Journal of Solution Chemistry (2020) 49:222–238
https://doi.org/10.1007/s10953-020-00954-8

Atomic-Level Topological Indices for Prediction of the Infinite Dilution Activity Coefficients of Oxo Compounds in Water

Fariba Safa[1]

Received: 6 July 2019 / Accepted: 11 February 2020 / Published online: 19 February 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract
The atomic-level *AI* topological indices and the modified *Xu* (^{m}Xu) index were utilized for quantitative structure–property relationship (QSPR) modeling of the infinite dilution activity coefficients of 108 oxo compounds in water at 298.15 K. Stepwise multiple linear regression (SMLR) analysis using the topological descriptors resulted in a model with $R^2 = R^2_{adj} = 0.9904$, SE = 0.3769, F = 1267.1 and an average relative error of 4.89%. The selected descriptors were then used to develop an artificial neural network (ANN) model for the activity coefficients. Findings of the study indicated that a 7–8-1 ANN trained by Levenberg–Marquardt algorithm results in the improved predictions, especially in view of a decrease as large as 47.24% in the average relative error compared to the SMLR model. The *AI* indices with a total contribution of 81.43% showed the dominant role of the atomic groups of the oxo compounds in determination of their activity coefficients at infinite dilution in water.

Keywords Infinite dilution activity coefficient · Quantitative structure–property relationship · Atomic-level topological indices · Modified *Xu* index · Artificial neural network

1 Introduction

The infinite dilution activity coefficient, γ^{∞}, is an important thermodynamic property of both practical and theoretical interest. The parameter provides insight into the kinds of physical and chemical intermolecular forces involved in the solute–solvent interactions, which is useful for estimating aqueous solubility and selecting solvents in many industrial processes including high purity extraction, azeotropic rectification, chemical separations and environmental pollution control [1]. Moreover, study of the infinite dilution activity coefficients is of great value in investigating the thermodynamic behavior of dilute aqueous solutions, developing new thermodynamic models [2] as well as calculating the solubility of solids in supercritical gases [3], excess enthalpies [4] and Henry constants [5]. Many practical implications in environmental, chemical and biochemical processes [6] and

✉ Fariba Safa
Safa@iaurasht.ac.ir

[1] Department of Chemistry, Rasht Branch, Islamic Azad University, Rasht, Iran

extensive applications in commercially important products like pharmaceuticals, coatings and paints [7, 8] make it essential to study γ^{∞} values. Gas–liquid chromatography [9], the dilutor method [10] and differential ebulliometry [11] are the most common methods of measuring γ^{∞} values. Nevertheless, for the reasons like safety, cost and technical availability, prediction of the infinite dilution activity coefficients by QSPR modeling [12–15] as an alternative method is of great importance.

Since 1947 when Wiener reported the first application of graph theory to QSPR modeling, many graph theoretical topological descriptors possessing high prediction potency of various physicochemical properties have been developed [16, 17]. Using the descriptors, useful information is obtained about molecular features such as size, shape, branching, symmetry, as well as the atom and bond types without the need for optimizing the geometry of molecules [18]. Atomic-level indices are highly efficient topological descriptors, which allow estimation of the individual contributions of the molecular fragments and atomic groups to the properties of chemical compounds. One of the most important descriptors of this type called atom-type-based *AI* indices was introduced by Ren [19]. In addition to describing the structure of a molecule at the atomic-level, the descriptors encode the structural environment of each atom-type in the molecule. The atomic-level *AI* indices combined with the bulk property topological descriptor of *Xu* [20, 21] (or ^{m}Xu [22]) showed satisfactory linear correlations to various properties such as molecular total surface area, enthalpies of vaporization, Pitzer's acentric factor, water solubility, narcosis activity, etc. [23–26]. Ren also employed the indices in linear regression modeling of the quantitative structure-retention relationship (QSRR) of aldehydes and ketones on gas chromatographic columns [27]. Moreover, Panneerselvam et al. used the atomic-level indices to predict the boiling points of trialkyl phosphates [28]. Recently, our group has successfully applied the atom-type-based topological descriptors combined with MLR technique for estimating the standard formation enthalpies of acyclic alkanes [29] and normal boiling points of esters [30]. The group has also reported the first application of ANN modeling in a QSRR study of monomethylalkanes using the topological indices [31].

In this work, the benefits of the atomic-level *AI* and ^{m}Xu topological descriptors in QSPR modeling of the infinite dilution activity coefficients of a group of oxygen containing organic compounds in water are illustrated. The structure–property relationships were investigated by SMLR and ANN modeling techniques. Additionally, the role of atomic groups affecting the activity coefficients of the studied molecules was characterized. As far as the author is aware, this is the first report on QSPR modeling of the infinite dilution activity coefficient of organic compounds using the atomic-level *AI* topological indices.

2 Method

2.1 Data Set and Topological Descriptors

The experimental values of the room temperature activity coefficients at infinite dilution for 108 oxygen containing organic compounds in water were taken from the literature [32–39]. Table 1 lists the data set including C_2–C_{18} linear and branched alcohols, ketones, ethers, esters, aldehydes and carboxylic acids with ln γ^{∞} values in the range of 1.32–23.34.

The topological indices for the studied molecules were calculated during the following steps [27]:

Table 1 Data set, topological indices and predicted values of the infinite dilution activity coefficients in water at 298.15 K

