Journal of the Serbian Chemical Society
JSCS-info@shd.org.rs • www.shd.org.rs/JSCS
J. Serb. Chem. Soc. *86* (7–8) 685–698 (2021)
JSCS–5454
Original scientific paper

New structure-based models for the prediction of normal boiling point temperature of ternary azeotropes

ZOHREH FARAMARZI[1], FATEMEH ABBASITABAR[2]*, HOSSEIN JALALI JAHROMI[1] and MAZIAR NOEI[1]

[1]*Department of Chemistry, Mahshahr Branch, Islamic Azad University, Mahshahr, Iran and* [2]*Department of Chemistry, Marvdasht Branch, Islamic Azad University, Marvdasht, Iran*

(Received 18 February, revised 6 April, accepted 13 April 2021)

Abstract: Recently, development of the QSPR models for mixtures has received much attention. The QSPR modelling of mixtures requires the use of the appropriate mixture descriptors. In this study, 12 mathematical equations were considered to compute mixture descriptors from the individual components for the prediction of normal boiling points of 78 ternary azeotropic mixtures. Multiple linear regression (MLR) was employed to build all QSPR models. Memorized_ACO algorithm was employed for subset variable selection. An ensemble model was also constructed using averaging strategy to improve the predictability of the final QSAR model. The models have been validated by a test set comprised of 24 ternary azeotropes and by different statistical tests. The resulted ensemble QSPR model had $R^2_{training}$, R^2_{test} and q^2 of 0.97, 0.95, and 0.96, respectively. The mean absolute error (*MAE*), as a good indicator of model performance, were found to be 3.06 and 3.52 for training and testing sets, respectively.

Keywords: QSPR modelling; azeotropic mixture; memorized ant colony optimization; mixture descriptors; ensemble multiple linear regression analysis.

INTRODUCTION

Azeotropes are among mixtures which exibit nonideal solution behaviour. They both present challenges and opportunities related to separation processes. Azeotropes cannot be separated easily, however, some certain separations would be carried out by means of the formation of an azeotrope. The ordinary distillation cannot be used directly to separate azeotropes due to the absence of an interphase compositional differential. Several special distillation methods including azeotropic distillation,[1–3] extractive distillation,[4,5] and pressure swing distillation[6] have been used for the azeotrope separation. The understanding of the azeotropic properties including the boiling point is important for the preliminary

* Corresponding author. E-mail: fabbasitabar@gmail.com
https://doi.org/10.2298/JSC210218035F

evaluations aimed at designing of the distillation process. With this background, the use of computer programs for accurate estimation of boiling points of azeotropes becomes highly effective.

There are a variety of tools available for the prediction of mixture behaviour, including group contribution methods (*e.g.*, ASOG[7], UNIFAC and diverse modified UNIFAC versions[8,9]), methods derived from the regular solution theory,[10] linear solvation relationships (LSER),[11] statistical thermodynamics[12] and solvation thermodynamics.[13] These methods are time–consuming and expensive especially when dealing about ternary mixtures. The intermolecular interactions in these mixtures are very complex and cannot be precisely represented by quantum chemistry or molecular dynamics because the demand for CPU is still significatly high (taking many days or even weeks of the computation). Therefore, a better alternative seems to be the use of quantitative structure–property relationships (QSPR).

QSPR models using computational molecular descriptors have been developed to predict various molecular properties. Few QSPR studies have been carried out to estimate properties of mixtures.[14,15] Computing of mixture descriptors is the most important problem facing in the QSPR study of mixtures. One solution to this problem is the combination of the molecular descriptors calculated for the individual constitutes via a specific mathematical formula,[16-18] Previously, we developed QSPR models of the prediction of normal boiling points of binary azeotropes.[19] 22 different mixing rules were applied to compute mixture descriptors. It was shown that twelve mixing rules resulted in satisfied QSPR models. The aim of this study is to develop QSPR models for the prediction of the boiling points of ternary azeotropes. In order to compute the mixture descriptors, 12 different mathematical formulas are employed. To the best of our knowledge, this is the first report of the QSPR studied for ternary azeotropes in which mixture descriptors are calculated from those of individual constitutes.

MATERIALS AND METHODS

Normal boiling point data used in this study were taken from literature.[20] Experimental boiling points along with the molecular compositions of all azeotrope mixtures are given in Tables S-I and S-II of the Supplementary material to this paper. Azeotropic compositions of this data have been characterized by molar fractions. Molecular structures of compounds constituting each mixtures were drawn in 2D ChemDraw.[21] After creating the corresponding 3D structures in CS Chem3D Ultra, the geometry optimization was performed in the molecular operation software (MOE) using the AM1 procedure.[22] For each compound, to characterize the molecular structure, a total of 3475 descriptors including 251 MOE descriptors and 3224 Dragon descriptors were calculated[23] and then classified into 22 different descriptor blocks. These descriptors were firstly investigated to find and remove those having missing values or containing small variation. Next, the correlations between descriptors were calculated and from a set of highly correlated descriptors the one with the best correlation with the boiling

points of ternary azeotropes was maintained and the others were eliminated. After these pretreatments, 328 molecular descriptors were obtained for each individual constituent.

