

Using Artificial Neural Networks To Classify the Activity of Capsaicin and Its Analogues

M. Hosseini,* D. J. Maddalena, and I. Spence

Department of Pharmacology, The University of Sydney, Sydney, NSW 2006, Australia

Received May 27, 1997[®]

Back-propagation artificial neural networks (ANNs) were trained with parameters derived from different molecular structure representation methods, including topological indices, molecular connectivity, and novel physicochemical descriptors to model the structure–activity relationship of a large series of capsaicin analogues. The ANN QSAR model produced a high level of correlation between the experimental and predicted data. After optimization, using cross-validation and selective pruning techniques, the ANNs predicted the EC₅₀ values of 101 capsaicin analogues, correctly classifying 34 of 41 inactive compounds and 58 of 60 active compounds. These results demonstrate the capability of ANNs for predicting the biological activity of drugs, when trained on an optimal set of input parameters derived from a combination of different molecular structure representations.

INTRODUCTION

Capsaicin is the active ingredient of chilis. It has a wide range of biological effects on the cardiovascular, nervous, and respiratory systems, and some of its actions suggest that it, or an analogue, may be useful as an analgesic.^{1–4} The potential clinical use of its analgesic and peripheral antiinflammatory effects has attracted much attention and prompted investigations into the relationship between the structure of capsaicin analogues and their agonist activities.

The molecular structure of the capsaicin molecule, shown in Figure 1, is that of an amide (region B), a hydrophilic ring on one end (region A), and a lipophilic carbon chain on the other (region C). Analogues of capsaicin have the same basic structure but contain variations in one or more of the three regions.

Previous structure–activity relationship studies of capsaicin analogues^{1–3} have shown that each region makes an implicit contribution to activity. The following moieties were identified as being common to these regions: (i) a 3-methoxy-4-hydroxybenzyl ring in region A; (ii) a dipolar amide or thiourea in region B; (iii) a lipophilic octanyl or *p*-chlorophenethyl group in region C.

A recent QSAR study⁴ analyzed a large set of capsaicin analogues. Using the MULTICASE methodology⁴ as a predictive tool, they identified structural features responsible for activity, which they quantitatively weighted to derive a QSAR equation using multilinear regression (MLR) analysis. The QSAR equation predicted the log EC₅₀ of 70 compounds ($n = 70$), resulting in a training correlation $R = 0.72$ and standard deviation $SD = 0.77$, correctly classifying 51 of 52 active compounds and 3 of 18 inactive compounds.

In the present study, ANNs were used to explore the possibility of predicting the activity of capsaicin analogues from their structural makeup. As ANNs have successfully been used in the fields of chemistry, QSAR studies of biological systems, and the field of drug design,^{5–8} the main objective of this study was to develop a scheme for encoding relevant information from molecular structure into a format which is suitable for use in an ANN and to develop a QSAR

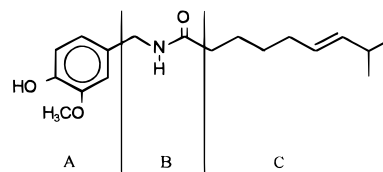


Figure 1. Structure of capsaicin.

model of the capsaicin analogues with predictive capabilities, which so far has been unattainable.

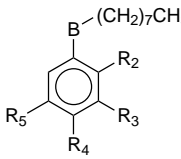
Data representation is an important aspect of any study involving ANNs as ANNs employ learning procedures to discern patterns in data that are encoded numerically and presented in the form of training examples. By extracting the relevant features from examples, the ANN develops an internal representation of the system, which allows predictions to be made. In order to present the ANN with as much useful information as possible from structure, we investigated a variety of structural parametrization methods. These were molecular connectivity,⁹ topological indices and charged indices,^{10,11} connection table theory,⁸ and a simple atomic decomposition of molecules into atom type.⁷ These were used in conjunction with novel physicochemical descriptors^{12,13} and some simple atomic descriptors developed in this study.

MATERIALS AND METHOD

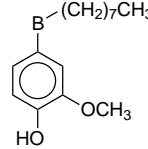
Experimental Data. A data set of 103 capsaicin analogues was used in the ANN training set, Table 1. Of these compounds 101 were obtained from a recent publication.⁴ These compounds have retained their original numbering for ease of back-referencing. Two “dummy” compounds (991 and 992) were developed to supplement the data for training purposes. These are referred to in the Results and Discussion.

Not all compounds available were used in the study. Of the original 123 compounds 20 were discarded owing to difficulties in encoding their structures. We are currently investigating new structural parametrization methods to represent these compounds for further analysis. Two other compounds (102 and 104) were discarded due to activity anomalies.

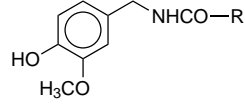
[®] Abstract published in *Advance ACS Abstracts*, October 15, 1997.

