

# A New Neural Network–Group Contribution Method for Estimation of Flash Point Temperature of Pure Components

Farhad Gharagheizi,<sup>\*,†</sup> Reza Fareghi Alamdari,<sup>‡</sup> and Mahmood Torabi Angaji<sup>†</sup>

Department of Chemical Engineering, Faculty of Engineering, University of Tehran, P.O. Box 11365-4563, Tehran, Iran, and Department of Chemistry, Faculty of Materials and Chemical Engineering, Malek-Ashtar University of Technology, Lavizan, Tehran, Iran

Received December 12, 2007. Revised Manuscript Received January 30, 2008

In the present study, a new collection of 79 functional groups are used to correlate flash point temperature (FP) of pure components. These functional groups construct an accurate neural network–group contribution correlation to estimate flash point of pure components. For developing the model, 1378 pure components of various chemical families are used. Therefore, the model can be utilized to estimate the FP of pure components without any basic limitations.

## 1. Introduction

Flash point (FP) temperature is defined as the lowest temperature at which a liquid produces enough vapor to ignite in air at atmospheric pressure when an ignition source such as an external flame, for instance, is applied under specified test conditions.<sup>1</sup> Above the FP, a liquid is capable of producing enough vapor to form a flammable mixture with air. This is important information for the safe transportation, storage, and use of flammable liquids.

Many correlations have been reported in the literature to estimate the FP of pure chemicals. These correlations have been extensively reviewed by Catoire and Naudet<sup>1</sup> and Vidal et al.<sup>2</sup> We can classify these correlations into three main classes.

Class 1 contains those correlations which need some other physical properties such as boiling point temperature, density, vapor pressure, critical properties, enthalpy of vaporization, and so on. From these classes of correlations, we can refer to the correlations presented by Prugh,<sup>3</sup> Fujii and Herman,<sup>4</sup> Patil,<sup>5</sup> Suzuki,<sup>6</sup> Satyarayana and Kakati,<sup>7</sup> Satyarayana and Rao,<sup>8</sup> Metcalfe and Metcalfe,<sup>9</sup> Hsieh,<sup>10</sup> and Catoire and Naudet.<sup>1</sup> These correlations have some important disadvantages. The accuracy of these correlations is directly related to the accuracy of the needed physical properties or methods used to estimate those physical properties. Furthermore, if only one of the needed properties is missing, no calculation can be performed to estimate the FP.

Class 2 contains the quantitative structure–property relationship (QSPR) correlations. In the QSPR methodology, many

molecular-based parameters which are called “molecular descriptors” are used.<sup>11–20</sup> Molecular descriptors are numeric characteristics of a pure component directly calculated from its molecular structure with special algorithms. Several molecular descriptors (usually less than 10 molecular descriptors) are then selected to correlate the desired property of pure components. There are several QSPR correlations to predict the FP of pure components such as the correlations proposed by Tetteh et al.,<sup>21</sup> Katritzky et al.,<sup>22,23</sup> and Gharagheizi and Alamdari.<sup>11</sup> These correlations are useful for prediction of the FP of pure components. The most important disadvantage of these correlations is the complex procedure of calculation of some molecular descriptors from the chemical structure. As a result, these correlations are not usually simple to apply.

Class 3 contains the well-known group contribution (GC) correlations. These correlations are simple to apply and their use in estimating many physical properties is preferred by scientists and engineers. Application of these methods for prediction of the FP of pure components is limited to the works of Albahri<sup>24</sup> and Pan et al.<sup>25</sup> The GC methodology presented by Albahri<sup>24</sup> has been presented to predict the FP of about 287 pure hydrocarbon components. Also, the model by Pan et al.<sup>25</sup> has been presented to estimate the FP of 92 pure alkanes. These

\* To whom correspondence should be addressed. Fax: +98 21 66957784. E-mail: fghara@ut.ac.ir.

<sup>†</sup> University of Tehran.

<sup>‡</sup> Malek-Ashtar University of Technology.

