# Prediction of Kovats Retention Indices of Some Aliphatic Aldehydes and Ketones on Some Stationary Phases at Different Temperatures Using Artificial Neural Network

**Elaheh Konoz[1], Mohammad H. Fatemi[2,*],** and **Razieh Faraji[2]**

[1]Department of Chemistry, Central Tehran Branch, Islamic Azad University, Tehran, Iran and [2]Department of Chemistry, Mazandaran University, Babolsar, Iran

## Abstract

**In this work, the Kovats retention indices of aliphatic ketones and aldehydes on four stationary phases at different temperatures are predicted. The data set consists of retention indices of 35 aldehydes and ketones on HP-1, HP-50, DB-210, and HP-Innowax stationary phase. The molecular descriptors that appear in this model are: path one connectivity index, fractional atomic charge weighted by partial positive surface area, and dipole moment, which are selected by stepwise multiple linear regression (MLR). The selected descriptors encode steric and electronic aspects of molecules. These descriptors, together with column temperature, are used as inputs of the constructed artificial neural network (ANN). The optimized network has 4-3-4 topology, in which its outputs are retention indices of molecules at four stationary phases at the desired temperature. Comparison between statistical results calculated for MLR and ANN models reveals that all statistics have improved considerably in the case of the ANN model. The improved statistics for the ANN would suggest the existence of a nonlinear relation between selected molecular descriptors and their retention in gas chromatography. Also, the simultaneous prediction of retention indices for aldehydes and ketones at four stationary phases at different temperatures using only three molecular descriptors shows the capability of the obtained ANN model.**

## Introduction

The Kovats retention indices in gas chromatography (GC) represent the retention behavior of a compound relative to a standard set of hydrocarbons, using a logarithmic scale (1). The identification of many compounds is often accomplished on the basis of a comparison between the GC retention of a standard sample and the suspected material. However, it is not always possible to obtain pure standard compounds for such a comparison; therefore, the development of a theoretical model for estimating the retention indices seems to be useful. Retention in GC is a phenomenon that primarily depends on the interactions of the solute molecules and the stationary phase. The forces associated with these interactions can be related to the geometric and topological structure and also to the electronic environments of the molecule. Intrinsic to chemistry is the concept that there are relationships between bulk properties of compounds and their molecular structures, which provide a connection between the macroscopic and the microscopic properties of the matter. Therefore, knowledge of the molecular structures is the key to understanding the properties and activities of molecules. Quantitative structure-property relationship (QSPR) is a mathematical method that relates the properties of a molecule to its structural features. This approach has been used to obtain simple models that explain and predict the chromatographic behavior of various classes of compounds. There are some reports about the applications of QSPR in chromatographic studies (2–6). Jurs and his group correlated the observed Kovats retention indices of sulfur vesicants by multiple linear regression techniques (MLR) using 9 descriptors in their models for different stationary phases (7). Dimov and Osman used a quantitative structure-retention relationship technique to relate the chromatographic retention of 38 iso-alkanes to their molecular structural features (8). Kang et al. successfully predicted the capillary GC retention indices of 100 polycyclic aromatic hydrocarbons (9). Also, O. Farkas and K. Héberger reported the construction of a linear model for the prediction of retention indices of some aliphatic alcohols (10). They used variable selection methods including ridge regressions, partial least square, pair-correlation method, forward selection, and best subset selection methods. They concluded that forward selection and best subset variable selection methods gave reliable results. There are some papers about the QSPR modeling of retention indices of oxo compounds. R.D.M.C. Amboni et al. reported the construction of a linear QSPR model for the estimation of retention indices of 54 oxo compounds on a low polarity stationary phase at 50°C, and also two separate MLR models for 30 oxo compounds on HP-50 and HP-1 stationary phases (11). They successfully used a semi-empirical topological index ($I_{ET}$) in their constructed QSPR models. In a similar work, B.S. Junkens et al. studied the application of these molecular indices in QSPR studies of the same data set at one temperature (50°C) (12). They successfully constructed four MLR models for prediction or Kovats retention indices on HP-1, HP-50, DB-210, and HP-Innowax

* Author to whom correspondence should be addressed: mhfatemi@umz.ac.ir.

stationary phases. Also, B. Ren investigated the quantitative correlation between the Kovats retention indices of 33 aldehydes and ketones and their atom-type-based AI topological indices on four stationary phases (HP-1, HP-50, DB-210, and HP-Innowax) at 50°C using four separate linear equations (13). They used 4–5 molecular descriptors in each model. The results of their studies indicate that the molecular size makes a dominant contribution to retention indices.

Over the last few years, the artificial neural network (ANN) has attracted increasing interest as a most promising method in classification and multivariate calibration problems, and also provides an interesting new approach to QSPR studies (14–18). Although in the MLR method, the analysis is limited to a certain number of possible interactions, more terms can be examined for interactions between features by the ANN. Also, ANNs are capable of recognizing non-linear relationships between inputs and outputs. In addition, the ANN can use qualitative as well as quantitative inputs, and also it does not require an explicit relationship between the inputs and the outputs. The main aim of the present work was to develop a QSPR model using an ANN to predict Kovats retention indices of aliphatic ketones and aldehydes on some stationary phases at different temperatures. In the first step, a combined MLR model was constructed, then an ANN was developed for inspection of non-linear relations between different parameters in the model, for the simultaneous modeling and prediction of the retention indices on different stationary phases.

## Methods

### Data Set

The data set of Kovats retention indices was taken from the values reported by Héberger et al. (19). The molecules in the data set are 35 aliphatic ketones and aldehydes, and are shown in Table I. The Kovats retention indices of all molecules included in the data set were obtained under the same conditions on four stationary phase, which are: dimethylpolysiloxane (HP-1), 50% phenylmethylpolysiloxane (HP-50), 50% Trifluoropropylmethylpolysiloxane (DB-210), and polyethylene glycol (HP-Innowax). The retentions of these compounds were measured at 50°C, 70°C, 90°C, and 110°C; therefore, the data set consists of 560 retention indices data. The Kovats retention indices fall in the range of 360.4 to 1055.1 for acetaldehyde and 5-nonanone on HP-1, 484.3 to 1178.8 for acetaldehyde, and 5-nonanone on HP-50, 630.4 to 1370 for acetaldehyde and 5-nonanone on HP- DB-210, and 715.8 to 1353 for acetaldehyde and 5-nonanone on HP-Innowax, respectively.

