71	3.42	0.10	0.49	4.73	0.20	0	0	0	0	1	0
235	3-Cl-4-$(CH_2)_2C_6H_3$-3′-Cl-2′-SO_2F	8.30	0.00	0.71	3.42	0.10	0.49	4.73	0.20	0	0	0	0	1	0
236	3-Cl-4-$(CH_2)_2C_6H_3$-2′-Cl-4′-SO_2F	8.33	0.00	0.71	3.42	0.10	0.49	4.73	0.20	0	0	0	0	1	0
237	3-Cl-4-$OCH_2C_6H_4$-4′-$SO_3C_6H_4$-2″-CF_3	8.33	0.00	0.71	4.09	0.10	0.49	7.19	0.10	1	0	0	1	0	0
238	3-$(CH_2)_4OC_6H_5$	8.35	0.00	3.61	0.00	0.10	4.52	0.10	–0.07	0	0	0	0	1	0
239	3-$(CH_2)_4C_6H_5$	8.35	0.00	3.66	0.00	0.10	4.37	0.10	–0.07	0	0	0	0	1	0
240	3-$(CH_2)_4C_6H_3$-4′-Cl-3′-SO_2F	8.37	0.00	4.42	0.00	0.10	5.81	0.10	–0.07	0	0	0	0	1	0
241	3-$(CH_2)_4C_6H_4$-4′-$NHCOCH_2Br$	8.38	0.00	3.24	0.00	0.10	6.47	0.10	–0.07	1	0	0	0	1	0
242	3-Cl-4-$OCH_2C_6H_4$-4′-$SO_3C_6H_4$-4″-CN	8.39	0.00	0.71	2.64	0.10	0.49	7.32	0.10	1	0	0	1	0	0
243	3-Cl-4-$OCH_2C_6H_4$-4′-$SO_3C_6H_4$-4″-OCH_3	8.40	0.00	0.71	3.19	0.10	0.49	7.47	0.10	1	0	0	1	0	0
244	3-Cl-4-$OCH_2C_6H_4$-4′-$SO_3C_6H_4$-4″-F	8.40	0.00	0.71	3.35	0.10	0.49	6.78	0.10	1	0	0	1	0	0
245	3-Cl-4-$OCH_2C_6H_4$-4′-$SO_3C_6H_4$-2″-OCH_3	8.40	0.00	0.71	3.19	0.10	0.49	7.47	0.10	1	0	0	1	0	0
246	3-$(CH_2)_4C_6H_4$-3′-$NHCOCH_2Br$	8.41	0.00	3.24	0.00	0.10	6.47	0.10	–0.07	1	0	0	0	1	0
247	3-Cl-4-$OCH_2C_6H_4$-4′-$SO_3C_6H_4$-3″-CH_3	8.44	0.00	0.71	3.77	0.10	0.49	7.25	0.10	1	0	0	1	0	0
248	3-Cl-4-$OCH_2C_6H_4$-4′-$SO_3C_6H_4$-3″-F	8.46	0.00	0.71	3.35	0.10	0.49	6.78	0.10	1	0	0	1	0	0
249	3-Cl-4-$OCH_2C_6H_4$-4′-$SO_3C_6H_4$-3″-OCH_3	8.52	0.00	0.71	3.19	0.10	0.49	7.47	0.10	1	0	0	1	0	0
250	3,4-Cl_2	8.54	0.00	0.71	0.71	0.10	0.49	0.60	0.60	1	0	0	0	0	0
251	3-Cl-4-$OCH_2C_6H_4$-4′-$SO_3C_6H_4$-2″-Cl	8.62	0.00	0.71	3.92	0.10	0.49	7.29	0.10	1	0	0	1	0	0
252	3-Cl-4-$OCH_2C_6H_4$-4′-$SO_3C_6H_4$-4″-$CON(CH_3)_2$	8.62	0.00	1.71	1.70	0.10	0.49	8.59	0.10	1	0	0	1	0	0
253	3-Cl-4-$OCH_2C_6H_4$-4′-$SO_3C_6H_4$-4″-$CON(CH_3)_2$	8.63	0.00	0.71	1.70	0.10	0.49	8.59	0.10	1	0	0	1	0	0
254	3-Cl-4-$OCH_2C_6H_4$-4′-$SO_3C_6H_4$-2″-CN	8.70	0.00	0.71	2.64	0.10	0.49	7.32	0.10	1	0	0	1	0	0
255	3-Cl-4-$OCH_2C_6H_4$-4′-$SO_3C_6H_4$-2″-F	8.74	0.00	0.71	3.35	0.10	0.49	6.78	0.10	1	0	0	1	0	0
256	3-Cl-4-$OCH_2C_6H_4$-4′-$SO_3C_6H_4$-3″-$CON(CH_3)_2$	8.76	0.00	0.71	1.70	0.10	0.49	8.59	0.10	1	0	0	1	0	0

[a] Reference 3.

Comparison of Neural Networks and Regression Models. The best NN, MLR, and MLRI models were determined for each dataset in Figure 5. These were compared for fitting biological activity surfaces and predicting activity on the basis of criteria described below.

A table with 256 compound rows and 21 columns corresponding to enzyme inhibitory activity, π_2, π_3, π_4, MR_2, MR_3, MR_4, $\Sigma\sigma_{3,4}$, their squares, and I_1–I_6 was constructed. Subsets of rows corresponding to different training and test sets (Figure 5) were selected for model development. MLR models were developed by determining linear combinations of physicochemical properties (π_2, π_3, π_4, MR_2, MR_3, MR_4, $\Sigma\sigma_{3,4}$, and their squares, but not indicator variables) that minimize the variance and maximize the correlation. MLRI models were similarly determined by including indicator variables in the analysis in addition to variables used in MLR. NN models were calculated with π_2, π_3, π_4, MR_2, MR_3, MR_4, and $\Sigma\sigma_{3,4}$ as independent variables. Their squares were not utilized as independent variables on the basis that, if these and other nonlinear variables contributed to reducing the variance, the neural network will automatically include their effects. Indicator variables were not utilized in NN models on the basis that they correspond to such nonlinear effects. For datasets in which R_2 = H, π_2 and MR_2 are constant, and hence were not used as independent variables.