Table 1. Experimental Activities and Structures of Capsaicin Analogues^a


compd no.	R ₂	R ₃	R ₄	R ₅	B	EC ₅₀ (μM)	classification
1	H	H	H	H	CH ₂ NHCO	> 100	+
2	H	OCH ₃	OH	H	CH ₂ NHCO	0.55	+
3	H	OCH ₃	OCH ₃	H	CH ₂ NHCO	6.41	+
4	H	OCH ₃	H	H	CH ₂ NHCO	> 100	+
5	H	H	OCH ₃	H	CH ₂ NHCO	> 100	+
6	OCH ₃	H	H	H	CH ₂ NHCO	> 100	+
7	H	OH	H	H	CH ₂ NHCO	> 100	+
9	H	OH	OH	H	CH ₂ NHCO	0.63	+
10	H	OCH ₃	H	OH	CH ₂ NHCO	> 100	+
11	OCH ₃	H	OH	OCH ₃	CH ₂ NHCO	> 100	+
12	OCH ₃	H	OH	H	CH ₂ NHCO	> 100	+
13	H	OCH ₃	OCH ₃	OCH ₃	CH ₂ NHCO	> 100	+
14	OH	OH	OH	H	CH ₂ NHCO	7.64	+
15	H	OCH ₃	SH	H	CH ₂ NHCO	> 100	+
16	H	OCH ₃	NO ₂	H	CH ₂ NHCO	7.91	—
24	H	OCH ₃	H	H	CH ₂ CONH	> 100	+
25	H	OCH ₃	OH	H	CH ₂ CONH	0.30	+
26	H	H	OH	H	CH ₂ CONH	6.50	+
28	H	OH	OH	H	CH ₂ CONH	0.41	+
29	H	CH ₃	OCH ₃	H	CH ₂ CONH	> 100	+
30	H	CH ₃	OH	H	CH ₂ CONH	> 100	+
31	H	CH ₂	OH	H	CH ₂ CONH	1.04	+
32	H	NH ₂	OH	H	CH ₂ CONH	> 100	—
33	H	NHCOCH ₃	OH	H	CH ₂ CONH	> 100	+
34	H	OC ₂ H ₅	OH	H	CH ₂ CONH	4.34	+
35	H	OH	OCH ₃	H	CH ₂ CONH	1.82	+
37	H	OCH ₃	OH	OCH ₃	CH ₂ CONH	> 100	—
39	OCH ₃	OCH ₃	OH	H	CH ₂ CONH	> 100	—

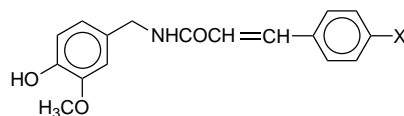


compd no.	B	EC ₅₀ (μM)	classification	compd no.	B	EC ₅₀ (μM)	classification
40	NHCO	4.48	+	53	CH=CHCONH (Z)	17.90	+
41	(CH ₂) ₂ NHCO	18.30	—	54	CH ₂ NHCONH	0.36	+
42	CH ₂ N(CH ₃)CO	> 100	+	55	CH ₂ NHCSNH	0.06	+
43	CH ₂ OCO	14.20	+	56	NHCSNH	2.57	+
44	CH ₂ NHCS	0.28	+	57	CH ₂ N(CH ₃)CSNH	> 100	+
45	CH ₂ NHSO ₂	1.32	+	58	CH ₂ NHCSN(CH ₃)	0.53	+
46	CONH	> 100	+	59	CH ₂ NHCSNHCO	> 100	+
47	(CH ₂) ₂ CONH	2.32	+	60	CH ₂ NHC(=NCN)NH	3.28	+
48	CH ₂ CONH	6.29	+	61	CH ₂ NHC(=CH-NO ₂)NH	> 100	+
49	CH(OH)CONH	1.16	+	62	CH=CHCO (E)	> 100	+
50	CH ₂ COO	0.67	+	63	CH ₂ CH ₂ CO	2.13	+
51	CH ₂ COS	1.17	+	64	CH ₂ COCH ₂	3.78	+
52	CH=CHCONH (E)	> 100	+	65	CH ₂ CH ₂ CH ₂	> 100	—

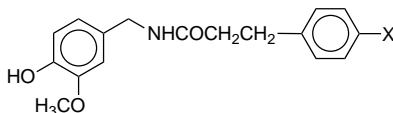


compd no.	R	EC ₅₀ (μM)	classification	compd no.	R	EC ₅₀ (μM)	classification
67	(CH ₂) ₆ CH(CH ₃) ₂	0.19	+	70	CH ₂ O(CH ₂) ₂ O(CH ₂) ₂ OCH ₃	> 100	+
68	(CH ₂) ₁₀ CO ₂ H	> 100	+	71	(CH ₂) ₁₆ CH ₃	> 100	—
991 ^b	(CH ₂) ₉ CO ₂ H	> 100	+	72	(CH ₂) ₇ CH=CH(CH ₇)CH ₃ (E)	0.36	+
992 ^b	(CH ₂) ₁₁ CO ₂ H	> 100	+	73	(CH ₂) ₇ CH=CH(CH ₇)CH ₃ (Z)	0.17	+
69	CH ₂ O(CH ₂) ₂ OCH ₃	> 100	+	74	CH ₂ Br	> 100	+
				75	CH ₂ Cl	> 100	+

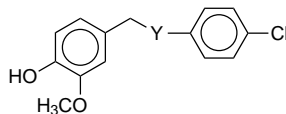
Table 1 (continued)



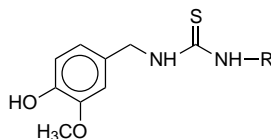
compd no.	X	EC ₅₀ (μM)	classification	compd no.	X	EC ₅₀ (μM)	classification
76	4-H (<i>E</i>)	11.8	+	82	4-N(CH ₃) (<i>E</i>)	4.39	+
77	4-Cl (<i>E</i>)	1.24	+	83	4-I (<i>E</i>)	0.35	+
78	4-Cl (<i>Z</i>)	50.1	+	84	4-NHCHO (<i>E</i>)	>100	+
79	4-NO ₂ (<i>E</i>)	4.58	+	85	2,4-Cl (<i>E</i>)	0.62	+
80	4-CN (<i>E</i>)	26.5	+	86	3-Cl (<i>E</i>)	2.58	+
81	4-Ph (<i>E</i>)	0.24	+				



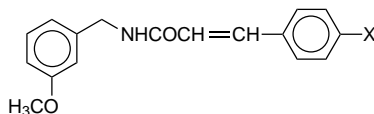
compd no.	X	EC ₅₀ (μM)	classification	compd no.	X	EC ₅₀ (μM)	classification
87	H	45.26	+	89	4-NO ₂	21.00	+
88	4-Cl	3.09	+	90	4-OH	>100	+



