J Therm Anal Calorim (2014) 116:529–538
DOI 10.1007/s10973-013-3465-9

A chemical structure-based model for estimating speed of sound in liquids

Yaser Bagheri-Chokami · Nasrin Farahani ·
Seyyed Alireza Mirkhani · Poorandokht Ilani-Kashkouli ·
Farhad Gharagheizi · Amir H. Mohammadi

Received: 17 June 2013 / Accepted: 4 October 2013 / Published online: 8 December 2013

Abstract Quantitative structure property relationship models for the speed of sound in liquids are developed based on molecular descriptors. A large dataset of 1,470 experimental data of speed of sound in 73 liquids is applied to derive the model. Twelve descriptors are selected by genetic function approximation to relate the speed of sound in liquids to their corresponding chemical structures. To capture the nonlinear nature of speed of sound in liquids, a model based on least-squared supported vector machine is also developed. The derived models are authenticated with several statistical validation techniques.

Electronic supplementary material The online version of this article (doi:10.1007/s10973-013-3465-9) contains supplementary material, which is available to authorized users.

Y. Bagheri-Chokami
Department of Chemical Engineering, South Tehran Branch,
Islamic Azad University, Tehran, Iran

N. Farahani
Department of Chemistry, Buinzahra Branch, Islamic Azad
University, Buinzahra, Iran

S. A. Mirkhani · P. Ilani-Kashkouli · F. Gharagheizi (✉)
Department of Chemical Engineering, Buinzahra Branch,
Islamic Azad University, Buinzahra, Iran
e-mail: fghara@ut.ac.ir; fghara@gmail.com

P. Ilani-Kashkouli · F. Gharagheizi · A. H. Mohammadi
Thermodynamics Research Unit, School of Engineering,
University of KwaZulu-Natal, Howard College Campus, King
George V Avenue, Durban 4041, South Africa

A. H. Mohammadi (✉)
Institut de Recherche en Génie Chimique et Pétrolier (IRGCP),
Paris Cedex, France
e-mail: a.h.m@irgcp.fr

Keywords Speed of sound · Database · QSPR model · LSSVM · Prediction · Liquids

Introduction

Speed of sound (u) is one of the important parameters in both acoustics as well as thermodynamics. Speed of sound and density are noticeable thermodynamic properties owing to their high level of experimental accuracy, which is at least one order of magnitude higher than the other quantities. Density is conventionally employed for the sake of modeling; however, recently attentions have been altered to the speed of sound regarding to the significant development of the rigorous measuring protocols in a wide array of temperature and pressure in fluid state. Quick and highly accurate measurement protocols for the speed of sound make it a reliable quantity to estimate other thermodynamic properties with high precision. All observable thermodynamic properties of a fluid phase can be directly obtained from the speed of sound by integration of partial differential equations which relate it with the other thermodynamic properties. This procedure offers promising predictions over conventional direct approaches owing to high accurate acoustic data.

In liquids, applying speed of sound data aligned with (p, ρ, T) data would offer an alternative approach to determine heat capacities instead of calorimetric method:

$$u^2 = \frac{1}{M}\left[\left(\frac{\partial \rho_{\mathrm{n}}}{\partial p}\right)_{\mathrm{T}} - \frac{T}{\rho_{\mathrm{n}}^2 C_{\mathrm{p,m}}}\left(\frac{\partial \rho_{\mathrm{n}}}{\partial T}\right)_{\mathrm{p}}^2\right]^{-1}. \tag{1}$$

Also, the combination of speed of sound with (p, ρ, T) data is the promising experimental way to determine the heat-capacity ratio γ and the isentropic compressibility κ_{s} of pure liquids:

