

## Development of QSAR Models To Predict and Interpret the Biological Activity of Artemisinin Analogues

Rajarshi Guha and Peter C. Jurs\*

152 Davey Laboratory - Chemistry, Penn State University, University Park, Pennsylvania 16802

Received February 5, 2004

This work presents the development of Quantitative Structure–Activity Relationship (QSAR) models to predict the biological activity of 179 artemisinin analogues. The structures of the molecules are represented by chemical descriptors that encode topological, geometric, and electronic structure features. Both linear (multiple linear regression) and nonlinear (computational neural network) models are developed to link the structures to their reported biological activity. The best linear model was subjected to a PLS analysis to provide model interpretability. While the best linear model does not perform as well as the nonlinear model in terms of predictive ability, the application of PLS analysis allows for a sound physical interpretation of the structure–activity trend captured by the model. On the other hand, the best nonlinear model is superior in terms of pure predictive ability, having a training error of 0.47 log RA units ( $R^2 = 0.96$ ) and a prediction error of 0.76 log RA units ( $R^2 = 0.88$ ).

### INTRODUCTION

Qinghao (*Artemisia annua*) is an herb that has been used for over 2000 years in Chinese medicinal practice to treat fevers.<sup>1</sup> In 1972 the active compound of this herb, artemisinin, was isolated and was demonstrated to have significant antimalarial activity.<sup>1</sup> This finding was significant because artemisinin is structurally very different from the standard family of antimalarial drugs, which are based on quinine and its synthetic analogues. Subsequent research led to derivatives<sup>1,2</sup> of artemisinin such as artemether, arteether, and artesunate (Figure 1). The artemisinin family of molecules has been extensively studied to elucidate its mechanism of action as an antimalarial and to develop more potent and selective antimalarial agents.<sup>3–6</sup> An essential feature of artemisinin (and analogous molecules) activity is hypothesized to be the presence of a peroxide bridge, which forms a bond with a high valence non-heme iron molecule, leading to generation of free radicals.<sup>4,5</sup>

A number of QSAR studies have also been reported for prescreening of prospective artemisinin analogues for antimalarial activity.<sup>7–15</sup> A number of these studies<sup>10–12</sup> have used comparative molecular field analysis (CoMFA)<sup>16,17</sup> as a tool to model the activity of artemisinin analogues in terms of active site binding. CoMFA is a 3D-QSAR technique that involves the alignment of a set of molecules in three-dimensional space. Once a suitable alignment is obtained, a steric or electrostatic field is constructed using a probe atom. The resultant field is then correlated with the reported activity values of the molecules. An example of this is the work presented by Avery et al.<sup>10</sup> in which they considered a data set of 211 artemisinin analogues. They performed PLS analyses of several CoMFA models built using a number of different training sets and a test set of 15 or 20 compounds,

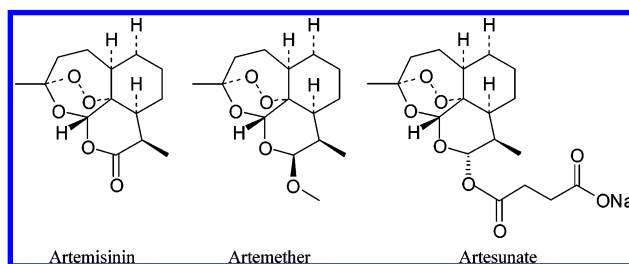


Figure 1. Artemisinin and derivatives.

depending on the size of the training set. Some of the models considered racemic compounds in the training set, whereas other models excluded them. For the former class of models the  $R^2$  values ranged from 0.82 to 0.88 with a  $q^2$  value of 0.72. For the latter class of models (in which the training set consisted of 157 molecules) they obtained  $R^2$  values ranging from 0.95 to 0.96 for the training set, while  $q^2$  values ranged between 0.68 and 0.73 during cross-validation.

The goal of this work is to use the data collected by Avery et al.<sup>10</sup> to develop 2D QSAR models using the ADAPT<sup>18,19</sup> methodology, which is not dependent on molecular alignments. ADAPT (Automated Data Analysis and Pattern Recognition Toolkit) has been shown to provide highly accurate QSAR models to predict biological activities of organic molecules.<sup>20–25</sup> For this study it will be shown that the best linear model provides an interpretation of the SAR trend present in the data set, while the neural network model provides superior predictive ability. These models could serve as a potential screening mechanism to identify useful artemisinin analogues from a large library of compounds.

### DATA SET

The total data set consisted of the 211 compounds reported by Avery et al.<sup>10</sup> For each molecule, the logarithm of the

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

relative activity (referred to as log RA), defined as

$$\log \text{RA} = \log \left( \frac{\text{IC}_{50} \text{ of artemisinin}}{\text{IC}_{50} \text{ of the analog}} \right) \times \log \left( \frac{\text{MW of analog}}{\text{MW of artemisinin}} \right)$$

was used as the dependent variable. However the data set contained a number of enantiomeric pairs. Since the ADAPT descriptors cannot differentiate between enantiomeric molecules, the member of each pair with the lowest log RA value was removed. This resulted in a data set of 179 molecules. One of the challenging aspects of this data set was that it contained several molecules with the same log RA of  $-4.0$ . The molecules with log RA values of  $-4.0$  were structurally diverse, thereby making model development a difficult task.

### METHODOLOGY

The ADAPT methodology involves several steps. The first step is to calculate molecular structure descriptors for the data set. The suite of descriptors calculated by ADAPT include geometric, topological, electronic, and hybrid descriptors. Geometric descriptors depend on accurate 3D geometries of the molecules. This class of descriptors encode various geometric features of the molecule such as moments of inertia,<sup>26</sup> molecular surface areas, and volumes.<sup>27</sup> Topological descriptors consider the molecular structure in terms of a mathematical graph, thus only 2-D representations are needed for descriptor generation. As a result this class of descriptors encode various topological invariants and features such as path lengths, connectivity indices,<sup>28–31</sup> and electrotopological state.<sup>32,33</sup> Finally, electronic descriptors characterize features such as HOMO and LUMO energies, electronegativity, and hydrogen bond formation.<sup>34</sup> Finally, hybrid descriptors are a combination of the above types. Examples of hybrid descriptors are the charged partial surface area descriptors<sup>35</sup> and hydrophobicity descriptors.<sup>36</sup>

In all, 299 descriptors were calculated for each structure in the data set. However, many of these descriptors were highly correlated or contained redundant information. Hence, the next step involved objective feature selection in which highly correlated and redundant descriptors were removed from the pool. This was achieved by two methods. First, the Pearson correlation coefficient was calculated for all pairs of descriptors in the pool. For two descriptors having an  $R^2$  value greater than a user specified cutoff (0.8 in this case), one of the two descriptors was rejected at random. Second, an identical test was carried out in which a descriptor was rejected if the values of the descriptor for more than a user specified percentage (80% in this case) of the molecules was identical. This resulted in a reduced pool of 65 descriptors which was used for model development.