No.	Compound	$\ln \gamma_{\exp}^{\infty}$	Topological descriptors							$\ln \gamma_{\mathrm{pred}}^{\infty}$		Ref
			^{m}Xu	$AI\,(-\mathrm{CH_3})$	$AI\,(-\mathrm{CH_2}-)$	$AI\,(>\mathrm{CH}-)$	$AI\,(-\mathrm{OH})$	$AI\,(>\mathrm{C}=)$	$AI\,(=\mathrm{O})$	SMLR	ANN	
1	2,2,3-Trimethyl-3-pentanol	6.95	3.6861	13.2455	3.2302	0	2.6255	0	0	6.68	6.88	32
2**	2,2-Dimethyl-1-butanol	6.61	2.8501	7.3461	5.5413	0	3.1859	0	0	5.25	6.15	32
3	2,2-Dimethyl-1-propanol	4.91	2.3147	6.5713	2.5237	0	3.0005	0	0	4.06	4.91	32
4*	2,2-Dimethyl-3-pentanol	6.66	3.3098	10.5782	3.2631	2.8256	2.6962	0	0	6.30	6.21	32
5	2,3-Dimethyl-2-butanol	4.88	2.7685	9.6955	0	2.8959	2.5387	0	0	5.08	4.86	32
6**	2,3-Dimethyl-2-pentanol	6.02	3.3098	10.5782	3.2631	2.8256	2.6962	0	0	6.30	6.21	32
7	2,3-Dimethyl-3-pentanol	5.96	3.2866	10.6984	3.0136	3.2191	2.5346	0	0	6.31	5.92	32
8*	2,4-Dimethyl-2-pentanol	6.16	3.3696	10.8624	2.3649	3.7867	2.7919	0	0	6.41	6.56	32
9	2,4-Dimethyl-3-pentanol	6.82	3.3483	10.9813	0	9.6574	2.6306	0	0	6.81	6.77	32
10**	2,4-Dimethyl-3-pentanone	7.01	3.3444	10.9213	0	6.8603	0	2.8559	2.5833	6.43	7.06	32
11*	2,6-Dimethyl-4-heptanone	9.08	4.4698	13.5180	5.8544	10.1938	0	3.5241	2.8207	9.13	9.28	32
12*	2-Butanol	3.26	1.8722	4.7186	2.3499	2.4062	2.4003	0	0	3.65	3.38	33
13	2-Butanone	3.24	1.8667	4.6780	2.3298	0	0	2.4775	2.3511	3.32	3.26	34
14*	2-Heptanone	7.24	3.6362	6.7171	12.8493	0	0	4.4677	3.2672	7.08	6.97	34
15	2-Hexanol	5.64	3.0680	6.0808	8.8043	3.5587	2.9975	0	0	6.37	5.89	33
16	2-Hexanone	5.87	3.0648	6.0332	8.7367	0	0	3.6974	2.9336	5.85	5.73	34
17	2-Methyl sec-butyl methyl ether	6.11	2.8958	10.5059	0	5.7783	0	0	0	6.38	6.16	32
18**	2-Methyl-1-butanol	5.08	2.4456	5.0155	5.3690	2.5285	3.0508	0	0	4.82	4.86	32
19**	2-Methyl-1-propanol	3.89	1.9306	4.5662	2.7344	2.3274	2.6730	0	0	3.71	3.72	34
20**	2-Methyl-2-hexanol	6.49	3.4878	8.8268	9.5319	0	2.8224	0	0	6.85	6.74	32
21**	2-Methyl-2-pentanol	5.14	2.9037	7.9606	5.7274	0	2.5847	0	0	5.53	5.16	32
22	2-Methyl-3-pentanol	5.63	2.8986	8.1199	2.9953	5.7620	2.5847	0	0	5.88	5.54	32
23	2-Methyl-3-pentanone	5.89	2.8944	8.0685	2.9753	3.0576	0	2.7778	2.5365	5.51	5.84	32
24**	2-Nonanone	9.70	4.7284	8.1073	22.9849	0	0	6.2039	3.9823	9.48	9.50	35

Table 1 (continued)

No.	Compound	$\ln \gamma^{\infty}_{exp}$	Topological descriptors							$\ln \gamma^{\infty}_{pred}$		Ref
			^{m}Xu	$AI\,(-CH_3)$	$AI\,(-CH_2-)$	$AI\,(>CH-)$	$AI\,(-OH)$	$AI\,(>C=)$	$AI\,(=O)$	SMLR	ANN	
25**	2-Pentanol	4.57	2.4789	5.3999	5.2799	2.9199	2.6804	0	0	5.01	4.62	33
26**	2-Pentanone	4.54	2.4746	5.3558	5.2374	0	0	3.0214	2.6245	4.60	4.49	34
27	2-Propanol	2.03	1.2619	4.0133	0	2.0872	2.1748	0	0	2.30	2.22	34
28*	3,3-Dimethyl-2-butanol	5.43	2.7706	9.4777	0	2.8934	2.7823	0	0	5.02	5.32	32
29	3,3-Dimethyl-2-butanone	5.66	2.7650	9.6083	0	0	0	2.9991	2.5117	4.66	5.64	32
30	3-Ethyl-3-pentanol	5.94	3.3404	8.9209	8.8025	0	2.4741	0	0	6.60	5.95	32
31	3-Hexanol	5.85	3.0182	6.2678	9.0559	2.8691	2.6359	0	0	6.32	5.63	32
32*	3-Hexanone	6.02	3.0141	6.2246	8.9946	0	0	2.9735	2.5861	5.92	5.87	33
33	3-Methyl-1-butanol	5.34	2.4843	4.8703	5.2690	2.9135	3.2796	0	0	4.87	5.11	34
34**	3-Methyl-2-Butanone	4.43	2.3432	7.0799	0	2.6522	0	2.7623	2.5544	4.20	4.53	34
35**	3-Methyl-2-pentanol	5.66	2.9004	7.9133	2.9930	5.7577	2.8171	0	0	5.82	5.67	32
36	3-Methyl-2-pentanone	5.56	2.8967	7.8582	2.9706	2.6627	0	3.1896	2.7597	5.39	5.67	32
37	3-Methyl-3-hexanol	6.28	3.4089	8.8602	9.3582	0	2.5698	0	0	6.75	6.22	32
38*	3-Methyl-3-pentanol	4.85	2.8441	8.0011	5.5544	0	2.4483	0	0	5.43	4.83	32
39*	3-Pentanone	4.67	2.4355	5.4855	5.3411	0	0	2.6169	2.4262	4.61	4.59	34
40	4-Heptanone	7.41	3.5667	7.0137	13.4017	0	0	3.1691	2.6550	7.24	7.17	32
41	4-Methyl-2-pentanol	5.86	2.9377	8.1312	2.3363	6.4297	2.9960	0	0	5.85	5.88	32
42*	4-Methyl-2-pentanone	5.68	2.9486	7.9713	2.2946	3.3497	0	3.5063	2.8941	5.45	5.85	32
43	5-Methyl-2-hexanone	7.33	3.5343	8.8941	5.1825	4.1176	0	4.3254	3.2457	6.68	7.24	36
44*	5-Nonanone	9.98	4.6434	8.5364	24.2523	0	0	3.7437	2.8689	9.91	9.52	32
45	Acetaldehyde	1.37	0.5214	1.4553	0	0	0	0	1.5204	1.05	1.31	33
46	Acetone	1.95	1.2551	3.9751	0	0	0	2.1377	2.1288	2.03	2.03	34
47	Butanol	3.92	2.0022	2.6215	7.3457	0	2.8923	0	0	3.97	3.99	34
48	Butyl methyl ether	6.30	2.5917	6.0332	8.0729	0	0	0	0	5.97	6.24	32

Table 1 (continued)