Mixture descriptors

Commercial softwares are able to calculate the molecular descriptors for individual compounds. To perform QSPR studies for the property of mixtures, the mixture descriptors can be computed from the molecular descriptors of individuals using different approaches.[18,24] In this study, twenty–two different types of mixture descriptors obtained from the molecular descriptors of individuals were used.[19] This 22 formulas contained both non-linear and linear mathematical formulas (Table I). From these 22 mathematical formulas, only twelve ones could be employed for ternary mixtures and those that involved minus in their mathematical formulas were ignored. These twelve mixture descriptors were classified into three distinct blocks: 1) those that not require any experimental data; 2) those in which mole fraction data is needed; 3) those which require information about potential energy and/or mole fraction. Information about the potential energies were extracted from MOE software based on the following illustration: consider a ternary azeotrope contains three individuals, namely, A, B and C. The aim is to compare potential energy of a system contains ternary azeotrope, *e.g.*, ABC to three systems each of which contains only one of individuals. In order to make better comparison, the potential energy of ABC system is separately compared to AAA, BBB and CCC. If the potential energy of ABC is close, for example, to that of AAA, it is then assumed that A has the highest contribution in the stability of the ternary mixture and therefore A should be weighted

TABLE I. Types of mixture descriptors, description and their formulas; the parameters c, x and D refer to potential coefficient, mole fraction and molecular descriptor, respectively

Descriptor symbol	Formula	Description
Centroid	$(D_A+D_B+D_C)/3$	Mean of descriptors of pure compounds
fmol–sum	$x_AD_A+x_BD_B+x_CD_C$	Sum of weighted descriptors by mol fractions
sqr–fmol	$x_A^2D_A+x_B^2D_B+x_C^2D_C$	Sum of weighted descriptors by square of mol fractions
root–fmol	$\sqrt{x_A}D_A+\sqrt{x_B}D_B+\sqrt{x_C}D_C$	Sum of weighted descriptors by root of mol fractions
sqr–fmol–sum	$(x_AD_A+x_BD_B+x_CD_C)^2$	Square sum of weighted descriptors by mol fractions
norm–cont	$\sqrt{(x_AD_A)^2+(x_BD_B)^2+(x_CD_C)^2}$	Norm of mole fraction contribution descriptors
poten–sum	$c_AD_A+c_BD_B+c_CD_C$	Sum of weighted descriptors by potential coefficients
sqr–poten	$c_A^2D_A+c_B^2D_B+c_C^2D_C$	Sum of weighted descriptors by square of potential coefficients
root–poten	$\sqrt{c_A}D_A+\sqrt{c_B}D_B+\sqrt{c_C}D_C$	Sum of weighted descriptors by root of potential coefficients
sqr–poten–sum	$(c_AD_A+c_BD_B+c_CD_C)^2$	Square sum of weighted descriptors by potential coefficients
poten–norm–cont	$\sqrt{(c_AD_A)^2+(c_BD_B)^2+(c_CD_C)^2}$	Norm of potential coefficients contribution descriptors
fmol–poten–sum	$x_Ac_AD_A+x_Bc_BD_B+x_Cc_CD_C$	Sum of weighted descriptors by both mol fractions and potential coefficients

more than the other two individuals in the mixture descriptor calculation process. The potential energy coefficients, which show the contribution of each individual component in the stability of a ternary mixture, were obtained from Eqs. (1)–(3):

$$c_A = \frac{(E_{ABC} - E_{AAA})^2}{(E_{ABC} - E_{AAA})^2 + (E_{ABC} - E_{BBB})^2 + (E_{ABC} - E_{CCC})^2} \quad (1)$$

$$c_B = \frac{(E_{ABC} - E_{BBB})^2}{(E_{ABC} - E_{AAA})^2 + (E_{ABC} - E_{BBB})^2 + (E_{ABC} - E_{CCC})^2} \quad (2)$$

$$c_C = \frac{(E_{ABC} - E_{CCC})^2}{(E_{ABC} - E_{AAA})^2 + (E_{ABC} - E_{BBB})^2 + (E_{ABC} - E_{CCC})^2} \quad (3)$$

where E_{ABC}, E_{AAA}, E_{BBB}, and E_{CCC} are the total potential energies, kcal* mol^{-1}, for the systems including ABC, AAA, BBB and CCC, respectively. MNDO Hamiltonian gives rough estimates of these potential energies. A brief description along with the mathematical formulas for all types of mixture descriptors is shown in Table I.

Model construction and descriptor selection method

In this study, QSPR models for the prediction of the boiling points of ternary azeotropes were established by the multiple linear regression (MLR). MLR has several advantages including its simple form, fast construction, and easily interpretation. However, the need for an effective tool to select an appropriate subset of descriptors is the main disadvantage of MLR. In this study, the memorized ant colony optimization algorithm as a powerful descriptor selection tool (memorized_ACO) was used.[25] To read more about this algorithm, readers are referred to our previous papers.[19,25-27] The reliability of the final models, constructed based on the descriptors selected by memorized_ACO, algorithm would be investigated by using internal and external validations. Also, the statistical parameters of $R^2_{training}$, R^2_{test}, and q^2 are used to compare the models. $R^2_{training}$ and R^2_{test} indicate the squared correlation coefficients of training and test sets, respectively. q^2 is the squared correlation coefficient obtained *via* leave-one-out cross-validation (LOO-CV) procedure. Moreover, 5-fold cross-validation (5-CV),[27] Monte Carlo cross-validation (MCCV)[28] and double-cross validation (D-CV)[29] are considered for further validation of the resulted QSPR models.

