A Fuzzy ARTMAP-Based Quantitative Structure−Property Relationship (QSPR) for Predicting Physical Properties of Organic Compounds

Gabriela Espinosa,[†] Denise Yaffe,[‡] Alex Arenas,[§] Yoram Cohen,[‡] and Francesc Giralt*[,†]

Departament d'Enginyeria Química and Departament d'Enginyeria Informàtica i Matemàtica, Escola Tècnica Superior d'Enginyeria Química, Universitat Rovira i Virgili, Carretera de Salou, s/n 43006 Tarragona, Catalunya, Spain, and Department of Chemical Engineering, University of California, Los Angeles, California 90095-1592

A modified fuzzy ARTMAP neural-network-based QSPR for predicting normal boiling points, critical temperatures, and critical pressures of organic compounds was developed. Seven or eight molecular descriptors (the sum of atomic numbers; five valence connectivity indices; and the second-order kappa shape index, without or with the dipole moment) were used to describe the topological and electronic features of a heterogeneous set of 1168 organic compounds. Optimal training and testing sets were selected with fuzzy ART. The fuzzy ARTMAP models with eight descriptors as input provided the best predictive and extrapolation capabilities compared to optimal back-propagation models and group contribution methods. The absolute mean errors of predictions for the normal boiling point (1168 compounds), the critical temperature (530 compounds), and the critical pressure (463 compounds) were 2.0 K (0.49%), 1.4 K (0.24%), and 0.02 MPa (0.52%), respectively. A composite model for simultaneously estimating the three properties yielded similar results.

Introduction

Physical properties and thermodynamic data of organic compounds are needed for the design and operation of industrial chemical processes. Knowledge of basic parameters, such as phase transition temperatures and critical properties, is crucial in petrochemical processes, distillation, waste treatment, and environmental cleanup. However, experimental data and correlation methods for the above parameters are often limited.[1,2] Recent approaches to the estimation of physical properties have been based on quantitative structure−property relationship (QSPR) methods and artificial neural networks (ANNs),[3−14] and recent review of such methods can be found elsewhere.[15]

QSPR techniques assume that correlations exist between physicochemical properties and molecular structure. Accordingly, the QSPR approach attempts to establish simple mathematical relationships to describe the correlation of a given property for a set of compounds. QSPR analyses involve the selection of molecular descriptors to satisfactorily characterize different sets of compounds and the application of algorithms, such as partial least-squares or artificial neural networks, to build the QSPR model.[3−6,8−28,29]

The challenge in developing QSPRs is the design and/or selection of molecular descriptors that strongly correlate with the desired molecular property. QSPR development typically involves the selection of the dominant indices within a large pool of available molecular descriptors.[9−12,14−16,18−22,27] As the number of descriptors increases, the capability of regression analysis methods decreases because of the redundancy of information incorporated by the different descriptors. Techniques to minimize the problem, such as principal component analysis and partial least-squares regression, have been used in QSPR development. Notwithstanding, these methods require a priori knowledge of the analytical form of the QSPR model. As an alternative, artificial neural networks (ANNs) have gained popularity in recent years as a technique for developing quantitative structure−property relationship (QSPR) models. The advantage of ANNs over the regression analysis methods is their inherent ability to incorporate nonlinear relationships among chemical structures and physical properties.[10−12,30]

Various neural-network-based QSPR models have been reported in the literature for estimating boiling points of organic compounds.[3,4,6,9−13,16,17,19,22,31] For example, Hall and Story[6] reported a back-propagation neural network model for boiling points using atom-type E-state indices for the 19 atom types present in the data set of 298 organic compounds that they used. For their 19−5−1 neural network architecture, the absolute mean errors of the predictions were 4.57 K for the test set and 3.93 K for the overall data set. It should be noted that Hall and Kier[17] were the first to use the atom-type E-state indices to develop a regression-based QSPR model to predict the boiling points of 245 alcohols and alkanes with a standard deviation of 8 K over a temperature range of about 282−504 K. A variety of other QSPR/ANN models have also been developed by Jurs and co-workers[10−12,18] to estimate the boiling points of heterogeneous data sets of organic compounds, such

* Author to whom correspondence should be addressed. Tel.: +34977559638. Fax: +34977559621. E-mail: fgiralt@etseq.urv.es.

[†] Departament d'Enginyeria Química, Universitat Rovira i Virgili.

[‡] University of California, Los Angeles.

[§] Departament d'Enginyeria Informàtica i Matemàtica, Universitat Rovira i Virgili.

10.1021/ie0008068 CCC: $20.00
Published on Web 05/16/2001

as pyrans, pyrroles, furans, tetrahydrofurans, and thiophenes. In these models, chemical input descriptors were selected from a large pool of molecular descriptors using regression analysis[16] or artificial neural networks.[10−12,14,16,19−21]

Diverse methods have also been reported for the estimation of critical properties, which are of particular interest in the present study. One common approach has been to use other physical properties such as boiling points, parachor, and molar refraction as input information. Input descriptors have also used molecular descriptors that integrate structural information. For example, Egolf et al.[3,16] introduced the charge partial surface area (CPSA) descriptor to correlate the critical temperatures of 147 heterogeneous molecules. They fitted the data with a standard deviation of 11.9 K using eight descriptors. In a later study, Hall and Story[6] reduced the standard deviation for the above critical temperature correlation to 6.61 K for the same set of 147 compounds by employing 19 atom types of electrotopological state indices. Katriztky and Mu[20] obtained a standard deviation of 14.9 K for the same set of organic compounds using a set of molecular indices that included the cubic root of the gravitation index, the area-weighted surface charge of the hydrogen-bonding donor atoms, and the topographic electronic index.

The estimation of critical pressures on the basis of a back-propagation neural network model correlation was developed by Turner et al.[14] using a data set of 165 compounds (alcohols, ketones, esters, carboxylic acids, aldehydes, phenols, ethers, nitriles, and amines). The input descriptors for the critical temperature model included three topological descriptors (simple path 3 connectivity index, number of oxygens, and the secondary sp^3 carbons), two electronic descriptors (dipole moment and average charge on positively charge atoms), one charge partial surface area descriptor (total-charge-weighted partial positive surface area), and one hydrogen-bonding descriptor. The best back-propagation architecture based on a quasi-Newton algorithm was 8−4−1, which yielded root-mean-square errors of 7.33 K for the training set, 7.76 K for the validation set, and 9.85 K for an external set. For the critical pressure model, Turner et al.[14] used a set of input descriptors that included two topological descriptors (winner number and number of length-3 paths), one electronic descriptor (the charge on the most negative atom), two geometric indices (second major moment and first major moment of inertia), two charged partial surface area descriptors (partial positive surface area and weighted partial negative surface area), and one hydrogen-bonding descriptor (sum of surface area of acceptor). The QSPR model based on the best back-propagation architecture of 8−5−1 yielded predictions with reported root-mean-square errors of 1.41 atm for the training set, 1.35 atm for the validation set, and 2.39 atm for an external set.