Table 1 Predicted speed of sound values in studied liquids by GFA model

No.	Family	AARD/%	Temperature range/K	$c^{\mathrm{rep./pred.(linear)}}$ range/m s^{-1}	N
1	1,1-Difluoroethane	31.2	278.00–298.80	430.09–525.54	3
2	1-Chlorohexane	19.6	293.15–373.15	704.57–1,047.36	12
3	1-Chlorononane	12.0	293.15–373.15	849.04–1,191.84	17
4	1-Hydroxyhexane	5.4	293.00–573.00	367.46–1,377.31	20
5	1-Iodoheptane	9.3	293.15–373.15	890.29–1,233.08	17
6	2,2,4-Trimethylhexane	2.0	243.00–313.00	1,040.40–1,366.57	15
7	2,3,4-Trimethylpentane	8.9	233.00–313.00	970.67–1,346.15	17
8	2,4-Dimethylpentane	3.2	233.00–313.00	960.11–1,335.59	17
9	2,5-Dimethylhexane	1.4	243.00–313.00	1,040.91–1,367.08	15
10	2-Hydroheptafluoropropane	9.2	293.22–373.99	71.90–417.74	18
11	2-Methylhexane	1.1	243.00–313.00	1,022.95–1,349.12	15
12	3-Bromohexane	12.2	213.00–523.00	358.12–1,613.04	10
13	3-Chloroheptane	15.1	233.00–473.00	395.33–1,391.34	9
14	HFC-365mfc	13.0	298.15–423.15	196.65–708.91	8
15	R-14	7.3	95.10–200.10	449.95–1,032.21	20
16	R-143a	13.7	288.11–344.04	147.03–393.17	15
17	Acetonitrile	1.4	240.00–470.04	602.23–1,553.78	27
18	Ammonia	7.1	200.20–270.00	1,467.87–1,813.45	15
19	Benzene	4.1	283.00–523.00	407.19–1,321.68	34
20	Bromobenzene	11.5	273.00–623.00	183.00–1,409.65	34
21	Butan-2-ol	9.8	183.00–503.00	491.18–1,840.93	24
22	Butane	5.7	199.79–360.00	681.08–1,425.54	19
23	Carbon dioxide	8.6	217.03–277.50	635.21–929.59	18
24	Cetane	3.0	293.15–368.15	1,103.38–1,426.02	7
25	Chlorotrimethylsilane	14.9	301.00–313.00	979.77–1,033.32	2
26	Cyclohexane	20.8	292.85–536.45	145.22–1,054.16	18
27	Diethyl ether	32.6	362.00–362.00	902.81–902.81	1
28	*dl*-2-Pentanol	5.2	293.00–543.00	366.03–1,293.16	17
29	Ethane	12.2	91.00–275.00	719.37–1,695.46	63
30	Ethyl silicon trichloride	18.7	499.00–507.00	72.08–97.13	2
31	Ethylbenzene	4.9	193.00–593.00	253.43–1,804.76	41
32	Formamide, *N*,*N*-dimethyl-	25.1	303.15–318.15	1,031.04–1,097.61	3
33	Freon 134a	2.0	294.80–328.10	362.77–510.36	8
34	Freon 160	30.0	259.15–285.15	712.78–834.96	12
35	Heptane	3.9	193.00–513.15	300.25–1,628.72	43
36	Hexane, 1-iodo-	8.6	293.15–373.15	850.31–1,193.11	17
37	Hexane, 2,2-dimethyl-	1.5	243.00–313.00	1,030.71–1,356.88	15
38	Hexane, 3-methyl-	2.3	243.00–313.00	1,020.41–1,346.58	15
39	Isooctane	4.0	233.00–313.00	1,056.74–1,432.22	17
40	Methane	9.5	91.00–160.00	1,043.18–1,436.17	48
41	Methanol	32.1	374.00–393.20	537.11–612.80	2
42	Methylbenzene	2.1	193.00–493.00	535.11–1,800.51	25
43	*n*-Butyl alcohol	6.1	193.00–533.00	399.49–1,787.41	27
44	*n*-Decane	7.7	313.15–593.15	193.54–1,165.06	15
45	*n*-Docosane	3.9	333.00–473.00	835.31–1,368.76	15
46	*n*-Dodecane	12.3	393.15–633.15	157.86–892.78	13
47	*n*-Hexane	3.6	183.00–473.15	374.57–1,627.84	40
48	*n*-Octane	4.4	313.15–533.15	283.73–1,091.91	12