- (1) Catoire, L.; Naudet, V. *J. Phys. Chem. Ref. Data* **2004**, *33*, 1083.
- (2) Vidal, M.; Rogers, W. J.; Holste, J. C.; Mannan, M. S. *Process Saf. Prog.* **2004**, *23*, 47.
- (3) Prugh, R. W. *J. Chem. Educ.* **1973**, *50*, A85.
- (4) Fujii, A.; Herman, E. *J. Saf. Res.* **1982**, *13*, 163.
- (5) Patil, G. S. *Fire Mater.* **1988**, *12*, 127.
- (6) Suzuki, T. *J. Chem. Eng. Jpn.* **1991**, *24*, 258–261.
- (7) Satyanarayana, K.; Kakati, M. C. *Fire Mater.* **1991**, *15*, 97.
- (8) Satyanarayana, K.; Rao, P. G. *J. Hazard. Mater.* **1992**, *32*, 81.
- (9) Metcalfe, E.; Metcalfe, A. E. M. *Fire Mater.* **1992**, *16*, 153.
- (10) Hsieh, H. Y. *Fire Mater.* **1997**, *21*, 277.

(11) Gharagheizi, F.; Alamdari, R. F. *QSAR Combin. Sci.*, in press; doi: 10.1002/qsar.200730110.

(12) Gharagheizi, F. *QSAR Combin. Sci.* **2008**, *27*, 165.

(13) Gharagheizi, F.; Fazeli, A. *QSAR Combin. Sci.*, in press; doi: 10.1002/qsar.200730020.

(14) Gharagheizi, F.; Alamdari, R. F. *Fullerenes, Nanotubes, Carbon, Nanostruct.* **2008**, *16*, 40.

(15) Gharagheizi, F. *e-Polymers* **2007**, No. 114.

(16) Gharagheizi, F. *Comput. Mater. Sci.* **2007**, *40*, 159.

(17) Gharagheizi, F.; Mehrpooya, M. *Energy Convers. Manage.* **2007**, *48*, 2453.

(18) Gharagheizi, F. *Chemom. Intell. Lab. Syst.* **2008**, *91*, 177.

(19) Gharagheizi, F. *Thermochim. Acta* **2008**, *469*, 8.

(20) Vatani, A.; Mehrpooya, M.; Gharagheizi, F. *Int. J. Mol. Sci.* **2007**, *8*, 407.

(21) Tetteh, J.; Suzuki, T.; Metcalfe, E.; Howells, S. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 491.

(22) Katritzky, A. R.; Petrunkhin, R.; Jain, R.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1521.

(23) Katritzky, A. R.; Stoyanova-Slavova, I. B.; Dobchev, D. A.; Karelson, M. *J. Mol. Graphics Modell.* **2007**, *26*, 529.

(24) Albahri, T. *Chem. Eng. Sci.* **2003**, *58*, 3629.

Table 1. Functional Groups Used To Develop NNGC Model<sup>a</sup>

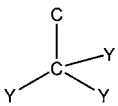
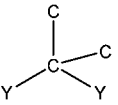
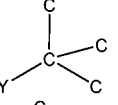
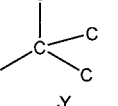
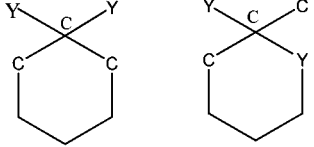
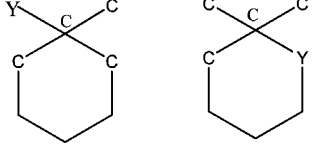
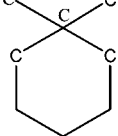
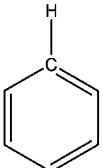
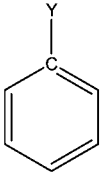
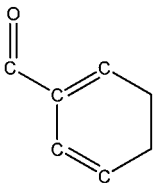
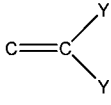
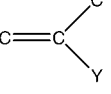
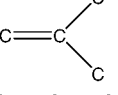
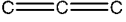
No.	ID	Chemical Structure	Comment
1	nCp		number of terminal primary C(sp3) Y = any terminal atom or heteroaromatic group (i.e. H, X, OH, NH2, etc.)
2	nCs		number of total secondary C(sp3) Y = H or any heteroatom
3	nCt		number of total tertiary C(sp3) Y = H or any heteroatom
4	nCq		number of total quaternary C(sp3)
5	nCrs		number of ring secondary C(sp3) Y = H or any heteroatom
6	nCrt		number of ring tertiary C(sp3) Y = H or any heteroatom
7	nCrq		number of ring quaternary C(sp3)
8	nCar	Sum of all the carbons belonging to any aromatic and heteroaromatic structure	number of aromatic C(sp2)
9	nCbH		number of unsubstituted benzene C(sp2)
10	nCb-		number of substituted benzene C(sp2) Y = carbon or any heteroatom
11	nCconj		number of non-aromatic conjugated C(sp2)
12	nR=Cp		number of terminal primary C(sp2) Y = any terminal atom or heteroaromatic group (i.e. H, X, OH, NH2, etc.)
13	nR=Cs		number of aliphatic secondary C(sp2) Y = H or any heteroatom
14	nR=Ct		number of aliphatic tertiary C(sp2)
15	n=C=		number of allenes groups