No.	Compound	$\ln\gamma_{\text{exp}}^{\infty}$	Topological descriptors							$\ln\gamma_{\text{pred}}^{\infty}$		Ref
			^{m}Xu	$AI\,(-CH_3)$	$AI\,(-CH_2-)$	$AI\,(>CH-)$	$AI\,(-OH)$	$AI\,(>C=)$	$AI\,(=O)$	SMLR	ANN	
49	Butyl pentanoate	9.87	5.1722	9.1983	27.3565	0	0	4.0526	3.4533	10.98	9.95	32
50	Butyraldehyde	3.88	1.9462	2.6726	4.8434	0	0	0	2.5485	4.03	3.93	34
51**	Decanol	12.38	5.3228	4.8823	40.8487	0	5.5306	0	0	11.89	12.51	32
52*	Diethyl ether	4.23	1.9916	5.2863	5.2206	0	0	0	0	4.58	4.56	34
53	Diisopropyl ether	6.44	2.9447	10.7622	0	6.7943	0	0	0	6.56	6.23	37
54	di-*n*-Butyl ether	10.76	4.2689	8.3087	24.0889	0	0	0	0	10.27	10.62	37
55	di-*n*-Propyl ether	7.74	3.1665	6.8009	13.2534	0	0	0	0	7.43	7.85	**37**
56*	Dodecanol	15.31	6.3214	5.6336	57.6129	0	6.4074	0	0	14.60	15.22	32
57	Ethanol	1.32	0.7243	1.8571	1.7618	0	2.0002	0	0	1.27	1.17	34
58	Ethyl acetate	4.18	2.4720	5.3184	2.9520	0	0	2.9988	2.6064	4.48	4.35	34
59	Ethyl butyrate	6.59	3.5648	6.9709	10.5320	0	0	3.1506	2.6409	7.09	6.98	34
60	Ethyl formate	3.86	1.9741	2.6119	2.1606	0	0	0	2.8424	3.88	3.87	34
61**	Ethyl isopropyl ether	5.30	2.4392	7.7949	2.6850	2.9970	0	0	0	5.49	5.69	32
62*	Ethyl pentanoate	7.96	4.1078	7.7489	15.2420	0	0	3.5161	2.7912	8.38	8.16	32
63*	Ethyl propanoate	5.54	3.0118	6.1840	6.4589	0	0	2.9546	2.5708	5.78	5.70	34
64	Ethyl propenoate	5.60	3.0229	3.0872	3.2664	0	0	2.8501	2.4869	5.89	5.60	32
65*	Ethyl propyl ether	5.55	2.5896	6.0436	8.8048	0	0	0	0	6.00	6.26	32
66	Ethylene oxide	1.83	0.5214	0	3.3023	0	0	0	0	1.71	1.84	38
67	Heptadecanol	21.30	8.6589	7.5105	111.7717	0	8.5978	0	0	21.69	21.27	39
68**	Heptanal	8.34	3.7353	3.7081	16.3574	0	0	0	4.0953	8.04	7.71	32
69**	Heptanol	8.09	3.7350	3.7540	20.9503	0	4.2139	0	0	7.91	8.11	33
70*	Hexadecanol	19.77	8.2070	7.1352	99.5401	0	8.1598	0	0	20.23	19.81	39
71	Hexanal	6.70	3.1755	3.3333	11.8094	0	0	0	3.6670	6.73	6.55	34
72	Hexanoic acid	6.40	3.7178	3.7277	13.1851	0	4.1832	4.6072	2.9247	6.19	6.42	32

Table 1 (continued)

No.	Compound	$\ln \gamma_{\text{exp}}^{\infty}$	Topological descriptors							$\ln \gamma_{\text{pred}}^{\infty}$		Ref
			^{m}Xu	$AI\,(-CH_3)$	$AI\,(-CH_2-)$	$AI\,(>CH-)$	$AI\,(-OH)$	$AI\,(>C=)$	$AI\,(=O)$	SMLR	ANN	
73	Hexanol	6.68	3.1763	3.3773	15.7166	0	3.7743	0	0	6.60	6.72	34
74**	Isobutyl acetate	6.74	3.5322	8.8466	2.5784	4.0924	0	4.2986	3.2276	6.54	7.19	34
75	Isobutyl methyl ether	6.09	2.4749	7.8677	2.2873	2.9310	0	0	0	5.54	5.75	32
76**	Isopentyl acetate	8.00	4.0988	9.7957	5.6901	4.9110	0	5.1717	3.5892	7.73	8.65	34
77**	Isopropyl acetate	5.28	2.9462	7.9257	0	3.3281	0	3.4834	2.8767	5.33	5.76	34
78	Isopropyl butyrate	8.03	4.0183	10.0132	7.4331	4.4791	0	3.3750	2.7530	8.03	8.12	32
79	Isopropyl propyl ether	7.09	3.0620	8.8363	6.4041	3.5780	0	0	0	7.01	7.29	32
80	Methyl acetate	3.12	1.8649	4.6363	0	0	0	2.4545	2.3300	3.20	3.10	34
81	Methyl butyrate	5.80	3.0133	6.1767	5.8658	0	0	2.9512	2.5681	5.76	5.67	33
82	Methyl formate	2.74	1.3706	2.1830	0	0	0	0	2.3522	2.62	2.54	34
83	Methyl hexanoate	8.29	4.1378	7.5865	14.0172	0	0	4.0484	3.0390	8.26	8.07	33
84	Methyl isobutyrate	5.73	2.8939	8.0096	0	3.0337	0	2.7571	2.5186	5.36	5.67	33
85	Methyl pentanoate	7.14	3.5827	6.8862	9.6502	0	0	3.4478	2.7843	7.02	6.89	33
86*	Methyl propanoate	4.47	2.4343	5.4399	2.6487	0	0	2.5958	2.4075	4.48	4.40	33
87*	Methyl propyl ether	4.88	1.9931	5.2786	4.7868	0	0	0	0	4.56	4.56	32
88	*n*-Butyl acetate	6.70	3.6340	6.6795	10.1048	0	0	4.4400	3.2491	6.93	6.87	34
89	*n*-Hexyl acetate	9.43	4.7268	8.0666	19.4449	0	0	6.1691	3.9624	9.30	9.47	34
90	Nonanal	11.23	4.8084	4.4575	27.4854	0	0	0	4.9519	10.69	11.22	32
91	Nonanol	11.03	4.8067	4.5064	33.5162	0	5.0920	0	0	10.55	11.02	32
92**	*n*-Pentyl acetate	8.08	4.1762	7.4803	13.9549	0	0	5.3657	3.6530	8.03	8.14	34
93	*n*-Propyl acetate	5.49	3.0623	5.9964	6.2806	0	0	3.6728	2.9160	5.72	5.60	34
94	*n*-Propyl formate	5.13	2.5949	2.9366	5.4773	0	0	0	3.2136	5.28	5.26	34
95	*n*-Propyl propanoate	6.99	3.6146	7.6227	10.5483	0	0	3.4154	2.7608	7.09	7.15	32
96	Octadecanol	23.34	9.1039	7.8858	124.7033	0	9.0357	0	0	23.18	23.32	39

Table 1 (continued)