A new approach for developing QSPR/ANN models was recently presented by Espinosa et al.[4] based on the cognitive classifier fuzzy ARTMAP. The approach, which was demonstrated for estimating the boiling points of aliphatic hydrocarbons, was shown to be superior to QSPRs obtained using a back-propagation neural network approach or other statistical correlations reported in the literature. The simple set of molecular descriptors used in the above study included four valence molecular connectivity indices ($^1\chi^v$, $^2\chi^v$, $^3\chi^v$, $^4\chi^v$), the second kappa shape index ($^2\kappa$), the dipole moment, and the molecular weight. The inclusion of the dipole moment proved to be particularly useful in distinguishing cis from trans isomers. It is interesting to note that, for a heterogeneous set of organic compounds, the prediction of boiling points with back-propagation or statistically based QSPR models does not improve when the input pattern includes thee-dimensional information (kappa index and dipole moment)[31] in addition to topological information (valence connectivities). This indicates the need to shift to fuzzy ARTMAP/QSPR models to take advantage of their associative memory and ability to classify and analyze noisy information with fuzzy logic and to avoid the plasticity−stability dilemma[32−37] of standard back-propagation architectures.

The purpose of the current study is to investigate the potential applicability of fuzzy ARTMAP/QSPR models to estimations of critical properties and boiling point temperatures for a heterogeneous set of compounds with a relatively simple set of molecular descriptors. The molecular descriptors include five valence molecular connectivity indices;[27,28] a second-order kappa shape index; the dipole moment; and the sum of atomic numbers, rather than the molecular weight used in a previous work.[4] The training and testing sets were built with a fuzzy ART classifier. The performance of the current model is compared with that of group contribution methods and other QSPR/ANN models. The predictive capabilities of the fuzzy ARTMAP/QSPR model are also illustrated with a composite model for the simultaneous estimation of the three properties considered.

Methodology

Data Set and Molecular Descriptors. The normal boiling points and critical temperatures and pressures for the 1168 compounds considered in this study were compiled from the literature.[38−40] This heterogeneous set includes normal and branched alkanes, branched and unbranched substituted alkanes, anilines, pyridines, aldehydes, amines, ketones, and aromatic compounds. The complete set was divided into four subsets, one for each of the properties T_b, T_c, and P_c, and another for the compounds with available information on all three properties. The list of all 1168 compounds considered is given in Table 1 of the Supporting Information, along with their physical properties and estimated values. This table also identifies with "tr" and "te" the compounds used for training or testing, respectively. The boiling points range from 111.6 to 771 K, the critical temperatures of 530 organic compounds from 190.5 to 926 K, and the critical pressures of 465 compounds from 8.95 to 1.02 MPa. The subset of 436 compounds common to the three data sets, i.e., compounds with information on all three of the properties considered, was used to develop a more global model and to assess the capabilities of the fuzzy ARTMAP system.

The molecular descriptors for each compound in included this study were calculated from knowledge of the chemical structure using Molecular Modeling Pro 3.01 (ChemSW Software Inc.). The first set of seven descriptors presented as input information to the network included topological information only and is labeled "tset". The tset descriptors consisted of five valence connectivity indices ($^0\chi$, $^1\chi^v$, $^2\chi^v$, $^3\chi^v$, $^4\chi^v$), the second-order kappa shape index,[26] and the sum of atomic numbers. The connectivity indices, which are based on local properties, are bond-additive quantities,

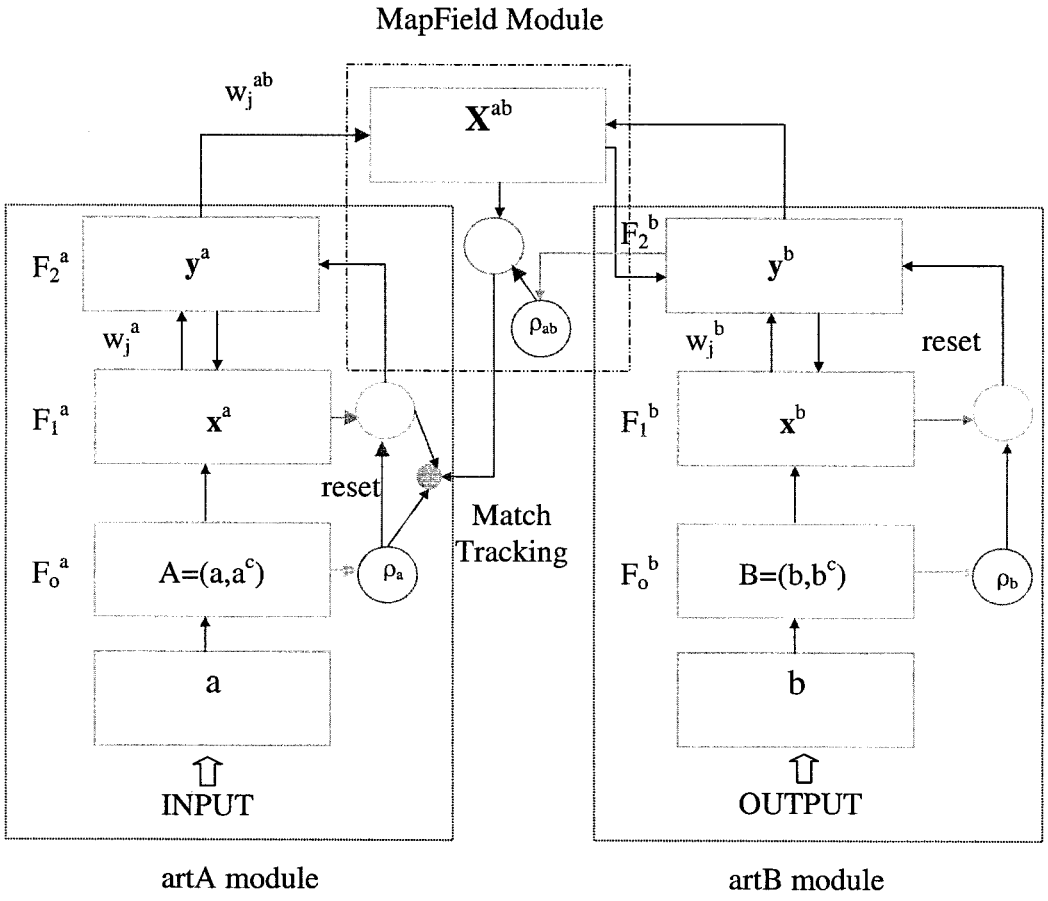

Figure 1. Block diagram of the fuzzy ARTMAP neural network architecture.

in which bonds of different kinds (primary, secondary, and tertiary carbon atoms) make different contributions to the index value. Substructures for a molecular skeleton are defined by the decomposition of the skeleton into fragments (hydrogen-suppressed). The connectivity index of each order is the valence-weighted counts of the substructure fragments. Therefore, features such as size, branching, unsaturation, heteroatom content, and cyclicity are encoded. The kappa-2 shape index describes the shape of the molecular graphs in terms of the related star graphs. Specifically, it counts paths of length two in the linear skeleton. The sum of atomic numbers completes the topological description of each compound. It is obvious that additional descriptors are needed to cover the increased variety of structural features that are introduced by the presence of heteroatoms. Although approaches that make use of a large set of complex topological descriptors will typically increase the level of molecular characterization,[3,12−21] it is necessary in most cases to take into account spatial interaction across the molecule. Therefore, as suggested by the recent study of Espinosa et al.,[4] to improve the ability to better distinguish among molecular structures, we included the dipole moment in this second set of eight descriptors (wset).