The criteria used for comparing surface fits are the correlation coefficient, R, coefficient of determination, R^2, defined in the usually way,[13] and the standard deviation of the error, σ_E, defined by

$$\sigma_E = \left\{\frac{1}{N}\sum_{i}^{N} E_i^2\right\}^{1/2}$$

$$E_i = \text{observed}\left(\log\frac{1}{C_i}\right) - \text{calculated}\left(\log\frac{1}{C_i}\right) \qquad (11)$$

where C_i is the concentration of compound i required for 50% DHFR inhibition and N is the number of points. Additionally, the number of outliers is also compared. An outlier is defined, somewhat arbitrarily, as a data point for which the absolute value of the error $\|E_i\|$ is greater than 0.8. This value is expected to be greater than experimental error for in vitro enzyme inhibition assays.

NN, MLR, and MLRI predictions were compared using two strategies. The first relied on cluster analysis[11,12] to split a parent dataset into a predetermined number of clusters from which training and test sets were chosen. The training set was obtained by randomly selecting a single point from each cluster. The remaining points constituted the test set. As shown in Figure 5, the 132 compounds tested on Walker's enzyme were split into two training/test combinations: 66/66 and 100/32. Similarly, the 113 compounds tested on L1210 enzyme were split into 57/56. The identities of the data points in the three training/test set combinations are shown in Table II. This procedure ascertains that the training set is well distributed in the subspace of independent variables and that every point in the test set has a points in training set in its vicinity. NN, MLR, and MLRI models were developed for the training subset and used to predict activities of the test subset. R, R^2, and σ_E and the number of outliers are compared in the Results.

The second strategy utilized the cross-validation procedure of Cramer et al.[14] The 132 point subset (Figure 5) was analyzed 132 times. Each time a different single data point was used as a test set for a training set consisting of the remaining 131 data points. NN, MLR, and MLRI models were developed for each training set and used to predicted the activity of the single point test set. The cross-validated r^2 was calculated by[14]

$$\text{cross validated } r^2 = \frac{\text{SD} - \text{press}}{\text{SD}} \qquad (12)$$

where SD is the variance of observed biological activity relative to its mean and press is the average squared errors of the 132 test sets. The corresponding values for NN,

Table II. Training and Test Sets Used for Comparing the Performance of MLR, MLRI, and NN

data sets[a]	data points[b]
66tr[c]	9, 11, 13, 14, 15, 16, 17, 20, 21, 23, 26, 28, 31, 33, 35, 39, 41, 43, 44, 50, 51, 56, 57, 59, 60, 63, 64, 65, 76, 85, 91, 96, 112, 116, 121, 122, 127, 134, 135, 139, 149, 152, 160, 166, 174, 177, 178, 180, 184, 185, 188, 194, 196, 198, 206, 218, 223, 232, 233, 242, 243, 246, 247, 250, 252, 256
66ts[c]	10, 19, 24, 25, 40, 42, 45, 47, 58, 66, 67, 70, 73, 77, 78, 79, 87, 88, 89, 90, 92, 93, 98, 104, 114, 117, 126, 128, 136, 137, 138, 142, 148, 150, 151, 159, 161, 167, 175, 176, 182, 183, 195, 200, 203, 204, 209, 210, 211, 216, 219, 220, 226, 228, 229, 230, 237, 241, 244, 245, 248, 249, 251, 253, 254, 255
100tr[c]	9, 11, 13, 14, 15, 16, 17, 20, 21, 23, 24, 25, 26, 28, 31, 33, 35, 39, 41, 42, 43, 44, 45, 47, 50, 51, 56, 57, 58, 59, 60, 63, 64, 65, 66, 67, 70, 73, 76, 77, 78, 85, 87, 88, 89, 91, 92, 93, 96, 98, 112, 114, 116, 117, 121, 122, 127, 134, 135, 139, 148, 149, 150, 152, 159, 160, 161, 166, 174, 175, 176, 177, 178, 180, 182, 184, 185, 188, 194, 195, 196, 198, 203, 204, 206, 209, 210, 218, 223, 226, 232, 233, 237, 242, 243, 246, 247, 250, 252, 256
32ts[c]	10, 19, 40, 79, 90, 104, 126, 128, 136, 137, 138, 142, 151, 167, 183, 200, 211, 216, 219, 220, 228, 229, 230, 241, 244, 245, 248, 249, 251, 253, 254, 255
57tr[d]	22, 27, 30, 32, 34, 38, 46, 48, 49, 55, 62, 68, 69, 71, 74, 80, 84, 86, 94, 101, 102, 105, 111, 113, 115, 119, 120, 123, 125, 130, 131, 132, 140, 141, 156, 158, 162, 165, 169, 170, 171, 179, 181, 189, 191, 193, 197, 199, 202, 205, 207, 213, 215, 222, 224, 225, 239
56ts[d]	18, 29, 52, 53, 54, 61, 72, 75, 81, 82, 83, 95, 97, 99, 100, 103, 105, 106, 107, 108, 110, 118, 124, 129, 133, 143, 144, 145, 146, 147, 153, 154, 155, 157, 163, 164, 168, 172, 173, 186, 187, 190, 192, 201, 208, 212, 214, 217, 221, 227, 231, 234, 235, 236, 238, 240

[a] The training and test sets in this column are those of Figure 5. [b] The numbers in this column correspond to those in column 1 of Table I. [c] Compounds with $I_1 = 1$ and $I_2 = 0$. [d] Compounds with $I_1 = 0$ and $I_2 = 0$.

MLR, and MLRI models are compared in the Results.

Computation Time. The algorithms of this section were coded and the models computed on a VAX 8800. Computation times varied depending on the number of input and hidden units as well as the size of the dataset. For example, training times for the 132 point dataset (Figure 5) with 5 input units were 40, 75, 140, and 390 s for the 2, 3, 4, and 6 hidden units models, respectively.