### Descriptors generation and regression analysis

Retention in GC is the result of competitive distribution of the solute between the mobile and stationary phase. The molecular structure and chemical properties of the solute and stationary phase determine the type and extent of the interactions between the solute and stationary phases. The differences between these interactions govern the retention behavior of the solute through the column. Due to the diversity of the molecules studied in this work, different molecular descriptors were calculated. These molecular descriptors were mainly computed using the CODESSA software. This software, developed by Katritzky's group, enables the calculation of a large number of quantitative descriptors based on molecular structural information (20–22) and codes this chemical information into mathematical forms. In the first stage, the structures of the molecules were drawn by the HyperChem 4.0 program (23) and exported in a file format suitable for the MOPAC program (24). Then the geometry optimization was performed with the semieperical quantum method AM1 (25) using the MOPAC 6.0. Then all geometries were fully optimized without symmetry restrictions. In all cases, frequency calculations have been performed in order to ensure that all the calculated geometries correspond to true minima. The HyperChem and MOPAC output files were used by the CODESSA program to calculate five classes of descriptors, including: constitutional (number of various types of atoms and bonds, number of rings, molecular mass, etc.); topological (Winner index, Randic indices, Kier-Hall shape indices, etc.); geometrical (moment of inertia, molecular volume, molecular surface area, etc.); electrostatic (minimum and maximum of partial charges, polarity parameters, charged partial surface area descriptors, etc.); and quantum chemical (reactivity indices, dipole moment, HOMO and LUMO energies, etc.). Because it is not possible to know a priori which descriptors are most relevant to the problem at hand, a comprehensive set of descriptors is usually employed, chosen based on experience, software availability, and computational cost. The heuristic MLR procedures available in the framework of the CODESSA program were used to perform a complete search for the best multi linear correlations with a multitude of descriptors. This procedures provide colinearity control (i.e., any two descriptors inter-correlated above 0.90 are never involved in the same model) and implement heuristic algorithms for the rapid selection of the best correlation, without testing all possible combinations of the available descriptors. The heuristic method of descriptor selection proceeds with a pre-selection of descriptors by eliminating (*i*) those descriptors that are not available for each structure, (*ii*) descriptors having a small variation in magnitude for all structure, (*iii*) descriptors that give an F-test's value below 1.0 in the one-parameter correlation, and (*iv*) descriptors whose t-values are less than the user-specified value, etc. This procedure orders the descriptors by decreasing correlation coefficient when used in one-parameter correlation coefficient. The next step involves correlation of the given property with (*i*) the top descriptor in the previous list with each of the remaining descriptors and (*ii*) the next one with each of the remaining descriptors, etc. The best pairs, as evidenced by the highest F-values in the two-parameter correlations, are chosen and used for further inclusion of descriptors in a similar manner. The heuristic method usually produces correlations 2–5 times faster than other methods with comparable quality. The rapidity of calculations from the heuristic method renders it as a suitable method of choice in practical research. Though MLR failed to obtain an appropriate QSPR model, the nonlinear relationship within the data was well incorporated into the model developed by the ANN. Descriptors that appeared in the combined MLR model for HP-1, HP-50, DB-210, and HP-Innowax stationary phase were used as inputs for the generated ANNs.