Table 1. Continued

No.	ID	Chemical Structure	Comment
16	nR#CH/X	$Y-C\equiv C$	number of terminal C(sp) Y = any terminal atom or heteroaromatic group (i.e. H, X, OH, NH <sub>2</sub> , etc.)
17	nR#C-	$Y-C\equiv C$	number of non-terminal C(sp) Y = C or any non-terminal heteroatom
18	nRNCO	$R-N=C=O$	number of isocyanates (aliphatic)
19	nArNCO	$Ar-N=C=O$	number of isocyanates (aromatic)
20	nRCOOH	$\begin{array}{c} HO \\   \\ C=O \\   \\ R \end{array}$	number of carboxylic acids (aliphatic)
21	nArCOOH	$\begin{array}{c} HO \\   \\ C=O \\   \\ Ar \end{array}$	number of carboxylic acids (aromatic)
22	nRCOOR	$\begin{array}{c} Y-O \\   \\ C=O \\   \\ R \end{array}$	number of esters (aliphatic) Y = Ar or R (not H) R = H or aliphatic group linked through C
23	nArCOOR	$\begin{array}{c} Y-O \\   \\ C=O \\   \\ Ar \end{array}$	number of esters (aromatic) Y = R or Ar
24	nRCONH <sub>2</sub>	$\begin{array}{c} H_2N \\   \\ C=O \\   \\ R \end{array}$	number of primary amides (aliphatic) R = H or aliphatic group linked through C
25	nRCONHR	$\begin{array}{c} Y-NH \\   \\ C=O \\   \\ R \end{array}$	number of secondary amides (aliphatic) Y = Ar or R (not H, not C = O) R = H or aliphatic group linked through C
26	nRCONR <sub>2</sub>	$\begin{array}{c} Y \\   \\ Y-N \\   \\ C=O \\   \\ R \end{array}$	number of tertiary amides (aliphatic) Y = Ar or R (not H, not C = O) R = H or aliphatic group linked through C
27	nRCOX	$\begin{array}{c} X \\   \\ C=O \\   \\ R \end{array}$	number of acyl halogenides (aliphatic)
28	nArCOX	$\begin{array}{c} X \\   \\ C=O \\   \\ Ar \end{array}$	number of acyl halogenides (aromatic)
29	nRCHO	$\begin{array}{c} H \\   \\ C=O \\   \\ R \end{array}$	number of aldehydes (aliphatic)
30	nArCHO	$\begin{array}{c} H \\   \\ C=O \\   \\ Ar \end{array}$	number of aldehydes (aromatic)
31	nRCO	$\begin{array}{c} R \\   \\ C=O \\   \\ R \end{array}$	number of ketones (aliphatic)
32	nArCO	$\begin{array}{c} Y \\   \\ C=O \\   \\ Ar \end{array}$	number of ketones (aromatic) Y = R or Ar