The molecules were divided into three sets: training, cross-validation (CV), and prediction sets. The training set was used to build both linear and nonlinear models so an accurate relationship could be found between structure and biological activity. The CV set was used during the development of nonlinear models to prevent neural network over training. This set is not required during the development of linear models, so the structures in the CV set were combined with the training set. Finally, the prediction set is a group of

**Table 1:** Statistics for the Best Linear Regression Model<sup>a</sup>

descriptor	beta	SE	<i>t</i>	<i>P</i>	VIF
constant	-60.5625	5.2834	-11.5	$<2 \times 10^{-16}$	
N7CH	-0.2148	0.0134	-16.1	$<2 \times 10^{-16}$	1.6
NSB-12	0.2238	0.0238	9.4	$<2 \times 10^{-16}$	1.3
WTPT-2	27.9391	2.6136	10.7	$<2 \times 10^{-16}$	1.4
MDE-14	0.1118	0.0247	4.5	$1.18 \times 10^{-5}$	1.5

<sup>a</sup> N7CH – number of seventh-order chains;<sup>28,29,31</sup> NSB-12 – number of single bonds; WTPT-2 – the molecular ID number<sup>39</sup> considering only carbon atoms; MDE-14 – the molecular distance edge vector,<sup>41</sup> considering only primary and quaternary atoms.

molecules that has not been used to develop the model and serve as to test the predictive ability of the model with unknown compounds. These sets were created using the activity binning method. This method consists of ranking the molecules according to their dependent variable value and then populating the training, CV, and prediction sets such that the range of the dependent variable is represented within each set. This method resulted in the training set containing 144 molecules, the cross-validation set 17 molecules, and the prediction set 18 molecules.

After objective feature selection, predictive models were generated by using a simulated annealing<sup>37</sup> or genetic algorithm<sup>38</sup> to search the descriptor space for optimal subsets of descriptors. The optimization routines were coupled with either a multiple linear regression routine or a computational neural network to find models. In the case of linear models, models were accepted if the *t* values for all the selected descriptors were greater than 4.0, thereby ensuring that the descriptors were statistically significant. For nonlinear models, the descriptor subsets selected by the genetic algorithm were fed to a three layer, fully connected, feed-forward neural network to test fitness. The best neural network models were those that minimized the cost function shown below:

$$\text{cost} = \text{TSET}_{\text{rms}} + 0.5 |\text{TSET}_{\text{rms}} - \text{CVSET}_{\text{rms}}|$$

After several of the top (low cost) models were obtained a more rigorous analysis was performed on each model to identify the optimal neural network parameters.

### RESULTS

**Linear Models.** The best linear model consisted of the four descriptors tabulated in Table 1. The first descriptor was the number of seventh-order chains.<sup>28,29,31</sup> A seventh-order chain is a series of seven atoms that contain at least one ring. For example cycloheptane would have one count of a seventh-order chain. Furthermore, a molecule such as 1-methylbenzene would also have one count of a single seventh-order chain because it consists of seven atoms, six of which form a ring. The second descriptor was the number of single bonds. The third descriptor was a weighted-path descriptor which is based on a modification of the molecular ID number<sup>39,40</sup> described by Randic in which the molecular ID number is divided by the total number of atoms in the molecule. Molecular ID numbers were designed to provide unique identification numbers based on topological path lengths but stressing more on local features. In data presented by Randic<sup>39</sup> a few general conclusions may be noted regarding the relation between molecular ID numbers and

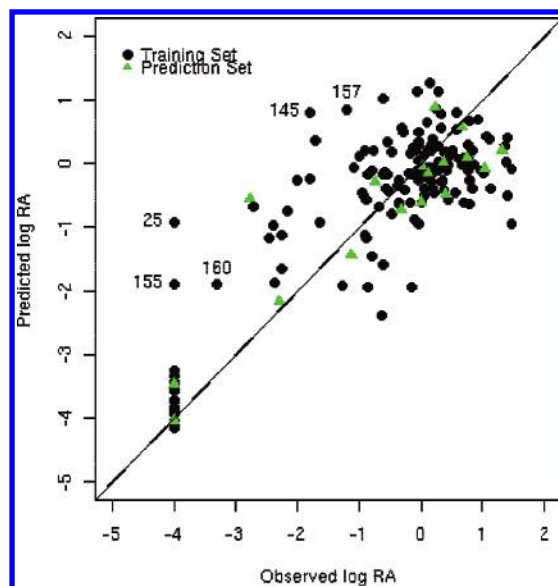
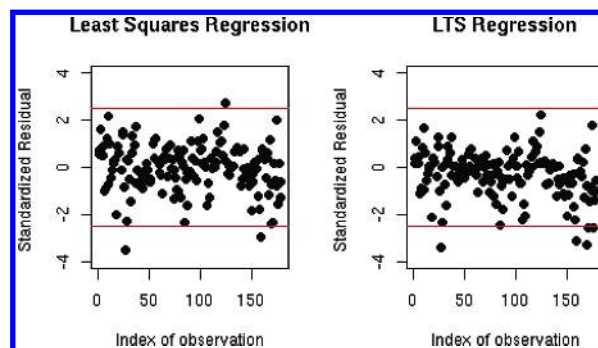
**Table 2:** Maximum and Minimum Values for the Descriptors Used in the Best Linear Model and the Dependant Variable

	dependent variable	N7CH	NSB	WTPT-2	MDE-14
maximum	1.47	36	37	2.13	19.99
minimum	-4.00	0	12	1.87	0.00

molecular structure. In the case of monocyclic systems, larger rings, larger chains, and increased substitution (for a given substituent) lead to higher values of the molecular ID number. Furthermore, for bicyclic structures, equalized branches as well as substitution at carbons of lower valency lead to higher values of the molecular ID number. In the original work, Randic<sup>39</sup> did not discuss the case of polycyclic molecules in depth. However, one may conclude that for the polycyclic structures used in this study a higher degree of branching (i.e., more substitutions) coupled with equalized branches in the ring system will lead to a higher value of the molecular ID number.

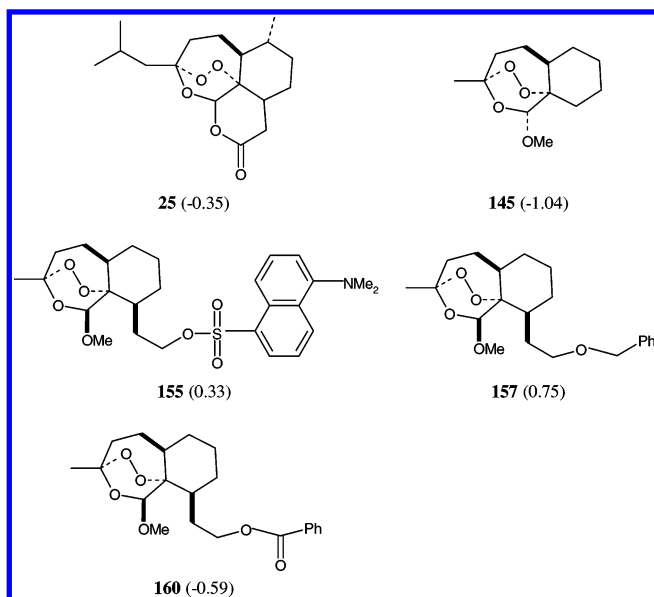
The final descriptor used in the model was a molecular distance edge vector,<sup>41</sup> denoted by  $\lambda$ . The MDE-14 descriptor is defined as the geometric mean of the topological path lengths between primary and quaternary carbons. As a result one may view this descriptor as characterizing the extent of side chains from the main body of a branched molecule. Thus the descriptor may also be correlated to molecular volume.<sup>41</sup> Consequently, molecules with a higher number of rings, longer side chains along with more substitution in the cyclic region will generally have higher values of MDE-14. As described in Liu et al.,<sup>41</sup> the descriptor provides good discrimination between structural isomers in a homologous data set. Table 2 shows the maximum and minimum values for all descriptors in the best linear model.