Table 1 continued

No.	Family	AARD/%	Temperature range/K	$c^{\text{rep./pred.(linear)}}$ range/m s^{-1}	N
49	*n*-Propane	8.8	90.00–325.21	689.03–1,897.45	49
50	*n*-Tetracosane	5.1	333.00–473.00	868.69–1,402.15	15
51	*n*-Tricosane	4.3	333.00–473.00	849.94–1,383.40	15
52	Nitrogen, diatomic	23.8	63.30–85.00	1,038.45–1,169.61	11
53	Nonane	7.8	323.00–578.00	198.70–1,088.07	22
54	Octamethylcyclotetrasiloxane	6.5	293.11–439.64	448.21–1,043.00	18
55	Oxygen	5.8	58.00–134.88	640.90–1,093.94	106
56	Pentafluoro(trifluoromethyl)benzene	21.2	293.00–353.00	450.37–711.60	4
57	Pentafluoroethane	10.2	293.10–335.50	139.74–326.85	7
58	Pentan-1-ol	5.0	293.00–553.00	377.80–1,333.18	23
59	Pentane	9.5	153.00–433.00	459.07–1,735.20	15
60	Pentane, 2,3-dimethyl-	1.1	233.00–313.00	1,045.09–1,420.56	16
61	Pentane, 2-methyl-	0.6	233.00–313.00	978.29–1,353.77	17
62	Pentane, 3-methyl-	1.5	233.00–313.00	987.60–1,363.08	17
63	Perfluoro-3-methyl-2-pentene	21.9	323.15–363.15	212.21–380.89	3
64	Perfluoroheptane	10.0	363.15–423.15	168.34–400.98	4
65	Perfluoroisononane	14.9	413.15–513.15	78.35–418.53	6
66	Perfluorooctane	15.2	383.15–483.15	80.33–440.89	6
67	Propene	3.9	88.00–281.70	978.80–2,003.91	44
68	Silane, dichlorodimethyl-	20.1	528.00–556.00	55.03–135.29	7
69	Trichloromonofluoromethane	9.4	162.68–457.72	129.00–1,439.19	158
70	Undecane	10.5	393.15–613.15	175.59–864.21	12
71	Water	22.8	452.57–646.47	369.05–914.90	11
72	Water-d2	20.9	578.22–643.23	376.07–532.05	7

Table 2 Model's descriptors

Item	Descriptor	Definition	Class	Reference
1	*AAC*	Mean information index on atomic composition	Information indices	[25]
2	Y_{index}	Balaban Y_{index}	Information indices	[26]
3	*SPH*	Spherosity	Geometrical descriptors	[27]
4	*Mor13m*	Signal 13/weighted by mass	3D-MoRSE descriptors	[28]
5	*E2v*	2nd component accessibility directional WHIM index/weighted by van der Waals volume	WHIM descriptors	[29]
6	*Ds*	D total accessibility index/weighted by I-state	WHIM descriptors	[29]
7	*HATS1m*	Leverage-weighted autocorrelation of lag 1/weighted by mass	GETAWAY descriptors	[30]
8	*RTp*	R total index/weighted by polarizability	GETAWAY descriptors	[30]
9	*nRCN*	Number of nitriles (aliphatic)	Functional group counts	–
10	*nHDon*	Number of donor atoms for H-bonds (N and O)	Functional group counts	–

$$u^2 = \frac{1}{\rho \kappa_S}, \tag{2}$$

$$u^2 = \frac{\gamma}{\rho \kappa_T}, \tag{3}$$

where

$$\kappa_S = \frac{1}{\rho}\left(\frac{\partial \rho}{\partial P}\right)_S, \tag{4}$$

$$\kappa_T = \frac{1}{\rho}\left(\frac{\partial \rho}{\partial P}\right)_T, \tag{5}$$

$$\gamma = \frac{C_p}{C_V}. \tag{6}$$

At higher pressures, (p, ρ, T) measurements are much more difficult and in this region sound speed measurements in liquids are probably of the greatest value [1].

In this communication, the quantitative structure property relationship (QSPR) methodology [2–8] is successfully applied for prediction of u for a wide array of liquids at the broad spectrum of temperatures.

Methodology

Data preparation

In this study, a comprehensive dataset of speed of sound comprising 1,470 data belongs to 73 liquids in a wide range of temperature (58–646.47 K) was extracted from ThermoData Engine [9]. In terms of reliability as well as the critical evaluation of the experimental data, ThermoData Engine would be one of the most promising options to collect experimental data.

Training and test set selection

Typically, in QSPR modeling, the compiled experimental database is split into two subsets: training set which is involved in model development and the test set used to assess the learning ability of the model from training set to produce reliable results for absent compounds. In this study, K-means clustering is applied to select training and test sets. K-means clustering is a method of cluster analysis, which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. As the rule of thumb, 20 % of collecting data was retained to test