Table 1. Continued

No.	ID	Chemical Structure	Comment
33	nC=O(OR) <sub>2</sub>		number of carbonate (-thio) derivatives Y = O or S
34	nRCNO		number of oximes (aliphatic) Y = H, Ar or R
35	nRNH <sub>2</sub>		number of primary amines (aliphatic) R = aliphatic group linked through C (not C = O)
36	nArNH <sub>2</sub>		number of primary amines (aromatic)
37	nRNHR		number of secondary amines (aliphatic)
38	nArNHR		number of secondary amines (aromatic) Y = Ar or R (not C = O)
39	nRNR <sub>2</sub>		number of tertiary amines (aliphatic) R = aliphatic group linked through C (not C = O)
40	nArNR <sub>2</sub>		number of tertiary amines (aromatic) Y = Ar or R (not C = O)
41	nN-N		number of N hydrazines Y = C or H
42	nRCN		number of nitriles (aliphatic)
43	nArCN		number of nitriles (aromatic)
44	nN <sup>+</sup>		number of positive charged N
45	nRNO <sub>2</sub>		number of nitro groups (aliphatic) R = H or aliphatic group linked through carbon
46	nArNO <sub>2</sub>		number of nitro groups (aromatic) R = aromatic group linked through carbon
47	nROH		number of hydroxyl groups R = aliphatic group linked through any atom
48	nArOH		number of aromatic hydroxyls Ar = aromatic group linked through any atom
49	nOHp		number of primary alcohols
50	nOHs		number of secondary alcohols

Table 1. Continued

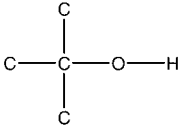
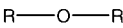
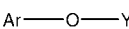
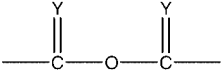
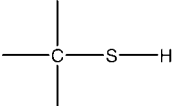
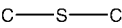

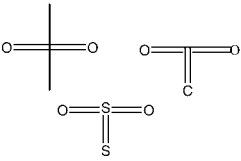
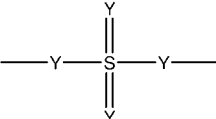
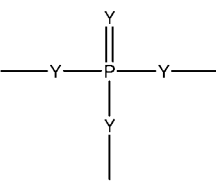
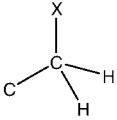
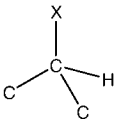
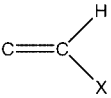
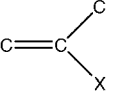
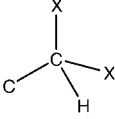
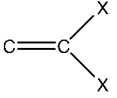
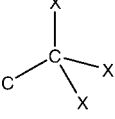
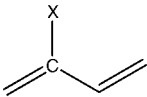







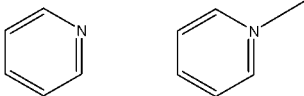
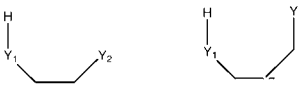
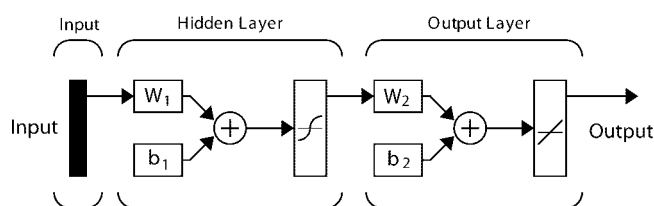
No.	ID	Chemical Structure	Comment
51	nOht		number of tertiary alcohols
52	nROR		number of ethers (aliphatic) R = aliphatic group linked through C (not C = O, not C # N)
53	nArOR		number of ethers (aromatic) Y = Ar or R (not C = O, not C # N)
54	nO(C=O)2		number of anhydrides (thio-) Y = O or S
55	nSH		number of thiols
56	nRSR		number of sulfides
57	nRSSR		number of disulfides
58	nS(=O)2		number of sulfones
59	nSO4		number of sulfates (thio- / dithio-) Y = O or S
60	nPO4		number of phosphates / thiophosphates Y = O or S
61	nCH2RX		number of CH2RX
62	nCHR2X		number of CHR2X
63	nR=CHX		number of R=CHX
64	nR=CRX		number of R=CRX
65	nCHRX2		number of CHRX2
66	nR=CX2		number of R=CX2
67	nCRX3		number of CRX3

Table 1. Continued

No.	ID	Chemical Structure	Comment
68	nArX	X—Ar	number of X on aromatic ring
69	nCconjX		number of X on exo-conjugated C
70	nAziridines		number of Aziridines
71	nOxiranes		number of Oxiranes
72	nPyrrolidines		number of Pyrrolidines
73	nOxolanes		number of Oxolanes
74	ntH-Thiophenes		number of tetrahydro-Thiophenes
75	nFuranes		number of Furanes
76	nThiophenes		number of Thiophenes
77	nPyridines		number of Pyridines
78	nHDon	Sum of the hydrogens linked to all of the Os and Ns in the molecule	number of donor atoms for H-bonds (N and O)
79	nHAcc		number of acceptor atoms for H-bonds (N, O, F) Y <sub>1</sub> = B, N, O, R, P, S Y <sub>2</sub> = N, O, F

<sup>a</sup> R = aliphatic chain linked through carbon unless otherwise stated. Ar = aromatic ring linked through carbon unless otherwise stated. X = halogen.