Results

Fitting Biological Activity Surface. In their analysis of the dataset of 256 DHFR inhibitors (I) using regression, Hansch and Silipo[3] noted that six indicator variables were needed in addition to the physicochemical properties of the R_2, R_3, and R_4 (I) to fit the dataset of Table I. Even so, it was only possible to fit 244 data points, leaving 12 outliers, which were excluded from the regression. Their analysis gave

$$A = 6.489 + 0.680\pi_3 - 0.118{\pi_3}^2 + 0.230\mathrm{MR}_4 - 0.0243{\mathrm{MR}_4}^2 + 0.238I_1 - 2.530I_2 - 1.991I_3 + 0.877I_4 + 0.686I_5 + 0.704I_6$$

$$N = 244,\ S = 0.377,\ R = 0.923 \qquad (13)$$

in which activity depends parabolically on hydrophobicity of R_3 and size of R_4. The coefficient of 0.238 of I_1 reflects that the inhibitory activity surfaces corresponding to the two enzyme systems are parallel and separated by a "vertical distance" of 0.238. This separation, if significant, could be due to a difference in the enzyme structure or to systematic differences in the assay procedure of the two enzymes. Furthermore, the contribution of R_2 to activity was found to be unrelated to its size, hydrophobicity, and electronic properties. That group lowers activity by a factor of approximately 340, as reflected by the coefficient of I_2 in the above equation. I_3–I_6 reflect a variety of chemical structural features related to size, flexibility, and reactivity of R_3 and R_4.

In order to avoid the possible complications of dealing with two separate enzyme surfaces, the datasets with 132 and 113 data points corresponding to DHFR from two different sources (Figure 5) were compared separately. The topology of the corresponding neural network with 4 hidden units is shown in Figure 6. As pointed out in the Methods indicator variables were not used in the NN models. The results of the NN, MLR, and MLRI models shown in Table III indicate that, for both datasets, the neural network models have higher R and R^2, lower σ_E, and fewer outliers.

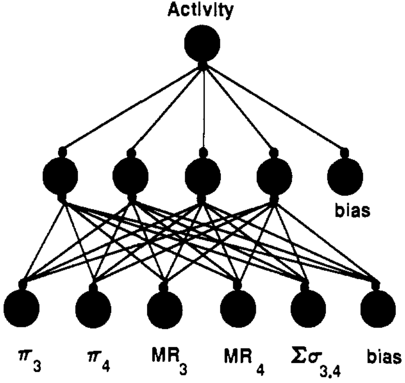


Figure 6. Topology of the neural network used in training different subsets of Table I.

Table III. Comparison of Neural Networks and Multiple Linear Regression with and without Indicator Variables for Fitting Biological Activity Surfaces

method[a]	no. of data points	data set[b]	R	R^2	σ_E	no. of outliers[c]
NN	256	A	0.922	0.850	0.374	12
MLR	256	A	0.703	0.494	0.686	61
MLRI	256	A	0.879	0.773	0.460	20
NN	245	B	0.891	0.794	0.339	10
MLR	245	B	0.494	0.244	0.651	41
MLRI	245	B	0.809	0.656	0.439	15
NN	132	C	0.903	0.815	0.385	4
MLR	132	C	0.622	0.387	0.701	23
MLRI	132	C	0.877	0.769	0.431	11
NN	113	D	0.892	0.796	0.236	1
MLR	113	D	0.517	0.268	0.447	5
MLRI	113	D	0.673	0.452	0.387	5

[a] NN = neural networks, MLR = multiple linear regression, without indicator variables; MLRI = multiple linear regression with indicator variables. [b] A all data points in Table I: B subset of Table I with $I_2 = 0$, designating $R_2(I) = H$; C subset of Table I with $I_2 = 0$ and $I_1 = 1$ (DHFR from Walker 256 leukemia tumors); D subset of Table I with $I_2 = 0$ and $I_1 = 0$ (DHFR from L1210 leukemia tumors). [c] An outlier is a data point for which the absolute value of the prediction error is greater than 0.8.

The two dataset were combined, and the resulting dataset of 245 compounds as well as the entire dataset of 256 compounds were analyzed using NN, MLR, and MLRI. In this case, however, it was felt that, due to possible difference of the active sites or differences in the assay procedure, the indicator variable I_1 that flags the two en-

Table IV. Comparison of Neural Networks and Multiple Linear Regression with and without Indicator Variables for Predicting Biological Activity

	no. of data points[a]		training set				test set			
method[b]	training set	test set	R	R^2	σ_E	no. of outliers[c]	R	R^2	σ_E	no. of outliers[c]
NN	100	32	0.913	0.833	0.358	6	0.897	0.804	0.372	2
MLR	100	32	0.616	0.380	0.690	18	0.558	0.312	0.707	7
MLRI	100	32	0.837	0.700	0.480	9	0.902	0.814	0.369	2
NN	66	66	0.919	0.844	0.397	3	0.820	0.672	0.431	5
MLR	66	66	0.627	0.393	0.780	15	0.527	0.277	0.671	13
MLRI	66	66	0.862	0.744	0.507	6	0.740	0.547	0.490	6
NN	57	56	0.962	0.926	0.147	0	0.721	0.511	0.341	1
MLR	57	56	0.656	0.426	0.413	3	0.373	0.139	0.461	2
MLRI	57	56	0.766	0.591	0.349	1	0.523	0.273	0.430	2

[a] Training and test sets were selected using cluster analysis. The 110/32 and 66/66 training/test combinations were subsets of the 132 data points that were tested on the enzyme from Walker 256 carcinoma (Table I, $I_1 = 1, I_2 = 0$). The 57/76 training/test combinations were subsets of the 113 data points that were tested on the enzyme from L1210 (Table I, $I_1 = 0, I_2 = 0$). [b] NN = neural networks, MLR = multiple linear regression, without indicator variables; MLRI = multiple linear regression with indicator variables. [c] An outlier is a data point for which the absolute value of the prediction error is greater than 0.8.

zymes should be retained. The network topology used was similar to that of Figure 6 with 5 hidden units and an additional input node corresponding to I_1. The results are shown in Table III.