The Fuzzy ARTMAP Neural System. A fuzzy ARTMAP neural system is capable of carrying out fast but stable online recognition, learning, hypothesis testing, and prediction of rare events or morphologically variable types of events. Moreover, such a system can adaptively establish many-to-one and one-to-many relationships in response to an arbitrary stream of analog or binary input patterns. A fuzzy ARTMAP combines a unique set of computational abilities that are needed to function autonomously in a changing world in which alternative models have not yet been achieved, i.e., it is capable of self-organizing stable recognition categories in response to arbitrary sequences of analog input patterns. It also offers a unique solution to the stability−plasticity dilemma that many popular autonomous learning systems, such as back-propagation neural networks, face when required long training causes either the loss of previous information or huge unstable networks.[32−36]

The fuzzy ARTMAP neural network developed in the present study consists of two fuzzy ART modules, artA and artB, as well as an inter-ART module, as shown in Figure 1. The fuzzy ART architecture was designed by Carpenter et al.[35] as a classifier for multidimensional data clustering according to a set of features. Module artA learns how to categorize the input patterns (molecular descriptors) presented to layer Fo^a with a vigilance parameter ρ_a, while artB develops categories (families) of the target patterns (physical properties) presented to layer Fo^b with a vigilance parameter ρ_b. The two modules work together and are linked by the map field module or internal controller of Figure 1, which controls the number of classes being created by associative learning. The controller is designed to create the minimum number of categories or hidden units needed to meet the accuracy criteria. When the molecular descriptors are presented, the artA module attempts the prediction through the map field of the category to which the current target belongs. If the prediction is correct, both modules learn their respective inputs by modifying the prototypes of the corresponding category. Otherwise, a match-tracking process starts. In this process, the vigilance parameter of the artA

module is raised by the minimum amount needed to cause mismatch with the current input. This mismatch subsequently triggers a search for a better new category. The process continues until the network either finds an artA category that predicts the category of the current target correctly or creates a new one and the corresponding map field, with the correctly learned current descriptor−physical property pair.

The fuzzy ARTMAP system was modified by Giralt et al.[37] to convert this neural classifier into a predictive system. In the current study, this modified cognitive architecture is used to establish QSPR models capable of predicting physical properties as in previous back-propagation models. A fuzzy ARTMAP incorporates predictive feedback to control the hypothesis-testing cycle and, thus, embodies characteristics of a self-organizing production system that is also goal-oriented, i.e., it is able to compress different sorts of information into distinct categories that can all be used to make the same prediction.

The expertise of the fuzzy ARTMAP system can be inferred by studying the rules that it follows to make predictions. For example, there can be multiple ways in which compounds with similar molecular descriptors (input vector) can be related to the same physical property (output). At any point in the learning process, the operator of the ARTMAP system can test how many of the recognized classes are related to a given value of the physical property. Within each recognition category, the prototype characterizes a particular rule or bundle of molecular descriptors that predicts this value of the physical property. Predictions are of an "if−then" nature, e.g., *if* the molecular structure has features close enough to a particular prototype, *then* it predicts the desired outcome. Many such rules coexist without mutual interference because of the competitive interactions whereby each hypothesis is compressed. Associative networks such as back-propagation often mix multiple rules because they do not have the competitive dynamics to separate them. The particular rule-based system of fuzzy ARTMAPs can also exhibit aspects of "creativity" because, although it is "supervised", it does not use the correct answers to directly force changes in its weights, as do supervised systems such as back-propagation. When answers are wrong for a given state of knowledge, a fuzzy ARTMAP tests new hypotheses until it discovers, on its own, new representations that are capable of predicting the correct answers.

The procedure followed in the current work to develop the back-propagation- and fuzzy-ARTMAP-based QSPRs for the prediction of boiling points, critical temperatures, and critical pressures (separately or together) was as follows. First, the data for each property were distributed between a train set and a test set following the methodology described in Figure 2. About 85% of the compounds in the entire set were selected for training by the fuzzy ART classifier to ensure that adequate information was provided to the system. The compounds were then classified on the basis of their molecular descriptors, either tset or wset, and the target physical property, which were all presented to fuzzy ART as an input vector. The selection of the training set by fuzzy ART yielded the lowest errors in the prediction of the test set compared to random partitioning of the complete data set for the two models considered. This is so because the clustering made by fuzzy ART allows the selection of a percentage of representative data, i.e., significant and redundant information per cluster.

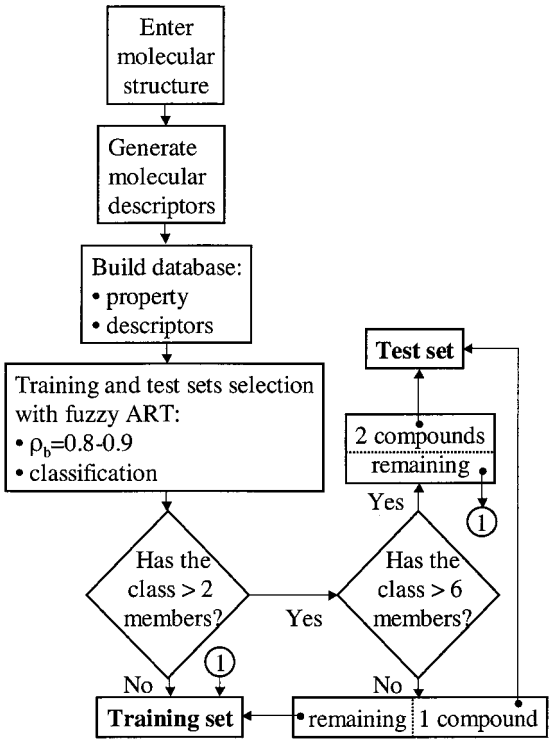

Figure 2. Flow diagram for the selection of training and test sets based on structural information using the fuzzy ART neural network.

Training of the fuzzy ARTMAP consisted of the presentation of the molecular descriptors and target property of the training set to modules artA and artB, respectively, to establish the input with the output categories and relate the two (F_1^{ab}). After training was completed, the hypothesis components of the artB module in Figure 1 (F_o^b and F_1^b) were disconnected, and an output in its category layer (F_2^b) was implemented;[35] the intent was to provide, through the map field module F_1^{ab}, a prediction for the target physical property for any input of descriptors presented to module artA. Additional information about fuzzy ART and fuzzy ARTMAP systems can be found elsewhere.[41−46]

Group Contribution Methods. To assess the performance of the present fuzzy ARTMAP models, its predictions were compared with those from the group contribution methods of Meissner[2] and Joback[39] for normal boiling points and critical temperatures, respectively. The method of Meissner correlates the boiling point T_b (K) with the molar refraction (R_D), the parachor (P), and a constant B whose value depends upon the chemical type.[2]

$$T_b = \frac{637 R_D^{1.47} + B}{P} \tag{1}$$

The Joback[39] method is based on a modification of Lydersen's method for the estimation of critical properties in which the following equation is used for the critical temperature T_c (K):

$$T_c = \frac{T_b}{[0.584 + 0.965\sum \Delta T - (\Delta T)^2]} \tag{2}$$

The ΔT values in eq 2 are given in Table 2-2 of Reid et al.[39]