RESULTS AND DISCUSSION

In the present work, twelve types of mixture descriptors were used to characterize the ternary azeotropes. These mixture descriptors are different in their mathematical formulas and all are calculated from the individuals' descriptors. The applicability of the proposed mixture descriptors in developing QSPR models for the prediction of boiling points of ternary azeotropes were investigated. Table I represents the required information about these mixture descriptors.

The comparison between different types of mixture descriptors would be easily performed, provided that a linear regression is applied. Consequently, MLR combined to the memorized_ACO algorithm was used to establish the

* 1 kcal = 4184 J

QSPR models. Fig. 1 shows the calculated q^2 values for MLR models developed for all types of mixture descriptors.

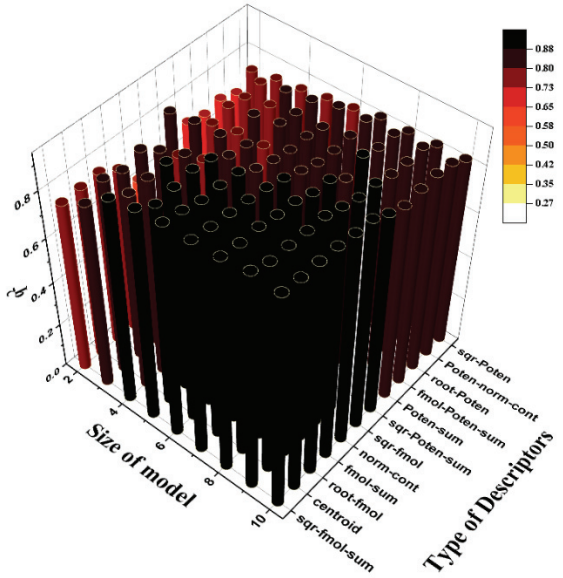

Fig. 1. q^2 values for QSPR models with different sizes constructed on the basis of different types of mixture descriptors.

It is worthy to note that MLR models with different sizes are generated for each type of mixture descriptor. That is, the momorized_ACO algorithm is run several times for each descriptor type while each time a specific number of descriptors is selected. MLR models of sizes 2–10 are investigated. Graphs related to $R^2_{training}$ and R^2_{test} are given in Figs. S-1 and S-2 of the Supplementary material. The statistical parameters for all types of QSPR models having different sizes are presented in Table S-III of the Supplementary material. More statistics for the each type of QSPR model can be found in Tables S-IV–S-XIII of the Supplementary material.

Analyzing Fig. 1 showed that all types of mixture descriptors proposed in this study are able in accessing predictive QSPR models. The values of $R^2_{training}$ and R^2_{test} of the best MLR models for each type of mixture descriptor were between 0.87 to 0.96 and 0.84 to 0.94, respectively. The value of q^2 was found to be in the range of 0.82–0.94. The training and the test sets had the values of MAE of 4.59 and 5.04 K, respectively, indicating the good prediction power of the best model. Based on the statistical parameters of the models, as it can be seen in Fig. 1, the goodness of the proposed mixture descriptors was as follows: sqr–fmol–sum > centroid > root–fmol > fmol–sum > norm–cont > sqr–fmol > sqr–poten–sum > poten–sum > fmol–poten–sum > root–poten > poten–norm–

–cont > sqr–poten. The best mixture descriptor in characterizing and developing predictive QSPR model for ternary azeotropic mixture was found to be sqr––fmol–sum. It should be emphasized again that all of the proposed mixture descriptors showed more or less similar results regarding the statistical parameters. For deep inspection, statistical parameters for QSPR models constructed on the basis of sqr–fmol–sum and sqr–poten are given in Tables II and III, respectively. The best QSPR model was constructed with six and seven descriptors for sqr––fmol–sum and sqr–poten descriptor types, respectively. Statistics of the resulted models for both types were good and found to be 0.96 and 0.87 for R^2_{training}, 0.94 and 0.84 for R^2_{test}, respectively. Statistical parameters for all other developed QSPR models are given in Tables S-IV–S-XIII.

TABLE II. Statistical parameters of the QSPR models based on sqr–fmol–sum descriptors; *F* stands for Fischer test

No. of descriptors	R^2_{training}	*RMSE*-Training	q^2	RMSE-CV	5-CV		MCCV		R^2_{test}	*RMSE*-Test	*F*
					q^2	*RMSE*	R^2_{training}	R^2_{test}			
2	0.79	9.37	0.76	9.98	0.75	10.15	0.79	0.73	0.74	11.29	94.42
3	0.85	7.83	0.82	8.63	0.82	8.70	0.85	0.82	0.85	6.70	95.46
4	0.92	5.72	0.91	6.24	0.90	6.38	0.92	0.86	0.93	4.16	142.36
5	0.95	4.70	0.93	5.34	0.93	5.49	0.95	0.92	0.92	4.51	169.46
6	**0.96**	**4.27**	**0.94**	**4.90**	**0.94**	**5.01**	**0.96**	**0.92**	**0.94**	**4.90**	**169.62**
7	0.96	3.93	0.94	4.87	0.94	5.07	0.96	0.93	0.94	4.53	169.12
8	0.96	3.84	0.95	4.61	0.94	4.83	0.97	0.93	0.95	3.83	152.15
9	0.98	3.07	0.96	3.81	0.96	4.04	0.98	0.96	0.97	3.64	209.62
10	0.97	3.39	0.96	4.26	0.95	4.48	0.97	0.93	0.97	2.96	150.12