Predicting Biological Activity. As described in the Methods two tests were used to compare the predictions of NN, MLR, and MLRI. In the first, the dataset of 132 compounds (Figure 5) was split into 100/32 and 66/66 training/test set combinations, and the dataset of 113 compounds was split into a 57/56 combination using cluster analysis. NN, MLR, and MLRI models were developed on the training set and used to predict the test set. Table IV compares the statistics for the three training/test set combinations. This table shows that NN models have enhanced predictive capabilities relative to MLR and MLRI in addition to its enhanced surface fits.

The second test utilized the cross-validation procedure described in the Methods. The cross-validated r^2 calculated for NN, MLR, and MLRI are 0.787, 0.30, and 0.640, respectively. These are consistent with the results of the cluster analysis tests.

Comparison of Neural Network and Regression Generated Biological Activity Surfaces. Equations 14 and 15 are the best fit MLR and MLRI equations for the training set of 100 data points (Figure 5).

$$A = 6.764 + \underset{(0.155)}{0.922\pi_3} - \underset{(0.041)}{0.135\pi_3{}^2} - \underset{(0.065)}{0.108\text{MR}_3} + \underset{(0.035)}{0.091\text{MR}_4}$$

$$N = 100,\ S = 0.705,\ R^2 = 0.380 \qquad (14)$$

$$A = 7.163 + \underset{(0.112)}{0.932\pi_3} - \underset{(0.031)}{0.167\pi_3{}^2} - \underset{(0.037)}{0.182\text{MR}_3} - \underset{(0.213)}{1.791I_3} + \underset{(0.188)}{0.642I_4} + \underset{(0.232)}{0.747I_5} + \underset{(0.198)}{0.556I_6}$$

$$N = 100,\ S = 0.498,\ R^2 = 0.700 \qquad (15)$$

These regression results suggest that activity depends parabolically on π_3 and linearly on MR_3 but is otherwise independent of π_4, MR_4, and $\Sigma\sigma_{3,4}$. Furthermore, a significant part of the variance is explained by the indicator variables I_3, I_4, I_5, and I_6, which select for specific R_3's and R_4's. In eq 15, the curvature of the π_3 parabola (−0.334) and the location of its maximum ($\pi_3 = 2.79$) as well as the slope of MR_3 are independent of the values of the other variables.

While it would be highly desirable to depict the results of the NN models with equations similar to 14 and 15 or with "drawings of the multidimensional response surfaces", this is not presently possible. An appreciation of the dependence of activity on properties in the NN model, however, is achievable by calculating the network output as a function of one independent variable while all the others are held constant. The results of such calculations are shown in Figure 7a–e.

Figure 7a depicts the dependence of activity on π_3 as the remaining independent variables (π_4, MR_3, MR_3, MR_4, $\Sigma\sigma_{3,4}$) are held constant at 30, 60, and 90% of their corresponding ranges in the dataset. This graph shows that, if the constant variables are held at 30% of their ranges, activity is a concave downward function of π_3 with a maximum at $\pi_3 \sim 1.5$. This resembles the parabolic dependence of activity on π_3 in eqs 14 and 15, although the location of the maximum is different and the form of the curve in Figure 7a is not formally parabolic but appears to be of higher order. However, unlike the regression models, Figure 7a shows that, if the constant variables are held at 60 and 90% of their ranges, the curve shifts to the right and elicits no maximum for $-2 < \pi_3 < 5.5$. Such shifts in functional dependence on one independent variable with values of the others is clear evidence of the ability of the network to elucidate couplings and interactions between the physicochemical properties. Such capability is not enjoyed by regression methods.

Figure 7c shows that activity is a nonlinearly decreasing function of MR_3. This is similar to the negative slope of MR_3 in eqs 14 and 15. However, unlike the regression results, the slopes of the curves in Figure 7c are sensitive to the value of all five independent variables. This is further evidence of the networks ability to elucidate intervariable couplings.

Furthermore, parts b, d, and e of Figure 7 show that activity has a strong functional dependence on π_4, MR_4, and $\Sigma\sigma_{3,4}$, which exhibit the same intervariable couplings as those described above. This is to be contrasted with the regression model which elicits no functional dependence on these three variables.

Discussion

The development of the field of quantitative structure–activity relationship is rooted in the knowledge that activity of chemical compounds is determined by their physical properties.[16] The commonly used functional form of eq 1 was inspired by prior observations that activity

[16] Hansch, C.; Muir, R. M.; Fujita, T.; Maloney, P. P.; Geiger, F.; Streich, M. *J. Am. Chem. Soc.* **1963**, *85*, 2817.