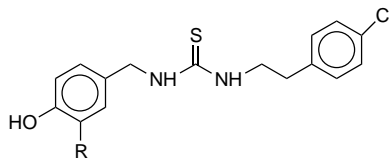
compd no.	Y	EC ₅₀ (μM)	classification	compd no.	Y	EC ₅₀ (μM)	classification
91	NHCOCH ₂	7.77	+	93	CONHCH ₂ CH ₂	0.66	+
92	NHCOC≡C	4.10	+				



compd no.	R	EC ₅₀ (μM)	classification	compd no.	R	EC ₅₀ (μM)	classification
94	H	>100	+	101	(CH ₂) ₁₁ CH ₃	0.16	+
95	(CH ₂) ₃ CH ₃	9.48	+	102	(CH ₂) ₁₅ CH ₃	>100	NA ^c
96	(CH ₂) ₄ CH ₃	1.96	+	103	(CH ₂) ₁₇ CH ₃	8.1	+
97	(CH ₂) ₅ CH ₃	0.18	+	104	(CH ₂) ₇ CH=CH(CH ₇)CH ₃ (<i>Z</i>)	>100	NA ^c
98	(CH ₂) ₆ CH ₃	0.29	+	106	CH(Ph) ₂	0.27	+
99	(CH ₂) ₈ CH ₃	0.1	+	107	C(Ph) ₃	>100	—
100	(CH ₂) ₉ CH ₃	0.11	+				



compd no.	X	EC ₅₀ (μM)	classification	compd no.	X	EC ₅₀ (μM)	classification
113	4-NHCHO (<i>E</i>)	>100	+	116	4-N(CH ₃) (<i>E</i>)	>100	+
114	4-NO ₂ (<i>E</i>)	>100	+	117	4-Cl (<i>E</i>)	>100	+
115	4-H (<i>E</i>)	>100	+	118	4-I (<i>E</i>)	>100	—



compd no.	R	EC ₅₀ (μM)	classification	compd no.	R	EC ₅₀ (μM)	classification
119	OCH ₃	0.06	+	120	OH	0.10	+

^a Classification is denoted by the following: “+” for a correct prediction; “—” for an incorrect prediction. ^b Dummy compounds. ^c NA = not used in data set.

The agonist activity available was the ability of the compound to stimulate the entry of Ca²⁺ into rat dorsal root

ganglia, measured as the concentration (μM) of compound needed to produce 50% (EC₅₀) of the maximal Ca²⁺ influx.⁴

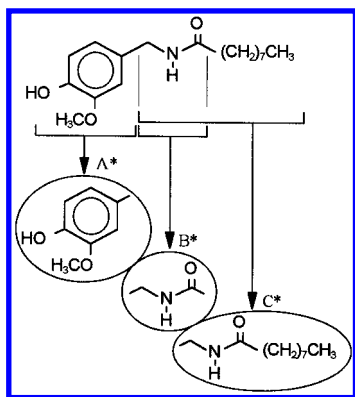


Figure 2. Substructures of compound 2 as defined by regions A*–C*.

These values ranged from $0.06 \mu\text{M}$ to $>100 \mu\text{M}$ implying a difference of 5 orders of magnitude between the EC_{50} of active and inactive compounds. $\log \text{EC}_{50}$ values were used.

Modeling Methods. Initial examination of the molecular structures of the compounds in the data set indicated that no one method of data representation could be used to model the whole molecule. To overcome this problem, the whole molecule was segmented into three regions for ease of analysis.

The structural models of the three regions used in this study contain substructures similar to regions A–C described in a previous paper.⁴ The three regions, termed regions A*–C*, are illustrated in Figure 2 and are as follows:

Region A* (similar to region A) is a single benzene ring, the base structure onto which a varying range of substituent groups such as hydroxy (–OH), methoxy (–OCH₃), etc., are attached to positions R₂, R₃, R₄, and R₅.

Region B* (similar to region B, connecting region A to C) is a branched chain fragment with a linear length of up to five atoms. These substructures contain atoms important for both structural integrity and bonding interaction.²

Region C* (a composite of regions B and C), the “tail-end” of the compound, is a molecule made up of both aliphatic and aromatic components of varying configurations. It has been suggested that the substructures within this region are responsible for lipophilicity.³

The QSAR models for the regions just described are as follows:

Region A*. The configuration of this region lent itself to conventional QSAR analysis using novel physicochemical descriptors. Five substituent principle properties¹² ω_1 – ω_5 which globally encode molecular bulk, electronic characteristics, and hydrogen bonding capabilities, respectively, were used, together with eight disjoint principle properties¹³ (DPP) s_1 , s_2 , e_1 , e_2 , l_1 , l_2 , h_1 , and h_2 which encode the two principle components of steric, electronic, lipophilic, and hydrogen bonding descriptor sets. The values of these substituent descriptors were assigned to each position R₂–R₅ around the benzene ring. Including the value $\log P$ (of the whole compound), a set of 53 input variables was generated for the ANN training set.

The substructures of 28 compounds (1–39) were selected from the complete data set for use in the training set for region A* using the following selection criterion and assumption: regions B and C must be kept constant, allowing variability in region A, and the amide and reverse amide moieties confer similar potencies¹ (cf. 2 vs 25). Figure 3

illustrates how region A* was parametrized into a format for processing by the ANN.

Region B*. It has been well-established that the spatial atomic arrangement of this region is important for activity.^{1–4} In order to extract as much information from this region while encoding its structural integrity, a combination of two data representation methods was employed.

The first method consists of information derived from a modified connection table representation⁸ of the molecule that includes the tabulation of hydrogen donor/acceptor capabilities of atoms. Figure 4(i) shows the substructure of compound 2 and the connection table numbering system.

The complete connection table, shown in Figure 4(ii), is an array of six parameters (A , c_1 , c_2 , b , h_a , h_d) by eight atoms, the maximum number of atoms in this region (compound 61). The first row of the table is generated as follows: the first atom, carbon, has an atomic number $A = 6$, has connection number $c_1 = 2$, is connected to atom 1, $c_2 = 1$, has a single bond, $b = 1$, is not a hydrogen acceptor, $h_a = 0$, and is not a hydrogen donor, $h_d = 0$. This procedure was repeated for all the (non-hydrogen) atoms in the region in a row by row fashion and was finished with two connection numbers (denoted by an asterisk) to encode the position of the last atom. For substructures with fewer than eight atoms, zeros were substituted in the rows of the respective non-existent atoms to fill the table.