The statistical details of this model are reported in Table 1. All the  $t$  values are significant with low  $p$  values which confirms the significance of each descriptor. The  $F$  statistic (on 4 and 157 degrees of freedom) for this model is 87.1 (compared to the critical value of 2.42 at the 0.05 level of significance) with a  $p$  value of less than  $2 \times 10^{-16}$ . The lowest partial  $F$  value for the coefficients was 20.5 (compared to a critical of the  $F$  distribution with 1 and 157 degrees of freedom of 3.90 at the 0.05 level of significance). Furthermore, the variance inflation factors are all less than 1.6, which indicates the absence of multicollinearities in the model. Thus the model is statistically valid. The RMSE for the training set was 0.86 ( $R^2 = 0.68$ ), and the RMSE for the prediction set was 0.78 ( $R^2 = 0.77$ ). Figure 2 shows a fit plot of observed versus calculated log RA values. As can be seen from Figure 2 there are several apparent outliers including a group of molecules having an observed log RA of -4.0, which are not well predicted. To detect outliers and to investigate whether the latter molecules are behaving as outliers or simply as leverage points, a least trimmed squares<sup>42</sup> (LTS) regression algorithm was employed using the R software package.<sup>43</sup> We used LTS rather than the usual least-squares regression to detect outliers due to the more robust nature of the LTS algorithm as a result of which it is able to differentiate between leverage points and true outliers to a better extent than ordinary least-squares regression using the LTS model a plot of the standardized residuals versus observation was generated. Figure 3 compares the plots of the standardized residuals versus the indices of the training set observations for the least-squares and LTS models. As

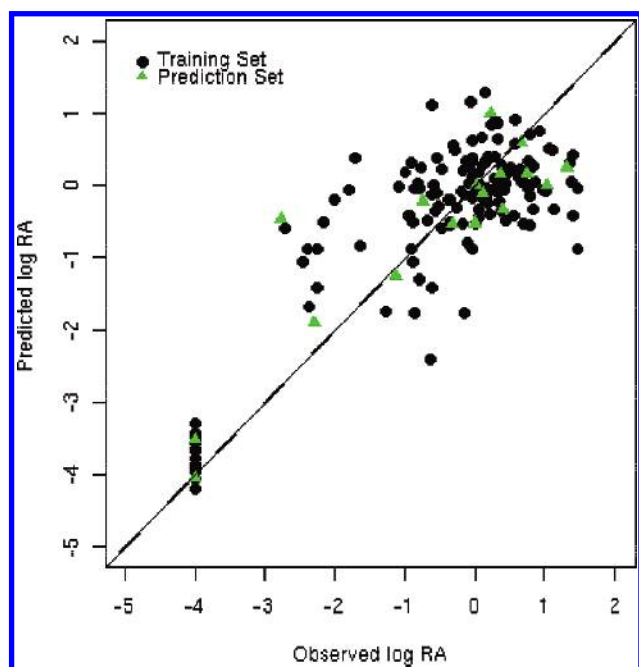
**Figure 2.** A plot of observed versus predicted relative activity values from the best linear model. The numbered points are the molecules that were considered to be outliers in the residual plot generated using LTS regression (see Figures 3 and 4).**Figure 3.** A comparison of standardized residuals versus indices of the training set observations using simple least squares and the more robust LTS algorithm.

is evident from the LTS residuals plot three observations appear to be distinct outliers with an additional two being borderline. The structures of the five molecules considered to be outliers are shown in Figure 4. It is interesting to note that, in general, the group of molecules with log RA values of -4.0 are not considered as outliers by the robust regression algorithm, thus demonstrating that this model was able to characterize these molecules well. However two members of the data set with log RA values of -4.0 were classified as outliers. Though there are no significant features of these two molecules that sets them apart from other members of the group, it may be noted that in the case of the outliers the peroxide linkage is surrounded by one or two hydroxyl groups. However it is not apparent as to how this would cause the model to classify them as outliers. Once the outliers were detected they were removed from the training set, and ordinary least-squares regression was carried out again. Figure 5 shows the results of a least-squares regression in which the molecules classified as outliers by the LTS model have been removed. The statistics of the resultant model are improved compared to the original least-squares model. The RMSE values for the training and prediction sets are both 0.77. The  $R^2$  value for the training has increased to 0.74 along with an  $F$  statistic (on 4 and 152 degrees of freedom) value





**Figure 4.** The structures of the outliers (and corresponding activity values) detected in the best linear model using LTS regression.



**Figure 5.** A plot of observed versus predicted log RA after outliers detected via LTS regression have been removed.

of 108.3 (compared to a critical value of 2.43 at the 0.05 level of significance). The lowest partial  $F$  value for the coefficients was 26.1 (compared to the critical value of the  $F$  distribution on 1 and 152 degrees of freedom of 3.90 at the 0.05 level of significance). The  $R^2$  for the prediction set using the new model was 0.77. At this point it is useful to note that in the original work, outlier removal was not carried out. Outlier detection and regeneration of the linear model in the current work was carried out mainly to increase the quality of the linear model for subsequent interpretation using the PLS. That is, linear models were investigated mainly for the purpose of providing interpretive ability as opposed to predictive ability (which is discussed later in the context of neural network models).

Finally, to ensure that the linear models were not due to chance correlations, the dependent variable for the training

**Table 3:** Summary of the PLS Analysis for the Best 4 Descriptor Type I Model

components	X variance	error SS	$R^2$	PRESS	$Q^2$
1	0.19	174.11	0.60	198.22	0.553
2	0.52	140.65	0.68	145.80	0.670
3	0.83	134.82	0.69	141.12	0.684
4	1.00	132.58	0.69	139.05	0.687

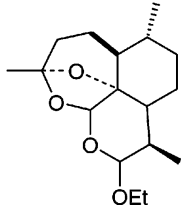
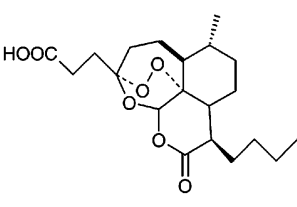
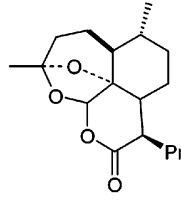
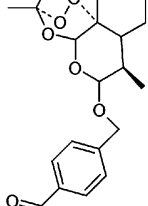
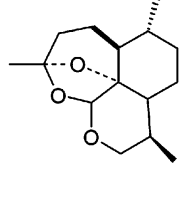
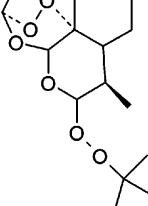
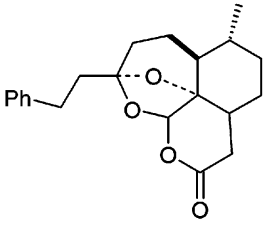
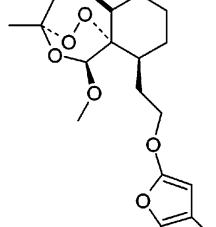
**Table 4:** X-Weights for the 4 Descriptor Linear Model

	component 1	component 2	component 3	component 4
N7CH	-0.68	-0.46	0.34	-0.47
NSB	0.65	-0.56	-0.14	-0.49
WTPT-2	0.27	0.54	0.67	-0.43
MDE-14	0.21	-0.44	0.65	0.59

set (with the outliers removed) was scrambled 100 times, and linear models were built with the randomized dependent variables. If a true QSAR relationship exists with the real dependent variable, results for the scrambling runs should be very poor. The average  $R^2$  for the 100 regressions was 0.02 with values ranging from 0.01 to 0.10. For the prediction set the average  $R^2$  was 0.21 with values ranging from 0.0003 to 0.68. Though a value of 0.68 does appear to be unnaturally high it should be noted that this occurred once in 100 randomized runs and that the next largest  $R^2$  value was 0.30. It may also be noted that the above results are in close accordance to the theoretically expected value of  $R^2$  for a model built from random variables. Thus, these results indicate that chance correlations played a minimal role (if any) during the model development stages.