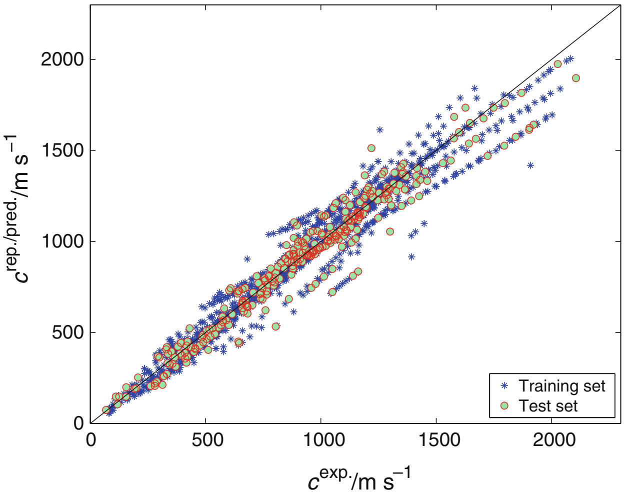

Fig. 1 Predicted speed of sound values by GFA model versus the experimental ones

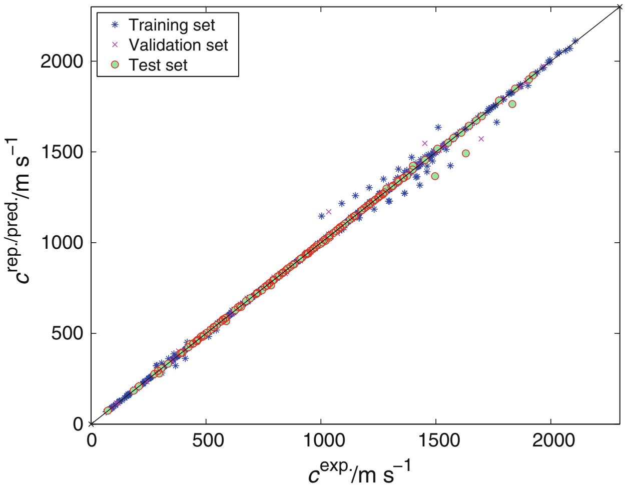

Fig. 3 Predicted speed of sound values by LSSVM model versus the experimental ones

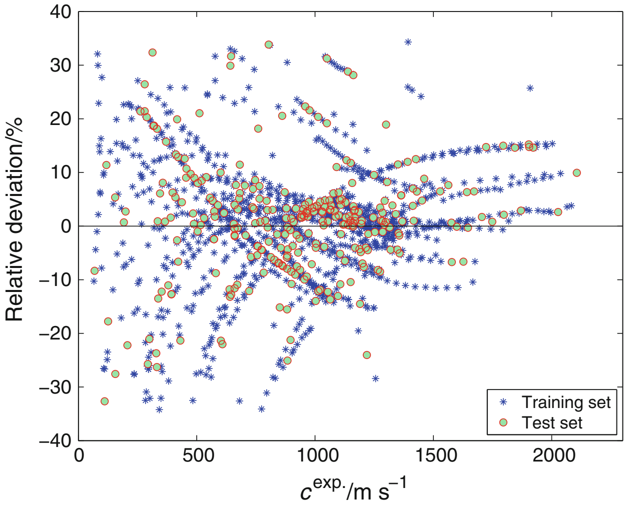

Fig. 2 Deviation of the predicted speed of sound by GFA model from experimental data

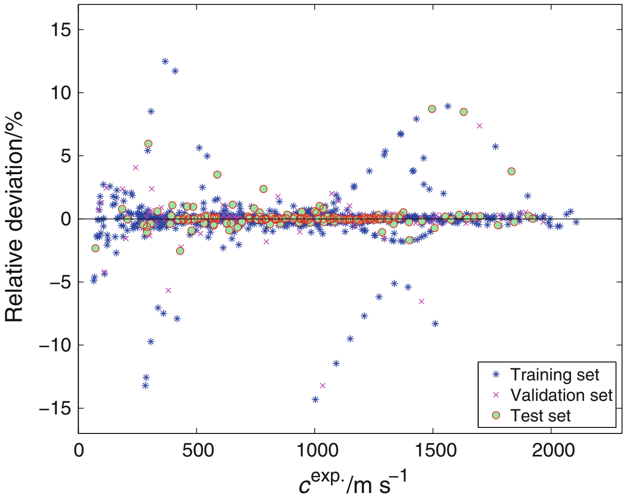

Fig. 4 Deviation of the predicted speed of sound by LSSVM model from experimental data