The second method employed a simple count of atom types⁷ with the inclusion of isomerism. Figure 4(iii) shows the tabulation of isomerism, where E/Z implies nonisomeric (or racemic) and E and Z implies their respective isomerism. Simple logic switches (1 and 0) were used to denote the isomerism of the molecule. The table shown in Figure 4(iv) encodes information relating to the frequency of atom types. The procedure for generating this table is started with a logic switch to denote the starting atom, either C or N, and continued with a count of atom types, where, for example, H indicates the number of hydrogen atoms and C3 indicates the number of carbon atoms with three connections.

Including the value of $\log P$, these two data representation methods generated 68 input variables for the ANN training set and were applied to the substructures of 28 compounds (2, 25, 40–65) which were selected from the complete data set on the basis that regions A and C remain constant with variability in region B. Figure 4 illustrates the parametrization of region B*.

Region C*. Klopman and Li⁴ used the computed $\log P$ values of the compounds as an indicator for predicting activity/inactivity. Although using this value alone had limited success, it highlighted the fact that lipophilicity is an important factor for compound activity. The variability in $\log P$ values is largely dependent on the tail-end of the compound.

With this fact in mind and with the knowledge that there must exist other unknown, and, thus, unavailable, physicochemical properties to represent this tail-end region, we explored the use topological indices as a way of generating information that, by chance, would correlate with these unknown physicochemical properties. The topological descriptors used in this study were connectivity indices^{9,10} and their linear combinations, charge indices¹¹ and their linear combinations, and a Wiener array¹⁴ which is an extension of the Wiener number principle. A computer program was developed for the purpose of generating these variables from

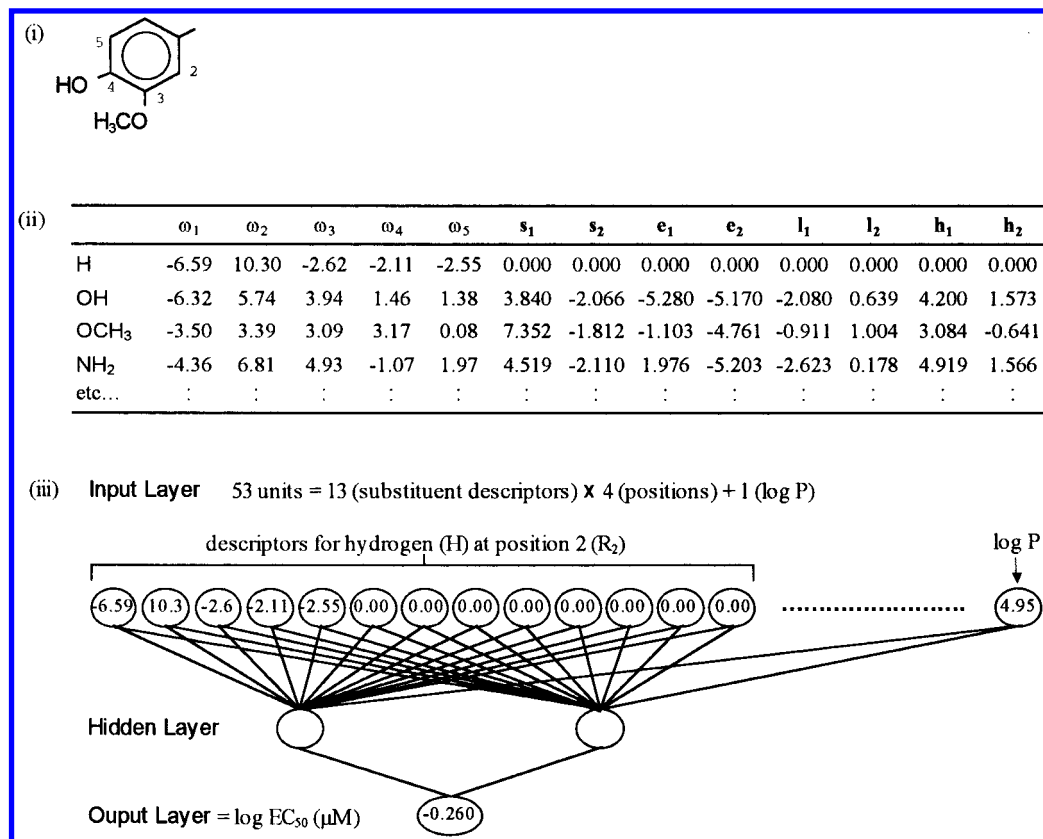


Figure 3. Parametization of region A*. (i) Substructure of compound 2 showing positions R_2 – R_5 . (ii) Substituent principle properties and DPP of substituents. (iii) Three layer ANN model of region A*.

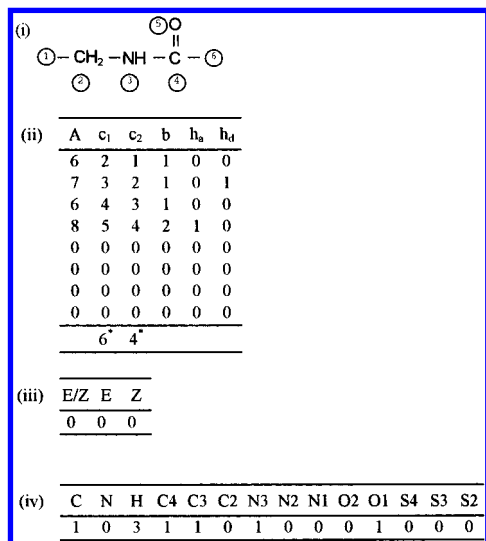


Figure 4. Variables used for the parametization of region B*. (i) Substructure of compound. (ii) Connection table representation of substructure: A = atomic number; c_1 = connection number of atom; c_2 = connection number of previous atom; b = bond type; h_a = hydrogen acceptor; h_d = hydrogen donor. (iii) Isomeric descriptors. (iv) Atom type descriptors.

connectivity data. A simple count of atom types⁷ and E/Z isomerism, as described above, were also encoded into the training set input data.