TABLE III. Statistical parameters of the QSPR models based on sqr–poten descriptors

No. of descriptors	R^2_{training}	*RMSE*-Training	q^2	*RMSE*-CV	5-CV		MCCV		R^2_{test}	*RMSE*-Test	*F*
					q^2	*RMSE*	R^2_{training}	R^2_{test}			
2	0.77	9.80	0.74	10.40	0.73	10.65	0.77	0.74	0.75	10.81	84.16
3	0.79	9.31	0.76	10.05	0.75	10.26	0.79	0.73	0.78	10.19	62.67
4	0.82	8.72	0.77	9.82	0.75	10.08	0.81	0.77	0.81	9.69	54.16
5	0.83	8.42	0.78	9.42	0.78	9.65	0.83	0.77	0.81	10.09	46.31
6	0.84	8.09	0.80	9.11	0.79	9.33	0.85	0.79	0.80	10.28	41.58
7	**0.87**	**7.44**	**0.82**	**8.63**	**0.81**	**8.91**	**0.87**	**0.82**	**0.84**	**11.31**	**42.46**
8	0.87	7.31	0.82	8.65	0.81	9.02	0.88	0.80	0.87	9.16	44.19
9	0.88	7.14	0.82	8.51	0.81	8.79	0.88	0.82	0.84	9.08	39.86
10	0.88	7.03	0.83	8.48	0.80	9.36	0.89	0.80	0.87	12.31	31.58

It is worth to further analyze the best QSPR model that has been constructed on the basis of sqr–fmol–sum. For this model that contained six descriptors, MAEs of training and test sets were 3.39 and 3.69 K, respectively. More statistical parameters of this QSPR model are given in Table IV. The model is statistically sound since all *t* values are significant, *t* stands for *t*-student value at

95 % confidence interval. Variance inflation factors (VIFs) calculated for considered descriptors in the model are all less than 10, indicating the lack of collinearities among them.[30] The external validation of the model was further verified using Q^2_{F1}, Q^2_{F2}, Q^2_{F3}, r^2_m and CCC (concordance correlation coefficient).[31,32] For an acceptable model, the Q^2_{F1}, Q^2_{F2}, and Q^2_{F3} values should be greater than 0.6, r^2_m greater than 0.5, and CCC greater than 0.85. The values of Q^2_{F1}, Q^2_{F2}, Q^2_{F3}, $r^2_{m,}$ and CCC for the sqr–fmol–sum based QSPR model were found to be 0.92, 0.90, 0.94, 0.91, and 0.97, respectively. The corresponding values for other types of QSPR models are tabulated in Table S-XIV. To test the risk of chance correlation in this QSPR model, *y*-randomization test was employed. In this test, the vector containing information about the boiling points of ternary azeotropes is scrambled and then a MLR model with the same descriptors used is developed.

TABLE IV. Statistics for the best QSPR model based on sqr–fmol–sum descriptors

Descriptor	Definition	beta	*t*	*p*-value	*VIF*
Constant	–	349.75	561.81	$< 10^{-32}$	–
ZM2V	Second Zagreb index by valence vertex degrees	8.59	8.50	5×10^{-11}	2.58
X0A	Average connectivity index chi–0	–7.22	–5.00	8×10^{-6}	5.29
E2e	2nd component accessibility directional WHIM index / weighted by Sanderson electronegativity	5.25	4.24	1×10^{-4}	3.88
HATS3u	Leverage-weighted autocorrelation of lag 3 / unweighted	3.49	4.72	2×10^{-5}	1.38
R1u+	R maximal autocorrelation of lag 1 / unweighted	15.69	9.54	1×10^{-12}	6.85
vsurf_W1	Hydrophilic volume at −0.2 kcal mol^{-1}	23.16	13.04	3×10^{-17}	7.99

This iterative algorithm usually repeats 100 times. Next, the squared correla¬ti¬on coefficients for the models produced in this test are collected in a vector and compared to that for the best MLR model. The results are shown as bar plot in Fig. 2.

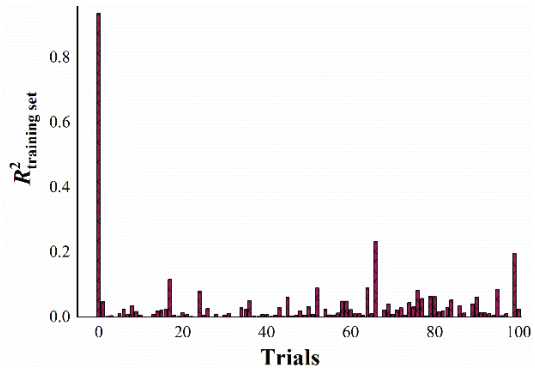

Fig. 2. *Y*-randomization test applied to the sqr–fmol–sum based QSAR model; the first bar shows the $R^2_{training}$ value for the original model.

Note that the first column represents R^2_{training} for the best QSPR model. It is clear that all of the models with scrambled property have poor quality. The same plots were provided for other types of mixture descriptors, given in Fig. S-3. These figures obviously showed that all the mixture descriptors proposed in this study had good ability to characterize mixture structures.