Table 3 Predicted speed of sound values in studied liquids by LSSVM model

No.	Family	AARD/%	Temperature range/K	$c^{\text{rep./pred.}}$ range/m s^{-1}	*N*
1	1,1-Difluoroethane	31.2	278.00–298.80	430.09–525.54	3
2	1-Chlorohexane	19.6	293.15–373.15	704.57–1,047.36	12
3	1-Chlorononane	12.0	293.15–373.15	849.04–1,191.84	17
4	1-Hydroxyhexane	5.4	293.00–573.00	367.46–1,377.31	20
5	1-Iodoheptane	9.3	293.15–373.15	890.29–1,233.08	17
6	2,2,4-Trimethylhexane	2.0	243.00–313.00	1,040.40–1,366.57	15
7	2,3,4-Trimethylpentane	8.9	233.00–313.00	970.67–1,346.15	17
8	2,4-Dimethylpentane	3.2	233.00–313.00	960.11–1,335.59	17
9	2,5-Dimethylhexane	1.4	243.00–313.00	1,040.91–1,367.08	15
10	2-Hydroheptafluoropropane	9.2	293.22–373.99	71.90–417.74	18
11	2-Methylhexane	1.1	243.00–313.00	1,022.95–1,349.12	15
12	3-Bromohexane	12.2	213.00–523.00	358.12–1,613.04	10
13	3-Chloroheptane	15.1	233.00–473.00	395.33–1,391.34	9
14	HFC-365mfc	13.0	298.15–423.15	196.65–708.91	8
15	R-14	7.3	95.10–200.10	449.95–1,032.21	20
16	R-143a	13.7	288.11–344.04	147.03–393.17	15
17	Acetonitrile	1.4	240.00–470.04	602.23–1,553.78	27
18	Ammonia	7.1	200.20–270.00	1,467.87–1,813.45	15
19	Benzene	4.1	283.00–523.00	407.19–1,321.68	34
20	Bromobenzene	11.5	273.00–623.00	183.00–1,409.65	34
21	Butan-2-ol	9.8	183.00–503.00	491.18–1,840.93	24
22	Butane	5.7	199.79–360.00	681.08–1,425.54	19
23	Carbon dioxide	8.6	217.03–277.50	635.21–929.59	18
24	Cetane	3.0	293.15–368.15	1,103.38–1,426.02	7
25	Chlorotrimethylsilane	14.9	301.00–313.00	979.77–1,033.32	2
26	Cyclohexane	20.8	292.85–536.45	145.22–1,054.16	18
27	Diethyl ether	32.6	362.00–362.00	902.81–902.81	1
28	*dl*-2-Pentanol	5.2	293.00–543.00	366.03–1,293.16	17
29	Ethane	12.2	91.00–275.00	719.37–1,695.46	63
30	Ethyl silicon trichloride	18.7	499.00–507.00	72.08–97.13	2
31	Ethylbenzene	4.9	193.00–593.00	253.43–1,804.76	41
32	Formamide, *N*,*N*-dimethyl-	25.1	303.15–318.15	1,031.04–1,097.61	3
33	Freon 134a	2.0	294.80–328.10	362.77–510.36	8
34	Freon 160	30.0	259.15–285.15	712.78–834.96	12
35	Heptane	3.9	193.00–513.15	300.25–1,628.72	43
36	Hexane, 1-iodo-	8.6	293.15–373.15	850.31–1,193.11	17
37	Hexane, 2,2-dimethyl-	1.5	243.00–313.00	1,030.71–1,356.88	15
38	Hexane, 3-methyl-	2.3	243.00–313.00	1,020.41–1,346.58	15
39	Isooctane	4.0	233.00–313.00	1,056.74–1,432.22	17
40	Methane	9.5	91.00–160.00	1,043.18–1,436.17	48
41	Methanol	32.1	374.00–393.20	537.11–612.80	2
42	Methylbenzene	2.1	193.00–493.00	535.11–1,800.51	25
43	*N*-butyl alcohol	6.1	193.00–533.00	399.49–1,787.41	27
44	*n*-Decane	7.7	313.15–593.15	193.54–1,165.06	15
45	*n*-Docosane	3.9	333.00–473.00	835.31–1,368.76	15
46	*n*-Dodecane	12.3	393.15–633.15	157.86–892.78	13
47	*n*-Hexane	3.6	183.00–473.15	374.57–1,627.84	40
48	*n*-Octane	4.4	313.15–533.15	283.73–1,091.91	12