Including the value of log P, a set of 108 input variables was generated by these data representation methods and was applied to the substructures of 44 compounds (2, 25, 55, 66–92, 94–101, 103, 106, 107, and 119) selected on the criterion that regions A and B remain constant with variability only in region C and the assumption that the amide and thiourea moieties confer similar potencies (cf. 2 vs 55 and 88 vs 119).

These input variables were entered into the input layer of the ANN for processing for region C*.

Neural Network Model. The studies were carried out using Bioactivnet,¹⁵ a three layer back-propagation ANN program running in Microsoft Windows 95 on a 586/120 MHz personal computer in conjunction with Microsoft Excel 5.0.

All ANNs were constructed with variable numbers of neurons in the input layer representing the input parameters, two neurons in the hidden layer and a single neuron in the output layer corresponding to log EC₅₀, the measured activity. All inputs were scaled to a range of 0.0–1.0 on the basis of the minimum and maximum values of the input source range. Inverse scaling was used for the outputs. Sigmoidal transfer functions were used in all layers.

In order to encode a distinct difference between the values of active and inactive compounds (>100 μ M) for training purposes in the ANN, the log EC₅₀ value of the inactive compounds was set at 3 (1000 μ M).

The ANNs were trained for a range of learning cycles varying between 7000 and 20 000, depending on the region being investigated and the number of parameters being used. In order to avoid overtraining, the number of learning cycles was set to a value to give the highest cross-validated correlation value with the lowest number of cycles.

The ability of the ANNs to predict results accurately was tested using the ($N - 1$), cross-validation technique,⁵ where for each given set of input parameters using N compounds, N ANNs were trained each using $N - 1$ compounds. At the end of the training run, each trained ANN was then used to predict the activity of the missing compound.

ANNs predict more accurately when they have been optimized,^{5,16} with the accuracy being gauged by R_{cv} (see

Table 2. log EC₅₀ Results of This Study^a

compd no.	actual	predicted	diff	compd. no.	actual	predicted	diff	compd no.	actual	predicted	diff
1	3.000	3.000	0.000	47	0.365	0.061	0.305	82	0.642	0.946	-0.304
2	-0.260	0.391	-0.650	48	0.799	1.466	-0.667	83	-0.456	-0.952	0.496
3	0.807	1.618	-0.811	49	0.064	0.792	-0.728	84	3.000	2.092	0.908
4	3.000	3.000	0.000	50	-0.174	-0.442	0.268	85	-0.208	-0.606	0.398
5	3.000	2.615	0.385	51	0.068	-1.152	1.220	86	0.412	1.147	-0.735
6	3.000	3.000	0.000	52	3.000	2.672	0.328	87	1.656	0.933	0.723
7	3.000	3.000	0.000	53	1.253	1.504	-0.251	88	0.490	1.265	-0.775
9	-0.201	0.366	-0.566	54	-0.444	-0.353	-0.091	89	1.322	0.902	0.420
10	3.000	3.000	0.000	55	-1.222	-0.819	-0.403	90	3.000	3.000	0.000
11	3.000	3.000	0.000	56	0.410	-0.174	0.584	91	0.890	0.667	0.223
12	3.000	3.000	0.000	57	3.000	2.876	0.124	92	0.613	0.510	0.103
13	3.000	3.000	0.000	58	-0.276	-0.287	0.011	93	-0.180	0.159	-0.340
14	0.883	1.708	-0.825	59	3.000	2.641	0.359	94	3.000	2.529	0.471
15	3.000	2.884	0.116	60	0.516	0.088	0.428	95	0.977	-0.323	1.299
16	0.898	2.459	-1.561	61	3.000	2.647	0.353	96	0.292	0.288	0.005
24	3.000	2.730	0.270	62	3.000	3.000	0.000	97	-0.745	-0.803	0.058
25	-0.523	-0.248	-0.275	63	0.328	1.243	-0.914	98	-0.538	-0.698	0.160
26	0.813	1.874	-1.061	64	0.577	0.524	0.053	99	-1.000	-0.810	-0.190
28	-0.387	-0.247	-0.140	65	3.000	1.905	1.095	100	-0.959	-0.875	-0.084
29	3.000	3.000	0.000	66	-0.523	-1.150	0.628	101	-0.796	-0.498	-0.298
30	3.000	2.729	0.271	67	-0.721	-0.044	-0.678	103	0.908	0.293	0.616
31	0.017	-0.992	1.009	68	3.000	2.548	0.452	106	-0.569	-0.397	-0.172
32	3.000	1.818	1.182	69	3.000	3.000	0.000	107	3.000	0.146	2.854
33	3.000	3.000	0.000	70	3.000	3.000	0.000	113	3.000	3.000	0.000
34	0.637	1.549	-0.911	71	3.000	0.427	2.573	114	3.000	3.000	0.000
35	0.260	0.615	-0.355	72	-0.444	0.776	-1.219	115	3.000	3.000	0.000
37	3.000	1.539	1.461	73	-0.770	0.271	-1.041	116	3.000	3.000	0.000
39	3.000	0.849	2.151	74	3.000	2.303	0.697	117	3.000	3.000	0.000
40	0.651	1.034	-0.383	75	3.000	3.000	0.000	118	3.000	0.520	2.480
41	1.262	2.642	-1.380	76	1.072	0.412	0.659	119	-1.222	-0.835	-0.387
42	3.000	2.130	0.870	77	0.093	0.664	-0.570	120	-1.000	-0.840	-0.160
43	1.152	0.720	0.432	78	1.700	0.200	1.500	991	3.000	2.035	0.965
44	-0.553	-0.029	-0.524	79	0.661	1.195	-0.535	992	3.000	2.479	0.521
45	0.121	0.505	-0.385	80	1.423	1.625	-0.201				
46	3.000	2.403	0.597	81	-0.620	1.366	-1.986				

^a Actual = the experimental log EC₅₀ values; predicted = the log EC₅₀ value predicted by the ANN using cross-validation techniques; diff = the difference between the Actual and Predicted values.

below). The optimum set of input parameters for use by the ANN was determined by the method of selective pruning,⁵ where, at the end of each ANN run, known as a pruning run, the inputs not significantly contributing to the prediction of the result were systematically eliminated to improve generalization. These input parameters were generally characterized by high cross-correlation values and/or by low internal weight values assigned by the ANN.