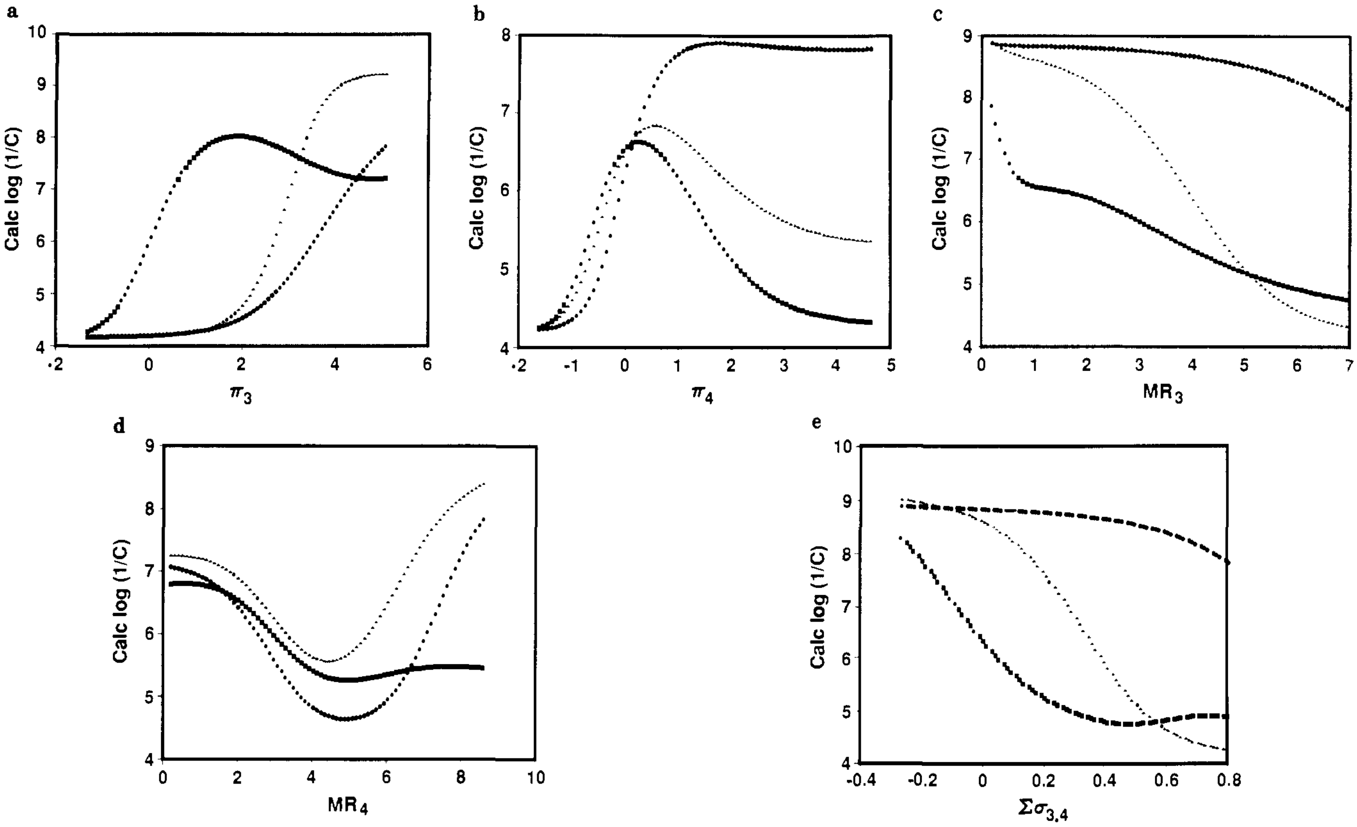


Figure 7. Neural network calculated activities as a function of π_3 (7a), π_4 (7b), MR$_3$ (7c), MR$_4$ (7d), and $\Sigma\sigma_{3,4}$ (7e). In each graph activity is calculated as a function of one variable while the remaining four are held constant at 30% (■), 60% (▲), and 90% (●) of their corresponding ranges in the dataset.

increases, reaches a maximum, then decreases as oil/water partition coefficient increases. A parabolic function was chosen to model this process. Later, this was shown to be consistent with passive permeation in biological tissue,[17] thereby giving a mechanistic foundation for correlating in vivo, but not necessarily in vitro, activities with eq 1. Familiarity with the parabolic form led to its successful utilization in modeling other physicochemical properties.

It is useful to temporarily digress and regard biological activity as a general mathematical function of physical properties, without preconceptions. Activity is measurable for any chemical compound. It is thus a bounded function. Although in principle it is conceivable that this function is discontinuous, no evidence exists to support this notion. Therefore, it is reasonable to expect that this function is analytic, having a Taylor series expressed by

$$f(x, y, ...) = \sum_{n_x=0}^{\infty}\sum_{n_y=0}^{\infty} \cdots \left\{\frac{\partial^{n_x+n_y+\ldots}f}{\partial x^{n_x}\partial y^{n_y}\ldots}\right\}_{x,y,\ldots=0} \frac{x^{n_x}y^{n_y}\ldots}{(n_x+n_y+\ldots)!} \qquad (16)$$

or, more explicitly

$$f(x, y, ...) = a_0 + a_1x + a_2y + a_3x^2 + a_4y^2 + a_5xy + a_6x^3 + a_7y^3 + a_8x^2y + a_9xy^2 + \ldots \qquad (17)$$

where f is biological activity and x and y are physical properties. Equation 1 is in fact a special case of eqs 16 and 17 in which the infinite sum is truncated at $n_x = n_y = 2$ and all cross-product terms are eliminated.

There have been many unpublished attempts enhance the capabilities of eq 1 by embellishing it with higher order and cross-product terms.[18] However, the staggering diversity of such terms can discourage the most stalwart. To the best of our knowledge, such terms are not used in practice. In the course of the current work, regression models including all second-order cross-products were attempted, resulting in insignificant improvements. Although the process of training a neural network does not explicitly invoke such higher order and cross-product terms, the results presented in the previous section clearly indicate that the networks are indirectly elucidating effects of such terms.

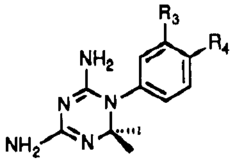
Historically, the transition from linear to nonlinear processing with neural networks occurred with the introduction of the hidden layer. Models calculated with the older perceptrons, or neural networks consisting solely of input and output layers, are equivalent to multiple linear regression models. Close inspection of eqs 2, 3, and 4 indicates that the output of each hidden unit is a nonlinear transformation of a specific linear combination of scaled independent variables. Furthermore, the network's output is a nonlinear transformation of a linear combination of the hidden units' output. Thus each hidden unit is a new variable that is a function of the original independent variables. Moreover, the output is a function of the variables in the "basis set" consisting of the totality of hidden units. Analysis of these equations indicates that the nonlinearity, specifically the number and severity of the "bends, bumps, and dips", in the resulting surfaces is related to the number of hidden units determined by minimizing the test-set variance (Figure 4). With too few hidden units, the network would be unable to extract all the relevant nonlinear features. Too many hidden units cause overfitting and "memorizing" individual data points rather than generalizing over the data set.