No.	Compound	$\ln\gamma_{\mathrm{exp}}^{\infty}$	Topological descriptors							$\ln\gamma_{\mathrm{pred}}^{\infty}$		Ref
			^{m}Xu	$AI\,(-\mathrm{CH_3})$	$AI\,(-\mathrm{CH_2}-)$	$AI\,(>\mathrm{CH}-)$	$AI\,(-\mathrm{OH})$	$AI\,(>\mathrm{C}=)$	$AI\,(=\mathrm{O})$	SMLR	ANN	
97	Octanal	9.02	4.2789	4.0828	21.5805	0	0	0	4.5236	9.36	9.08	35
98*	Octanol	9.56	4.2779	4.1303	26.8835	0	4.6531	0	0	9.23	9.55	32
99**	Pentadecanol	18.77	7.7479	6.7599	88.0084	0	7.7218	0	0	18.79	18.61	39
100*	Pentanal	5.39	2.5975	2.9583	7.9262	0	0	0	3.2383	5.41	5.34	34
101	Pentanoic acid	4.84	3.0682	3.3032	8.6429	0	2.9420	3.6533	2.9020	5.11	4.82	32
102	Pentanol	5.29	2.5996	2.9999	11.1819	0	3.3339	0	0	5.29	5.35	34
103**	Propanol	2.60	1.3798	2.2413	4.2067	0	2.4486	0	0	2.64	2.61	34
104	Propionaldehyde	2.56	1.3735	2.2057	2.0717	0	0	0	2.3781	2.73	2.73	34
105*	sec-Butyl methyl ether	5.71	2.4194	7.6030	2.7182	2.5591	0	0	0	5.43	5.63	32
106	tert-Butanol	2.48	1.7298	6.2544	0	0	2.2660	0	0	2.93	2.46	34
107	tert-Butyl methyl ether	4.73	2.3040	9.3401	0	0	0	0	0	4.68	4.69	34
108	Tetradecanol	17.50	7.2809	6.3845	77.1767	0	7.2837	0	0	17.37	17.50	39

*, **Referring to the molecules included in the validation and prediction sets, respectively

(i) Illustration of the hydrogen depleted structure of each molecule by the molecular graph.

(ii) Deriving the distance matrix, $D = [d_{ij}]_{n \times n}$, whose elements are the shortest path length between the atoms *i* and *j* in the molecular graph. Then, the sum over the column *i* (or row *j*) of the matrix was calculated to give the distance sum vector, $S = [s_i]_{n \times 1}$.

(iii) Coding of the graph by the vertex degree vector, $V = [v_i]_{n \times 1}$, whose elements are the number of connections to the atom *i*.

(iv) Calculation of *Xu* index for each graph by the following equation where the sum is over all *i* atoms in the molecular graph.

$$Xu = n^{1/2} \log_{10} \left(\sum_{i=1}^{n} v_i s_i^2 \Big/ \sum_{i=1}^{n} v_i s_i \right) \tag{1}$$

(v) Calculation of the atomic-level *AI* topological descriptors for each atom *i* belonging to *j*th atom-type in the graph as follows:

$$AI_i(j) = 1 + \varphi_i(j) = 1 + \left(v_i(j) s_i^2(j) \Big/ \sum_{i=1}^{n} v_i s_i \right) \tag{2}$$

where the perturbing term, $\phi_i(j)$, reflects the impact of the structural environment of the *i*th atom on the topological index value. The $AI_i(j)$ values of *m* atoms of the same type was then utilized to obtain the desired atom-type topological descriptor, *AI*(*j*), using the following equation:

$$AI(j) = \sum_{i=1}^{m} AI_i(j) = m + \left(\sum_{i=1}^{m} v_i(j)\, s_i^2(j) \Big/ \sum_{i=1}^{n} v_i\, s_i \right) \tag{3}$$

To modify the *Xu* and *AI* topological descriptors for differentiation of the heteroatoms (oxygen in the work) and multiple bonds, v_i values were replaced by the degree of vertex developed by Ren, v^{m}, [19]. The parameter is defined using the number of connections of the atom (δ), valence connectivity of Kier–Hall, δ^v [40], and the principal quantum number of the valence shell (*N*).

$$v^m = \delta + \frac{1}{(2/N)^2 \delta^v + 1} \tag{4}$$

where δ is calculated by the difference between the number of valence electrons and the number of hydrogens bonded to an atom.

2.2 Model Development

The quantitative structure–activity coefficient correlations for the studied compounds were firstly generated by SMLR using SPSS software [41]. In the modeling process, the room temperature values of $\ln \gamma^{\infty}$ for the oxo compounds (dependent variable) and the calculated topological descriptors (independent variables) were mathematically correlated by the following equation:

$$\ln \gamma^{\infty} = a_0 + a_1{}^{m}Xu + b_j AI(j) \tag{5}$$

where a_0 is a constant and the parameters a_1 and b_j are the contribution coefficients of ${}^{m}Xu$ index and jth AI index, respectively. The oxo compounds was randomly split into a training set with 86 molecules and a prediction set with 22 molecules, and the corresponding ln γ^{∞} values were utilized for developing the model and model validation, respectively. To select the best linear model, the coefficient of multiple determination (R^2), adjusted correlation coefficient (R^2_{adj}), Fisher-ratio (F) and standard error (SE) of the generated equations were compared. Moreover, the values of t-scores for the model coefficients indicating the level of significance of the topological indices in the model, along with the statistics of standard error for the coefficients, were evaluated to choose a high quality subset of the indices.

In the next step, a feed-forward back propagation ANN algorithm was written in MATLAB [42] to develop the nonlinear QSPR model. The theoretical explanations for the modeling technique can be found in the literature [43]. The nonlinear model was generated using inputs and targets, normalized in the range of -1 to $+1$ to achieve the minimum computational errors. The network included an input layer with N_i neurons equal to the number of topological indices selected by SMLR, a hidden layer with N_h neurons, and an output layer with 1 neuron representing the targets, i.e. the ln γ^{∞} values. The neurons were intercorrelated by the connections called weight (W) and bias (b) whose values were modified during the network training. In this work, the Levenberg–Marquardt algorithm was utilized to train the network because of its robustness and accuracy [44] and the performance function of root mean squared error ($RMSE$) was employed based on the following definition:

$$RMSE = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(\ln\gamma_{\mathrm{exp},j} - \ln\gamma_{\mathrm{pred},j}\right)^2} \tag{6}$$

where the subscripts exp and pred refer to the experimental and predicted values of the activity coefficients for jth molecule of N model compounds, respectively. Optimum N_h value was also determined during the training stage of ANN through search within the range defined by the criterion, ρ, (Eq. 7) and observing the variations in $RMSE$.

$$\rho = \frac{Number\ of\ compounds\ presented\ to\ the\ neural\ network}{Number\ of\ connections\ in\ the\ neural\ network} \tag{7}$$

where the lower limit of ρ value is adjusted at 1 to avoid memorizing the data by the neural network, and the upper limit should not exceed 3 due to the inability of ANN to generalize [45].