**Figure 1.** Schematic structure of the three-layer feed forward neural network (FFNN) used in this study.

methods are useful but their applications are limited to a particular group of materials.

In this work, a new neural network—group contributions (NNGC) method is presented to predict the FP of pure components. The most important aim of this study is to develop a universal model to estimate the FP of pure components.

## 2. Methodology

**2.1. Data Set.** For presenting a model, use of an accurate and precise data for the FP is needed. Therefore, the use of an evaluated database instead of other sources of data is preferred. One of the best evaluated databases presented for the FP of pure components, as well as other physical properties of pure

components, is DIPPR 801.<sup>26</sup> This database is recommended by the American Institute of Chemical Engineers (AIChE). The values of the FP of 1378 pure components were extracted from this database and were used in our work.

**2.2. Development of the New Group Contributions.** After providing the data set, the chemical structures of all 1378 components were analyzed and finally 79 functional groups were found to be useful to estimate the FP. The functional groups found and used in this study are extensively presented in Table 1.

These 79 functional groups and their availability in each of the 1378 pure components are presented as Supporting Information. These functional groups are used as input parameters for NNGC.

**2.3. Development of Neural Network—Group Contribution.** Neural networks are widely used in various scientific and engineering areas. These powerful tools are generally used to study complicated systems such as prediction of physical properties of pure components and mixtures based on their chemical structures.<sup>27</sup> The main advantage of neural network

(25) Pan, Y.; Jiang, J.; Wang, Z. *J. Hazard. Mater.* **2007**, 147, 424.

(26) Project 801, Evaluated Process Design Data, Public Release Documentation, Design Institute for Physical Properties (DIPPR), American Institute of Chemical Engineers (AIChE), 2006.

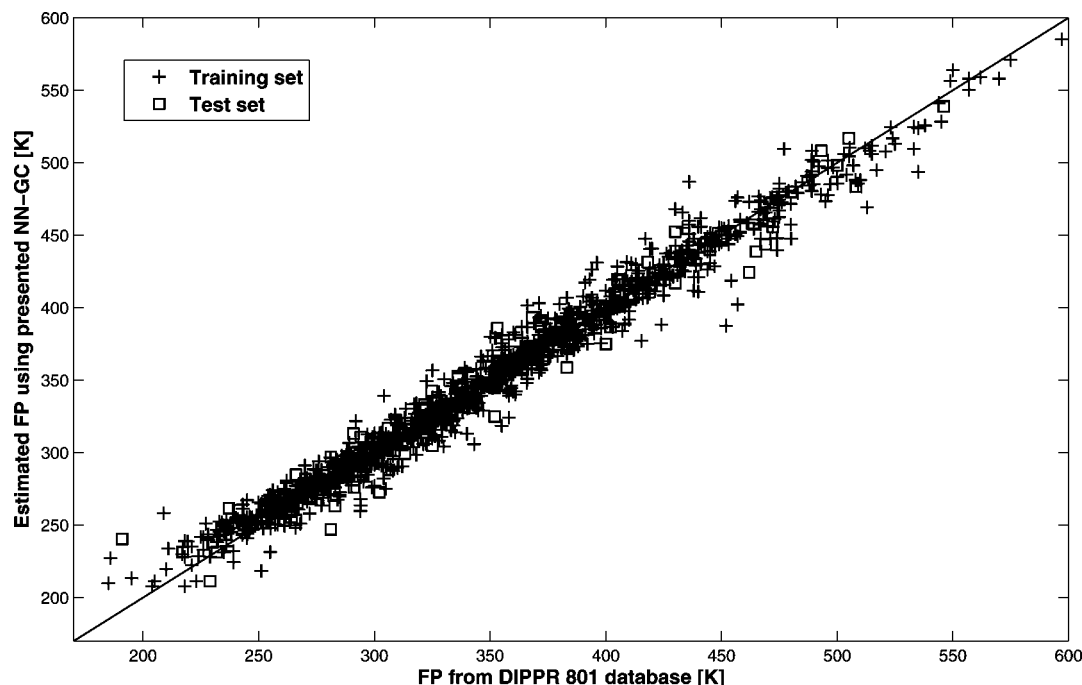


Figure 2. Comparison between the NNGC model and the DIPPR 801 data.