Table 1. QSPR Performance for Boiling Point Prediction of Heterogeneous Set of Organic Compounds Using Fuzzy ARTMAP and Back-Propagation Neural Networks

		average error		average error		std dev	
data set	records	K	%	K	%	K	%
Fuzzy ARTMAP model using $N, \mu, \chi(0-4), \kappa^2$							
all data	1168	2.00	0.49			8.61	2.09
training	1015			0.26	0.06	1.33	0.30
test	153			13.47	3.27	19.86	4.82
Fuzzy ARTMAP model using $N, \chi(0-4), \kappa^2$							
all data	1168	4.04	1.00			14.21	3.50
training	1015			1.61	0.40	7.23	1.79
test	153			20.47	5.03	30.20	7.42
Back-propagation model (8−12−1) using $N, \mu, \chi\ (0-4), \kappa^2$							
all data	1168	27.70	7.02			38.30	10.30
training	1015			28.70	7.20	39.60	10.80
test	153			20.80	5.40	26.70	10.20
Meissner's group contribution method							
all data	1168	28.70	6.95			38.37	10.51

Table 2. QSPR Performance for Critical Temperature Prediction of Heterogeneous Set of Organic Compounds Using Fuzzy ARTMAP and Back-Propagation Neural Networks

		average error		average error		std dev	
data set	records	K	%	K	%	K	%
Fuzzy ARTMAP model using $N, \mu, \chi(0-4), \kappa^2$							
all data	530	1.37	0.24			5.59	0.96
training	461			0.27	0.05	0.21	0.04
test	69			8.51	1.45	13.29	2.30
Fuzzy ARTMAP model using $N, \chi(0-4), \kappa^2$							
all data	530	2.05	0.35			8.94	1.48
training	461			0.32	0.06	0.69	0.12
test	69			13.84	2.30	21.62	3.58
Back-propagation model (8−12−1) using $N, \mu, \chi(0-4), \kappa^2$							
all data	530	30.20	5.60			29.20	6.00
training	463			31.40	5.80	30.30	6.30
test	68			21.60	3.90	18.90	3.90
Joback's group contribution method							
all data	530	19.50	3.40			23.90	4.10

Results and Discussion

Four sets of compounds were used in the present study: (1) the complete set of 1168 organic compounds for which boiling points were available, (2) a subset of 530 compounds with their corresponding critical temperatures, (3) a subset of 463 compounds for the critical pressure model development, and (4) a composite subset of 436 compounds for which all three properties were available. The above data sets were used to develop and evaluate the different fuzzy-ARTMAP-based QSPR models and to compare their performances with results from previous studies. The models for each property were tested with the two sets of input molecular descriptors, tset and wset, described in the previous methodology section. The complete data set, neural predictions with wset, and estimates obtained from group contribution methods are included in the Supporting Information.

The performance of the four fuzzy ARTMAP neural network/QSPR models is summarized in Tables 1−4, along with predictions obtained with an 8−12−1 back-propagation model and group contribution methods. Each table contains the absolute error and standard deviation and the corresponding relative errors between predictions and the complete data set, training set, and testing set for each model.

Boiling Point. The training of the fuzzy ARTMAP neural network model for boiling points was carried out with the 1015 compounds of the complete data set that are assigned the identifier tr for this property. The average absolute errors of the predictions (see Table 1) for the complete data set and the training and test sets, using the wset descriptors, are 2.0 K (0.49%), 0.26 K (0.06%), and 13.5 K (3.3%), respectively, with corresponding standard deviations of 8.6 K (2.1%), 1.3 K (0.30%), and 19.9 K (4.8%). These absolute errors increased to 4.0 K (1%), 1.6 K (0.4%), and 20.5 K (5.0%), respectively, for the model developed with the tset descriptors, the corresponding standard deviations being 14.2 K (3.5%), 7.2 K (1.8%), and 30.2 K (7.4%). Clearly, the three-dimensional information, which is added by the dipole moment into the wset descriptor set, improves the performance of the boiling point QSPRs developed with the fuzzy ARTMAP network. This fuzzy ARTMAP network system is able to establish appropriate weighted links for the nonlinear relations that exist between the input (descriptors) and output (physical property) during both the training and testing stages.

Predictions of boiling points using an optimal back-propagation 8−12−1 architecture and the wset descriptors (Table 1) fit the boiling point data with a significantly higher absolute mean error of 27.7 K (7.0%) and a standard deviation of 38.3 K (10.3%). The corresponding absolute mean errors for the training and testing sets are 28.7 K (7.2%) and 20.8 K (5.4%), with corresponding standard deviations of 39.6 K (10.8%) and 26.7 K (10.2%). An optimal back-propagation 7−14−1 architecture using the tset descriptors yielded predictions with an absolute mean error of 28 K (7.1%) and a standard deviation of 28.8 K (7.6%). The absolute mean errors for the training and testing sets are 30.0 K (8.0%) and 23.7 K (5.2%), respectively, with corresponding standard deviations of 29.3 K (8.1%) and 27.2 K (5.8%). It should be noted that the dimension of the training and test sets selected by fuzzy ART change from the wset to the tset descriptors, explaining the differences in relative error reported above.

Figure 3 shows that the performance of Meissner's group contribution method, with an absolute mean error of 28.7 K (6.95%) (Figure 3b), is not as good as that of back-propagation (Figure 3a) with the wset descriptors. The significant improvement in the predictions of the fuzzy ARTMAP model relative to those of the back-propagation neural-network-based model (both with the wset descriptors) and the Meissner model is evident in Figure 3, parts a and b, respectively, for the complete data set. The plots show that the fuzzy ARTMAP/QSPR

Table 3. QSPR Performance for Critical Pressure Prediction of Heterogeneous Set of Organic Compounds Using Fuzzy ARTMAP and Back-Propagation Neural Networks

		average error		average error		std dev	
data set	records	MPa	%	MPa	%	MPa	%
Fuzzy ARTMAP model using $N, \mu, \chi(0-4), \kappa^2$							
all data	463	0.02	0.52			0.08	2.48
training	409			0.00	0.01	0.00	0.02
test	54			0.13	4.31	0.19	6.02
Fuzzy ARTMAP model using $N, \chi(0-4), \kappa^2$							
all data	463	0.02	0.65			0.09	2.91
training	409			0.00	0.08	0.04	0.77
test	54			0.15	4.90	0.21	6.87
Back-propagation model (8−10−1) using $N, \mu, \chi(0-4), \kappa^2$							
all data	463	0.31	7.71			0.39	7.38
training	409			0.33	7.96	0.40	7.58
test	54			0.19	6.00	0.21	5.43

Table 4. QSPR Performance for Boiling Point, Critical Temperature, and Critical Pressure Predictions of Heterogeneous Sets of Organic Compounds Using Fuzzy ARTMAP and Back-Propagation Neural Networks