The presence of any outlier was checked through defining applicability domains (AD) of QSPR models. AD is a theoretical working space within which the property of a new compound can be accurately predicted. Several QSPR models with different type of descriptors that are capable to characterized ternary azeotropic mixtures were developed. Each QSPR model has been constructed on the basis of its own descriptor type, they would have different ADs. Different strategies have been introduced to define AD. The standardization approach reported by Roy et al. is the simplest one that was used in this study.[28,33] In this approach, the descriptor matrix of training set is autoscaled. The maximum and the minimum values of the autoscaled descriptors for each mixture are detected. If the minimum value exceeds 3, that compound should be treated as outlier. If the maximum value exceeds 3 and the minimum value is lower than 3, then a new measure (l) is to be computed for that compound by the following equation:

$$l = \bar{d} + 1.28 s_{\text{d}} \tag{4}$$

where $\bar{d}$ and s_{d} are the average and standard deviation of descriptor values for a compound, respectively. If the computed l value is more than 3, then that compound is an outlier.

No outliers within training set were detected for all types of QSPR models except for that constructed on the basis of sqr–poten–sum. Two mixtures (mixture numbers 6 and 39) were highlighted as outliers for the sqr–poten–sum model. Mixtures 6 and 39 contain water/carbon disulfide/ethanol and water/ethanol/hexane, respectively. The QSPR models had more complex behaviours regarding to the detected outs of domain. All samples in the test set were found to be within applicability domain for the root–sum based QSPR model whilst for other cases, one or two samples were detected out of domain. Mixture 7 was found to be out of applicability domain for seven QSPR models. This mixture contains water, carbon disulfide, and acetone. It is interesting that there are two mixtures in the data set in which carbon disulfide participated. They are mixtures 6 and 7. The former belongs to the training set and the later to the test set. Both of them were detected as outliers.

Plots of the actual versus predicted boiling points by all types of QSPR models for training and test sets are depicted in Fig. 3. According to this figure, a good correlation can be found between the experimental boiling points of ternary azeotropes and the predicted ones for all types of QSPR models developed in this study.

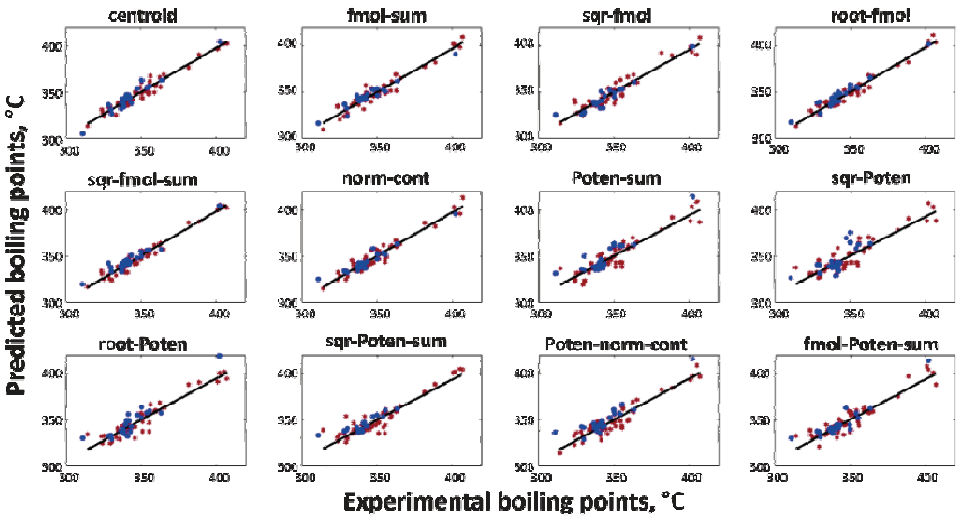

Fig. 3. Plot of observed *vs.* predicted boiling points by different QSPR models.

It was shown previously that prediction by an ensemble model would be more accurate than a single model.[19,34,35] Through a simple averaging strategy, an ensemble model was constructed using all of the QSPR models. R^2_{training}, R^2_{test} and q^2 of the ensemble model were found to be 0.96, 0.94, and 0.95, respectively. MAEs of training and test sets for the ensemble model were found to be 3.06 and 3.52, respectively. In comparison to the sqr–fmol–sum based QSPR model, that has been found to be the best QSPR model, the prediction by the ensemble model was improved.

A fit plot of experimental *versus* predicted boiling points of ternary azeotropes by the ensemble model for the training and test sets is shown in Fig. 4. The two outliers can be identified by visual inspection; one belongs to the training set

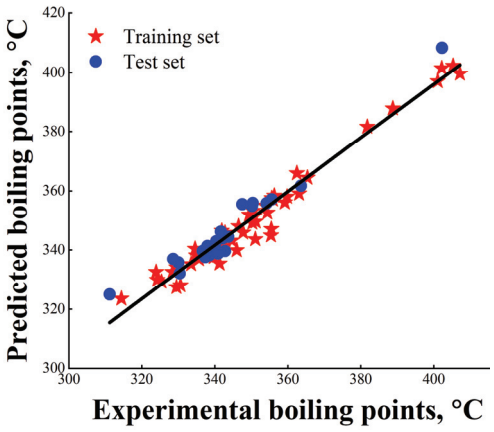

Fig. 4. Plot of observed *vs.* predicted boiling points by the ensemble QSPR model.

and the other to the test set. The constituent molecules of these outliers are water/ /carbon disulfide/acetone and water/carbon disulfide/ethanol. Both of them contain carbon disulfide. Since only these outliers contain carbon disulfide, maybe carbon disulfide can cause these mixtures to be outliers. Carbon disulfide has a linear structure and is a nonpolar compound with low basicity and zero acidity properties. Other statistical parameters of the ensemble model can be found in Table S-XIV.