Notwithstanding nonlinear processing capabilities and enhanced predictions, the goal of the QSAR program is

[17] Penniston, J. T.; Beckett, L.; Bentley, D. L.; Hansch, C. *Mol. Pharmacol.* **1969**, *5*, 333.

[18] Hansch, C. Personal communication.

Table V. Outliers in the Training Set of 100 Compounds (Figure 5: 100tr) Tested on DHFR from Walker 256 Carcinoma



			obsd log	residuals[a]											
	no.[b]	R	$1/C$	MLR[c]	MLRI[d]	NN	π_3	π_4	MR_3	MR_4	$\Sigma_{\sigma 3,4}$	I_3	I_4	I_5	I_6
I	15	4-CH=CHCONHC$_6$H$_4$-4′-SO$_2$F	5.19	−2.04	−0.16	−0.10	0.00	1.99	0.10	5.22	−0.01	1	0	0	0
	17	4-CH(Ph)CH$_2$CONHC$_6$H$_4$-4′-SO$_2$F	5.74	−1.70	0.39	0.26	0.00	3.53	0.10	7.59	−0.09	1	0	0	0
	206	4-CH$_2$C$_6$H$_5$	8.05	1.02	0.16	0.28	0.00	2.01	0.10	3.00	−0.09	0	0	1	0
	223	4-(CH$_2$)$_4$OC$_6$H$_4$-4′-SO$_2$F	8.14	0.90	0.25	−0.03	0.00	4.62	0.10	5.37	−0.17	0	0	1	0
II	9	4-CONHC$_6$H$_4$-4′-SO$_2$F	4.68	−2.46	−0.67	−0.21	0.00	1.50	0.10	4.23	0.36	1	0	0	0
	11	4-C$_6$H$_5$	4.70	−2.28	−0.65	−0.14	0.00	1.96	0.10	2.54	−0.01	1	0	0	0
	51	3-C$_6$H$_5$	6.85	−0.94	0.75	−0.48	1.96	0.00	2.51	0.10	0.06	1	0	0	0
	14	4-CN	5.14	−1.67	−2.00	−0.06	0.00	−0.57	0.10	0.63	0.66	0	0	0	0
	25	4-CH$_2$CH(CH$_2$CH$_2$Ph)CONHC$_6$H$_4$-4′-SO$_2$F	6.20	−1.33	−0.94	−0.41	0.00	4.23	0.10	8.52	−0.17	0	0	0	0
	28	4-OCH$_2$CONMe$_2$	6.26	−0.73	−0.88	−0.37	0.00	−1.36	0.10	2.58	−0.27	0	0	0	0
	182	3-Cl-4-O(CH$_2$)$_2$NHCONHC$_6$H$_4$-4′-SO$_2$F	7.92	0.94	0.78	0.28	0.71	2.22	0.49	5.77	0.10	0	0	0	0
	194	4-OCH$_2$CONHC$_6$H$_4$-3′-SO$_2$F	8.00	0.80	0.86	0.25	0.00	1.61	0.10	4.91	−0.27	0	0	0	0
	209	4-CH$_2$CONHC$_6$H$_4$-3′-SO$_2$F	8.06	0.88	0.92	0.34	0.00	1.31	0.10	4.69	−0.17	0	0	0	0
	250	3,4-Cl$_2$	8.54	1.19	0.89	0.06	0.71	0.71	0.49	0.60	0.60	0	0	0	0
III	233	3-(CH$_2$)$_2$C$_6$H$_4$-4′-NHCOCH$_2$Br	8.26	0.68	0.10	0.87	2.29	0.00	5.55	0.10	−0.07	0	0	1	0
	20	3-CONHC$_6$H$_4$-4′-SO$_2$F	5.96	−1.42	0.35	−1.23	1.50	0.00	4.33	0.10	0.35	1	0	0	0
IV	35	4-CH(CH$_2$NHCOCH$_2$Br)(CH$_2$)$_3$C$_6$H$_5$	6.52	−0.87	−0.62	−0.90	0.00	2.94	0.10	7.03	−0.17	0	0	0	0
	24	4-OCH$_2$CON(Me)C$_6$H$_5$	6.17	−1.00	−0.97	−0.95	0.00	0.12	0.10	4.55	−0.27	0	0	0	0
	31	3-CH(CH$_2$NHCOCH$_2$Br)(CH$_2$)$_3$C$_6$H$_5$	6.37	−1.20	−0.83	−1.03	2.94	0.00	6.94	0.10	−0.07	0	0	0	0

[a] Residual = obsd (log $1/C$) − calc (log $1/C$). [b] Numbers in this column represent the entry in Table I. [c] Activity calculated using eq 14. [d] Activity calculated using eq 15.

efficient and facile biological activity optimization based on a mathematical model that relates it to physicochemical properties. In practice, such models are utilized in two ways. Most commonly, biological activities of entries in a functional group database are calculated with the model. Those predicted to have optimal activities are identified and pursued. Less frequently, however, novel functional groups or molecular structures that incorporate optimal physicochemical features are designed, synthesized, and tested. The practical usefulness of any QSAR method depends on the ease of accurately calculating activity from properties as well as providing a clear understanding of the quantitative and qualitative relationship between properties and activity.

In regression-based methods, the relationship between activity and properties is expressed by linear equations such as 15. With neural nets on the other hand, activity is expressed in terms of the nonlinear functions of eqs 2–4. Calculating activities of a functional group database using either method is straightforward.

De Novo design of functional groups is possible with the aid of regression-based models. The physicochemical variables in regression equations provide insights into the forces that control biological activity. Thus π_3 and MR_3 terms in eq 15 suggest that the binding site of the 3 substituent is hydrophobic and size limited. Indicator variables on the other hand reflect the effect of certain structural features on the magnitude of activity. Quite frequently they are responsible for explaining a major part of the variance. However, they provide little, if any, insight into the forces controlling activity. For example, it is not clear why the features flagged by I_3 in eq 15 decrease activity whereas those flagged by I_4, I_5, and I_6 enhance it. That notwithstanding, the optimal physicochemical and structural features indicated by regression equations such as 15 can be readily incorporated into novel functional groups.