**Table I. Kovats Retention Indices of Ketones (1–19) and Aldehydes (20–35) which were Used as Data Set***

| Compound | A1 | A2 | A3 | A4 | B1 | B2 | B3 | B4 | C1 | C2 | C3 | C4 | D1 | D2 | D3 | D4 | TBP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Acetone | 469.7 | 469.3 | 469.4 | 469.4 | 470.2 | 606.3 | 603.7 | 601.9 | 600.7 | 792.9 | 799.6 | 807.7 | 816.6 | 835.0 | 837.5 | 840.8 | 329.4 |
| 2-Butanone | 574.7 | 575.0 | 574.8 | 574.8 | 575.8 | 711.6 | 710.9 | 709.7 | 710.4 | 882.1 | 889.8 | 898.5 | 906.8 | 919.8 | 923.7 | 927.8 | 352.8 |
| 3-Methyl-2-butanone | 639.9 | 640.5 | 641.5 | 641.5 | 642.8 | 767.1 | 767.2 | 767.3 | 767.8 | 943.3 | 951.8 | 960.9 | 970.7 | 949.4 | 954.4 | 959.6 | 367.7 |
| 3-Pentanone | 675.4 | 675.5 | 675.8 | 675.8 | 676.9 | 809.8 | 809.6 | 809.6 | 810.4 | 960.8 | 968.1 | 976.1 | 984.1 | 996.9 | 1001.5 | 1006.3 | 374.9 |
| 2-Pentanone | 665.4 | 665.7 | 666.1 | 666.1 | 666.9 | 799.3 | 798.6 | 798.6 | 799.3 | 973.9 | 981.8 | 991.1 | 1000.7 | 996.2 | 1000.7 | 1005.6 | 375.2 |
| 2,2-Dimethyl-3-butanone | 691.8 | 693.3 | 695.5 | 695.5 | 697.4 | 808.0 | 808.5 | 809.7 | 811.1 | 992.0 | 1000.9 | 1010.9 | 1021.2 | 968.5 | 974.3 | 980.2 | 379.2 |
| 4-Methyl-2-pentanone | 720.1 | 720.7 | 721.7 | 721.7 | 722.9 | 841.8 | 841.6 | 842.0 | 842.5 | 1027.1 | 1035.3 | 1045.1 | 1054.7 | 1025.2 | 1029.0 | 1033.9 | 390.0 |
| 3-Methyl-2-pentanone | 733.5 | 734.9 | 736.8 | 736.8 | 739.0 | 859.6 | 860.5 | 862.2 | 864.2 | 1036.1 | 1045.7 | 1056.0 | 1066.8 | 1033.9 | 1040.8 | 1047.6 | 391.2 |
| 2,4-Dimethyl-3-pentanone | 777.7 | 779.4 | 781.1 | 781.1 | 783.1 | 881.0 | 882.2 | 883.4 | 885.3 | 1038.5 | 1046.3 | 1054.5 | 1064.4 | 1014.8 | 1020.8 | 1026.9 | 397.7 |
| 3-Hexanone | 764.1 | 764.4 | 765.0 | 765.0 | 766.0 | 893.6 | 893.8 | 894.5 | 895.2 | 1048.4 | 1055.7 | 1064.2 | 1073.0 | 1068.0 | 1073.3 | 1078.9 | 398.2 |
| 2-Hexanone | 767.0 | 767.3 | 768.0 | 768.0 | 769.0 | 901.3 | 901.6 | 902.1 | 903.0 | 1081.5 | 1090.2 | 1100.2 | 1110.4 | 1097.2 | 1102.2 | 1107.6 | 404.1 |
| 4-Heptanone | 851.8 | 852.8 | 853.7 | 853.7 | 855.1 | 976.1 | 976.6 | 977.2 | 978.4 | 1134.5 | 1142.6 | 1151.1 | 1160.3 | 1139.4 | 1145.0 | 1151.1 | 417.2 |
| 5-Methyl-2-hexanone | 835.4 | 836.1 | 837.2 | 837.2 | 838.4 | 964.3 | 964.7 | 965.1 | 966.1 | 1161.3 | 1170.0 | 1181.1 | 1192.2 | 1156.1 | 1160.9 | 1166.3 | 417.2 |
| 3-Heptanone | 864.9 | 865.4 | 866.6 | 866.6 | 867.3 | 994.4 | 994.8 | 995.4 | 996.9 | 1153.6 | 1161.2 | 1170.6 | 1180.0 | 1167.2 | 1173.2 | 1179.9 | 420.2 |
| 2-Heptanone | 867.5 | 867.9 | 868.7 | 868.7 | 869.9 | 1000.1 | 1002.8 | 1003.3 | 1004.2 | 1184.3 | 1193.4 | 1203.9 | 1214.6 | 1195.8 | 1201.9 | 1207.6 | 424.6 |
| 2-Methyl-3-heptanone | 917.5 | 918.7 | 920.1 | 920.1 | 921.6 | 1031.9 | 1032.9 | 1034.4 | 1035.6 | 1194.2 | 1202.0 | 1210.4 | 1220.1 | 1178.7 | 1185.0 | 1191.0 | 431.2 |
| 5-Methyl-3-heptanone | 921.7 | 922.9 | 924.5 | 924.5 | 926.1 | 1041.5 | 1042.3 | 1043.7 | 1045.4 | 1206.9 | 1214.2 | 1224.9 | 1234.8 | 1200.1 | 1205.5 | 1213.0 | 432.5 |
| 3-Octanone | 964.8 | 965.4 | 966.0 | 966.0 | 967.3 | 1094.9 | 1095.4 | 1096.3 | 1097.4 | 1255.5 | 1264.0 | 1272.9 | 1282.6 | 1265.5 | 1271.4 | 1277.9 | 440.2 |
| 5-Nonanone | 1051.4 | 1052.4 | 1053.5 | 1053.5 | 1055.1 | 1175.4 | 1176.1 | 1177.3 | 1178.8 | 1342.6 | 1351.2 | 1360.4 | 1370.3 | 1334.1 | 1340.6 | 1347.3 | 461.6 |
| Acetaldehyde | 360.4 | 360.6 | 360.4 | 360.4 | 360.9 | 487.9 | 485.3 | 484.3 | 484.9 | 630.4 | 636.3 | 641.5 | 649.2 | 715.8 | 716.6 | 717.8 | 294.0 |
| Propanal | 472.7 | 472.7 | 473.0 | 473.0 | 473.6 | 604.5 | 603.1 | 601.1 | 601.3 | 739.4 | 746.2 | 753.0 | 761.6 | 808.8 | 810.4 | 812.9 | 322.0 |
| Acroleine | 462.8 | 462.6 | 462.9 | 462.9 | 463.5 | 603.3 | 603.9 | 603.8 | 605.6 | 743.7 | 751.4 | 760.3 | 768.4 | 867.0 | 869.4 | 871.8 | 326.2 |
| Isobutanal | 540.3 | 540.9 | 541.6 | 541.6 | 543.0 | 660.5 | 659.6 | 658.3 | 659.4 | 803.7 | 810..8 | 819.5 | 828.4 | 830.4 | 836.8 | 838.6 | 336.2 |
| Butanal | 571.1 | 571.9 | 572.5 | 572.5 | 573.3 | 702.5 | 702.9 | 702.6 | 703.3 | 843.1 | 851.1 | 859.9 | 869.0 | 894.8 | 897.8 | 901.8 | 348.9 |
| Trimethylacetsldehyde | 581.7 | 583.3 | 584.5 | 584.5 | 586.1 | 681.0 | 680.9 | 681.6 | 682.1 | 841.6 | 849.7 | 858.2 | 867.1 | 822.6 | 826.7 | 831.1 | 350.7 |
| Isovaleralehyde | 635.0 | 636.4 | 637.5 | 637.5 | 639.2 | 757.6 | 758.0 | 758.7 | 760.1 | 912.8 | 922.2 | 932.4 | 943.0 | 936.0 | 940.3 | 945.6 | 365.7 |
| Methylmutanal | 645.3 | 646.9 | 648.2 | 648.2 | 650.0 | 767.5 | 767.9 | 768.6 | 770.5 | 913.3 | 922.0 | 932.0 | 941.9 | 931.2 | 936.7 | 942.8 | 365.7 |
| Yaleraldehyde | 674.4 | 675.2 | 676.2 | 676.2 | 677.3 | 807.4 | 807.7 | 808.5 | 809.5 | 953.8 | 963.0 | 972.9 | 983.3 | 998.1 | 1002.8 | 1007.4 | 376.2 |
| 2-Butenal | 623.4 | 624.5 | 625.8 | 625.8 | 627.3 | 787.7 | 789.6 | 791.6 | 794.3 | 967.2 | 980.2 | 993.0 | 1006.4 | 1061.5 | 1069.3 | 1077.1 | 377.7 |
| 3,3-Dimethylbutanal | 689.1 | 691.4 | 694.0 | 694.0 | 697.0 | 803.7 | 805.5 | 807.3 | 809.3 | 978.4 | 989.2 | 1000.7 | 1012.3 | 968.6 | 974.8 | 981.4 | 378.1 |
| 2-Ethylbutanal | 742.1 | 744.0 | 746.3 | 746.3 | 748.9 | 862.7 | 864.8 | 866.8 | 869.1 | 1009.6 | 1019.3 | 1030.6 | 1041.5 | 1018.0 | 1025.3 | 1032.7 | 391.2 |
| Hexanal | 776.5 | 777.2 | 778.5 | 778.5 | 780.0 | 909.9 | 910.5 | 911.4 | 912.6 | 1059.3 | 1068.6 | 1079.4 | 1090.1 | 1098.3 | 1104.0 | 1110.0 | 401.2 |
| Heptanal | 877.2 | 878.7 | 880.0 | 880.0 | 881.7 | 1009.8 | 1011.4 | 1013.3 | 1013.8 | 1162.7 | 1173.2 | 1183.9 | 1194.8 | 1199.6 | 1205.3 | 1211.5 | 426.0 |
| 2-Ethylhexanal | 933.2 | 935.4 | 937.5 | 937.5 | 940.1 | 1049.3 | 1051.4 | 1053.2 | 1055.4 | 1205.4 | 1217.4 | 1228.6 | 1240.8 | 1197.8 | 1205.9 | 1213.1 | 434.1 |
| Octanal | 977.8 | 979.7 | 981.2 | 981.2 | 983.0 | 1110.9 | 1111.5 | 1112.7 | 1114.4 | 1265.5 | 1275.1 | 1287.6 | 1299.8 | 1298.8 | 1306.1 | 1313.0 | 444.2 |