At the end of each pruning run, the average training set correlation (R_T) and standard error (SE_T) and the cross-validation correlation (R_{CV}) and standard error (SE_{CV}) between the predicted and known data sets were calculated. Once the calculated value of R_{CV} for the run reached a maximum, the set of input parameters, hence the ANN model for the region being analyzed, was defined as optimized.

Three separate ANNs were initially constructed to analyze regions A*, B*, and C*, respectively. Upon optimization of each ANN, their respective sets of input parameters were amalgamated to produce a single ANN model to represent the whole compound. This was then used to analyze the complete data set of 103 compounds.

RESULTS AND DISCUSSION

Table 2 lists the overall results of this study, showing the experimental log EC₅₀ values, the predicted log EC₅₀, and their difference. The predicted log EC₅₀ values are based on cross-validation data. Any compound with a log EC₅₀ value greater than 2 (>100 μ M) was classified as inactive.

Table 3.

(a) Overall Results of the Study ^a							
	R_T	SE_T	R_{CV}	SE_{CV}	accuracy, %	R_{CV}^*	SE_{CV}^*
region A*	0.984	0.254	0.899	0.617	89	0.970	0.351
region B*	0.990	0.194	0.930	0.519	96	0.945	0.474
region C*	0.954	0.457	0.827	0.856	93	0.894	0.647
whole compound	0.961	0.420	0.854	0.796	91	0.926	0.571
(b) Comparison of the Results from This Study and Those of Ref 4 ^b							
	n	R_T	SE_T	R_{CV}	SE_{CV}	active	inactive
this study	103	0.961	0.420	0.854	0.796	58/60	36/43
other study ^d	70	0.720	0.770			51/52	3/18
							77

^a R_T = mean correlation and SE_T = standard error between known and predicted activity during training; R_{CV} = mean correlation and SE_{CV} = standard error between known and predicted based on ($N - 1$) cross-validation studies. An asterisk indicates the mean correlation and standard error of compounds when incorrectly classified compounds (outliers) were removed. ^b n = number of compounds. Active 58/60 implies that 58 out of 60 active compounds were classified correctly.

Table 1 shows which compounds were classified correctly, where a plus sign (+) implies a correct classification.

Table 3a summarizes the overall results, including the correlations and standard errors of regions A*–C*. The ANN model representing the whole compound was trained on 34 input variables representing 103 compounds ($\rho = 2.6$). It accurately classified 36 of 43 inactive compounds (including dummy compounds 991 and 992) and 58 of 60 active

Table 4. Squared Cross-Correlation Table ($R^2 > 0.10$) of the Optimal Parameter Set for Region A*^a

	$R_2\omega_1$	R_2l_1	$R_3\omega_1$	$R_3\omega_4$	$R_4\omega_1$	R_4l_1	R_5l_2	R_5h_2	$\log P$
$R_2\omega_1$	1.00								
R_2l_1	0.40	1.00							
$R_3\omega_1$	—	—	1.00						
$R_3\omega_4$	—	—	—	1.00					
$R_4\omega_1$	—	—	—	—	1.00				
R_4l_1	—	—	—	—	—	1.00			
R_5l_2	—	—	—	—	—	—	1.00		
R_5h_2	—	—	—	—	—	—	—	1.00	
$\log P$	—	—	—	—	0.10	0.20	—	—	1.00

^a $R_2\omega_1$ implies the descriptor ω_1 at position R_2 .

compounds with a training set correlation of $R_T = 0.961$ and cross-validation correlation of $R_{CV} = 0.854$ between experimental and predicted data. These results are a considerable improvement when compared to the work of previous authors.⁴ Table 3b shows a comparison of the results from this study and the work of other authors.⁴

The compounds used in the ANN training sets for determining the optimum set of input parameters for regions A*–C* did not include eight compounds (93, 113–118, and 120). These compounds were introduced to the whole compound ANN without any previous training, forcing the ANN to predict their activity based on previous experience. The ANN accurately classified seven of them.

The ANN model for region A* initially consisted of 53 input variables for representation of 28 compounds. The information content of these inputs was sufficient for the ANN to learn and to form correlations, resulting in a training set correlation of $R_T = 0.992$ and cross-validation correlation of $R_{CV} = 0.786$. After six pruning runs, the ANN was optimized and the input variables were reduced to nine, yielding an improvement in generalization and hence prediction of activity. Although this improvement led to an increase in $R_{CV} = 0.899$, the ANN still had difficulty in correctly classifying three compounds (inactive, 32, 39; active, 26).

All parameters in the optimum input for region A* were considered to be independent by cross-correlation; see Table 4. The highest squared cross-correlation value between any two parameters was between $R_2\omega_1$ and R_2l_1 (0.40). However, removal of either of them caused a significant reduction in both training and cross-validation correlations, suggesting that both had important contributions to make.

Of the remaining nine inputs, the ANN indicated a strong dependence (i.e., a high weight value) on ω_1 , the substituent principle property that encodes molecular bulk, and l_1 , the DPP descriptor representing lipophilicity, most notably at positions R_2 , R_3 , and R_4 . At position R_5 , the ANN favoured l_2 and h_2 . It is interesting to note that the ANN showed a dependence on $\log P$. This would mean that the subtle differences in $\log P$ values caused by the different structural configurations of region A* correlates with activity, on the condition that region B contains an amide or reverse amide moiety and region C contains an octanoyl group.

The ANN model for region B* initially consisted of 68 input variables, representing 28 compounds, resulting in correlation values $R_T = 0.996$ and $R_{CV} = 0.593$. After 9 successive pruning runs, the ANN was optimized and the number of inputs was reduced to 12, yielding correlation values $R_T = 0.990$ and $R_{CV} = 0.930$. The ANN had difficulty classifying one compound (active, 41).