Table 3 continued

No.	Family	AARD/%	Temperature range/K	$c^{\text{rep./pred.}}$ range/m s^{-1}	*N*
49	*n*-Propane	8.8	90.00–325.21	689.03–1,897.45	49
50	*n*-Tetracosane	5.1	333.00–473.00	868.69–1,402.15	15
51	*n*-Tricosane	4.3	333.00–473.00	849.94–1,383.40	15
52	Nitrogen, diatomic	23.8	63.30–85.00	1,038.45–1,169.61	11
53	Nonane	7.8	323.00–578.00	198.70–1,088.07	22
54	Octamethylcyclotetrasiloxane	6.5	293.11–439.64	448.21–1,043.00	18
55	Oxygen	5.8	58.00–134.88	640.90–1,093.94	106
56	Pentafluoro(trifluoromethyl)benzene	21.2	293.00–353.00	450.37–711.60	4
57	Pentafluoroethane	10.2	293.10–335.50	139.74–326.85	7
58	Pentan-1-ol	5.0	293.00–553.00	377.80–1,333.18	23
59	Pentane	9.5	153.00–433.00	459.07–1,735.20	15
60	Pentane, 2,3-dimethyl-	1.1	233.00–313.00	1,045.09–1,420.56	16
61	Pentane, 2-methyl-	0.6	233.00–313.00	978.29–1,353.77	17
62	Pentane, 3-methyl-	1.5	233.00–313.00	987.60–1,363.08	17
63	Perfluoro-3-methyl-2-pentene	21.9	323.15–363.15	212.21–380.89	3
64	Perfluoroheptane	10.0	363.15–423.15	168.34–400.98	4
65	Perfluoroisononane	14.9	413.15–513.15	78.35–418.53	6
66	Perfluorooctane	15.2	383.15–483.15	80.33–440.89	6
67	Propene	3.9	88.00–281.70	978.80–2,003.91	44
68	Silane, dichlorodimethyl-	20.1	528.00–556.00	55.03–135.29	7
69	Trichloromonofluoromethane	9.4	162.68–457.72	129.00–1,439.19	158
70	Undecane	10.5	393.15–613.15	175.59–864.21	12
71	Water	22.8	452.57–646.47	369.05–914.90	11
72	Water-d2	20.9	578.22–643.23	376.07–532.05	7

the model and the remaining was applied for model derivation [10]. For LSSVM model derivation, 80-10-10 % of data points split into training-validation-test sets, respectively. This selection like the previous one is performed by *k*-means clustering.

Calculation of descriptor

Prior to the descriptor calculation, the optimization of 3D structures of present compounds is required. The well-known Dreiding Force field [11] implemented by Chemaxon's JChem software was applied to optimize 3D structures in this study. About 3,000 descriptors from 22 diverse classes of descriptors are calculated by Dragon software [12]. These 22 classes of descriptors are Constitutional descriptors, Topological indices, Walk and path counts, Connectivity indices, Information indices, 2D autocorrelations, Burden Eigen values, Edge-adjacency indices, Functional group counts, Atom-centered fragments, Molecular properties, topological charge indices, Eigenvalue-based indices, Randic molecular profiles, geometrical descriptors, RDF descriptors, 3D-MORSE descriptors, WHIM descriptors, GETAWAY descriptors, charge descriptors, 2D binary fingerprint, and 2D frequency finger print.

Descriptors that could not be calculated for a compound are excluded completely from the list. Next, the pair correlation for each binary group of descriptors is performed. For correlation greater than 0.9, one of the descriptors is omitted randomly.

Subset variable selection

Genetic function approximation (GFA) was successfully implemented for subset variable selection in this study. GFA originally developed by Rogers and Hopfinger [13] is the fusion of two seemingly distinctive algorithms: multivariate adaptive regression splines algorithm [14] and genetic algorithm [15]. One of the promising features of the GFA is to evolve a series of models instead of one model. In addition, by utilizing the Friedman's LOF scoring function in GFA, derived models are prone to overfitting with better predictions. In this study, population and the number of maximum generations are set to 100 and 5,000, respectively. The value of mutation probability is set to 1.5 in this study. To study the nonlinear nature of speed of sound, the LSSVM is also practiced for the sake of model derivation.

Result and discussion

Linear model

The final linear model derived by GFA for estimation of speed of sound in liquids contains 12 descriptors as follows:

$$\begin{aligned} c = {} & 1564.5671(\pm 15.2855) - 6.5480(\pm 0.1143)T \\ & + 0.0034(\pm 0.0001)T^2 + 438.7373(\pm 12.9701)AAC \\ & + 12.8962(\pm 2.2292)Y_{\text{index}} + 455.4282(\pm 12.5541)SPH \\ & - 336.8196(\pm 7.5202)Mor13m + 432.8956(\pm 22.0591)E2v \\ & - 504.0771(\pm 16.3251)Ds - 837.9355(\pm 25.5316)HATS1m \\ & + 69.1189(\pm 1.6293)RTp + 635.6081(\pm 20.0058)NRCN \\ & + 193.9080(\pm 6.4533)nHDon \end{aligned} \tag{7}$$

$R^2 = 0.949$; $R^2_{\text{adj}} = 0.948$; $n_{\text{Training}} = 1176$; $n_{\text{Test}} = 294$; $F = 21652.59$; $Q^2 = 0.938$; $Q^2_{\text{boot}} = 0.948$; $Q^2_{\text{ext}} = 0.952$ $a(R^2) = -0.019$; $\Delta K = 0.974$; $\Delta Q = 0$; $R^{\text{P}} = 0$; $R^{\text{N}} = 0.996$