		average error		average error		std dev	
data set	records	K/MPa	%	K/MPa	%	K/MPa	%
Fuzzy ARTMAP boiling point model using N, μ, $\chi(0-4)$, κ^2							
all data	436	1.01	0.28			4.72	1.43
training	386			0.02	0.01	0.08	0.02
test	50			8.31	2.31	11.32	3.57
Fuzzy ARTMAP critical temperature model using N, μ, $\chi(0-4)$, κ^2							
all data	436	1.34	0.24			5.83	1.06
training	386			0.02	0.00	0.01	0.00
test	50			11.10	2.01	13.40	2.43
Fuzzy ARTMAP critical pressure model using N, μ, $\chi(0-4)$, κ^2							
all data	436	0.02	0.63			0.11	2.74
training	386			0.00	0.01	0.00	0.00
test	50			0.18	5.23	0.27	6.29
Fuzzy ARTMAP boiling point model using N, $\chi(0-4)$, κ^2							
all data	436	1.59	0.44			7.81	2.33
training	376			0.21	0.06	1.30	2.01
test	50			11.79	3.29	19.66	5.98
Fuzzy ARTMAP critical temperature model using N, $\chi(0-4)$, κ^2							
all data	436	2.07	0.38			9.03	1.75
training	376			0.30	0.06	2.01	0.37
test	50			15.09	2.81	21.66	4.28
Fuzzy ARTMAP critical pressure model using N, $\chi(0-4)$, κ^2							
all data	436	0.02	0.63			0.09	2.30
training	376			0.00	0.12	0.04	0.88
test	50			0.15	4.44	0.19	4.75
Back-propagation boiling point model (8-15-3) using N, μ, $\chi(0-4)$, κ^2							
all data	436	19.45	5.48			22.54	6.32
training	376			20.95	5.92	23.63	6.63
test	50			10.03	2.72	9.51	2.58
Back-propagation critical temperature model (8−15−3) using N, μ, $\chi(0-4)$, κ^2							
all data	436	27.65	5.23			30.46	5.99
training	376			29.84	5.64	31.92	6.28
test	50			13.90	2.62	11.83	2.38
Back-propagation critical pressure model (8−15−3) using N, μ, $\chi(0-4)$, κ^2							
all data	436	0.31	7.64			0.40	6.99
training	376			0.33	8.03	0.42	7.23
test	50			0.18	5.39	0.18	4.68

model is superior throughout the large range of boiling point temperatures covered, with just a few compounds with errors higher than the standard deviation. A reasonable lesson to be learned from the above comparison is that the ability of fuzzy ARTMAP to partition the input feature space allows for the development of more general QSPR models compared to global activation models such as back-propagation. In fact, only six compounds were misclassified when wset was used, with four of them considered to be outliers because their absolute mean errors were greater than 15%. An examination of the Supporting Information indicates that the compounds misclassified by the fuzzy ARTMAP are cyclopentene, pyridine, methyl cyclohexane, 3-ethyl pentane, 2-methyl-3-ethyl pentane, and 4-chlorophenol.

The above misclassifications can be better understood by considering the functioning of the fuzzy ART preclassification and fuzzy ARTMAP classification−prediction systems. The fuzzy ART assigns cyclopentene to a class formed by cyclic compounds with 5−6 carbon atoms during the preclassification stage, which is carried out with a vigilance parameter $\rho_a = 0.9$. This compound, which is the only cyclic hydrocarbon with a double bond and five carbon atoms present in the data set, was not selected for training according to the procedure described in Figure 2. The fuzzy ARTMAP was subsequently trained with the higher vigilance parameter $\rho_b = 0.997$ to capture the greater details of the distribution of physical properties. The new prototype classes formed by artA, therefore, represented classes with fewer members. As a result, none of the new prototype classes, which contain narrower structural information, include the above-mentioned specific characteristics of the cyclopentene, which was then mapped into an alcohol class during the generalization stage. The reason for the misclassification of pyridine is that molecules with nitrogen and oxygen cannot be distinguished during generalization with the wset information. As reported by Randic,[27] the valence connectivity indices are not able to capture the differences between some heteroatoms such as nitrogen and oxygen. Similar arguments apply to methyl cyclohexane, 3-ethyl pentane, and 2-methyl-3-ethyl pentane, which are clustered with aromatics. The maximum error (24%) is observed for 4-chlorophenol, the sixth misclassified compound, which was assigned to an aromatic class without oxygen atoms. It should be noted that the numbers of misclassifications and outliers increased to 20 and 18, respectively, when the tset descriptors were used in the fuzzy ARTMAP model. A maximum error of 37% was obtained with tset for methyl cyclohexane, which was classified into a naphthalene family.

Critical Temperature. The critical temperature data set (see Supporting Information) consists of 530 heterogeneous compounds, of which 461 are in the training set for neural network model building with the wset descriptors, while the remaining 69 are for testing. Table 2 shows that the fuzzy ARTMAP model with the

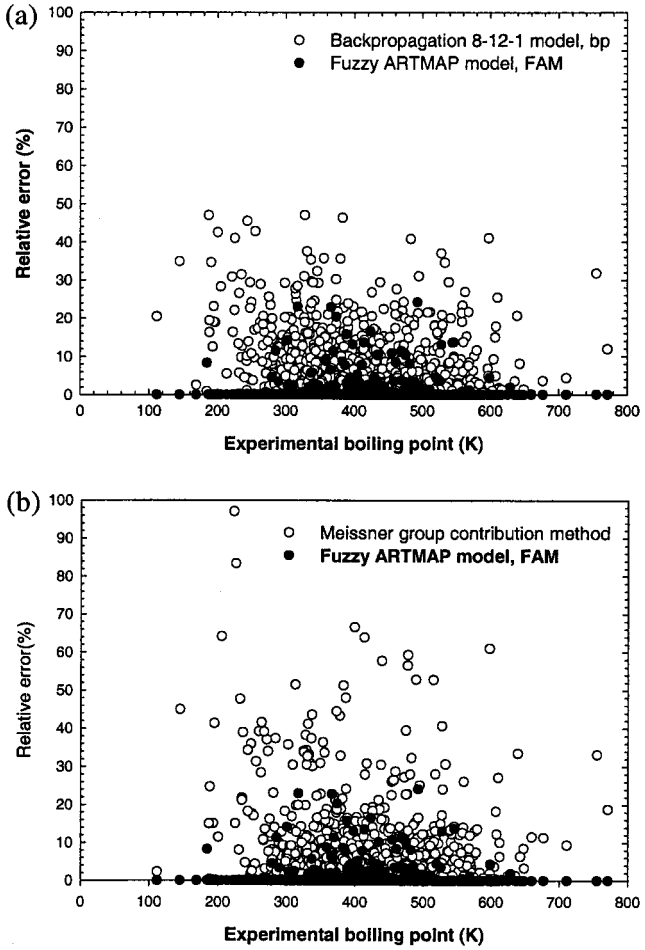

Figure 3. Comparison of the relative errors of the boiling point fuzzy ARTMAP model with (a) the 8−12−1 back-propagation model and (b) Meissner's group contribution method for the complete data set.

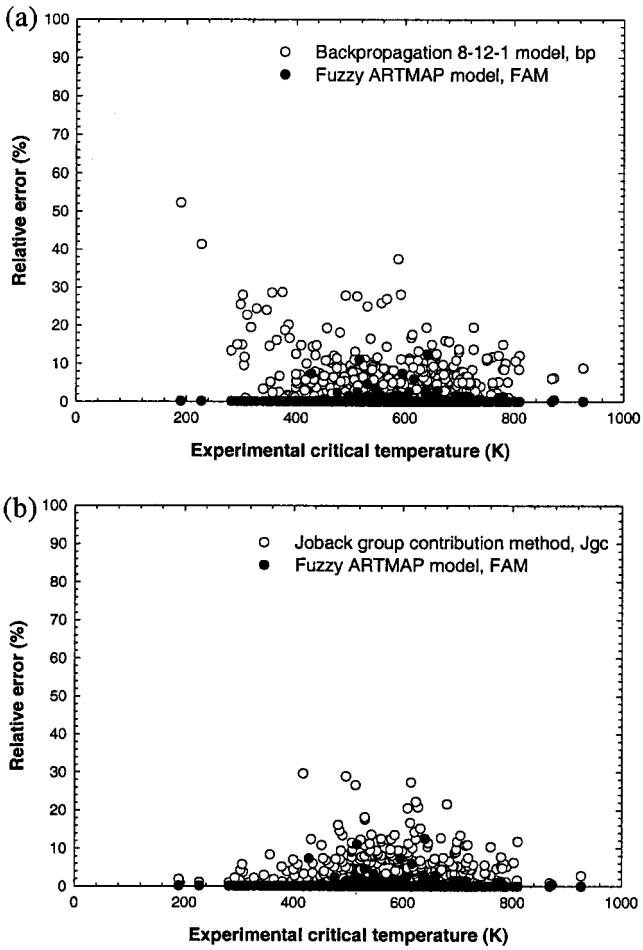

Figure 4. Comparison of the relative errors of the critical temperature fuzzy ARTMAP model with (a) the 8−12−1 back-propagation architecture and (b) Joback's group contribution method for the complete data set.