* Notes: A1 = HP-1, 50°C; A2 = HP-1, 70°C; A3 = HP-1, 90°C; A4 = HP-1, 110°C; B1 = HP-50, 50°C; B2 = HP-50, 70°C; B3 = HP-50, 90°C; B4 = HP-50, 110°C; C1 = DB-210, 50 °C; C2 = DB-210, 70°C; C3 = DB-210, 90°C; C4 = DB-210, 110°C; D1 = HP-Innowax, 50°C; D2 = HP-Innowax, 70°C; D3 = HP-Innowax, 90°C; D4 = HP-Innowax, 110°C.

**Table II. ANN Predicted (Calculated) Values of the Retention Indices***

| Compound | A1 | A2 | A3 | A4 | B1 | B2 | B3 | B4 | C1 | C2 | C3 | C4 | D1 | D2 | D3 | D4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Acetone | 463.87 | 465.71 | 467.7 | 469.8 *t* | 595.89 | 598.83 | 601.85 | 605 *t* | 798.7 | 803.22 | 807.69 | 812.2 *t* | 829.83 | 836.28 | 842.6 | 848.8 *t* |
| 2-Butanone | 560.3 *p* | 564.21 | 568.3 *t* | 572.61 | 696.6 *p* | 702.06 | 707.6 *t* | 713.35 | 878.2 *p* | 891.37 | 904.4 *t* | 917.37 | 907.1 *p* | 918.9 | 930.5 *t* | 941.87 |
| 3-Methyl-2-butanone | 625.46 | 630.5 *p* | 635.3 *p* | 639.69 | 757.35 | 762.1 *p* | 766.6 *p* | 770.79 | 942.2 | 949.3 *p* | 955.8 *p* | 961.8 | 971.47 | 976.3 *p* | 980.9 *p* | 985.28 |
| 3-Pentanone | 652.07 *t* | 660.18 | 668.08 | 675.74 | 786.1 *t* | 794.27 | 802.23 | 909.97 | 904.4 *t* | 971.07 | 986.85 | 1002.1 | 981.9 *t* | 992.87 | 1003.61 | 1014.09 |
| 2-Pentanone | 657.56 | 661.12 | 664.84 | 668.83 | 789.56 | 792.97 | 796.57 | 800.38 | 984.4 | 989.8 | 995.44 | 1001.2 | 999.99 | 1003.83 | 1007.72 | 1011.66 |
| 2,2-Dimethyl-3-butanone | 676.52 | 679.41 | 682.2 *p* | 685.0 *p* | 796.15 | 799.18 | 802.2 *p* | 805.1 *p* | 968.3 | 973.54 | 978.6 *p* | 983.6 *p* | 978.44 | 982.48 | 986.5 *p* | 990.5 *p* |
| 4-Methyl-2-pentanone | 718.06 | 720.92 | 723.7 *t* | 726.27 | 840.54 | 843.83 | 847.0 *t* | 850.12 | 1024.6 | 1031.9 | 1039 *t* | 1045.7 | 1019.98 | 1026.15 | 1032.2 *t* | 1038.19 |
| 3-Methyl-2-pentanone | 720.8 *p* | 723.54 | 728.2 *t* | 728.7 *t* | 846.2 *p* | 849.28 | 804.6 *t* | 855.3 *t* | 1015 *p* | 10221 | 1029 *t* | 1036 *t* | 1032.2 *p* | 1037.89 | 1043.5 *t* | 1048.9 *t* |
| 2,4-Dimethyl-3-pentanone | 771.55 | 774.7 *t* | 778.03 | 781.45 | 880.5 | 884.5 *t* | 888.75 | 893.16 | 1037.6 | 1047 *t* | 1057.6 | 1068.1 | 1016.87 | 1026 *t* | 1035.76 | 1045.89 |
| 3-Hexanone | 751.0 *p* | 757.21 | 763.19 | 768.94 | 882.9 *p* | 889.29 | 895.47 | 901.5 | 1033 *p* | 1043.9 | 1054.7 | 1065.3 | 1062.9 *p* | 1071.32 | 1079.8 | 1088.34 |
| 2-Hexanone | 767.38 | 772.25 | 776.95 | 781.5 *t* | 903.3 | 908.55 | 913.68 | 918.7 *t* | 1091.7 | 1101.7 | 1111.5 | 1121 *t* | 1105.46 | 1113.21 | 1120.92 | 1128.6 *t* |
| 4-Heptanone | 844.12 | 847.33 | 850.64 | 854.1 *t* | 973.83 | 977.78 | 981.9 | 982.2 *t* | 1147.1 | 1154.2 | 1161.6 | 1169 *t* | 1157.51 | 1164.2 | 1171.2 | 1178.5 *t* |
| 5-Methyl-2-hexanone | 828.95 | 832.2 *p* | 835.42 | 838.71 | 959.6 | 963.4 *p* | 967.27 | 971.26 | 1144.6 | 1152 *p* | 1159.6 | 1167.2 | 1149.09 | 1155.5 *p* | 1162.09 | 1168.85 |
| 3-Heptanone | 848.19 | 851.4 | 854.7 *p* | 858.05 | 979.9 | 983.88 | 988.0 *p* | 992.26 | 1155.2 | 1161.9 | 1169 *p* | 1176.4 | 1167.86 | 1174.37 | 1181.1 *p* | 1188.19 |
| 2-Heptanone | 866.4 *t* | 869.81 | 873.17 | 876.57 | 1000 *t* | 1004.2 | 1008.3 | 1012.5 | 1184 *t* | 1190.5 | 1196.8 | 1203.3 | 1193 t | 1199.28 | 1205.39 | 1211.65 |
| 2-Methyl-3-heptanone | 913.03 | 916.5 *t* | 920.2 | 924.1 *p* | 1025.1 | 1029.6 *t* | 1034.4 | 1039.6 *p* | 1201.2 | 1209 *t* | 1217.5 | 1226 *p* | 1182.35 | 1191 *t* | 1200.44 | 1210.2 *p* |