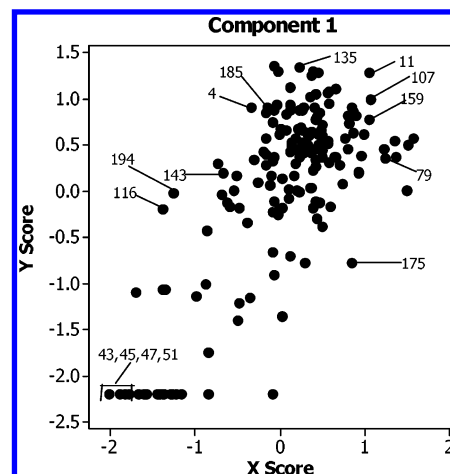
The last step in the analysis of the best linear model involved a partial least-squares analysis to provide an interpretation of the SAR trend captured by the model. The technique described by Stanton<sup>44</sup> enables one to extract information regarding SAR trends captured by a linear model. The PLS analysis of the 4-descriptor model was carried out with the help of the Minitab<sup>45</sup> software package (using a leave-one-out cross-validation scheme). The analysis indicated that the number of optimal components was four, thus the model was not overfitted. A summary of the statistics for the four components is provided in Table 3. Table 4 shows the X-weights for the four valid components. Each PLS component is a linear combination of the four descriptors used in the model. Thus, the X-weights represent the contribution (or relative importance) of each descriptor within a given component. However, as can be seen from Table 3, component 4 explains less than 0.5% of the total variance ( $Q^2$ ) explained by the model, and so the following discussion only considers the first three components.

From Table 4 it is seen that in component 1, the most highly weighted descriptors are NSB and N7CH. The coefficient for NSB is positive indicating that a larger number of single bonds is correlated with a higher activity value. On the other hand, the negative coefficient of N7CH indicates that smaller values of this descriptor are correlated with higher values of activity. This trend can be seen in molecules 43, 45, 47, and 51 all of which are inactive and have correspondingly high values for the descriptor N7CH. Molecules 11, 79, 107, and 159 are relatively active and have correspondingly small values for the N7CH descriptor. The structures of these molecules are compared in Figure 6, and their positions are marked on the score plot for component

Inactive	Active
	
<b>43</b> (-4.00)	<b>11</b> (1.36)
	
<b>45</b> (-4.00)	<b>79</b> (-0.07)
	
<b>47</b> (-4.00)	<b>107</b> (0.92)
	
<b>51</b> (-4.00)	<b>159</b> (0.58)

**Figure 6.** A comparison of the more active and less active compounds described using component 1. The value of log RA is provided within parentheses.

1 (Figure 7). The feature common to the less active molecules is the fact that they all have an ether linkage bridging the seven-member ring, whereas the more active molecules contain a peroxide linkage. As a result of the presence of the ether linkage, the number of seventh-order chains (i.e., a contiguous series of seven atoms containing a ring structure) increases. From Figure 6 it appears that molecule 51 is similar in size to molecule 107 and thus appears to invalidate the size trend described above. However molecule 51 does not contain an endoperoxide group but does have a number of ether linkages. The absence of the endoperoxide group is responsible for the low activity of this molecule, even though it is similar in size to the active molecules shown in Figure 6. This is further confirmed by the fact that a number of molecules in the data set containing a peroxide linkage but lacking the endoperoxide group showed very low activities (with log RA values around -4.0). Since active



**Figure 7.** The score plot for component 1.

compounds contain the endoperoxide group, this trend supports the theory that antimalarial activity of artemisinins depends on the presence of this group to form a high valence non-heme iron oxo species<sup>4,5</sup> and is evidence for the fact that the model has been able to capture an important feature of the data sets in the context of antimalarial activity.

The other highly weighted descriptor in component 1 is NSB, the number of single bonds. This descriptor is very simplistic in nature and essentially characterizes the size of the molecule. Since the weight of the descriptor in the PLS model is positive, higher activity is correlated with a larger number of single bonds, indicating that larger molecules will tend to have higher activity, all other factors being equal. This trend can be seen in the log RA values for compounds **11**, **79**, **107**, and **159** (which are generally larger and have higher log RA values) and compounds **43**, **45**, **47**, and **51** (which are generally smaller due to lack of large side chains and have lower log RA values). As can be seen from the score plot for component 1 (Figure 7) the upper left (overestimated) and lower right (underestimated) regions of the plot are not significantly populated. Thus it appears that component 1 has been able to capture the majority of information regarding the molecules. This is also confirmed by the fact that component 1 explains 60% of the variance (out of a total of 69.7%).

A similar analysis is performed with component 2. From the score plot (Figure 8) it is seen that it accounts for some molecules that component 1 underestimated. For example molecules **116** and **143** are predicted correctly as more active, whereas in component 1 they were underestimated. The most highly weighted descriptors in component 2 are NSB and WTPT-2. In contrast to component 1 NSB is now negatively weighted indicating smaller values correlate with higher activities. This component is correcting for larger molecules that might not be active. As a result it moves molecule **175** from its position in the score plot for component 1 to a position closer to the lower left quadrant thus compensating for the overestimation by component 1. As described previously, the WTPT-2 descriptor essentially characterizes the branched nature of a molecule, with the presence of larger rings, balanced branches (in the case of bicyclic molecules), longer chains, and increased substitution (for a given substituent) leading to higher values of the molecular ID number. The molecules in the upper right quadrant of the score plot for component 2 (Figure 8) such as **91** and **92**

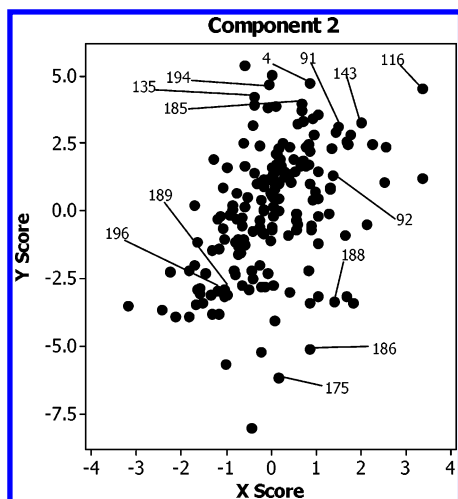


Figure 8. The score plot for component 2.