By contrast, the closed analytic forms of eqs 2–4 do not give vivid insights into the relationships between activity and physical properties. Regression-like equations that describe NN surfaces, although desirable, are not currently avaialble. Some understanding of the relationship between biological activity and physical properties may be gleaned from Figure 7 which shows "cuts" in the corresponding surface. Each curve depicts the variation of activity with one property while the others are held constant, and the shape variation among the three curve in each part of Figure 7a–e manifests the strong intervariable coupling. Such figures describe the local, but not global, structure–activity relationships, somewhat limiting the model-based design of novel functional groups or chemical structures. Efforts to alleviate this limitation are currently underway.

In addition to enhanced predictions, the Results showed that, for the current dataset, NN models circumvent indicator variables. The latter are used to augment size, hydrophobicity, and electronic parameters in regression models. Typically, initially one attempts to correlate activity with the usual physicochemical properties. Data points are individually inspected to identify structural features common to outliers, which are then flagged with indicators variables and regression reattempted. This requires several modeling iterations and considerable time particularly with large datasets. By contrast, outliers are not separately handled in NN models. The procedure outlined in the Methods is followed using size, hydrophobicity, and electronic properties but not indicator variables as inputs. The network with the number of hidden units that minimizes the variance of a test set is the optimal model. Such models account for many MLRI and most MLR outliers.

The difference between NN and regression models' handling of outliers is related to the shapes of the corresponding surfaces. In regression models, the best surface over the space of physicochemical properties is first determined and has one of the shapes in Figure 1. Data points too far above or below it are designated outliers. An indicator variable creates a new surface above or below the former but parallel to it, that passes through correspondingly displaced outliers. NN surfaces corresponding to few hidden units and large test-set variance are akin to, although not identical with, regression surfaces over physicochemical properties, and outliers persist. Increasing the number of hidden units introduces localized deformations: elevations and depressions, that pass through outliers. Close inspection of the resulting models reveals that if NN surfaces could be expressed as polynomials, such deformations would correspond to high-order nonlinear and cross-product terms. Elucidating the NN based physicochemical effects that correspond to indicator variables such as I_3–I_6 in eqs 13 and 15 is highly desirable but unfortunately not possible at present.

Following the suggestion of one of the reviewers, Table V, which lists the outliers of the 100 data point training set in the different models, was constructed. All points in that table are MLR outliers. They are grouped into four groups: points that are well predicted by MLRI and NN models (I) and outliers in one (II, III) or both methods (IV). The structures of outliers in the different methods do not have obvious trends.

Introduction of hidden units is tantamount to fixing the number of bonds whose weights are the adjustable parameters of the model. In a fully connected network, each hidden unit with the exception of the bias unit is linked to all input and output units. The number of bonds in a neural network is calculated by

$$P = (I + 1)H + (H + 1)O$$

where P, I, H, and O are the numbers of adjustable parameters, input, hidden, and output units, respectively. As in any modeling method that relies on adjustable parameters, the ratio

$$\rho = \frac{\text{no. of datapoints}}{P}$$

is critical. In regression models, it is desirable to have at least five data points for each adjustable parameter or coefficient in the regression equation. No corresponding rule of thumb exists for NN models except that ρ must be >1. In the examples of this paper, the optimal NN models have $1.8 < \rho < 2.2$. Models with $\rho > 2.2$ were unable to extract all the relevant features and gave poor predictions. Ones in which ρ approached 1 overfitted the training set and were unable to accurately predict the test set. In other unpublished work in which the surfaces are nearly linear, optimal predictions are obtained with models having $\rho >$ 5.

To summarize, the key strengths of neural network relative to regression models for the dataset in hand are enhanced prediction accuracy and accounting for most outliers without indicator variables. On the other hand, these models could benefit from developments that improve the understanding of global structure–activity relationships. The longer computational times required for developing neural network relative to regression models is somewhat inconveniencing but not limiting.

Our experience with neural modeling of datasets from diverse backgrounds and origins including in vitro enzyme inhibitory activity, whole organism in vivo biological activities, as well as in environmental models not related to biological activity optimization indicates that, if the surfaces are linear or nearly so, the predictions of regression and neural models are close. Such linear models are frequently obtained, as may be expected, when the range of variation of independent variables is narrow. Although outliers in general are well accounted for with neural models, occasionally, albeit infrequently, they persist even in cases where the experimental data is of demonstrably high quality.

Acknowledgment. We wish to acknowledge the help of Mrs. Valeria Linn for her expert help in preparing some of the diagrams and Mr. Evan Hunt for some of the coding.