The data used as the training and prediction sets for ANN modeling were the same as those employed for developing SMLR model. However, the training set was randomly divided into two sets with 64 and 22 data as the training and validation sets, respectively, to reveal the overtraining of ANN by tracking $RMSE$ values as a function of epoch number. Additionally, different combinations of the linear, logarithmic sigmoid and hyperbolic tangent sigmoid transfer functions were utilized for the hidden and output layers to achieve the best architecture for the neural network.

3 Results and Discussion

3.1 Quantitative Structure–Activity Coefficient Relationships

As mentioned, the ^{m}Xu index in combination with AI topological indices was firstly correlated to ln γ^{∞} data using SMLR modeling technique. The best SMLR model obtained for the activity coefficients is as follows:

$$\begin{aligned}\ln\gamma^{\infty} = {} & 0.408(\pm 0.171) + 2.148(\pm 0.117)^{m}Xu - 0.073(\pm 0.033)AI\left(-CH_3\right) \\ & + 0.054(\pm 0.007)AI\left(-CH_2-\right) + 0.090\,(\pm 0.025)AI(>CH-) - 0.326(\pm 0.041)AI(-OH) \\ & - 0.123(\pm 0.038)AI(>C=) - 0.244(\pm 0.062)AI(=O) \\ N = 86 \quad & \mathrm{R}^2 = \mathrm{R}^2_{\mathrm{adj}} = 0.9904 \quad F = 1267.1 \quad SE = 0.3769\end{aligned} \tag{8}$$

The statistics of R^2 show that the developed model explains ~ 99% of the variances in the activity coefficients. Moreover, R$^2_{\mathrm{adj}}$ equals R^2 indicating a high significance level of the model and the F-value implies that the relationship described by SMLR equation is significant with a certainty of 99.99%. Values of the t-scores for the model coefficients are 18.291, − 2.225, 7.658, 3.615, − 7.872, − 3.241 and − 3.951, respectively, proving that all the selected topological descriptors are significant to the model developed for ln γ^{∞} data. Investigation of the prediction power of the linear model was graphically done using the residual plot shown in Fig. 1. As illustrated, the residuals with a range of − 1.36 to + 1.11, did not follow a normal distribution around the average error of zero. Moreover, the average relative deviation (ARD) obtained for the model was 4.89%, suggesting that the present model can not make sufficiently accurate predictions for the

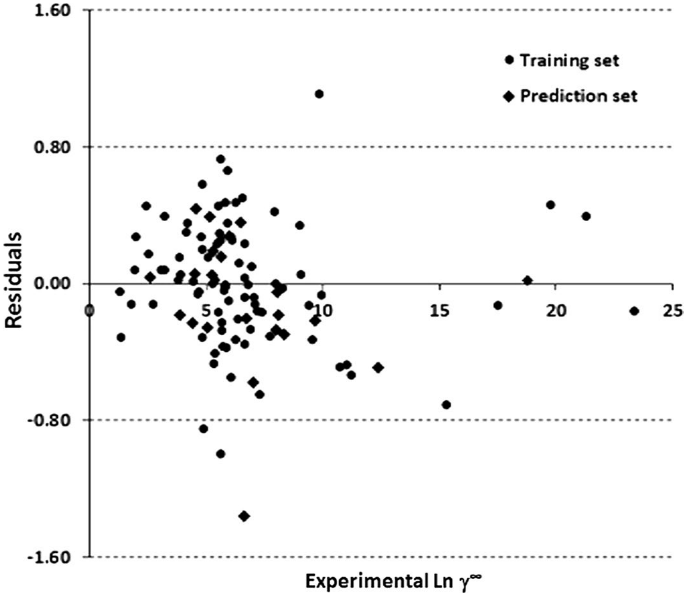

Fig. 1 The plot of the residuals resulted from MLR model vs. experimental ln γ^{∞} values of the oxygen containing organic compounds

desired data. Values of ${}^{m}Xu$ and AI indices entered in the SMLR model and the predicted activity coefficients are listed in Table 1.

In order to achieve a more accurate model, ANN modeling of the ln γ^{∞} values was also examined. Therefore, the topological descriptors were used as inputs for developing the nonlinear model. To find the best ANN architecture, different combinations of the transfer functions were applied and the optimum N_h value of the generated networks was sought within the range of 3–9, based on Eq. 7. By comparing the results (not shown), the best statistical quality was found to belong to an ANN with 7–8-1 topology and the transfer functions of tansig-linear for the hidden-output layers. Table 2 gives the characteristics of the best ANN model found for predicting ln γ^{∞} values of the studied compounds. As shown, the nonlinear model could predict ln γ^{∞} values for the three sets of training, validation and prediction with $R^2 > 0.99$ and $RMSE < 0.3$. Optimum values of the weights and biases for the proposed ANN model are presented in Table 3. Additionally, Fig. 2 indicates the values of mean squared error (MSE) against the epoch number for the training, validation and prediction sets. Obviously, the best performance of the ANN was achieved at epoch 26 with a MSE value of 0.0718. The ANN predicted ln γ^{∞} values for the model molecules are given in Table 1, and the quality of the predictions is graphically illustrated in Figs. 3 and 4. According to Fig. 3, the nonlinear model offers satisfactory efficiency to correlate the activity coefficients to the topological indices as judged from the good agreement between the data points and the straight line indicating perfect predictions. Moreover, the residual plot shown in Fig. 4 shows that there is no systematic error in the developed model, and the residuals ranging from -0.63 to $+0.71$, are considerably smaller than those obtained by SMLR. To further illustrate the efficiency of proposed ANN model for prediction of ln γ^{∞} values, the residuals reported by He and Zhong [46] as well as Estrada et al. [47] for the same compounds are also shown in the figure. The data clearly indicate the narrower range of the residuals resulted from the ANN model compared to those previously obtained by the researchers. It was also found that the average relative deviation of the nonlinear model (2.58%) was not only 47.24% lower than the developed SMLR model, but also 33.16% lower than ADR reported by He and Zhong, and 65.08% lower than the report of Estrada et al. Relatively small deviations of ANN predicted activity coefficients from the experimental values prove the superiority of the nonlinear model over the previous regression models developed for ln γ^{∞} of the oxo compounds.