Table 2. Statistical Parameters of the Obtained NNGC Model

statistical parameter	value
Training Set	
$R^2$	0.9767
average absolute deviation	7.898
standard deviation error	10.985
root mean square error	10.967
$n$	1241
Test Set	
$R^2$	0.9661
average absolute deviation	9.94
standard deviation error	13.139
root mean square error	13.236
$n$	137
Training Set + Test Set	
$R^2$	0.9757
average absolute deviation	8.101
standard deviation error	11.198
root mean square error	11.206
$n$	1378

modeling is that complex, nonlinear relationships can be modeled without any assumptions of the form of the model.<sup>27</sup> The theoretical basis of neural networks has been extensively presented elsewhere.<sup>28–30</sup>

Three-layer feed forward neural networks (FFNNs) with the sigmoidal (hyperbolic tangent) transfer function are one of the most widely used types of neural networks that have found many applications in the prediction of physical properties.<sup>27</sup> These types of neural networks are available in the Neural Network Toolbox in MATLAB software (Mathworks Inc. software) and we selected this type of neural networks for developing our model.

The schematic structure of the three-layer FFNN used in this study is shown in Figure 1. Extensive explanations about the FFNN used in this study have been presented in our previous works.<sup>13–15</sup>

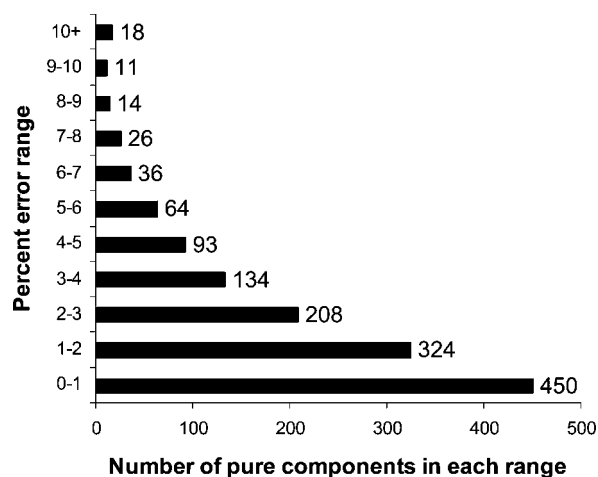


Figure 3. Percent errors obtained by using the presented NNGC model and the number of pure components in each range.

Table 3. Comparison between the Presented NNGC Model and Previous Models<sup>a</sup>

model	$R^2$	AE	$s$	$N$
Suzuki et al. <sup>6</sup>	0.9351	10.3	13.52	400
Tetteh et al. <sup>21</sup>	0.9326	10.2	13.1	400
Katritzky et al. <sup>22</sup>	0.902	—	16.1	271
Katritzky et al. <sup>23</sup>	0.878	13.9	—	758
Gharagheizi and Alamdari <sup>11</sup>	0.9669	10.2	12.7	1030
the NNGC	0.9757	8.101	11.206	1378

<sup>a</sup>  $R^2$ , AE,  $s$ , and  $n$  are squared correlation coefficients, average absolute error, root mean square error, and number of pure components used to develop models, respectively.

The FFNN has an input and an output. Input of the FFNN is the number of occurrences of the 79 functional groups in each pure component, and output of the FFNN is the estimated the FP by this FFNN.

Usually, all inputs and outputs of FFNN are normalized between  $-1$  and  $+1$  to decrease the calculation's error. We normalized inputs and outputs by means of the minimum and the maximum values of each molecular descriptor in the input matrix.

(27) Taskinen, J.; Yliruusi, J. *Adv. Drug Delivery Rev.* **2003**, 55, 1163.