To understand the predictive power of the resulted QSPR models better, the obtained models were compared with those published previously. Up to now, to the best of our knowledge, there is only one reported QSPR study for the prediction of boiling points of ternary azeotropes.[14] In this paper, the universal solvation equation and artificial neural network to predict boiling points of ternary azeotropes was used. MLR was also used for QSPR model development based on the universal solvation equation. Unfortunately, the MLR model over the whole data set of 78 azeotropes and no test set was considered successful. So, for better comparison the MLR models on the same training set were developed. The resulted statistical parameters along with those reported previously are given in Table V. The $R^2_{training}$ value for training set were reported 0.873 which was increased to 0.909 after the removal of two outliers. Since we used multiple linear regression with few independent variables, this comparison shows the preference of our models to the previously reported, in which non-linear fitting with large number of independent variables have been used.[14,36,37]

TABLE V. Comparison of statistics for the QSPR models with the previously reported

Ref.	Model	No. of descriptors	$R^2_{training}$	q^2	F	*RMSE*--Training	*RMSE*--CV
This study	fmol_sum	5	0.928	0.913	185.76	5.17	5.67
	sqr_fmol	6	0.910	0.890	119.54	5.79	6.39
	root_fmol	8	0.954	0.939	177.18	4.15	4.75
	sqr_fmol_sum	6	0.949	0.939	220.46	4.35	4.77
	norm_cont	6	0.932	0.912	162.37	5.02	5.72
	poten_sum	6	0.854	0.825	69.24	7.36	8.07
	sqr_poten	6	0.810	0.758	50.37	8.41	9.49
	root_poten	7	0.875	0.844	69.78	6.82	7.62
	sqr_poten_sum	5	0.856	0.836	85.85	7.31	7.81
	poten_norm_cont	6	0.865	0.832	75.94	7.08	7.90
	fmol_poten_sum	5	0.868	0.837	94.85	7.00	7.77
Oliferenk *et al.*[14]	Universal solvation equation	10	0.873	–	–	–	–

Details of the best model

The mathematical equation of the best QSPR model developed for each type of mixture descriptor is given in Table S-XIV. The best single QSPR model for

the prediction of boiling points of ternary azeotropes was found to be the one constructed on the basis of the sqr–fmol–sum mixture descriptor. This model consists of six descriptors including ZM2V, X0A, E2e, HATS3u, R1u+ and vsurf_W1>

$$T = 349.75 + 8.59(\pm 1.01)\text{ZMV2} - 7.22(\pm 1.44)\text{XOA} + 5.25(\pm 1.24)\text{E2e} + \\ + 3.49(\pm 0.74)\text{HATS3u} + 15.69(\pm 1.64)\text{R1u} + 23.16(\pm 1.78)\text{vsurf\_W1} \quad (5)$$

The vsurf_W1 has positive influence on the boiling points of ternary azeotropes since it pertinent to polarity, which means that the individual constitutes with more polarity result in a mixture with higher boiling point.[38] X0A which is corresponded to the molecular branching properties is a topological descriptor.[39] It has negative sign in Eq. (5), indicating that the more branching the individuals, the lower the boiling point of the resulted azeotrope.

ZM2V is the second Zagreb index by valence degrees. ZM2V is related to molecular branching. However, the sign of this descriptor in the QSPR model is positive because not only the molecular branching but also the valence degrees affect the amount of this descriptor. For example, ZM2V for 1-heptene and *n*-heptane were calculated as 26 and 20, respectively. The positive sign of ZM2V indicated that individual constitutes with unsaturated bonds cause to increase the boiling point of the resulted azeotrope.

E2e corresponds to 2nd component accessibility directional WHIM index/ /weighted by atomic Sanderson electronegativities. WHIM descriptors are built in such a way as to capture relevant 3D structural knowledge about molecular size, symmetry, shape, and distribution of atoms in relation to reference frames. Since E2e has positive sign in Eq. (5), it is concluded that by increasing the size and symmetry of the molecule, along with the presence of electronegative atoms, the boiling point increases.

HATS3u and R1u+ correspond to leverage–weighted autocorrelation of lag 3/unweighted and to R maximal autocorrelation of lag 1/unweighted, respectively. Both belong to the GETAWAY category. GETAWAYs are geometrical descriptors encode information on structural fragments and, to some extent, account for the information on molecular shape and size and for the specific atomic properties, as well. It was shown that they also encode information about the effective position of fragments and/or substituents in the molecular space.[28,40] Both of them have positive effects on the boiling points of ternary azeotropes.

CONCLUSIONS

Between mixtures, azeotropes exhibit non-ideal behaviour. Azeotropes cannot be separated using conventional distillation due to the lack of a combined phase differential. The estimation of boiling points of azeotropes for initial evaluation is very effective in the design of the distillation process.