wset descriptors predicts the critical temperatures for the complete set with an absolute mean error of 1.4 K (0.24%) and a standard deviation of 5.6 K (0.96%). When only structural information is presented to the network (tset), the mean error increased to 2.1 K (0.96%), and the standard deviation increased to 8.9 K (1.5%), as also reported in Table 2. Similar trends, but with larger errors, are observed for the test compounds when either the tset or the wset of descriptors are used. However, the absolute mean prediction errors for the test set decreased from 13.8 K (2.3%) to 8.5 K (1.5%) when the dipole moment was considered.

As in the case of the boiling point results, the inclusion of the dipole moment in the wset descriptors reduces the number of misclassifications and outliers. We note that, although the two compounds 2-ethyl-1-hexanol and *n*-dipropylamine are misclassified with the wset descriptors, the errors are too low (<15%) to consider them as outliers. The former is misclassified as belonging to the aniline class, and its critical temperature is predicted with the largest relative error of 12.5%. The inverse behavior is observed for *n*-dipropylamine, as it is misclassified as belonging to the ether class. In contrast to the results for the wtest-based model, twelve compounds were misclassified with the tset descriptors. Again, it appears that compounds containing nitrogen cannot be differentiated from others containing diverse oxygen functional groups. Consistent with the results obtained with the wset descriptors, 2-ethyl-1-hexanol also presents the largest relative error (12.5%) with the tset descriptors and is also classified with the anilines. The present results indicate that, for a critical temperature mode, the inclusion of the dipole moment does not provide sufficient additional input information to classify this compound or any of the outliers (Figure 4) with a close prototype class created during the training phase. The above suggests that additional input information (perhaps quantum-type descriptors) might be needed to develop a more accurate critical temperature model.

A comparison of the critical temperature predictions of the fuzzy ARTMAP model with those of the optimal 8−12−1 back-propagation neural network, both with wset, and Joback's group contribution method[39] is shown in Figure 4, parts a and b, respectively, and is also provided in the Supporting Information. The back-propagation model predicts the critical temperatures with an absolute mean error of 30.2 K (5.6%) for the entire data set (see Supporting Information) and with a standard deviation of 29.2 K (6.0%). The absolute mean errors for the training and testing sets are 31.4 K (5.8%) and 21.6 K (3.9%), respectively, with corresponding standard deviations of 30.3 K (6.3%) and 18.9 K (3.9%). The highest errors correspond to the smallest molecules, e.g., ≤2 carbon atoms, with a largest observed error of 52.2% for methane. Hall and Kier[17]

already observed this behavior when predicting boiling points of alkanes and alcohols from structural information only.

The improved performance of the fuzzy ARTMAP model relative to Joback's group contribution method for estimating critical temperatures is clearly evident in Figure 4b. We note that Joback's method reduces to a linear equation, and thus, the advantage of neural network models over linear or conventional methods is an important advantage. The group contribution estimates have a mean error 19.5 K (3.4%) for the entire critical temperature data set, with a standard deviation of 23.9 K (4.1%). The maximum relative error of 29.6% corresponds to isobutylene. Overall, the fuzzy ARTMAP model is more accurate than the group contribution method and has better generalization capabilities over the whole range of temperatures covered.

Critical Pressure. Fuzzy ARTMAP and back-propagation QSPRs were developed and tested using a subset of 463 compounds (listed in the Supporting Information). The fuzzy ARTMAP model, which was trained with 409 compounds selected with the fuzzy ART classifier, predicted the critical pressures for the complete data set of 463 compounds with absolute mean errors of 0.02 MPa (0.52%) and 0.02 MPa (0.65%) for wset and tset, respectively, and with corresponding standard deviations of 0.08 MPa (2.5%) and 0.09 MPa (2.9%) (Table 3). Only four compounds were misclassified when the wset descriptors were used. The critical pressure of *trans*-4-methyl-2-pentene was predicted with the maximum relative error of 34.9%. This alkane was misclassified as an aromatic cluster. We note that isobutyl propionate is also an outlier, but this is not because of a wrong classification. In this case, there is no clear pattern of misclassification, as was the case for critical temperature. It appears that there is a slight tendency to group compounds with the same number of carbon atoms independently of the functional groups involved. The number of misclassified compounds increased from four to 10 when the tset descriptors were used as input. In this case, the maximum relative error of 33.6% was associated with 3-methyl-heptane, which was also misclassified into an aromatic family of compounds. Because the numbers of both misclassifications and outliers increased for the tset-based model, with the same pattern of failed classifications observed for the wset-based model, it can be concluded that the dipole moment provides some additional information, but not enough to improve predictions and generalization.

The performance of the fuzzy ARTMAP critical pressure model was compared with that of the optimal 8−10−1 back-propagation architecture, as illustrated in Figure 5. The mean relative error for the back-propagation model was 0.31 MPa (7.7%), with a standard deviation of 0.39 MPa (7.4%) for the entire data set, as reported in Table 3. The absolute mean errors for the training and testing sets of the back-propagation model were 0.33 MPa (8.0%) and 0.19 MPa (6.0%), respectively, with corresponding standard deviations of 0.40 MPa (7.6%) and 0.21 MPa (5.4%). It is interesting to note that the error in the test set for the fuzzy-ARTMAP-based model is lower than the error in the training set for the back-propagation model. The largest individual error obtained with this model corresponds to *trans*-4-methyl-2-pentene. Overall, the corresponding errors and deviations of the fuzzy ARTMAP model given in Table 3 for critical pressures are lower than those for the back-propagation model.

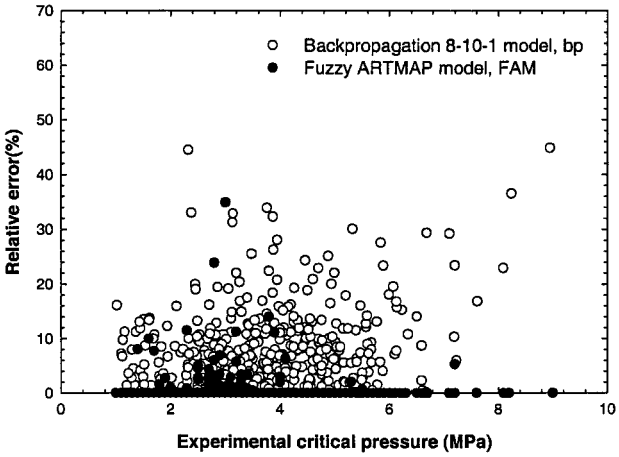

Figure 5. Comparison of the relative errors of the critical pressure fuzzy ARTMAP model with the 8−10−1 back-propagation architecture for the complete data set.