(28) Bishop, C. M. *Neural networks for pattern recognition*; Oxford University Press: London, 1995.

(29) Fausett, L. V. *Fundamentals of neural networks*; Prentice Hall: New York, 1994.

The values of  $W_1$ ,  $W_2$ ,  $b_1$ , and  $b_2$  are obtained by minimization of an objective function which is commonly the sum of squares error between the outputs of neural network (estimated FP of pure components) and the target values (FP of pure components from data set). This minimization is usually performed by the Levenberg–Marquardt algorithm. This algorithm is rapid and accurate in the process of training neural networks.<sup>28–30</sup>

In most cases, the number of neurons in the hidden layer ( $n$ ) is fixed; then it is attempted to produce a neural network that can predict the target values as accurately as expected. This work is then repeated till the best neural network is obtained. In many cases, especially in three-layer FFNNs, it is better that, as a complementary work, the number of neurons in the hidden layer is optimized according to the accuracy of the obtained FFNN.

### 3. Results and Discussion

In the first step of constructing FFNN, the main data set is divided into two data sets: the first for training the network and the second for testing it. Neural networks are good at fitting functions and there is proof that a simple neural network can fit any data set very well.<sup>30</sup> As a result, for checking the predictive power of the neural network and also for prevention of overfitting, use of the test set is needed. The test set is only used for checking the produced neural network and is not used to train it. In this study, 90% of the main data set (1241 pure components) is used for the training set and 10% of the data set (137 pure components) is used for the test set. The components are randomly selected. The status of every pure component (belonging to the training set of the test set) used in our study can be found in the Supporting Information. Also, the upper and lower FP used in training set, and test are respectively 597, 185, 517, and 218 K.

Using our program, a FFNN was constructed. Then, the number of neurons of its hidden layer was optimized. The optimized structure of the obtained FFNN is 79-9-1 (the optimized number of neurons obtained is equal to 9). The estimated values of the FP for all 1378 pure components available in the data set are presented as Supporting Information. Also, the estimated values of these components in comparison to the data set values are shown in Figure 2. Our program in m-file format, and also the obtained FFNN in mat-file format, are available from the corresponding author (F. Gharagheizi) by e-mail.

The statistical parameters of the obtained FFNN are presented in Table 2.

The obtained results presented in Table 2 show that the squared correlation coefficient, average absolute deviation, standard deviation error, and root-mean-square error of the presented NNGC model over all 1378 pure components are 0.9757, 8.101, 11.198, and 11.206, respectively.

Also, the results show that the maximum percentage error obtained is about 26% and only the percent error of 19 pure components of 1378 pure components used is greater than 10%. These results are shown in detail in Figure 3. As shown, the percentage error of 450 pure components is less than 1%.

Comparison between the obtained model and previous models is not possible, because the data sets that previous researchers have been used are smaller than the data set used in this study. But, a comparison between models which their parameters are calculated only from chemical structures was performed and the results are presented in Table 3.

The comparison shows that the obtained NNGC is better than all of them.

### 4. Conclusion

In this study, a new NNGC was presented for prediction of the FP of pure components. The needed parameters of the model are the number of occurrence of each of the 79 functional groups in each molecule. Of course, most of the pure components do not have all of these 79 functional groups. These parameters can only be calculated from chemical structure of every pure component. Also, calculation of these parameters is easy. Therefore, the presented model is simple to use. In this study, 1378 pure components were used; thus, the obtained model does not have any basic applicability limit. The highest applicability limit of the presented model is the range of FP of pure components used to develop the model. The minimum FP used in this study is 185 K, and the maximum FP used in this study is 597 K. No further evaluation beyond this range was performed. As a result, the applicability range of the obtained model is between these two values. Also, although, the model seemed to be effective enough, it still cannot effectively distinguish the isomeric compounds.

**Supporting Information Available:** Table listing names of 1378 pure components used in this study and their FP extracted from DIPPR 801, and obtained using the model; also, the number of occurrences of each of the 79 functional groups in each molecule. This information is available free of charge via the Internet at <http://pubs.acs.org>.

EF700753T

(30) Hagan, M.; Demuth, H. B.; Beale, M. H. *Neural network design*; International Thompson Publishing: Tampa, FL, 2002.