In the present study, new and reliable boiling point prediction models for the ternary azeotropes were developed, based on different types of mixture descriptors. The mixture descriptors were calculated according to the molecular descriptors of the individual pure components involved in the mixture and their molar fraction using 12 different mixing rules. The subset of mixture descriptors was selected using memorized_ACO algorithm. Moreover, the multiple linear regression as a simple technique with easy interpretation was also employed to develop models based on selected mixture descriptors. The best QSPR model was found to be the one constructed on the basis of sqr–fmol–sum mixture descriptor. Due to the good predictive power of all models, all of them were considered in the construction of an ensemble model through a simple averaging strategy. The ensemble model prediction improved when compared to the best QSPR model. The validity of the proposed models was examined by the applying of internal and external validations together with the different statistical analysis. The results showed that the proposed models based on the mixed rules have a good ability to predict the boiling points of 78 ternary azeotropes studied. The proposed models for the prediction of the boiling points of ternary azeotropes were obtained using only the information about the molar fraction and the molecular structure of pure components and without any other information, including the boiling points of pure components of azeotropic mixtures.

SUPPLEMENTARY MATERIAL

Additional data are available electronically at the pages of journal website: https://www.shd-pub.org.rs/index.php/JSCS/index, or from the corresponding author on request.

Acknowledgments. This manuscript was extracted from the PhD Thesis of Zohreh Faramarzi. The financial support of this work, by Mahshahr branch–Islamic Azad University, is greatly appreciated.

ИЗВОД

НОВИ МОДЕЛИ ЗА ПРЕДВИЂАЊЕ ТЕМПЕРАТУРЕ НОРМАЛНИХ ТАЧКИ КЉУЧАЊА ТЕРНЕРНИХ АЗЕОТРОПА ЗАСНОВАНИ НА СТРУКТУРИ

ZOHREH FARAMARZI[1], FATEMEH ABBASITABAR[2], HOSSEIN JALALI JAHROMI[1] и MAZIAR NOEI[1]

[1]*Department of Chemistry, Mahshahr Branch, Islamic Azad University, Mahshahr, Iran и* [2]*Department of Chemistry, Marvdasht Branch, Islamic Azad University, Marvdasht, Iran*

У новије време је много пажње привукао развој QSPR модела за смеше. QSPR моделовање смеша захтева коришћење одговарајућих дескриптора смеша. У овој студији је разматрано 12 математичких једначина за израчунавање дескриптора смеша из појединачних компоненти у циљу педвиђања нормалних тачки кључања за 78 тернерних азеотропских смеша. Вишеструка линеарна регресија (MLR) коришћена је за прављење свих QSPR модела. „Memorized_ACO“ алгоритам примењен је за одабир подскупа варијабли. Такође је конструисан ансамбл модел коришћењем стратегије упросечивања да би се побољшала моћ претсказивања коначног QSAR модела. Модели су валидирани тест скупом који садржи 24 тернерна азеотропа и путем различних статистичких тестова. Добијени QSPR ансамбл модел је имао R^2_{training}, R^2_{test}, и q^2 од 0,97, 0,95, односно 0,96.

Средња апсолутна грешка (*MAE*) је добар индикатор ваљаности модела и нађено је да је 3,06 и 3,52 за пробне тестове.

(Примљено 18. марта, ревидирано 6. априла, прихваћено 13. априла 2021)

REFERENCES

1. M. B. Franke, *Comput. Chem. Eng.* **89** (2016) 204 (https://doi.org/10.1016/j.compchemeng.2016.03.027)
2. Q.-K. Le, I. J. Halvorsen, O. Pajalic, S. Skogestad, *Chem. Eng. Res. Des.* **99** (2015) 111 (https://doi.org/10.1016/j.cherd.2015.03.022)
3. W. Li, L. Zhong, Y. He, J. Meng, F. Yao, Y. Guo, et al., *Ind. Eng. Chem. Res.* **54** (2015) 7668 (https://doi.org/10.1021/acs.iecr.5b00572)
4. Y. Wang, S. Liang, G. Bu, W. Liu, Z. Zhang, Z. Zhu, *Ind. Eng. Chem. Res.* **54** (2015) 12908 (https://doi.org/10.1021/acs.iecr.5b03666)
5. B. ZareNezhad, N. Hosseinpour, *Energy Convers. Manage.* **50** (2009) 1491 (https://doi.org/10.1016/j.enconman.2009.02.016)
6. Y. Tavan, S. Shahhosseini, *Energy Technol.* **4** (2016) 424 (https://doi.org/10.1002/ente.201500287)
7. K. Tochigi, D. Tiegs, J. Gmehling, K. Kojima, *J. Chem. Eng. Jpn.* **23** (1990) 453 (https://doi.org/10.1252/jcej.23.453)
8. S. M. Hosseini, M. M. Alavianmehr, D. Mohammad-Aghaie, F. Fadaei-Nobandegani, J. Moghadasi, *J. Ind. Eng. Chem.* **19** (2013) 769 (https://doi.org/10.1016/j.jiec.2012.10.013)
9. J. Gmehling, J. Li, M. Schiller, *Ind. Eng. Chem. Res.* **32** (1993) 178 (https://doi.org/10.1021/ie00013a024)
10. M. J. Hait, C. L. Liotta, C. A. Eckert, D. L. Bergmann, A. M. Karachewski, A. J. Dallas, et al., *Ind. Eng. Chem. Res.* **32** (1993) 2905 (https://doi.org/10.1021/ie00023a064)
11. S. Yousefinejad, F. Honarasa, H. Montaseri, *RSC Adv.* **5** (2015) 42266 (https://doi.org/10.1039/C5RA05930E)
12. A. Klamt, F. Eckert, *Fluid Phase Equilib.* **172** (2000) 43 (https://doi.org/10.1016/s0378-3812(00)00357-5)
13. D. E. Nanu, T. W. De Loos, *Mol. Phys.* **102** (2004) 235 (https://doi.org/10.1080/00268970410001655871)
14. A. A. Oliferenko, P. V. Oliferenko, J. S. Torrecilla, A. R. Katritzky, *Ind. Eng. Chem. Res.* **51** (2012) 9123 (https://doi.org/10.1021/ie202550v)
15. A. R. Katritzky, I. B. Stoyanova-Slavova, K. Tämm, T. Tamm, M. Karelson, *J. Phys. Chem., A* **115** (2011) 3475 (https://doi.org/10.1021/jp104287p)
16. I. Oprisiu, S. Novotarskyi, I. V. Tetko, *J. Cheminform.* **5** (2013) 4 (https://doi.org/10.1186/1758-2946-5-4)
17. V. Zare-Shahabadi, M. Lotfizadeh, A. R. A. Gandomani, M. M. Papari, *J. Mol. Liq.* **188** (2013) 222 (https://doi.org/10.1016/j.molliq.2013.09.037)
18. T. Gaudin, P. Rotureau, G. Fayet, *Ind. Eng. Chem. Res.* **54** (2015) 6596 (https://doi.org/10.1021/acs.iecr.5b01457)
19. Z. Faramarzi, F. Abbasitabar, V. Zare-Shahabadi, H. J. Jahromi, *J. Mol. Liq.* **296** (2019) 111854 (https://doi.org/10.1016/j.molliq.2019.111854)
20. Y. Demirel, *Thermochim. Acta* **339** (1999) 79 (https://doi.org/10.1016/s0040-6031(99)00211-7)
21. *ChemDraw Ultra 6.0 and Chem3D Ultra*, Cambridge Soft Corporation, Cambridge, MA
22. *MOE*, Chemical Computing Group Inc., Montreal (http://www.chemcomp.com)