Table 2 Characteristics of the best artificial neural network generated for prediction of ln γ^{∞} data

Topology	7–8-1	
Transfer function	Hidden layer	Hyperbolic tangent sigmoid
	Output layer	Linear
R^2	Training set	0.9988
	Validation set	0.9949
	Prediction set	0.9924
RMSE	Training set	0.1325
	Validation set	0.2680
	Prediction set	0.2933

Table 3 Optimum weights and biases of the developed ANN for the infinite dilution activity coefficient of the oxo compounds

N_h	Input layer weights							Input biases	Hidden layer	
	^{m}Xu	AI (−CH$_3$)	AI (−CH$_2$−)	AI (>CH−)	AI (−OH)	AI (>C=)	AI (=O)		Weights	Bias
1	− 1.6634	− 0.8662	0.4296	0.3679	0.5219	0.1061	− 0.4070	2.1311	− 0.5896	
2	− 0.7369	1.2638	− 0.1202	− 0.6785	− 0.1449	− 0.8768	0.8803	1.8269	0.2576	
3	− 0.3864	− 0.1191	− 0.8415	− 0.2011	1.9320	− 0.5921	0.7240	− 0.8001	0.9971	
4	− 1.4622	0.0350	0.8689	0.1966	− 0.1073	0.0406	− 2.7090	− 0.6267	0.2013	1.0168
5	1.9309	0.4581	0.4086	− 0.2434	− 1.6265	− 0.2776	0.2099	0.1809	0.5313	
6	− 0.1452	2.1790	0.2426	0.4455	− 0.3181	− 0.2867	− 0.9895	0.0547	0.1155	
7	0.5315	− 0.7335	− 0.3658	0.6853	− 0.0875	1.1191	− 0.1811	1.1991	0.4131	
8	0.2251	− 0.6965	1.7255	− 0.7082	− 1.1729	− 0.1518	− 0.0197	− 1.6649	0.9817	

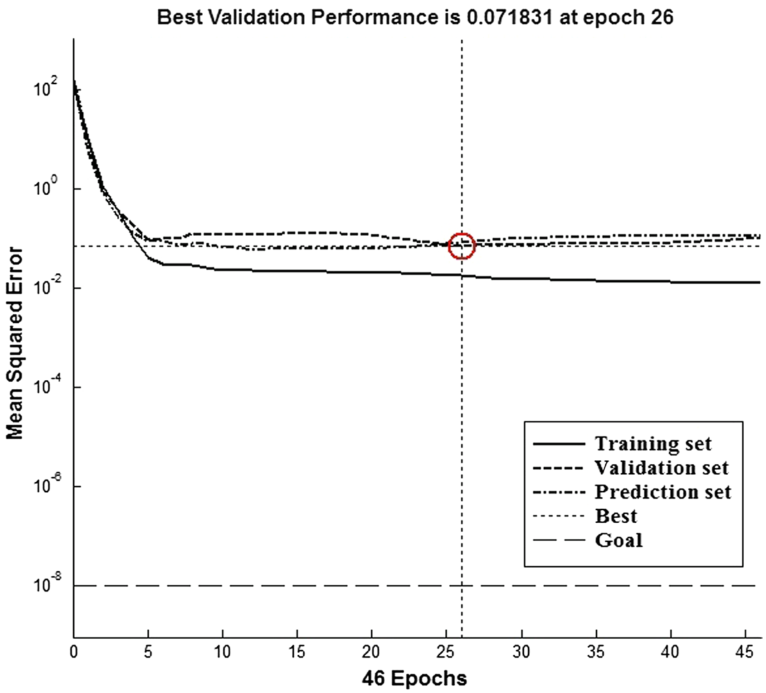

Fig. 2 Mean squared errors of the training, validation and prediction sets versus epoch numbers for the best ANN topology found for estimation of ln γ^{∞} values

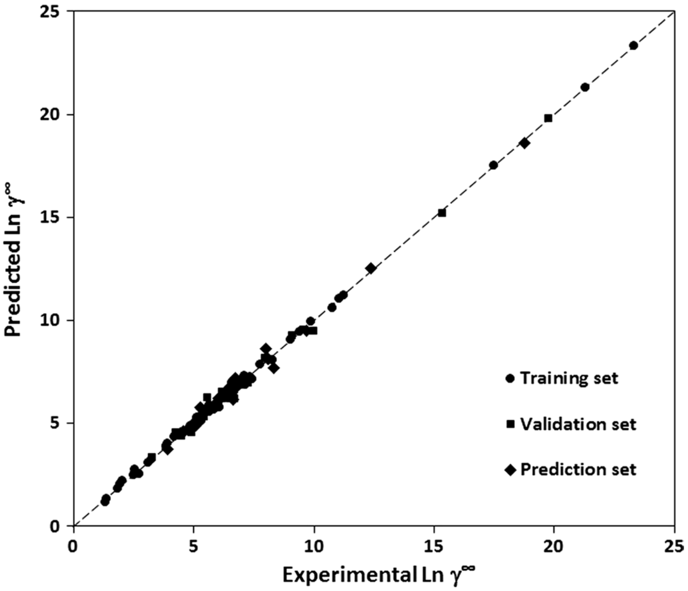

Fig. 3 Predicted versus experimental ln γ^{∞} values based on the developed ANN model

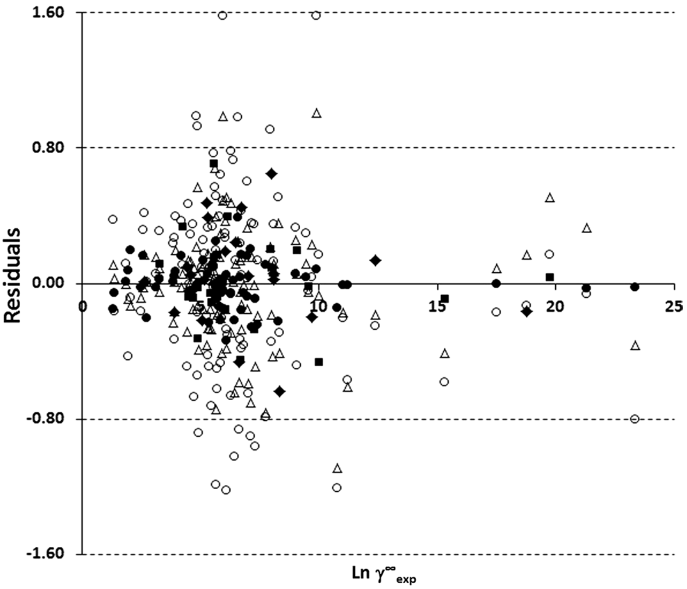

Fig. 4 Comparison of the residuals obtained from ANN model for the (filled circle) training, (filled square) validation and (filled diamond) prediction sets used in this work, with those reported by (open triangle) He and Zhong [46], and (open circle) Estrada et al. [47]