| 5-Methyl-3-heptanone | 913.32 | 917.46 | 921.85 | 926.5 | 1027.7 | 1032.9 | 1038.4 | 1044.3 | 1209.4 | 1217.9 | 1226.9 | 1236.2 | 1190.79 | 1200.26 | 1210.14 | 1220.36 |
| 3-Octanone | 958.69 | 962.4 *p* | 966.3 | 970.4 | 1085.5 | 1090.3 *p* | 1095.4 | 1100.8 | 1270.5 | 1276 *p* | 1282.2 | 1288.2 | 1267.75 | 1274.3 *p* | 1280.94 | 1287.6 |
| 5-Nonanone | 1043.9 *t* | 1048.7 | 1053.7 | 1058.7 | 1171 *p* | 1176.2 | 1182.6 | 1188.4 | 1337 *p* | 1341.3 | 1345.1 | 1348.6 | 1336.5 *p* | 1340.91 | 1344.99 | 1348.77 |
| Acetaldehyde | 365.68 | 368.7 *p* | 372.06 | 375.7 | 486.82 | 489.8 *p* | 492.9 | 496.37 | 618.62 | 621.7 *p* | 625.1 | 628.9 | 715.75 | 717.5 *p* | 719.28 | 721.23 |
| Propanal | 450.16 | 455.8 *t* | 461.61 | 467.6 | 581.5 | 586.95 *t* | 592.5 | 598.24 | 749.43 | 756.5 *t* | 763.65 | 770.9 | 818.74 | 823.0 *t* | 827.28 | 831.53 |
| Acroleine | 465.3 *t* | 467.85 | 470.71 | 473.85 | 607.2 *t* | 610.57 | 614.2 | 618.06 | 742.9 *t* | 747.37 | 752.14 | 757.3 | 861.7 t | 867.31 | 872.97 | 878.67 |
| Isobutanal | 534.06 | 538.33 | 542.4 *p* | 546.2 | 648.89 | 652.99 | 656.9 *p* | 660.6 | 802.11 | 808.12 | 813.8 p | 819.4 | 828.99 | 832.84 | 836.6 *p* | 840.22 |
| Butanal | 564.82 | 570.59 | 576.09 | 581.3 *t* | 691.15 | 695.57 | 701.74 | 706.7 *t* | 841.05 | 749.15 | 856.8 | 864.0 *t* | 894.54 | 899.4 | 904.1 | 908.6 *t* |
| Trimethylacetsldehyde | 569.3 | 571.6 | 573.87 | 576.1 *p* | 677.3 | 679.84 | 682.38 | 684.9 *p* | 844.45 | 848.7 | 852.9 | 857.1 *p* | 839.17 | 842.81 | 846.47 | 850.2 *p* |
| Isovaleralehyde | 640.62 | 643.48 | 646.3 *p* | 649.02 | 758.3 | 761.28 | 764.2 *p* | 767.1 | 928.78 | 933.49 | 938.26 | 942.58 | 938.7 | 942.45 | 946.1 *p* | 949.82 |
| Methylmutanal | 640.7 *t* | 643.61 | 646.4 | 649.14 | 758.4 *t* | 761.43 | 764.36 | 767.25 | 928.9 *t* | 933.67 | 938.1 *p* | 942.75 | 938.9 *t* | 942.65 | 946.34 | 950.01 |
| Valeraldehyde | 670.49 | 673.95 | 677.3 *t* | 680.4 | 797.6 | 801.16 | 806.0 *t* | 807.88 | 961.17 | 967.75 | 974.0t | 979.98 | 995.06 | 999.82 | 1004.4 *t* | 1008.93 |
| 2-Butenal | 608.8 *p* | 612.6 *t* | 616.77 | 621.23 | 774.9 *p* | 779.3 *t* | 784.0 | 788.9 | 982.3 *p* | 988.9 *t* | 995.79 | 1002.9 | 1072.5 *p* | 1078T | 1083.51 | 1089.01 |
| 3,3-Dimethylbutanal | 687.62 | 690.9 p | 694.12 | 697.4 | 798.74 | 802.4 *p* | 806.0 | 809.7 | 982.92 | 989.6 *p* | 996.2 | 1002.9 | 961.26 | 966.99 *p* | 972.52 | 978.74 |
| 2-Ethylbutanal | 740.22 | 743.4 t | 746.48 | 749.58 | 859.68 | 863.2 *t* | 866.7 | 870.13 | 1024.8 | 1032 *t* | 1039.9 | 1047.2 | 1033.88 | 1039.7 *t* | 1045.58 | 1051.4 |
| Hexanal | 768.86 | 771.9 | 77493 | 777.94 | 899.05 | 902.56 | 906.1 | 909.54 | 1067.9 | 1075.8 | 1083.5 | 1091.1 | 1095.09 | 1101.11 | 1107.08 | 1113.03 |
| Heptanal | 875.66 | 878.44 | 881.4 *t* | 884.4 | 1007.9 | 1011.3 | 1014.9 *t* | 1018.7 | 1177.3 | 1184.2 | 1191 *t* | 1198.7 | 1199.14 | 1204.92 | 1210.9 *t* | 1217.09 |
| 2-Ethylhexanal | 926.87 | 930.69 | 934.73 | 938.9 *p* | 1043.1 | 1047.9 | 1053.0 | 1058.3 *p* | 1207.1 | 1216.6 | 1226.3 | 1236 *p* | 1209.49 | 1218.15 | 1227.08 | 1236.2 *p* |
| Octanal | 974.50 | 977.01 | 979.8 *t* | 982.8 *p* | 1105.9 | 1109.4 | 1113.1 *t* | 1117.2 *p* | 1273.0 | 1277.8 | 1283 *t* | 1288 *p* | 1286.04 | 1290.71 | 1295.6 *t* | 1300.6 *p* |