Registry No. 1, 20285-54-5; 2, 20285-47-6; 3, 20285-55-6; 4, 20285-46-5; 5, 20292-77-7; 6, 20285-44-3; 7, 20285-56-7; 8, 20285-45-4; 9, 50566-91-1; 10, 16048-35-4; 11, 4653-78-5; 12, 3932-29-4; 13, 135006-55-2; 14, 17711-68-1; 15, 17794-14-8; 16, 50581-68-5; 17, 17730-69-7; 18, 31362-31-9; 19, 17794-21-7; 20, 15422-08-9; 21, 19159-91-2; 22, 50699-41-7; 23, 20285-47-6; 24, 50575-19-4; 25, 17730-70-0; 26, 10161-71-4; 27, 20092-22-2; 28, 50581-75-4; 29, 20092-20-0; 30, 135006-56-3; 31, 19159-89-8; 32, 50866-22-3; 33, 10161-69-0; 34, 20092-21-1; 35, 19159-90-1; 36, 20285-53-4; 37, 20285-57-8; 38, 31191-36-3; 39, 19161-83-2; 40, 19160-11-3; 41, 50581-63-0; 42, 50574-80-6; 43, 50575-17-2; 44, 50574-70-4; 45, 19159-88-7; 46, 135006-57-4; 47, 50581-80-1; 48, 31362-32-0; 49, 31191-28-3; 50, 50575-02-5; 51, 52980-48-0; 52, 21970-37-6; 53, 30937-83-8; 54, 50568-38-2; 55, 30937-81-6; 56, 50581-20-9; 57, 64706-33-8; 58, 19161-91-2; 59, 50575-03-6; 60, 50581-74-3; 61, 31444-54-9; 62, 20110-67-2; 63, 17711-74-9; 64, 19161-96-7; 65, 19161-97-8; 66, 50574-86-2; 67, 50574-79-3; 68, 28392-79-2; 69, 30885-62-2; 70, 17754-25-5; 71, 31191-27-2; 72, 28392-86-1; 73, 50581-83-4; 74, 30885-58-6; 75, 31191-34-1; 76, 17730-72-2; 77, 15422-10-3; 78, 17730-73-3; 79, 50568-39-3; 80, 30885-68-8; 81, 30885-70-2; 82, 28392-87-2; 83, 30855-24-4; 84, 31191-32-9; 85, 19160-13-5; 86, 24892-10-2; 87, 50574-69-1; 88, 50574-98-6; 89, 50574-87-3; 90, 17730-66-4; 91, 50575-26-3; 92, 50575-41-2; 93, 50574-85-1; 94, 24892-11-3; 95, 24892-14-6; 96, 15422-09-0; 97, 17794-15-9; 98, 50575-15-0; 99, 30855-19-7; 100, 135006-58-5; 101, 30885-76-8; 102, 20110-90-1; 103, 28392-85-0; 104, 17794-16-0; 105, 30885-72-4; 106, 30855-22-2; 107, 20110-65-0; 108, 30855-21-1; 109, 28392-82-7; 110, 30855-14-2; 111, 20110-66-1; 112, 14052-52-9; 113, 31191-26-1; 114, 47732-14-9; 115, 47723-35-3; 116, 50574-84-0; 117, 50575-51-4; 118, 30855-20-0; 119, 31191-31-8; 120, 31191-33-0; 121, 50575-14-9; 122, 50574-67-9; 123, 30885-74-6; 124, 30855-23-3; 125, 50699-31-5; 126, 17730-67-5; 127, 19161-89-8; 128, 50575-18-3; 129, 31191-24-9; 130, 31191-25-0; 131, 30855-13-1; 132, 15422-13-6; 133, 31066-08-7; 134, 19223-16-6; 135, 17818-13-2; 136, 50574-84-0; 137, 19161-90-1; 138, 19161-87-6; 139, 19161-86-5; 140, 50567-38-9; 141, 24892-31-7; 142, 19160-07-7; 143, 24892-12-4; 144, 22083-85-8; 145, 31362-34-2; 146, 50567-39-0; 147, 24892-92-0; 148, 15422-12-5; 149, 19223-11-1; 150, 48223-06-9; 151, 38559-98-7; 152, 17794-18-2; 153, 20110-69-4; 154, 28392-83-8; 155, 30855-25-5; 156, 47698-53-3; 157, 16048-34-3; 158, 25292-67-5; 159, 13351-02-5; 160, 1492-81-5; 161, 38559-94-3; 162, 50699-39-3; 163, 31362-29-5; 164, 31362-25-1; 165, 20091-76-3; 166, 50567-07-2; 167, 19160-09-9; 168, 22083-84-7; 169, 31191-29-4; 170, 31444-55-0; 171, 20110-68-3; 172, 135006-59-6; 173, 50567-12-9; 174, 50575-13-8; 175, 50575-71-8; 176, 50651-37-1; 177, 50575-31-0; 178, 50575-04-7; 179, 24892-89-5; 180, 17720-23-9; 181, 20091-77-4; 182, 22083-83-6; 183, 19160-08-8; 184, 10161-70-3; 185, 19161-92-3; 186, 24892-98-6; 187, 17077-18-8; 188, 24892-98-6; 189, 31191-30-7; 190, 31362-35-3; 191, 31362-23-9; 192, 30855-18-6; 193, 31362-33-1; 194, 17794-23-9; 195, 50575-75-2; 196, 4038-61-3; 197, 24892-90-8; 198, 48227-67-4; 199, 50699-33-7; 200, 47698-51-1; 201, 31362-27-3; 202, 14052-52-9; 203, 17794-22-8; 204, 17730-68-6; 205, 28392-77-0; 206, 36076-92-3; 207, 135006-60-9; 209, 19160-12-4; 210, 47732-13-8; 211, 38560-04-2; 212, 50699-35-9; 213, 22877-78-7; 214, 24988-55-4; 215, 24892-95-3; 216, 17794-17-1; 217, 31362-28-4; 218, 50575-83-2; 219, 16156-69-7; 220, 48227-67-4; 221, 24947-06-6; 222, 25292-66-4; 223, 28392-78-1; 224, 4086-50-4; 225, 50699-37-1; 226, 36983-62-7; 227, 31362-30-8; 228, 50567-40-3; 229, 47723-40-0; 230, 38560-09-7; 231, 22796-71-0; 232, 38559-99-8; 233, 16156-73-3; 234, 31362-22-8; 235, 31362-21-7; 236, 31362-17-1; 237, 38560-05-3; 238, 24892-87-3; 239, 14357-85-8; 240, 31444-56-1; 241, 19161-80-9; 242, 38560-08-6; 243, 38589-51-4; 244, 38560-00-8; 245, 38560-07-5; 246, 16156-72-2; 247, 38560-03-1; 248, 38560-01-9; 249, 38560-06-4; 250, 13344-99-5; 251, 38666-63-6; 252, 38559-95-4; 254, 38560-10-0; 255, 38560-02-0; 256, 38559-96-5; dihydrofolate reductase, 9002-03-3.