Table 4 Statistical parameters of LSSVM model

Statistical parameter	
Training set	
R^{2a}	0.999
Average absolute relative deviation[b]	0.4
Standard deviation error[c]	13.27
Root mean square error[d]	13.27
N^{e}	1,158
Validation set	
R^2	0.998
Average absolute relative deviation	0.6
Standard deviation error	18.09
Root mean square error	18.03
N	144
Test set	
R^2	0.998
Average absolute relative deviation	0.5
Standard deviation error	17.33
Root mean square error	17.46
N	144
Total	
R^2	0.999
Average absolute relative deviation	0.5
Standard deviation error	14.27
Root mean square error	14.27
N	1,446

[a] $R^2 = 1 - \sum_{i=1}^{N} \frac{(\text{Calc.}(i) - \text{Exp.}(i))^2}{(\text{Calc.}(i) - \text{Average}(\text{Exp.}(i)))^2}$

[b] $\%\,\text{AAD} = \frac{100}{N-n} \sum_{i=1}^{N} \frac{|\text{Calc}(i) - \text{Exp}(i)|}{\text{Exp}(i)}$

[c] $\text{Std} = \frac{1}{N} \sqrt{\sum_{i=1}^{N} (\text{Calc.}(i) - \text{Average}(\text{Calc.}(i))^2}$

[d] $\text{RMSE} = \left(\frac{\sum_{i=1}^{N} (\text{calc.}(i) - \text{Exp.}(i))^2}{n} \right)^{\frac{1}{2}}$

[e] Number of experimental data

n Number of model parameters

Table 1 demonstrates the GFA predicted values of speed of sounds in the studied liquids (u is in m s^{-1} unit). The definitions of molecular descriptors in Eq. 7 are enlisted in Table 2. Figure 1 illustrates the predicted speed of sound values versus experimental data. As it can be seen, the majority of points are located in the vicinity of the bisection of graph. This indicates the reasonable agreement between GFA predicted values versus experimental ones. Relative deviations of GFA predicted values from experimental data are depicted in Fig. 2.

Nonlinear model

For the sake of nonlinear modeling, LSSVM was successfully implemented in this study. LSSVM is a member of large machine-learning family namely support vector machine (SVM) which profoundly based on the seeking of an optimal separating hyperplane to minimize expected generalization error in the feature space. The detailed

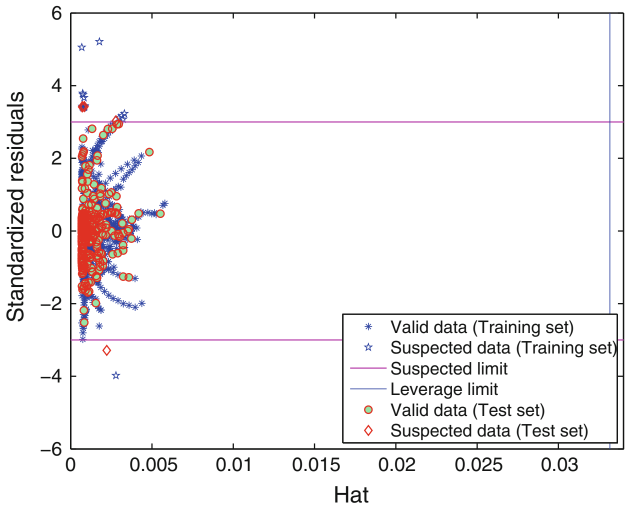

Fig. 5 Williams graph of the developed model

mathematical explanation of the optimization problem treated by LSSVM approach is not provided here and can be found in detail in mentioned references [16–18]. To implement the original SVM algorithm to handle nonlinear problem, radial basis function is defined as the kernel function. The objective of the definition of the kernel function is to map the data into the higher dimensional feature space in order to increase computational power. The simulated annealing optimization method is actuated to find the proper combination of the LSSVM parameters, namely (γ, σ^2) considering the minimum mean squared error of leave-one-out (LOO) cross-validation of the training set as the optimal condition. The twelve descriptors selected for linear model by GFA were introduced as inputs to LSSVM for the nonlinear model derivation. The obtained parameters of the final model are described as follows: $\gamma = 14649.98$, $\sigma^2 = 0.5602$.

Figure 3 shows the LSSVM predicted values versus the experimental speeds of sound. As it is vividly clear in this figure, the great improvement of prediction is achieved by employing LSSVM instead of GFA. Besides, deviation of predicted LSSVM values from experimental ones is depicted in Fig. 4. The significant reduction of deviation of predicted values by LSSVM model in comparison with GFA model is apparent in this figure. Table 3 demonstrates the LSSVM predicted values of speed of sounds in the studied liquids. Statistical parameters of LSSVM model are enlisted in Table 4.