* *p* is referring to prediction set and *t* is referring to test set other data set are used in stationary training set; The conditions (stationary phase and temperature) and notations are shown in Table I.

### ANN

An ANN is a biologically inspired computer program designed to learn from data in a manner of emulating the learning pattern in the brain. Most ANN systems are very complex high-dimension processing systems. Training of the ANN can be performed using the back-propagation algorithm. In order to train the network using the back-propagation algorithm, the differences between the ANN output and its desired value are calculated after each training iteration and the values of weights and biases modified using these error terms. A detailed description of the theory behind a neural network has been adequately described elsewhere (26–28). In the present work, an ANN program was written in FORTRAN 77 in our laboratory. This network was feed-forward fully connected with three layers with sigmoidal transfer functions. The inputs of this network are descriptors, which are selected by a stepwise MLR feature selection technique. The value of each input was divided into its mean value to bring them into dynamic range of the sigmoid transfer function of the network. The initial values of weights were randomly selected from a uniform distribution that ranged between –0.3 and +0.3, and the initial values of biases were set at one. These values were optimized during the network training. The back-propagation algorithm was used to train the network. Before training, the network parameters were optimized. These parameters are: number of nodes in the hidden layer, weights and biases learning rates, and the momentum. Procedures for the optimization of these parameters were reported in our previous papers (29–32). Then the optimized network was trained using a training set for the adjustment of weight and bias values. It is known that a neural network can become over-trained. An over-trained network has usually perfectly learned the stimulus pattern it has seen, but cannot give an accurate prediction for unseen stimuli, and is no longer able to generalize. There are several methods for overcoming this problem. One method is to use a test set to evaluate the prediction power of the network during its training. In this method, after each 1000 iterations, the network was used to calculate the retention indices of molecules included in the test set. To maintain the predictive power of the network at a desirable level, training was stopped when the value of error for the test set started to increase. Because the test error is not a good estimate of the generalization error, the prediction potential of the model was evaluated on a third set of data, named the prediction set. The compounds in the prediction set were not used during the training process and were reserved to evaluate the predictive power of the ANN model.

## Results

Tables I and II show the data set and corresponding ANN predicted values of retention indices of all molecules studied in this work, respectively. Table III shows the best-combined MLR model, which contains three molecular descriptors that encode the structural features of molecules, polarity of stationary phase, and column temperature. These molecular descriptors are: path one molecular connectivity index ($^1\chi$), fractional atomic charge weighted by partial positive surface area (FPSA3), and dipole moment (dp). Examination of the descriptors included in the combined model reveals that they encode different aspects of the molecular structures. These parameters mainly show the topological and electronic characteristics of these molecules. From the values of the mean effect that calculated for each descriptor, it can be concluded that the most important descriptor in the model is path one connectivity index ($^1\chi$), with the largest mean effect. The appearance of this descriptor in the model reveals the contribution of steric interaction on molecular retention in GC. The positive mean effect for this descriptor indicates that the molecule with a higher value of $^1\chi$ will have a higher retention. This relationship may be explained thus: that the magnitude of intermolecular interaction between the solute and the stationary phase is directly related to the size of the molecule (33). Another descriptor was FPSA3; this quantum chemical descriptor clearly indicates the importance of charge distribution of the solute on their retention in GC due to electronic interaction. The remaining molecular descriptor was dipole moment of solute, which has an electronic nature. The positive sign of the mean effects for these two descriptors reveals that an increase in these descriptors causes an increase in molecular retention due to an increase in solute-stationary phase interactions. Other parameters in the MLR model are polarity of stationary phases and column temperature, which indicate the contribution of experimental conditions in the combined MLR model.

The next step was to construct the ANN model. The data used in ANN was a matrix containing eight columns: three molecular descriptors, temperature, and retention index in four stationary phases. Because we have the retention data for 35 molecules at four different temperatures, this matrix has 140 rows. The data matrix is randomly divided into three groups (training, test, and prediction set), each of which consists of 100, 20, and 20 rows, respectively. The training set was used to train the network, the test set was used to avoid overtraining, and the prediction set was used to evaluate the