All parameters in the optimum input for region B* were considered to be independent by cross-correlation: see Table 5. The highest squared cross-correlation was between P_2b and E_B (0.46) and P_2b and C4 (0.35). However, removal of any one of them or any combination of them was detrimental to the overall performance of the ANN, so they were kept.

Of the remaining 12 inputs, the ANN showed a strong dependence on variables b , h_a , and h_d , the descriptors encoding bond type and hydrogen acceptor and donor capabilities, at the second, third, fourth, fifth, and seventh atoms in region B*. These results would indicate that the first atom in region B* and any atom past the fifth has little, if no, contribution to activity, and of the remaining atoms in between, their spatial distance from region A* and their molecular composition including features such as =O, =S (hydrogen acceptors), and –NH– (hydrogen donor) are significantly required for activity, supporting the ideas proposed by Walpole et al.^{1–3} and Klopman and Li⁴ relating the atomic makeup of region B and activity.

The ANN model for region C* was initially comprised of 108 inputs, resulting in correlation values $R_T = 0.991$ and $R_{CV} = 0.633$. After 18 successive pruning runs, the ANN was optimized and the number of inputs was reduced to 13, yielding correlation values $R_T = 0.954$ and $R_{CV} = 0.827$. The ANN had difficulty classifying three compounds (inactive, 71, 84, and 107).

All parameters in the optimum input for region C* were considered to be independent by cross-correlation, see Table 6. The highest squared cross-correlation was between Δ^3J and Δ^5J (0.39) and $f^4\chi_{PC}$ and O2 (0.36). However, removal of any one of them or any combination of them was detrimental to the overall performance of the ANN, so they were kept.

The reason why the ANN had difficulty predicting certain compounds lies partly in the way that ANNs learn. ANNs, like any other type of artificial intelligence system that learns by the way of examples, requires a minimum amount of training data similar enough in content to learn from and, thus, to predict from. If the data are unique, it will not contribute any value to the training set and its ability of prediction. This “uniqueness” was found in most of the compounds that the ANN could not accurately predict.

One possible method for testing and overcoming this problem of uniqueness of data was to create dummy data. By way of example, the original ANN model for region C* could not correctly identify compound 68 as inactive. The ANN model was then modified to include two dummy compounds (991 and 992) with similar activities, $EC_{50} > 100 \mu M$ (inactive). The ANN as a result classified all three compounds correctly as inactive.

In this case, creating compounds 991 and 992, $(CH_2)_9CO_2H$ and $(CH_2)_{11}CO_2H$, respectively, analogous to compound 68, and assuming their EC_{50} values to be $> 100 \mu M$ (inactive) could be justified, as both Walpole et al.^{1–3} and Klopman and Li⁴ clearly expressed that any hydrophilic moieties, such as a carboxylic group, in region C destroy activity.

The ANN had more difficulty predicting inactive compounds than active compounds. The most probable reason for this lies mainly in the fact that the inactive compounds had a cut-off value for EC_{50} (> 100) rather than a determined value. Because of this, the ANN was trained on partially false values and, as a consequence, formed partially false

Table 5. Squared Cross-Correlation Table ($R^2 > 0.10$) of the Optimal Parameter Set for Region B*^a

	P _{2b}	P _{2h_a}	P _{3h_a}	P _{3h_b}	P _{4A}	P _{5b}	P _{5h_b}	P _{7h_a}	E _B	C4	S2	S1
P _{2b}	1.00											
P _{2h_a}	—	1.00										
P _{3h_a}	—	—	1.00									
P _{3h_b}	—	0.12	—	1.00								
P _{4A}	—	—	—	0.10	1.00							
P _{5b}	—	—	0.14	—	—	1.00						
P _{5h_b}	—	—	0.19	—	0.14	0.14	1.00					
P _{7h_a}	—	—	—	—	—	—	0.16	1.00				
E _B	0.46	—	—	—	—	—	—	—	1.00			
C4	0.35	—	—	—	—	—	—	—	0.16	1.00		
S2	—	—	—	—	0.12	—	—	—	—	—	1.00	
S1	—	—	—	—	0.32	—	—	—	—	—	—	1.00

^a P_{5b} implies the descriptor *b* for the fifth atom in the region B* substructure.**Table 6.** Squared Cross-Correlation Table ($R^2 > 0.10$) of the Optimal Parameter Set for Region C*^a

	³ χ _{C^v}	⁴ χ _{C^v}	<i>f</i> ³ χ _P	<i>f</i> ⁴ χ _{PC}	Δ ⁴ <i>G</i>	Δ ³ <i>J</i>	Δ ⁵ <i>J</i>	W ₁ - C	N3 ⁺	O2	S1	E/Z _c	Z _c
³ χ _{C^v}	1.00												
⁴ χ _{C^v}	—	1.00											
<i>f</i> ³ χ _P	0.17	—	1.00										
<i>f</i> ⁴ χ _{PC}	—	—	—	1.00									
Δ ⁴ <i>G</i>	0.13	—	—	—	1.00								
Δ ³ <i>J</i>	—	—	—	—	—	1.00							
Δ ⁵ <i>J</i>	—	—	—	—	—	0.39	1.00						
W ₁ - C	—	—	0.18	—	—	—	—	1.00					
N3 ⁺	—	—	0.11	—	0.14	—	—	—	1.00				
O2	—	—	—	0.36	—	—	—	—	—	1.00			
S1	—	—	0.15	0.19	0.12	—	—	—	—	—	1.00		
E/Z _c	0.18	—	0.16	—	0.25	—	—	—	—	—	0.20	1.00	
Z _c	—	—	—	—	—	—	—	—	—	—	—	0.10	1.00

^a For explanation of variables, refer to refs 9–11.

internal representations of the compounds to make predictions from. This not only affected the correct classification of the inactive compounds but also had an impact on the predicted EC₅₀ values of the active compounds that it classified correctly.

Fortunately, ANNs have the robustness for handling noisy data, and in most cases, the ANN could “see” through this partially false value to correctly classify most of the inactive compounds. For two inactive compounds (32 and 65) that were incorrectly classified as active, intuition would say otherwise when looking at their predicted EC₅₀ values (80 and 66 μM, respectively).