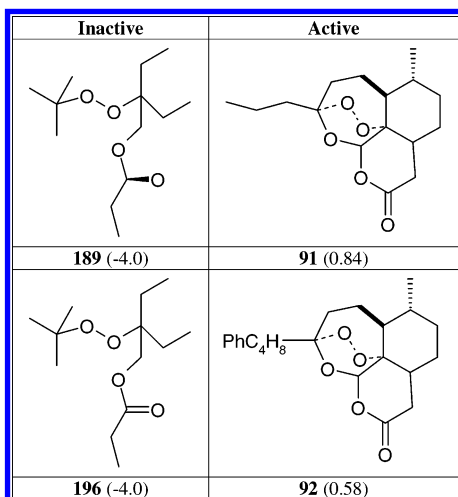


Figure 9. A comparison of the more active and less active compounds described using component 2. The value of log RA is provided within parentheses.

support these trends. They have relatively low values for NSB. They also have higher values of the WTPT-2 descriptor, which can be ascribed to the longer side chains. Both compounds are relatively active. In comparison, compounds **189** and **196** have low values for the WTPT-2 descriptor (which may be due to the absence of side chains) and are predicted as inactive by component 2. The structures of these molecules are compared in Figure 9. However, in general component 2 does not predict the inactive molecules very well (since the lower left quadrant is relatively unpopulated) thereby demonstrating the importance of component 2 in predicting the active compounds.

In component 3 the molecules that were not accounted for by components 1 and 2 are now correctly predicted as active (**135**, **185**, **194**, and **4**) as can be seen from the score plot of component 3 (Figure 10). In addition compound **186** is also more accurately predicted thus correcting for the overestimation by component 2. To some extent component 3 makes up for the over- or underestimations made by components 1 and 2. For this component the most significant descriptors are WTPT-2 and MDE-14. As mentioned previously molecules with a higher number of rings, longer side chains along with more substitution in the cyclic region (especially bridging carbons) will generally have higher values of MDE-14. Taken with the positive sign of the weight

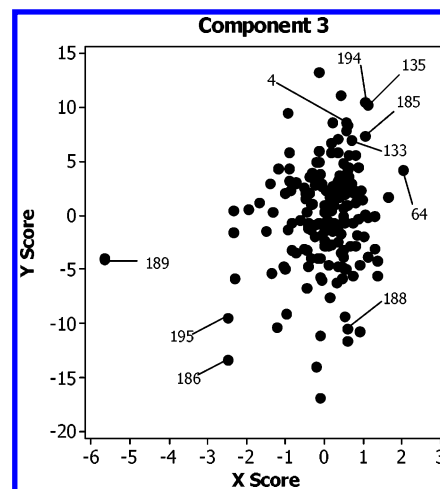


Figure 10. The score plot for component 3.

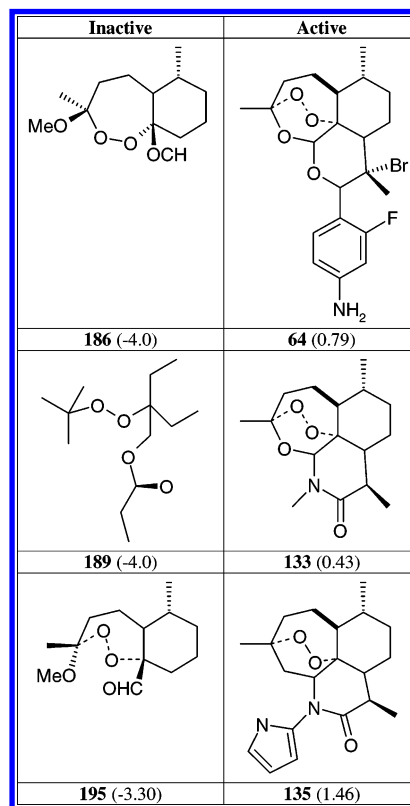


Figure 11. A comparison of the more active and less active compounds described using component 3. The value of log RA is provided within parentheses.

for MDE-14 we may conclude that molecules with extended side chains coupled with substitution in the ring system (essentially, larger molecules) would exhibit higher activities, a trend also seen with NSB in component 1. The molecules present in the upper right (**133**, **135**, and **64**) satisfy these trends and can be seen to have longer side chains as well as increased substitution on the rings compared to the inactive molecules. The molecules with smaller values of MDE-14 (**186**, **189**, and **195**) are predicted as inactive compounds. The structures of these molecules are compared in Figure 11.

It should be noted that a PLS analysis provides a guideline regarding the interpretation of the descriptors in the model and does not provide exact quantitative descriptions of descriptor contributions. Furthermore the analysis is restricted

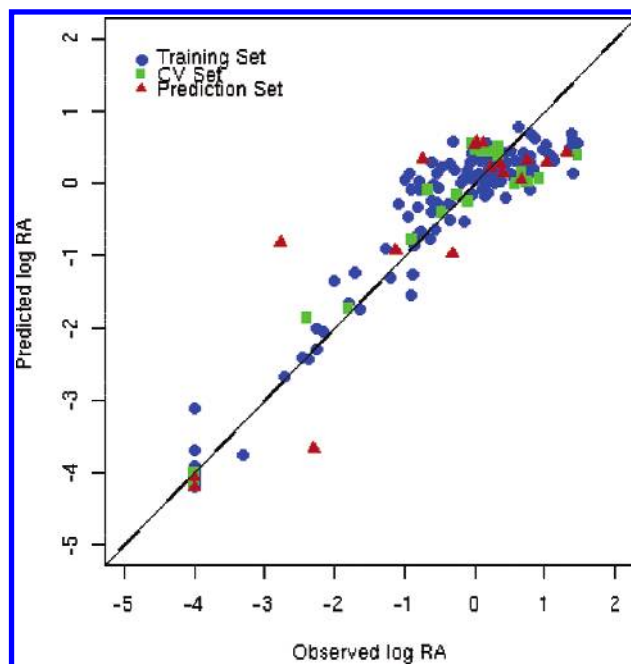
**Table 5.** Summary of the Various Nonlinear CNN Models Generated

architecture	$R^2$			RMSE			cost
	TSET	CV set	PSET	TSET	CV set	PSET	
3-2-1	0.81	0.79	0.81	0.66	0.66	0.70	0.67
7-4-1	0.90	0.89	0.80	0.49	0.48	0.76	0.51
7-5-1	0.91	0.92	0.81	0.47	0.42	0.70	0.50
10-5-1	0.96	0.94	0.88	0.42	0.47	0.76	0.44

to the descriptors present in the best model. In this case, *best* implies best statistical quality and not necessarily the presence of meaningful descriptors.

**Nonlinear Models.** Several of the best nonlinear models selected by the genetic algorithm were analyzed to a greater extent in order to find the best neural network parameters. The four best models were further investigated by systematically varying the network architecture. The results of the best four models are summarized in Table 5. The best model has a 10-5-1 architecture and contains the following descriptors: KAPPA-6,<sup>46-48</sup> NDB, MOMI-4,<sup>26</sup> N7CH,<sup>28,29,31</sup> MOLC-8,<sup>30,49</sup> WTPT-5,<sup>39</sup> MDE-12,<sup>41</sup> MDE-13,<sup>41</sup> ELEC, and FPSA-3.<sup>35</sup> The KAPPA-6 descriptor belongs to a class of descriptors termed Kier shape descriptors,  $^m\kappa$ . These descriptors are defined by the number of vertices and paths of length  $m$  ( $1 < m < 3$ ) in a hydrogen depleted molecular graph. KAPPA-6 is the atom corrected version of the  $^3\kappa$  descriptor and thus accounts for heteroatoms in addition to carbons. The values of the  $^3\kappa$  descriptor are generally larger when molecular branching is absent or when branching occurs at the extremities of a molecular graph, and thus this descriptor characterizes the centrality of branching in a molecule.<sup>50</sup> The NDB is simply the count of double bonds in the molecule. MOMI-4 is the ratio of the X and Y components of the principle moment of inertia of the molecule. Thus, this descriptor provides information about the shape of a molecule in the XY plane. WTPT-5 is a modification of the molecular ID number<sup>39</sup> (described above) described by Randic which only considers nitrogen atoms. The MOLC-8 descriptor is the path cluster-4 molecular connectivity index and measures the degree of branching in a structure. The descriptors MDE-12 and MDE-13 are the molecular distance edge vectors between primary and secondary and primary and tertiary carbons, respectively. The ELEC descriptor is simply the electronegativity of the molecule. The value of electronegativity is taken as the mean of the HOMO and LUMO energies. The FPSA-3 descriptor belongs to a class of hybrid descriptors termed CPSA<sup>35</sup> descriptors. These combine partial charge and surface area information for a molecule resulting in a holistic description of polar surface area features. FPSA-3 is defined as the atom weighted partial positive surface area divided by the total molecular surface area.