**Table III. Specification of Combined Multiple Linear Regressions Model**

| Descriptor | Notation | Coefficient | Mean effect |
|---|---|---|---|
| Path one connectivity index | $^1\chi$ | 193.752 ± (3.637) | 61.280 |
| Fractional atomic charge weighted partial positive surface area | FPSA3 | 5748.301 ± (907.088) | 7.289 |
| Dipole moment | dp | 41.749 ± (11.357) | 4.228 |
| Polarity of column | P | 0.160 ± (0.004) | 50.301 |
| Temperature | T | 0.199 ± (0.1236) | 1.812 |
| Constant 1 | | –437.578 ± (66.770) | –7.539 |

prediction power of the ANN model. Before training the network, the parameters of the number of nodes in the hidden layer, weights and biases learning rates, and momentum values were optimized. Table IV shows the architecture and specification of the optimized network. The optimized ANN has 4-3-4 topology. The ANN inputs are three molecular descriptors and column temperatures, while each node in the output layers represents the retention index (RI) of the molecule of interest on one stationary phase (four outputs nodes represent the RI on four stationary phases). Thus, entering one row of molecular descriptors ($^1\chi$, FPSA3, dp, and temperature), this network simultaneously predicts the retention indices of this molecule in HP-1, HP-50, DB-210, and HP-Innowax columns. After optimization of the ANN parameters, the network was trained for the adjustment of weight and bias values. In order to evaluate the predictive power of the ANN model, the trained network was used to calculate the retention indices for molecules in the prediction set. The overall average of relative error between ANN calculated and experimental values of retention indices were 7.2%. To compare the applied chemometric methods of MLR and ANN in predicting the retention indices, some statistics for these models are calculated and are shown in Table V. As can be seen from this table, all statistics have improved considerably in the case of the ANN model. The MLR standard error values of 47.441, 46.848, and 50.473 were obtained for training, test, and prediction sets, respectively, which should be compared with corresponding values of 8.047, 9.881, and 10.479 for the ANN model. Also, the standard error of the ANN model is comparable with the level of experimental uncertainty in the determination of retention indices in GC. Figure 1 shows a plot of the calculated versus the experimental values of RIs for the prediction set. A correlation coefficient of 0.999 for this plot confirms the ability of the ANN model to simultaneously predict RIs of four columns at different temperatures. The residuals of the ANN calculated values of the RI are plotted against their experimental values in Figure 2. The propagation of the residuals on both sides of the zero line indicates that no systematic error exists in the development of the neural network. In a comparison of the present model with other previously reported models (11–13), it was concluded that the present QSPR model is able to simultaneously calculate the retention indices of aliphatic aldehydes and ketones of four stationary phases at different temperatures using three molecular descriptors and only one ANN model.

**Table IV. Architecture and Specification of the ANN Model**

| Value | Parameter |
|---|---|
| 4 | No. of nodes in the input layer |
| 3 | No. of nodes in the hidden layer |
| 4 | No. of nodes in the output layer |
| 0.7 | Weights learning rate |
| 0.5 | Biases learning rate |
| 0.5 | Momentum |
| Sigmoid | Transfer function |

**Table V. The Statistical Parameters Obtained Using the ANN and MLR Models***

| | MLR | | | ANN | | |
|---|---|---|---|---|---|---|
| Group | F | SE | R | F | SE | R |
| Training | 6832.428 | 47.441 | 0.972 | 40851.45 | 8.047 | 0.999 |
| Test | 1184.900 | 46.848 | 0.969 | 31754.99 | 9.881 | 0.999 |
| Prediction | 1321.585 | 50.473 | 0.972 | 33991.69 | 10.479 | 0.999 |

## Conclusion

In this work, MLR and ANN are used as feature mapping techniques for the prediction of the retention indices of some aliphatic aldehydes and ketones. The optimized 4-3-4 ANN model with three molecular descriptors appearing in the MLR

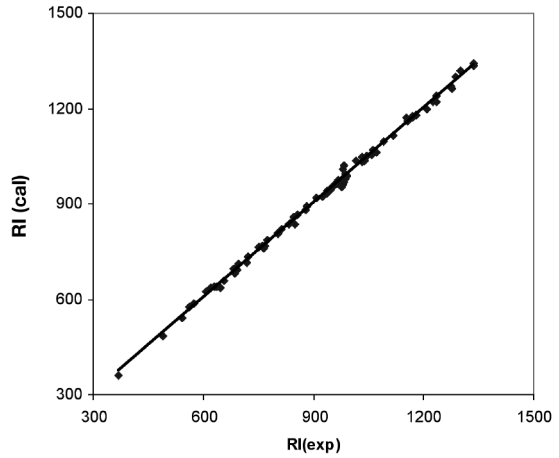


**Figure 1.** Plot of the calculated retention indices against the experimental value.

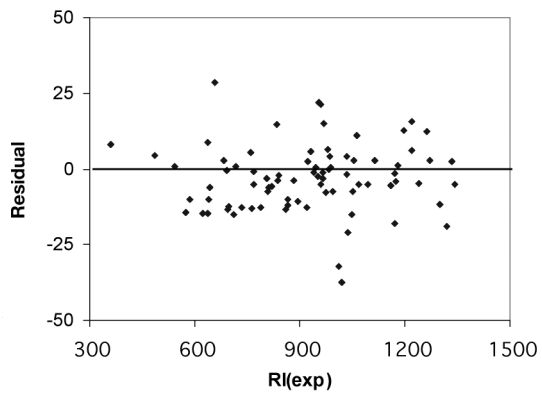


**Figure 2.** Plot of the residuals versus experimental values of retention indices.

model as its inputs showed a remarkable improvement over the linear model. The capability of this model in the simultaneous prediction of retention indices for aldehydes and ketones at four stationary phases in different temperatures adds another dimension of application of the ANN in QSPR studies. Also, the appearance of path one connectivity indices, dipole moment, and fractional atomic charge weighted by partial positive surface area in the model indicate the importance of the steric and electronic interactions in the molecular retention in GC.