Multiple Predictions. The capabilities of the fuzzy ARTMAP system were further evaluated on the basis of the simultaneous prediction of boiling point temperatures, critical temperatures, and critical pressures using both the wset and tset molecular descriptors. A subset of 436 organic compounds for which the above three properties were available was used. The fuzzy ART classified 386 compounds for training the fuzzy ARTMAP cognitive system, resulting in 383 categories using the wset descriptors. This means that the cognitive system was able to differentiate nearly every compound in the training set on the basis of simple molecular information. The average absolute errors for the complete data set with the wset and test descriptors given in Table 4 are, respectively, (a) 1.0 K (0.28%) and 1.6 K (0.44%) for boiling point temperatures, (b) 1.3 K (0.24%) and 2.1 K (0.38%) for critical temperatures, (c) 0.02 MPa (0.63%) and 0.02 MPa (0.63%) for critical pressures. The corresponding standard deviations for the wtest and test descriptor-based models were, respectively, (a) 4.7 K (1.43%) and 7.8 K (2.3%) for boiling points, (b) 5.8 K (1.1%) and 9.03 K (1.8%) for critical temperatures, (c) 0.1 MPa (2.3%) and 0.1 MPa (2.7%) for critical pressures. The inclusion of the dipole moment in the input information resulted in a trend similar to that observed for the individual property-specific models: it improved predictions of boiling points and critical temperatures but had a significantly lesser effect on the estimation of critical pressures. The composite back-propagation model with the optimal 8−15−3 architecture yielded larger errors for the three properties, as shown in Table 4.

The prediction errors for each of the three properties, based on the composite fuzzy ARTMAP model, appear to be similar to the errors reported for the property-specific models (Tables 1−3). This behavior indicates that the increased demand imposed on the fuzzy ARTMAP system by the simultaneous prediction of three properties (instead of a single property) is compensated by the additional information that the three properties together provide through the artB module to the map field during training. In other words, it is more robust to map the molecular descriptors of each compound with three physical properties than with one. The extra information is not a problem for a neural classifier because it can both correlate multiple inputs with

multiple outputs and also adequately treat the nonlinear relationships that exist between the molecular structure and the target properties. As a consequence, the composite model is also superior in performance relative to group contribution methods that predict only single properties with linear equations.

Finally, we note that four compounds were misclassified during the test phase for the wset-based model. The compound 1,5-hexadiene was misclassified with 2,5-dioxane, isobutylene was misclassified with tributylamine, 2-methyl-2-butanol was misclassified with methyl *tert*-butyl ether, and cyclobutane was misclassified with 2-methyl butane. The misclassifications produced some outliers in the test set. For the boiling point predictions, isobutylene was the only outlier, with a relative error of 20.9%; however, its estimated critical temperature and pressure were within the respective standard deviations. For 2-methyl-2-butanol, the relative errors were higher than the average standard deviations for the three properties but smaller than 15%. For the critical pressure, the maximum error of 32% was associated with cyclobutane, for which the critical temperature and boiling point errors were lower than the average errors. The highest relative error for the critical temperature estimation was 9.4% for dimethyl oxalate. This error and similar ones for any of the three properties correspond to compounds with few members in their classes. This situation is the result of the larger number of classes that have to be considered to properly predict the three properties together. In contrast with the results of the wtest-based model, 11 compounds were misclassified when the tset descriptors were used. As in the wtest-based model, the maximum relative error for boiling points (35%) also corresponded to cyclobutane.

Conclusions

The results obtained with the fuzzy ART classifier and the fuzzy ARTMAP cognitive system show that it is possible to establish a reasonably accurate quantitative structure−property relationships for heterogeneous compounds based on a simple set of molecular descriptors. The fuzzy ARTMAP neural-network-based QSPR models proved to be of higher accuracy than back-propagation models for similar input information as well as traditional group contribution methods. The addition of the dipole moment to the simple set of seven descriptors (sum of atomic numbers, five valence connectivity indexes, second-order kappa shape index) improved the accuracy of the models for boiling points and critical temperatures but had a marginal effect on the critical pressure predictions. For the fuzzy-ARTMAP-based QSPRs, overall absolute mean errors for the respective complete data sets were 2 K (0.49%) for the boiling points of 1168 compounds, 1.4 K (0.24%) for the critical temperatures of 530 compounds, and 0.02 MPa (0.52%) for the critical pressures of 463 compounds. Moreover, the fuzzy ARTMAP was shown to be capable of simultaneously predicting boiling points and critical temperatures and pressures without loss of performance.

The current fuzzy ARTMAP system did misclassify some compounds. However these misclassifications, mostly involving compounds containing nitrogen and oxygen, occurred during the generalization of the different fuzzy ARTMAP models. A potential reason for the misclassification is that the misclassified compounds are indistinguishable with the input information provided to the system, even though they have different physical properties. Apparently, the present input information (molecular descriptors) does not contain sufficient three-dimensional molecular information to encompass all possible complex functional groups. To improve the reliability and range of applicability of the current fuzzy ARTMAP/QSPR approach, future work will focus on the use of quantum similarity descriptors to account for the differences between functional groups.

Acknowledgment

The authors are grateful for the financial support received from the CIRIT "Programa de Grups de Recerca Consolidats de la Generalitat de Catalunya", Project 1998SGR-00102, and from the "Dirección General de Investigación Científica y Técnica", Projects PB96-1011 and PPQ2000-1339. Partial support from the UCLA Center for Environmental Risk Reduction is also acknowledged.

Supporting Information Available: Table listing the normal boiling points and critical temperatures and pressures for all 1168 compounds considered in this study. This material is available free of charge via the Internet at http://pubs.acs.org.

Literature Cited

(1) Joback, K.; Reid, R. Estimation of Pure Component Properties from Group Contributions. *Chem. Eng. Commun.* **1987**, 233.

(2) Lymman, W.; Reehl, W.; Rosenblatt, D. *Handbook of Chemical Property Estimation Methods*, 3rd ed.; American Chemical Society: Washington, D.C., 1990.

(3) Egolf, L.; Jurs, P. Prediction of Boiling Points of Organic Heterocyclic Compounds Using Regression and Neural Network Techniques Structure. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 616.

(4) Espinosa, G.; Yaffe, D.; Cohen, Y.; Arenas, A.; Giralt, F. Neural Network Based Quantitative Structural Property Relations (QSPRs) for Predicting Boiling Points of Aliphatic Hydrocarbons. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 859.

(5) Gakh, A.; Gakh, E.; Sumpter, B.; Noid, D. Neural Network−Graph Theory Approach to the Prediction of the Physical Properties of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 832.

(6) Hall, L.; Story, C. Boiling Point and Critical Temperature of a Heterogeneous Data Set: QSAR Atom Type Electrotopological State Indices Using Artificial Neural Networks. *J. Chem. Inf. Compt. Sci.* **1996**, *36*, 1004.

(7) Lee, M.; Chen, J. Fluid Property Predictions with the Aid of Neural Networks. *Ind. Eng. Chem. Res.* **1993**, *32*, 995.

(8) Raymond, J.; Rogers, T. Molecular Structure Disassembly Program (MOSDAP): A Chemical Information to Automate Structure-Based Physical Property Estimation. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 463.

(9) Simamora, P.; Miller, A.; Yalkowsky, S. Melting Point and Normal Boiling Point Correlations: Applications to Rigid Aromatic Compounds. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 437.

(10) Stanton, D. T.; Jurs, P. C. Development and Use of Charged Partial Surface Area Structural Descriptors for Quantitative Structural−Property Relationship Studies. *Anal. Chem.* **1990**, *62*, 2323.