At this point we would like to point out that the descriptors selected for the best CNN model are distinct from the descriptors used in the best linear model. The reason underlying this behavior is due to the different selection criteria that are used to include descriptors in the respective models from the reduced pool (which was the same for both linear and nonlinear models). In the case of a linear model, subsets of descriptors that lead to models with high values of the  $t$  statistic are preferentially selected. In the case of CNN models, descriptor subsets that lead to minimization of the cost function described previously are preferentially

**Figure 12.** A plot of observed versus predicted log RA produced from the best nonlinear CNN model using a 10-5-1 architecture.

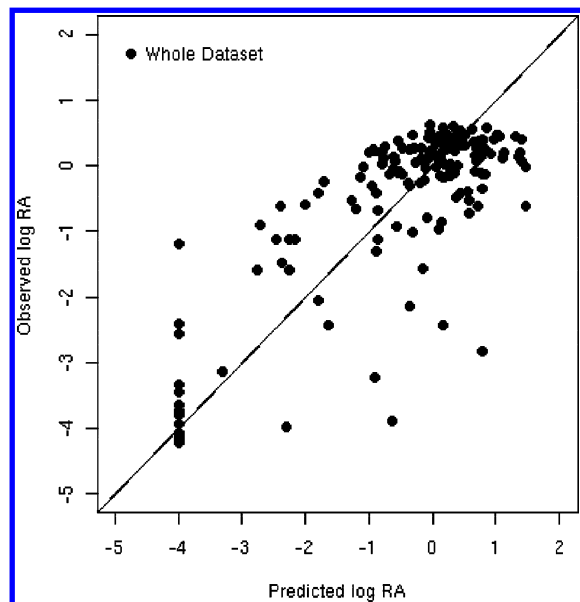
selected. At the same time one must consider that fact that different combinations of descriptors may be equally valid in describing a SAR trend. Furthermore, since a linear model is, by definition, restricted to capturing linear relationships, it cannot be used to investigate subsets of descriptors which when combined nonlinearly might better describe a SAR trend. This is evidenced by using the descriptors from the best CNN model in a linear model. The resultant residuals are very high, and the predictive power of such a model is very poor and as a result a linear model would not have considered this subset of descriptors.

The plot of observed versus calculated values is shown in Figure 12. The RMSE values for the training, CV, and prediction sets were 0.42, 0.47, and 0.76, respectively. It is important to note that the group of compounds with log RA's of -4.0 have been predicted relatively well. The  $R^2$  values for the training, CV, and prediction sets were 0.96, 0.94, and 0.88, respectively. To ensure that the behavior of the model was independent of the composition of the training, CV, and prediction sets the nonlinear model described above was regenerated using a leave  $n\%$  out procedure. In this procedure the molecules are ranked according to the values of their activity and then grouped, the number of groups being determined from the percentage left out. Next, an equal number of empty groups are populated by selecting molecules from each group of the ranked data set such that the whole range of activity is evenly distributed across the groups. These groups are then used to create the training, CV, and prediction sets. Essentially for  $n$  groups, the first group is the prediction set, the second group is the CV set, and the remaining  $n-2$  groups constitute the training set. These sets were then used to build and validate a CNN with a 10-5-1 architecture. In the next step, the last group was made the prediction set, the first group was the CV set, and the remaining groups constitute the new training set. The CNN was rebuilt and validated using these sets. This process is repeated such that each group acts as the prediction set once. As a result the entire data set is predicted once. The



**Table 6.** Summary of the Statistics Generated by a 3 Round Leave 14% out Procedure Using the Best Nonlinear CNN Model (10–5–1 Architecture)

	RMSE			$R^2$		
	training set	CV set	prediction set	training set	CV set	prediction set
mean	0.44	0.59	0.89	0.91	0.85	0.69
SD	0.05	0.10	0.16	0.01	0.06	0.11

**Figure 13.** A plot showing the predicted versus observed log RA values for the whole data set using the best nonlinear CNN model (10–5–1 architecture). This result was obtained by a leave 14% out procedure which was run 3 times giving 3 predictions for each member of the data set. The average value was taken as the final predicted value.

whole process is repeated so that each member of the data set is predicted multiple times, and the final reported value for a molecule is the average of all the predictions for that molecule.

In this study a leave 14% out procedure was used and repeated 3 times. Thus each molecule in the data set was predicted three times, and the final reported value was the average of these predictions. The results of this procedure are summarized in Table 6, and a plot of the observed versus predicted values can be seen in Figure 13. As can be seen the RMSE values for the cross-validation and prediction have degraded compared to the original CNN model. Similar behavior is seen for the  $R^2$  values. However, the standard deviations for these values over all the runs is quite low for the training and CV sets indicating that the model trains consistently. However when comparing these results to the original CNN model, only the statistics for the prediction set should be compared. The degradation of RMSE and  $R^2$  values of the prediction set is to be expected as the model is trained on different sets of molecules at each stage. The leave 14% out procedure used here gives us a more realistic view of the behavior of this model when biases due to QSAR set composition are removed.

To further investigate the effect of QSAR set composition, a CNN model with a 10–5–1 architecture was generated using random sets (i.e., the training, CV, and prediction sets were selected randomly from the data set). The results of this model are shown in Table 7. Though the RMSE's and

**Table 7.** Results of a Nonlinear CNN Model (10–5–1 Architecture) Using Randomly Generated Training, Cross-Validation, and Prediction Sets<sup>a</sup>

RMSE			$R^2$		
TSET	CV set	PSET	TSET	CV set	PSET
0.41	0.53	0.68	0.93	0.91	0.81

<sup>a</sup> The descriptors used were the same as those for the best nonlinear CNN model.

**Table 8.** Results of a Nonlinear CNN Model (10–5–1 Architecture) Using a Scrambled Dependent Variable<sup>a</sup>

RMSE			$R^2$		
TSET	CV set	PSET	TSET	CV set	PSET
1.50	1.40	1.60	0.09	0.08	0.01

<sup>a</sup> The descriptors used were the same as those for the best nonlinear CNN model.

$R^2$  values for the training and CV sets are similar to the average values for the leave 14% out based model, the prediction set performance is significantly degraded. This observation could simply be explained by considering that a poor combination of QSAR sets was created. However, an alternative explanation is that due to random set generation, the full range of activities is not properly represented in the training, CV, and prediction sets.