## References


1. L.S. Ander, P.C. Jurs, and P.A. Edwards. Quantitative structure-retention relationship studies of odor-active aliphatic compounds with oxygen-containing functional groups. *Anal. Chem.* **62:** 2676–2684 (1990).
2. Zhihua Lin, Shushen Lin, and Zhiliang Li. Molecular modeling of quantitative structure retention relationship studies: Retention behavior of polychlorinated dibenzofurans on gas chromatographic stationary phases of varying polarity by a novel molecular distance edge vector. *J. Chromatogr. Sci.* **40:** 7–13 (2002).
3. M.H. Fatemi. Simultaneous modeling of the Kovats retention indices on OV-1 and SE-54 stationary phases using artificial neural networks. *J. Chromatogr. A* **955:** 273–280 (2002).
4. E.C.Y. Chan, W.L. Tan, P.C. Ho, and L.J. Fang. Modeling Caco-2 permeability of drugs using immobilized artificial membrane chromatography and physicochemical descriptors. *J. Chromatogr. A* **1072:** 159–168 (2005).
5. Q.S. Xu, D.L. Massart, Y.Z. Liang, and K.T. Fang. Two-step multivariate adaptive regression splines for modeling a quantitative relationship between gas chromatography retention indices and molecular descriptors. *J. Chromatogr. A* **998:** 155–167 (2003).
6. A.G. Fragkaki, M.A. Koupparis, and C.G. Georgakopoulos. Quantitative structure–retention relationship study of α-, β1-, and β2-agonists using multiple linear regression and partial least-squares procedures. *Anal. Chim. Acta.* **512:** 165–171 (2004).
7. T.F. Woloszyn and P.C. Jurs. Quantitative structure-retention relationship studies of sulfur vesicants. *Anal. Chem.* **64:** 3059–3063 (1992).
8. N. Dimov and A. Osman. Selection of molecular descriptors used in quantitative structure-gas chromatographie retention relationships : II. Isoalkanes and alkenes. *Anal. Chim. Acta.* **323:** 15–25 (1996).
9. J. Kang, C. Cao, and Z. Li. Quantitative structure–retention relationship studies for predicting the gas chromatography retention indices of polycyclic aromatic hydrocarbons: Quasi-length of carbon chain and pseudo-conjugated system surface. *J. Chromatogr. A* 799: 361–367 (1998).
10. O. Farkas and K. Heberger. Comparison of ridge regression, partial least-squares, pairwise correlation, forward- and best subset selection methods for prediction of retention indices for aliphatic alcohols. *J. Chem. Inf. Model.* **45:** 339–346 (2005).
11. R.D.M.C. Amboni, B.S. Junkes, R.A. Yunes, and V.E.F. Heinzen. Quantitative structure-property relationship study of chromatographic retention indices and normal boiling points for oxo compounds using the semi-empirical topological method. *J. Mol. Struct.* (Theochem.) **586:** 71–80 (2002).
12. B.S. Junkes, R.D.M.C. Amboni, R.A. Yunes, and V.E.F. Heinzen. Application of semi-empirical topological index in quantitative structure-chromatographic retention relationship studies of aliphatic ketones and aldehydes on stationary phases of different polarity. *J. Braz. Chem. Soc.* **15:** 183–189 (2004).
13. B. Ren. Atom-type-based AI topological descriptors for quantitative structure-retention index correlations of aldehydes and ketones. *Chemom. Intell. Lab. Syst.* **66:** 29–39 ( 2003).
14. S.P. Niculescu. Artificial neural networks and genetic algorithms in QSAR. *J. Mol. Struc.*( Theochem.) **622:** 71–83 (2003).
15. T.L. Chiu and S.S. So. Development of Neural Network QSPR Models for Hansch Substituent Constants. 1. Method and Validations. *J. Chem. Inf. Comput. Sci.* **44:** 147–153 (2004).
16. R. Zhang, A. Yan, M. Liu, H. Liu, and Z. Hu. Application of artificial neural networks for prediction of the retention indices of alkylbenzenes. *Chemom. Intell. Lab. Syst.* **45:** 113–120 (1999).
17. Karuna Katsuwon, Kornkanok Aryusuk, and Kanit Krisnangkura. Prediction of gas chromatographic retention times of esters of long chain alcohols and fatty acids. *J. Chromatogr. Sci.* **44:** 148–154 (2006).
18. J. Zupan and J. Gasteiger. Neural networks: A new method for solving chemical problems or just a passing phase? *Anal. Chim. Acta.* **248:** 1–30 (1992).
19. K. Héberger and M. Görgényi. Principal component analysis of Kováts indices for carbonyl compounds in capillary gas chromatography. *J. Chromatogr. A.* **845:** 21–31 (1999).
20. A.R. Katritzky, V.S. Lobanov, and M. Karelson. *Comprehensive Descriptors for Structural and Statistical Analysis.* Reference Manual, Ver. 2.0, 1994.
21. A.R. Katritzky, V.S. Lobabanov, and M. Karelson. QSPR: The correlation and quantitative prediction of chemical and physical properties from structure. *Chem. Soc. Rev.* **24:** 279–287 (1995).
22. A.R. Katritzky, V S. Lobabanov, and M. Karelson. QSPR as a means of predicting and understanding chemical and physical properties in terms of structure. *Pure & Appl. Chem.* **69:** 245–248 (1997).
23. Hyperchem 4.0. Hypercube, 1994.
24. J.P.P. Stewart. MOPAC 6.0, *Quantum Chemistry Program Exchange,* QCPE, No. 455, Indiana University, Bloomington, IN, 1989.
25. M.J.S. Dewar, E.G. Zoebisch, E.F. Healy. and J.J.P. Stewart. Development and use of quantum mechanical molecular models 76. AM1: a new general-purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **107:** 3902–3909 (1985).
26. S. Haykin. *Neural Network.* Prentice-Hall, Englewood Cliffs, NJ, 1994, pp. 145–187.
27. M.T. Beal, H.B. Hagan, and M. Demuth. *Neural Network Design.* PWS, Boston, MA, 1996, pp. 75–92.
28. N.K. Bose and P. Liang. *Neural Network Fundamentals.* McGraw-Hill, New York, 1996, pp. 241–250.
29. M. Jalali-Heravi and M.H. Fatemi. Simulation of mass spectra of noncyclic alkanes and alkenes using artificial neural network. *Anal. Chim. Acta.* **415:** 95–103 (2000).
30. M. Jalali-Heravi and M.H. Fatemi. Prediction of flame ionization detector response factors using an artificial neural network. *J. Chromatogr. A* **825:** 161–169 (1998).
31. M.H. Fatemi. Prediction of the electrophoretic mobilities of some carboxylic acids from theoretically derived descriptors. *J. Chromatogr. A* **1038:** 231–237 (2004).
32. M. Jalali-Heravi and M.H. Fatemi. Prediction of thermal conductivity detection response factors using an artificial neural network. *J. Chromatogr. A* **897:** 227–235 (2000).
33. M. Randic. The structural origin of chromatographic retention data. *J. Chromatogr. A* **161:** 1–14 (1978).


Manuscript received August 23, 2006;
Revision received June 6, 2007.