CONCLUSION

The results have shown that ANNs¹⁵ can be used to predict the activity of drugs from calculable information derived from structure and available physicochemical descriptors. A data set of 101 of capsaicin analogues was analyzed and an ANN QSAR model was developed that accurately classified the activity of 91% of the compounds with a high degree of precision. Included in the data set were eight compounds that had not been used for training the ANN. Seven of these compounds were classified correctly.

The process of developing the ANN QSAR model for the whole compound demonstrated the utility of applying combinations of different data representation methods to substructures of molecules that would otherwise be too difficult to analyze as a whole. By separation of the capsaicin analogues into analyzable segments, regions A*–C* and application of an optimal combination of each data representation method to each region, a single ANN QSAR model for the complete molecule was developed from the amalgamation of the optimized descriptor sets of each region.

The flexibility of segmenting a compound into analyzable segments proved especially useful for analyzing region C*. When region C*, consisting of region C only (refer to Introduction and refs 3 and 4), was initially encoded^{7,9–11,14} for analysis, the information content of the data proved to be insufficient for the ANN to form any correlations from. By creation of region C* as a composite of regions B and C (refer to Introduction and refs 2–4), this problem was overcome.

The application of physicochemical parameters and DPP descriptors to model the substructures of region A* has been shown to be successful. These descriptors can have the potential of broadly describing a pharmacophore of the capsaicin–receptor interaction site, given a more complete data set for representation. The use of connection table representations to model the substructures of region B* has also been shown to be very successful, in so far as the optimized set of descriptors closely defined a model of the receptor site as proposed by other researchers.² With more work, this method may be refined to develop a complete pharmacophoric representation of the region. Region C* proved to be more difficult to model. Although the information content implicit in the descriptor sets was sufficient to model the region with success, more work will be required to develop methods of extracting information, such as structural shapes, from the optimized descriptors.

We are currently investigating new methods of data representation to model the complete set of 123 capsaicin analogues which in its generic form will have the capability of being extended to other classes of drugs.

REFERENCES AND NOTES

- (1) Walpole, C. S. J.; Wrigglesworth, R.; Bevan, S.; Campbell, E. A.; Dray, A.; James, I. F.; Perkins, M. N.; Reid, D. J.; Winter, J. Analogues of Capsaicin with Agonist Activity as Novel Analgesic Agents; Structure-Activity Studies. Part 1. The Aromatic "A-Region". *J. Med. Chem.* **1993**, *36*, 2362-2372.
- (2) Walpole, C. S. J.; Wrigglesworth, R.; Bevan, S.; Campbell, E. A.; Dray, A.; James, I. F.; Perkins, M. N.; Reid, D. J.; Winter, J. Analogues of Capsaicin with Agonist Activity as Novel Analgesic Agents; Structure-Activity Studies. Part 2. The Amide-Bond "B-Region". *J. Med. Chem.* **1993**, *36*, 2373-2380.
- (3) Walpole, C. S. J.; Wrigglesworth, R.; Bevan, S.; Campbell, E. A.; Dray, A.; James, I. F.; Perkins, M. N.; Reid, D. J.; Winter, J. Analogues of Capsaicin with Agonist Activity as Novel Analgesic Agents; Structure-Activity Studies. Part 3. The Hydrophobic Side-Chain "C-Region". *J. Med. Chem.* **1993**, *36*, 2381-2389.
- (4) Klopman, G.; Li, J. Quantitative structure-agonist activity relationship of capsaicin analogues. *J. Comput. Aided. Mol. Des.* **1995**, *9*, 283-294.
- (5) Maddalena, D. J.; Johnston, J. A. R. Prediction of Receptor Properties and Binding Affinity of Ligands to Benzodiazepines/GABA_A Receptors Using Artificial Neural Networks. *J. Med. Chem.* **1995**, *38*, 715-724.
- (6) Maddalena, D. J. Applications of Artificial Neural Networks to Quantitative Structure Activity Relationships. *Expert Opin. Ther. Pat.* **1996**, *6*, 239-251.
- (7) Burden, F. R. Using Artificial Neural Networks to Predict Biological Activity from Simple Molecular Structural Considerations. *Quant. Struct.-Act. Relat.* **1996**, *15*, 7-11.
- (8) Elrod, D. W.; Maggiora, G. M. Applications of Neural Networks in Chemistry. 1. Prediction of Electrophilic Substitution Reactions. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 477-484.
- (9) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure Activity Analysis*. Research Studies Press (John Wiley and Sons): Letchworth, Hertfordshire, England, 1986.
- (10) Galvez, J.; Garcia-Domenech, R. Topological Approach to Drug Design. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 272-284.
- (11) Galvez, J.; Garcia-Domenech, R.; Salabert, M. T.; Soler, R. Charge Indexes. New Topological Descriptors. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 520-525.
- (12) Van de Waterbeemd, H.; El Tayar, N.; Carrupt, P. A.; Testa, B. Pattern recognition study of QSAR substituent descriptors. *J. Comput.-Aided Mol. Des.* **1989**, *3*, 111-132.
- (13) Van de Waterbeemd, H.; Costantino, G.; Clementi, S.; Crciani, G.; Valigi, R. Disjoint Principle Properties of Organic Substituents. *Chemometric Methods in Molecular Design*; Methods and Principles in Medicinal Chemistry, Vol. 2; VCH: Weinheim, Germany, 1995; pp 103-112.
- (14) Melssen, W. J.; Smits, J. R. M.; Daalmans, G. J.; Kateman, G. Using Molecular Representations in Combination with Neural Networks. A Case Study: Prediction of HPLC Retention Index. *Comput. Chem.* **1994**, 157-172.
- (15) Bioactivnet is a back-propagation ANN generation program available from AiMaze (desm@mail.usyd.edu.au), 14 Birch Place, Kirrawee, NSW 2232, Australia.
- (16) Maddalena, D. J.; Johnston, G. A. R. Use of Artificial Neural Networks as Receptor Pharmacophores for Flexible Molecules. *J. Mol. Graph.*, in press.

CI9700384