Finally to test whether the results described above could have been due to chance, Monte Carlo runs were carried out in which the dependent variable is scrambled and models are built using the scrambled dependent variable. The architecture was maintained at 10–5–1, and the descriptors used were those found in the best nonlinear model above. As can be seen from Table 8, the RMSE errors increase significantly with a corresponding decrease in the  $R^2$  values. This appears to indicate chance correlations did not play a significant role in the results described above.

## DISCUSSION AND CONCLUSION

This work presents both linear and nonlinear models to predict antimalarial activity for a set of 179 artemisinin analogues. The goal of the project was to create QSAR models, which were both interpretable as well as having good predictive ability. The linear regression model was found to be statistically valid, and the PLS routine enabled an investigation of the effects of each descriptor in the model. That is, it was possible to isolate the action of the individual descriptors and explain specific SAR trends captured by the descriptors (interpretability).

The nonlinear models were developed based mainly on their pure predictive ability. The nonlinear model presented both superior predictive ability as well as a relatively simple neural network architecture. Interpretation of neural network models is difficult due to the black box nature of the neural network algorithm. Methods exist for a probabilistic interpretation of neural network classification models.<sup>51,52</sup> Techniques also exist to extract rules and decision trees from CNN regression models.<sup>53,54</sup> However these methods do not allow for a clear interpretation of descriptor contributions, as is available from a PLS analysis of a linear model and are not easily combined with the ADAPT methodology. Finally



randomization tests showed that the possibility of chance correlations (if any) in the best models was low.

It may be noted that in both types of models the descriptors themselves are not necessarily amenable to simple physical interpretation. The ADAPT methodology seeks the most *information rich* subset of descriptors for a given model. In many cases the members of the resultant subset do not have simple physical meaning but rather contribute information to the statistical model. That is, many descriptors calculated by ADAPT are not designed to necessarily provide a simple physical description of a molecule. Instead they extract information that in many cases may be of a more abstract (such as graph theoretical) nature but provide *information* about a molecule. An attempt was made to introduce more meaningful descriptors into the models by replacing some of the selected descriptors with other correlated (and physically meaningful) descriptors from the reduced pool. In all cases, the resultant models performed poorly in comparison to the best models reported in this work. Clearly, the inclusion of physically meaningful descriptors does not necessarily lead to a better model.

A direct comparison with the original work is not feasible as the model development process in this study was different. However the results of the PLS analysis indicate that in terms of  $q^2$ , the current linear model performs comparably to the original PLS model using 157 docked compounds described by Avery.<sup>10</sup> The PLS analysis was also able to provide an interpretation of the contributions of the individual descriptors in describing the overall activity of the majority of the molecules. One aspect of model interpretability would be a ranking of descriptor contributions. However the PLS technique does not allow a global ranking of individual descriptors since each PLS component is a linear combination of all the descriptors in the model. Thus such a ranking of descriptor contributions is only valid within a given component. One disadvantage of the current methodology was the inability to consider enantiomeric pairs compared to the original work in which the CoMFA methodology was able to handle such pairs. At the same time the current methodology does not involve the problem of alignments inherent in the CoMFA approach and furthermore was able to avoid making any assumptions regarding bioactive conformations.

Finally, though the linear model in this work does not exhibit significant predictive ability when compared to the models described by Avery,<sup>10</sup> it does provide interpretability. Coupled with the good predictive ability of the neural network model developed in this study we believe that these models would perform well as rapid screening tools to uncover new and more potent antimalarial drugs.

## REFERENCES AND NOTES

- (1) Haynes, R. K.; Vonwiller, S. C. From Qinghao, Marvelous Herb of Antiquity, to the Antimalarial Trioxane Qinghaosu — and Some Remarkable New Chemistry. *Acc. Chem. Res.* **1997**, *30*, 73–79.
- (2) Klayman, D. L. *Science* **1985**, *228*, 1049.
- (3) Kamchonwongpaisan, S.; Meshnick, S. R. The Mode of Action of the Antimalarial Artemisinin and its Derivatives. *Gen. Pharmac.* **1996**, *27*, 587–592.
- (4) Posner, G. H.; Cumming, J. N.; Ploypradith, P.; Oh, C. H. Evidence for Fe(IV) O in the Molecular Mechanism of Action of the Trioxane Antimalarial Artemisinin. *J. Am. Chem. Soc.* **1995**, *117*, 5885–5886.
- (5) Posner, G. H.; Park, S. B.; Gonzalez, L.; Wang, D.; Cumming, J. N.; Klindinst, D.; Shapiro, T. A.; Bachi, M. D. Evidence for the Importance of High Valent Fe O and of a Diketone in the Molecular Mechanism of Action of Antimalarial Trioxane Analogues of Artemisinin. *J. Am. Chem. Soc.* **1996**, *118*, 3537–3538.
- (6) Robert, A.; Meunier, B. Is alkylation the main mechanism of action of the antimalarial drug artemisinin? *Chem. Soc. Rev.* **1998**, *27*, 273–274.
- (7) Avery, M. A.; McLean, G.; Edwards, G.; Ager, A. Structure activity relationships of peroxide based artemisinin antimalarials. *Biol. Act. Nat. Prod.* **2000**, 121–132.
- (8) Woolfrey, J. R.; Avery, M. A.; Doweiko, A. M. Comparison of 3D quantitative structure–activity relationship methods: analysis of the in vitro antimalarial activity of 154 artemisinin analogues by hypothetical active site lattice and comparative molecular field analysis. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 165–181.
- (9) Avery, M. A.; Alvim-Gaston, M.; Woolfrey, J. R. Synthesis and structure–activity relationships of peroxidic antimalarials based on artemisinin. *Adv. Med. Chem.* **1999**, *4*, 125–217.
- (10) Avery, M. A.; Alvim-Gaston, M.; Rodrigues, C. R.; Barreiro, E. J.; Cohen, F. E.; Sabnis, Y. A.; Woolfrey, J. R. Structure Activity Relationships of the Antimalarial Agent Artemisinin. The Development of Predictive In Vitro Potency Models Using CoMFA and HQSAR Methodologies. *J. Med. Chem.* **2002**, *45*, 292–303.
- (11) Tommuphean, S.; Kokpol, S.; Parasuk, V.; Wolschann, P.; Winger, R. H.; Liedl, K. R.; Rode, B. M. Comparative molecular field analysis of artemisinin derivatives: ab initio versus semiempirical optimized structures. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 397–409.
- (12) Avery, M. A.; Gao, F.; Wesley, C. K. M.; Mehrotra, S.; Milhous, W. K. Structure–activity relationships of the antimalarial agent artemisinin. 1. Synthesis and comparative molecular field analysis of C-9 analogues of artemisinin and 10-deoxoartemisinin. *J. Med. Chem.* **1993**, *36*, 4264–4275.
- (13) Cheng, F.; Shen, J.; Luo, X.; Zhu, W.; Gu, J.; Ji, R.; Jiang, H.; Chen, K. Molecular docking and 3-D-QSAR studies on the possible antimalarial mechanism of artemisinin analogues. *Bioorg. Med. Chem.* **2002**, *10*, 2883–2891.
- (14) Girones, X.; Gallegos, A.; Carbo-Dorca, R. Modeling Antimalarial Activity: Application of Kinetic Energy Density Quantum Similarity Measures as Descriptors in QSAR. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1400–1407.
- (15) Tommuphean, S.; Parasuk, V.; Kokpol, S. QSAR Study of Antimalarial Activities and Artemisinin-Heme Binding Properties Obtained from Docking Calculations. *Quant. Struct.-Act. Relat.* **2000**, *19*, 475–483.
- (16) Cramer III, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). I. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.
- (17) Cramer III, R. D.; Patterson, D. E.; Bunce, J. D.; Frank, I. E. Crossvalidation, Bootstrapping and Partial Least Squares Compared with Multiple Regression in Conventional QSAR Studies. *Quant. Struct.-Act. Relat. Pharmacol., Chem. Biol.* **1988**, *7*, 18–25.
- (18) Jurs, P. C.; Chou, J. T.; Yuan, M. Studies of Chemical Structure Biological Activity Relations Using Pattern Recognition. In *Computer Assisted Drug Design*; Olsen, E. C., Christoffersen, R. E., Eds.; American Chemical Society: Washington, DC, 1979.
- (19) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer Assisted Studies of Chemical Structure and Biological Function*; Wiley: New York, 1979.
- (20) Bakken, G. A.; Jurs, P. C. Classification of Multidrug-Resistance Reversal Agents Using Structure-Based Descriptors and Linear Discriminant Analysis. *J. Med. Chem.* **2000**, *43*, 4534–4541.
- (21) Kauffman, G. W.; Jurs, P. C. QSAR and *k*-Nearest Neighbor Classification Analysis of Selective Cyclooxygenase-2 Inhibitors Using Topologically-Based Numerical Descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1553–1560.
- (22) Patankar, S. J.; Jurs, P. C. Prediction of IC<sub>50</sub> Values for ACAT Inhibitors from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 706–723.
- (23) Stanton, D. T.; Murray, W. J.; Jurs, P. C. Comparison of QSAR and Molecular Similarity Approaches for a Structure–Activity Relationship Study of DHFR Inhibitors. *Quant. Struct.-Act. Relat.* **1993**, *12*, 239–245.
- (24) Mattioni, B. C.; Kauffman, G. W.; Jurs, P. C.; Custer, L. L.; Durham, S. K.; Pearl, G. M. Predicting the Genotoxicity of Secondary and Aromatic Amines Using Data Subsetting to Generate a Model Ensemble. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 949–963.
- (25) Mattioni, B. E.; Jurs, P. C. Prediction of Dihydrofolate Reductase Inhibition and Selectivity Using Computational Neural Networks and Linear Discriminant Analysis. *J. Mol. Graph. Model.* **2003**, *21*, 391–419.
- (26) Goldstein, H. *Classical Mechanics*; Addison-Wesley: Reading, MA, 1950.
- (27) Pearlman, R. S. Molecular Surface Areas and Volumes and Their Use in Structure/Activity Relationships. In *Physical Chemical Properties*