3.2 Structural Interpretation of the Infinite Dilution Activity Coefficients

To assess the role of structural characteristics of the studied compounds which determine the ln γ^{∞} values, the relative importance of each topological descriptor (IM_{j}) was calculated using the connection weights of the developed ANN by the following equation [48]:

$$IM_j = \frac{\sum_{m=1}^{m=N_h}\left(\left(|W_{jm}^{ih}|/\sum_{k=1}^{N_i}|W_{jm}^{ih}|\right)\times|W_{mn}^{ho}|\right)}{\sum_{k=1}^{k=N_i}\left\{\sum_{m=1}^{m=N_h}\left(\left(|W_{km}^{ih}|/\sum_{k=1}^{N_i}|W_{km}^{ih}|\right)\times|W_{mn}^{ho}|\right)\right\}} \quad (9)$$

where the superscripts i, h and o for the weights refer to the input, hidden and output layers of the network, and the superscripts k, m and n refer to the corresponding neurons, respectively. Calculated values of IM_{j} for the topological indices are presented in Fig. 5. Obviously, the $^{\mathrm{m}}Xu$ index characterizing the molecular size [20] had the maximum contribution indicating the dominant role of the molecular size in determining the ln γ^{∞} values. The atomic-level AI topological indices with an overall contribution of 81.43% showed highly considerable role in determination of the activity coefficients of the oxo compound. Among the indices, AI ($-CH_3$) had a contribution of 17.59% indicating that the degree of branching was nearly as effective as bulkiness of the molecule in determining the ln γ^{∞} values. The functional groups of –OH and = O as an indication of the molecular polarity were in the next ranks in view of contribution to the activity coefficients. Moreover, relatively large IM_j values for AI (> CH–) and AI (–CH$_2$–) proved the dominant role of the position of branching in the molecule as well as the hydrophobic interactions in determining ln γ^{∞} values, respectively. The descriptor of AI (> C =) with IM_j = 9.90% was also important in determination of the activity coefficients for the studied oxo compounds. According to

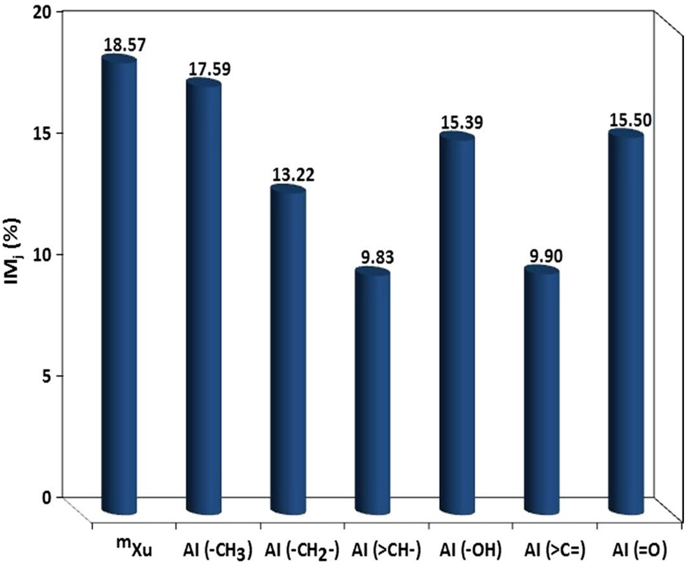

Fig. 5 Relative importance of the topological descriptors entered in the generated ANN model

the results, the topological indices entered in the proposed ANN model allowed to achieve beneficial insights about the contribution and role of the structural characteristics affecting ln γ^{∞} of the oxygen containing organic compounds.

4 Conclusion

In the study, the atomic-level *AI* topological indices combined with ^{m}Xu index were employed for SMLR and ANN modeling of the room temperature activity coefficients at infinite dilution for a group of oxygen containing organic compounds. The results showed that a 7–8-1 ANN was superior over the linear model in predicting the activity coefficients. Obtaining an average relative deviation of 2.58%, which is significantly lower than those previously reported using the regression models, validated the prediction power of the nonlinear ANN model generated for the desired data. Among the topological indices, ^{m}Xu and *AI* (–CH$_3$) were found to be the most important descriptors affecting ln γ^{∞} values indicating the major role of the molecular bulkiness and degree of molecular branching in determining the activity coefficient values. The findings of the study suggest the atomic-level topological indices combined with the neural network modeling as a promising choice to achieve improved prediction results in QSPR study of the infinite dilution activity coefficients of the oxo compounds in water at 298.15 K.

References

1. Hsieh, C.M., Lin, S.T.: Prediction of 1-octanol–water partition coefficient and infinite dilution activity coefficient in water from the PR+COSMOSAC model. Fluid Phase Equilib. **285**, 8–14 (2009)