(11) Stanton, D.; Egolf, L.; Jurs, P.; Hicks, M. Computer-Assisted Prediction of Normal Boiling Points of Furans, Tetrahydrofurans and Thiophenes. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 301.

(12) Stanton, D.; Egolf, L.; Jurs, P.; Hicks M. Computer-Assisted Prediction of Normal Boiling Points of Pyrans and Pyrroles. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 306.

(13) Tetteh, J.; Suzuki, T.; Metcalfe, E.; Howells, S. Quantitative Structure−Property Relationships for the Estimation of Boiling Point and Flash Point Using a Radial Basis Function Neural Network. *J. Chem. Inf. Compt. Sci.* **1999**, *39*, 491.

(14) Turner, B.; Costello, C.; Jurs, P. Prediction of Critical Temperatures and Pressures of Industrially Important Organic Compounds from Molecular Structure. *J. Chem. Inf. Compt. Sci.* **1998**, *38*, 639.

(15) Katritzky, A. R.; Maran, U.; Lobanov, V. S.; Karelson, M. Structurally Diverse Quantitative Structure−Property Relationship Correlations of Technologically Relevant Physical Properties. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1.

(16) Egolf, L.; Jurs, P. Prediction of Boiling Points and Critical Temperatures of Industrially Important Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1996**, *34*, 947.

(17) Hall, L.; Kier, L. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological and Valence State Information. *J. Chem. Inf. Comput.* Sci. **1995**, *35*, 1039.

(18) Jurs, P. Prediction of Chemical Properties of Organic Compounds from Molecular Structure. 214th ACS National Meeting, Las Vegas, NV, Sept 7−11, 1997.

(19) Katritzky, R.; Mu, L.; Lobanov, V. Correlation of Boiling Points with Molecular Structure. 1. A Training Set of 298 Diverse Organics and a Test Set of 9 Simple Inorganics. *J. Phys. Chem.* **1996**, *100*, 10400.

(20) Katrizky, F.; Mu, L. Relationships of Critical Temperatures to Calculated Molecular Properties. *J. Chem. Inf. Compt. Sci.* **1998**, *38*, 293.

(21) Katritzky, A.; Mu, L. A QSPR study of the solubility of gases and vapors in water. *J. Chem. Comput. Sci.* **1996**, *36*, 1162.

(22) Krzyzaniak, J.; Myrdal, P.; Simamora, P.; Yalkowsky S. Boiling Point and Melting Point Prediction for Aliphatic, Non-Hydrogen-Bonding Compounds. *Ind. Eng. Chem. Res.* **1995**, *34*, 2530.

(23) Medir, M.; Giralt, F. Correlation of Activity Coefficients of Hydrocarbons in Water at Infinity Dilution with Molecular Parameters. *AIChE J.* **1982**, *28*, 341.

(24) Needham, D.; Wei, I.; Seybold, P. Molecular Modeling of the Physical Properties of the Alkanes. *J. Am. Chem. Soc.* **1988**, *110*, 4186.

(25) Pogliani, L. Molecular Modeling by Linear Combinations of Connectivity Indexes. *J. Phys. Chem.* **1995**, *99*, 925.

(26) Pogliani, L. Modeling with Special Descriptors Derived from a Medium-Sized Set of Connectivity Indices. *J. Phys. Chem.* **1996**, *100*, 18065.

(27) Randic, M.; Trinajstic, N. Comparative Structure−Property Studies: The Connectivity Basis. *J. Mol. Struct.* **1993**, *284*, 209.

(28) Randic, M.; Dobrowolski, J. Optimal Molecular Connectivity Descriptors for Nitrogen-Containing Molecules. *Int. J. Quantum Chem.* **1998**, *70*, 1209.

(29) Wikel, H. J. The Use of Neural Networks for Variable Selection in QSAR. *Bioorg. Med. Chem. Lett.* **1993**, *3*, 645.

(30) Bünz, P.; Braun, B.; Janowsky, R. Application of Quantitative Structure−Performance Relationship and Neural Network Models for the Prediction of Physical Properties from Molecular Structure. *Ind. Eng. Chem. Res.* **1998**, *37*, 3043.

(31) Espinosa, G.; Arenas, A.; Giralt, F. Prediction of Boiling Points of Organic Compounds from Molecular Descriptors by Using Artificial Neural Networks. In *Fundamentals of Molecular Similarity*; Carbó-Dorca, R., Ed.; Kluwer Academic: New York, 2001.

(32) Carpenter, A.; Grossberg, S. A Massively Parallel Architecture for a Self-Organizing Neural Pattern Recognition Machine. *Comput. Vision, Graphics, Image Process.* **1987**, *37*, 54.

(33) Carpenter, A.; Grossberg, S. The ART of Adaptive Pattern Recognition by a Self-Organizing Neural Network. *Computer* **1988**, 77.

(34) Carpenter, G. A.; Grossberg, S.; Marcuzon, N.; Reynolds, J. H.; Rosen, D. B. Fuzzy ARTMAP: A Neural Network Architecture for Incremental Supervised Learning of Analog Multidimensional Maps. *IEEE Trans. Neural Networks* **1992**, *3*, 698.

(35) Carpenter, G. A.; Grossberg, S.; Marcuzon, N.; Rosen, D. B. Fuzzy ART: Fast Stable Learning and Categorization of Analog Patterns by an Adaptive Resonance System. *Neural Networks* **1991**, *4*, 759.

(36) Carpenter, G.; Grossberg, S. A Self-Organizing Neural Network for Supervised Learning, Recognition, and Prediction. *IEEE Commun. Mag.* **1992**, 38.

(37) Giralt, F.; Arenas, A.; Ferre-Giné, J.; Rallo R. The Simulation and Interpretation of Turbulence with a Cognitive Neural System. *Phys. Fluids* **2000**, *12*, 1826.

(38) Daubert, T. E.; Danner, R. P. *Data Compilation Tables of Properties of Pure Compounds*; Design Institute for Physical Property Data, AICHE: New York, 1984.

(39) Reid, R. C.; Prausnitz J. M.; Sherwood, T. K. *The Properties of Gases and Liquids*; 3rd ed.; McGraw-Hill: New York, 1987.

(40) *POC (Properties of Organic Compounds)*, Personal Edition, version 5.1; CRC Press: Boca Raton, FL, 1996.

(41) Carpenter, G. A.; Grossberg, S. *Pattern Recognition by Self-Organizing Neural Networks*; MIT Press: Cambridge, MA, 1991.

(42) Carpenter, G. A.; Grossberg, S. *Neural Network for Vision and Image Processing*; MIT Press: Cambridge, MA, 1992.

(43) Bartfai, B. Hierarchical Clustering with ART Neural Networks. *Proc. IEEE Int. Conf. Neural Networks* **1994**, *2*, 940.

(44) Bartfai, B. On the Match Tracking Anomaly of ARTMAP Neural Network. *Neural Networks* **1996**, *9*, 295.

(45) Bartfai, B. An ART-Based Modular Architecture for Learning Hierarchical Clusterings. *Neurocomputing* **1996**, *13*, 31.

(46) Bartfai, B. An Adaptive Resonance Theory-Based Neural Network capable of Learning via Representational Redescription. *Proc. IEEE Int. Joint Conf. Neural Networks* **1998**, 1137.

Received for review September 14, 2000
Accepted March 25, 2001

IE0008068