- of Drugs; Yalkowsky, S. H., Sinkula, A. A., Valvani, S. C., Eds.; Marcel Dekker: New York, 1980.
- (28) Kier, L. B.; Hall, L. H.; Murray, W. J. Molecular connectivity I: Relationship to local anesthesia. *J. Pharm. Sci.* **1975**, *64*.
- (29) Kier, L. B.; Hall, L. H. Molecular connectivity VII: Specific treatment to heteroatoms. *J. Pharm. Sci.* **1976**, *65*, 1806–1809.
- (30) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.
- (31) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure Activity Analysis*; John Wiley & Sons: 1986.
- (32) Kier, L. B.; Hall, L. H. An Electrotopological-State Index for Atoms in Molecules. *Pharm. Res.* **1990**, *7*, 801–807.
- (33) Kier, L. B.; Hall, L. H. *Molecular Structure Description. The Electrotopological State*; Academic Press: London, 1999.
- (34) Vinogradov, S. N.; Linnell, R. H. *Hydrogen Bonding*; Van Nostrand Reinhold: New York, 1971.
- (35) Stanton, D. T.; Jurs, P. C. Development and Use of Charged Partial Surface Area Structural Descriptors in Computer Assisted Quantitative Structure Property Relationship Studies. *Anal. Chem.* **1990**, *62*, 2323–2329.
- (36) Mattioni, B. E. The Development of QSAR Models for Physical Property And Biological Activity Prediction of Organic Compounds. Ph.D. Chemistry, The Pennsylvania State University, University Park, 2003.
- (37) Sutter, J. M.; Dixon, S. L.; Jurs, P. C. Automated Descriptor Selection For Quantitative Structure–Activity Relationships Using Generalized Simulated Annealing. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 77–84.
- (38) Goldberg, D. E. *Genetic Algorithms in Search Optimization & Machine Learning*; Addison-Wesley: Reading, MA, 2000.
- (39) Randic, M. On molecular identification numbers. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 164–175.
- (40) Randic, M. Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1972**, *97*, 6609.
- (41) Liu, S.; Cao, C.; Li, Z. Approach to estimation and prediction for normal boiling point (nbp) of alkanes based on a novel molecular distance edge (mde) vector, lambda. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 387–394.
- (42) Rousseeuw, P. J.; Leroy, A. M. *Robust Regression and Outlier Detection*; Wiley Series in Probability and Mathematical Statistics; John Wiley & Sons: 1987.
- (43) R Development Core Team 2003: *R: A Language and Environment for Statistical Computing*. <http://www.R-project.org>.
- (44) Stanton, D. T. On The Physical Interpretation of QSAR Models. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1423–1433.
- (45) *Minitab*, v. 14; Minitab Minitab Inc.: State College, PA.
- (46) Kier, L. B. A Shape Index From Molecular Graphs. *Quant. Struct.-Act. Relat. Pharmacol., Chem. Biol.* **1985**, *4*, 109–116.
- (47) Kier, L. B. Shape Indices for Orders One and Three From Molecular Graphs. *Quant. Struct.-Act. Relat. Pharmacol., Chem. Biol.* **1986**, *5*, 1–7.
- (48) Kier, L. B. Distinguishing Atom Differences in a Molecular Graph Index. *Quant. Struct.-Act. Relat. Pharmacol., Chem. Biol.* **1986**, *5*.
- (49) Balaban, A. T. Highly Discriminating Distance Based Topological Index. *Chem. Phys. Lett.* **1982**, *89*, 399–404.
- (50) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Mannhold, R., Kubinyi, H., Timmerman, H., Eds.; Methods and Principles in Medicinal Chemistry; Weinheim, 2000.
- (51) Gupta, A.; Park, S.; Sluwa, M. L. Generalized Analytic Rule Extraction for Feedforward Neural Networks. *IEEE Trans. Know. Data. Eng.* **1999**, *11*, 985–991.
- (52) Ney, H. On the Probabilistic Interpretation of Neural Network Classifiers and Discriminative Training Criteria. *IEEE Trans. Pat. Anal. Mach. Intell.* **1995**, *17*, 107–119.
- (53) Schmitz, G. P. J.; Aldrich, C.; Gouws, F. S. ANN-DT: An Algorithm for Extraction of Decision Trees from Artificial Neural Networks. *IEEE Trans. Neural. Net.* **1999**, *10*, 1392–1401.
- (54) Setiono, R.; Leow, W. K.; Zurada, J. M. *Extraction of Rules from Artificial Neural Networks for Nonlinear regression*; CiteSeer. [citeseer.nj.nec.com/494713.htm](http://citeseer.nj.nec.com/494713.htm).

CI0499469