2. Kojima, K., Zhang, S., Hiaki, T.: Measuring methods of infinite dilution activity coefficients and a database for systems including water. Fluid Phase Equilib. **131**, 145–179 (1997)
3. Cheng, J.S., Tang, M., Chen, Y.P.: Correlation of solid solubility for biological compounds in supercritical carbon dioxide: comparative study using solution model and other approaches. Fluid Phase Equilib. **194–197**, 483–491 (2002)
4. Shen, S., Nagata, I.: Prediction of excess enthalpies of ketone–alkane systems from infinite dilution activity coefficients. Themochim. Acta **258**, 19–31 (1995)
5. Morton, D.W., Young, C.L.: Henry's law constants and infinite dilution activity coefficients of C2–C8 hydrocarbons in phenylalkanes. J. Chem. Thermodyn. **28**, 895–904 (1996)
6. Sandler, S.I.: Infinite dilution activity coefficients in chemical, environmental and biochemical engineering. Fluid Phase Equilib. **116**, 343–353 (1996)
7. Lindvig, T., Hestkjær, L.L., Hansen, A.F., Michelsen, M.L., Kontogeorgis, G.M.: Phase equilibria for complex polymer solutions. Fluid Phase Equilib. **194–197**, 663–673 (2002)
8. Kolář, P., Shen, J.W., Tsuboi, A., Ishikawa, T.: Solvent selection for pharmaceuticals. Fluid Phase Equilib. **194–197**, 771–782 (2002)
9. Dohnal, V., Ondo, D.: Refined non-steady-state gas–liquid chromatography for accurate determination of limiting activity coefficients of volatile organic compounds in water: application to C_1–C_5 alkanols. J. Chromatogr. A **1097**, 157–164 (2005)
10. Krummen, M., Gmehling, J.: Measurement of activity coefficients at infinite dilution in *N*-methyl-2-pyrrolidone and *N*-formylmorpholine and their mixtures with water using the dilutor technique. Fluid Phase Equilib. **215**, 283–294 (2004)
11. Dallinga, L., Schiller, M., Gmehling, J.: Measurement of activity coefficients at infinite dilution using differential ebulliometry and non-steady-state gas-liquid chromatography. J. Chem. Eng. Data **38**, 147–155 (1993)
12. Xu, J., Zhang, H., Wang, L., Ye, W., Xu, W., Li, Z.: QSPR analysis of infinite dilution activity coefficients of chlorinated organic compounds in water. Fluid Phase Equilib. **291**, 111–116 (2010)
13. Atabati, M., Zarei, K., Borhani, A.: Predicting infinite dilution activity coefficients of hydrocarbons in water using ant colony optimization. Fluid Phase Equilib. **293**, 219–224 (2010)
14. Xu, J., Wang, L., Wang, L., Zhang, H., Xu, W.: Predicting infinite dilution activity coefficients of chlorinated organic compounds in aqueous solution based on three-dimensional WHIM and GETAWAY descriptors. J. Solution Chem. **40**, 118–130 (2011)
15. Zarei, K., Atabati, M.: Prediction of infinite dilution activity coefficients of halogenated hydrocarbons in water using classification and regression tree analysis and adaptive neuro-fuzzy inference systems. J. Solution Chem. **42**, 516–525 (2013)
16. Xiao, F., Peng, G., Nie, C., Wu, Y., Dai, Y.: Quantum topological method studies on the thermodynamic properties of polychlorinated phenoxazines. J. Mol. Struct. **1074**, 679–686 (2014)
17. Atabati, M., Emamalizadeh, R.: A Quantitative structure property relationship for prediction of flash point of alkanes using molecular connectivity indices. Chin. J. Chem. Eng. **21**, 420–426 (2013)
18. Mamy, L., Patureau, D., Barriuso, E., Bedos, C., Bessac, F., Louchart, X., Martin-Laurent, F., Miege, C., Benoit, P.: Prediction of the fate of organic compounds in the environment from their molecular properties: a review. Crit. Rev. Environ. Sci. Technol. **45**, 1277–1377 (2015)
19. Ren, B.: Novel atom-type AI indices for QSPR studies of alcohols. Comput. Chem. **26**, 223–235 (2002)
20. Ren, B., Chen, G., Xu, Y.: A novel topological index for QSPR/QSAR study of organic compounds. Acta Chim. Sin. **57**, 563–571 (1999). **(in Chinese)**
21. Ren, B., Xu, Y., Chen, G.: Estimation of heat capacity of complex organic compounds by a novel topological index. J. Chem. Eng. China **50**, 280–286 (1999). **(in Chinese)**
22. Ren, B.: Novel atomic-level-based AI topological descriptors: application to QSPR/QSAR modeling. J. Chem. Inf. Comput. Sci. **42**, 858–868 (2002)
23. Ren, B.: Application of novel atom-type AI topological indices in the structure-property correlations. J. Mol. Struct. (THEOCHEM) **586**, 137–148 (2002)
24. Ren, B.: Atom-type-based AI topological descriptors: application in structure-boiling point correlations of oxo organic compounds. J. Chem. Inf. Comput. Sci. **43**, 1121–1131 (2003)
25. Ren, B.: Atomic-level-based AI topological descriptors for structure–property correlations. J. Chem. Inf. Comput. Sci. **43**, 161–169 (2003)
26. Ren, B.: Application of novel atom-type AI topological indices to QSPR studies of alkanes. Comput. Chem. **26**, 357–369 (2002)
27. Ren, B.: Atom-type-based AI topological descriptors for quantitative structure–retention index correlations of aldehydes and ketones. Chemom. Intell. Lab. Syst. **66**, 29–39 (2003)

28. Panneerselvam, K., Antony, M.P., Srinivasan, T.G., Rao, P.R.V.: Estimation of normal boiling points of trialkyl phosphates using retention indices by gas chromatography. Thermochim. Acta **511**, 107–111 (2010)
29. Safa, F., Yekta, M.: Quantitative structure–property relationship study of standard formation enthalpies of acyclic alkanes using atom-type-based AI topological indices. Arab. J. Chem. **10**, 439–447 (2017)
30. Osaghi, B., Safa, F.: QSPR study on the boiling points of aliphatic esters using the atom-type-based AI topological indices. Rev. Roum. Chim. **64**, 183–189 (2019)
31. Safdel, F., Safa, F.: Atom-type-based AI topological indices for artificial neural network modeling of retention indices of monomethylalkanes. J. Chromatogr. Sci. **57**, 1–8 (2019)
32. Sutter, J.M., Jurs, P.C.: Adoption of Simulated Annealing to Chemical Optimization Problems. In: J.H. Kalivas (ed.), Elsevier, Amsterdam (1995)
33. Mitchell, B.E., Jurs, P.C.: Prediction of infinite dilution activity coefficients of organic compounds in aqueous solution from molecular structure. J. Chem. Inf. Comput. Sci. **38**, 200–209 (1998)
34. Dallas, A.J.: Fundamental solvatochromic and thermodynamic studies of complex chromatographic media. PhD Thesis. University of Minnesota, Minneapolis, MN (1995)
35. Li, J.: Solvatochromic and thermodynamic studies of retention in gas chromatography and gas-liquid equilibria. PhD Thesis. University of Minnesota, Minneapolis, MN (1992)
36. Sutter, J.M., Dixon, S.L., Jurs, P.C.: Automated descriptor selection for quantitative structure-activity relationships using generalized simulated annealing. J. Chem. Inf. Comput. Sci. **35**, 77–84 (1995)
37. Li, J., Dallas, A.J., Eikens, D.I., Carr, P.W., Bergmann, D.L., Hait, M.J., Eckert, C.A.: Measurement of large infinite dilution activity coefficients of nonelectrolytes in water by inert gas stripping and gas chromatography. Anal. Chem. **65**, 3212–3218 (1993)
38. Wessel, M.D.: PhD Thesis, Pennsylvania State University, University Park, PA (1996)
39. Yaws, C.L., Yang, H., Hopper, J.R., Hansen, K.C.: Organic chemicals: water solubility data. Chem. Eng. **97**, 115–116 (1990)
40. Kier, L.B., Hall, L.H.: Molecular connectivity in structure–activity studies. Research Studies Press, Letchworth (1986)
41. SPSS for Windows.: Statistical package for IBM PC, Release 20.0, SPSS Inc., https://www.spss.com.
42. MATLAB R2013a, The Math Works Inc., https://www.mathworks.com
43. Agatonovic-Kustrin, S., Beresford, R.: Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. J. Pharm. Biomed. Anal. **22**, 717–727 (2000)
44. Marquardt, D.: An algorithm for least-squares estimation of nonlinear parameters. J. Soc. Ind. Appl. Math. **11**, 431–441 (1963)
45. Andrea, T.A., Kalayeh, H.: Applications of neural networks in quantitative structure-activity relationships of dihydrofolate reductase inhibitors. J. Med. Chem. **34**, 2824–2836 (1991)
46. He, J., Zhong, Ch.: A QSPR study of infinite dilution activity coefficients of organic compounds in aqueous solutions. Fluid Phase Equilib. **205**, 303–316 (2003)
47. Estrada, E., Díaz, G.A., Delgado, E.J.: Predicting infinite dilution activity coefficients of organic compounds in water by quantum-connectivity descriptors. J. Comput. Aided Mol. Des. **20**, 539–548 (2006)
48. Aleboyeh, A., Kasiri, M.B., Olya, M.E., Aleboyeh, H.: Prediction of azo dye decolorization by UV/H_2O_2 using artificial neural networks. Dyes Pigm. **77**, 288–294 (2008)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.