23. R. Todeschini, V. Consonni, M. Pavan, *Dragon Software Version 2.1*, Chemometrics and QSAR Research Group, Milano, 2002
24. E. N. Muratov, E. V. Varlamova, A. G. Artemenko, P. G. Polishchuk, V. E. Kuz'min, *Mol. Inform.* **31** (2012) 202 (https://doi.org/10.1002/minf.201100129)
25. F. Abbasitabar, V. Zare-Shahabadi, *SAR QSAR Environ. Res.* **23** (2011) 1 (https://doi.org/10.1080/1062936x.2011.623316)
26. B. Hemmateenejad, M. Shamsipur, V. Zare-Shahabadi, M. Akhond, *Anal. Chim. Acta* **704** (2011) 57 (https://doi.org/10.1016/j.aca.2011.08.010)
27. V. Zare-Shahabadi, *Med. Chem. Res.* **25** (2016) 2787 (https://doi.org/10.1007/s00044-016-1666-z)
28. F. Abbasitabar, V. Zare-Shahabadi, *Chemosphere* **172** (2017) 249 (https://doi.org/10.1016/j.chemosphere.2016.12.095)
29. D. Baumann, K. Baumann, *J. Cheminform.* **6** (2014) 47 (https://doi.org/10.1186/s13321-014-0047-1)
30. D. L. Massart, B. G. M. Vandeginste, L. M. C. Buydens, S. De Jong, P. J. Lewi, J. Smeyers-Verbeke, *Handbook of Chemometrics and Qualimetrics Part A*, Elsevier, Amsterdam, 1997, pp. 286–288
31. S. Saaidpour, *Phys. Chem. Res.* **4** (2016) 61 (https://doi.org/10.22036/pcr.2016.11759)
32. F. Abbasitabar, V. Zare-Shahabadi, *Drug Res (Stuttgart)* **67** (2017) 476 (https://doi.org/10.1055/s-0043-108553)
33. K. Roy, S. Kar, P. Ambure, *Chemom. Intell. Lab. Syst.* **145** (2015) 22 (http://dx.doi.org/10.1016/j.chemolab.2015.04.013)
34. X. Bian, P. Diwu, Y. Liu, P. Liu, Q. Li, X. Tan, *J. Chemom.* **32** (2018) e2940 (https://doi.org/10.1002/cem.2940)
35. V. Zare-Shahabadi, F. Abbasitabar, M. Akhond, M. Shamsipur, *J. Braz. Chem. Soc.* **24** (2013) 1561 (http://dx.doi.org/10.5935/0103-5053.20130197)
36. S. Ma, S. Li, *Ind. Eng. Chem. Res.* **52** (2013) 543 (https://doi.org/10.1021/ie302909b)
37. A. A. Oliferenko, P. V. Oliferenko, J. S. Torrecilla, A. R. Katritzky, *Ind. Eng. Chem. Res.* **52** (2013) 545 (https://doi.org/10.1021/ie3033125)
38. S. Guariento, M. Tonelli, S. Espinoza, A. S. Gerasimov, R. R. Gainetdinov, E. Cichero, *Eur. J. Med. Chem.* **146** (2018) 171 (https://doi.org/10.1016/j.ejmech.2018.01.059)
39. O. Deeb, B. Hemmateenejad, *Chem. Biol. Drug Des.* **70** (2007) 19 (https://doi.org/10.1111/j.1747-0285.2007.00528.x)
40. R. Todeschini, V. Consonni, R. Mannhold, H. Kubinyi, G. Folkers. *Molecular Descriptors for Chemoinformatics: Volume I: Alphabetical Listing / Volume II: Appendices, References*, Wiley, New York, 2009, pp. 17–20.