Applicability domain (AD) [19]

To test the reliability of the predicted responses, the AD of the derived model is investigated. AD is a theoretical spatial domain defined by molecular descriptors as well as by both training and test sets. The AD objective is to investigate whether the test and training sets share the same domain or not. This is crucial since prediction outside of the AD might be erroneous.

In this study, Williams graph generated from Hat indices is used to investigate AD. Hat indices are calculated based on Hat matrix (H) with the following definition:

$$H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}, \tag{8}$$

where X is a two-dimensional matrix comprising n compounds (rows) and k descriptors (columns). The diagonal elements of H are leverages or hat values (h_{i}) of the chemicals in the descriptor space.

Williams graph shows the correlation of hat values and standardized cross-validated residuals (R). A warning leverage ($h^* = 0.0337$)—blue vertical line—is generally fixed at $3n/p$, where n is number of training compounds and p is the number of model variables plus one. The leverage of 3 is considered as a cutoff value to accept the points that lay ±3 (two horizontal red lines) standard deviations from the mean (to cover 99 % normally distributed data).

The AD is located in the region of $0 \le h \le 0.0337$ and $-3 \le R \le +3$. The prediction within this region is considered valid. As it is clearly illustrated in this figure, the majority of test and training compounds are located in this region. There are 24 points that wrongly predicted by the model ($3 < R$ or $R < -3$), however, their hat values lie in the domain of AD. This erroneous prediction could probably be attributed to wrong experimental data rather than the molecular structure [20]. Figure 5 depicts the Williams graph of the studied model.

The absolute relative deviation is defined as follows:

Table 5 Validation techniques

Item	Validation technique	Interpretation	Numerical value	Reference
1	R^2	Its high value suggests the reliability of the model	0.9486	–
2	F test	Its high value suggests the reliability of the model	21,652.59	[31]
3	Q^2	Statistical parameter of well-known internal validation technique namely leave one out (LOO). Its closeness to the R^2 suggests the model is reliable	0.9484	[32]
4	R^2_{adj}	Its closeness to the R^2 suggests the reliability of the model	0.9485	[32]
5	RQK	Satisfying four proposed constraints guarantees that the model is prone to being a chance correlation. To insure model reliability, all the four constraints should have values equal or greater than zero	$R^{\mathrm{P}} = 0$ $R^{\mathrm{N}} = 0.996$ $\Delta K = 0.974$ $\Delta Q = 0$	[17, 33–36]
6	Boot strap (Q^2_{Boot})	Its closeness to the R^2 suggests the reliability of the model	0.9482	[37]
7	Y-Scrambling (a)	Examine the immunity of the model to the chance correlation by studying the model response to the shuffled prediction set. Near zero values of a reflecting nonchance correlation	−0.019	[38]
8	External validation technique (Q^2_{ext})	Its closeness to the R^2 suggests the reliable model	0.952	[39]

$$\mathrm{ARD}\% = \left(\frac{1}{N_\mathrm{p}}\right)\sum_{i=1}^{N_\mathrm{p}}\left|\frac{c_\mathrm{exp} - c_\mathrm{calc}}{c_\mathrm{exp}}\right|, \tag{9}$$

where N_p is the number of total points. The GFA-driven model shows that for 72 studied liquids the mean ARD % is 10.4 % with maximum deviation of 34.2 %.The highest error associated with GFA model belongs to water at $T = 452.57$. However, this point is located at the wrongly predicted area with the high chance of being wrong experimental data. 32.7 % of the estimated speed of sound was within absolute deviation of 0.00–3.00 %, 19.9 % was within 3.001–6.00 %, 16.5 % was within 6.001–10.00 %, 8.2 % was within 10.001–13.00 %, and only 11.1 % was within 13.001–20 % and 11.6 % was within 20.1–34.2 %. The results obtained by the nonlinear model present that 98.4 % of the estimated speeds of sounds were within absolute deviation of 0.00–3.00 %, and merely 1.6 % of the predicted value have the error higher than 3 %.

The applied validation techniques as well as their interpretations are shown in Table 5. The readers can find the detailed statistical procedures of the mentioned techniques from previous works of the authors [2, 7, 8, 18, 21–24]. The results of validation techniques indicate that the derive model is not only an accurate one but also prone to being a chance-correlated model.

Conclusions

In the light of highly accurate measurement protocols, the application of speed of sound to correlate thermodynamic properties of liquids received many attentions in the recent decade. The ease of measurement as well as highly precise mensuration make speed of sound a reliable option to replace arduous (p, ρ, T) measurement at high pressures. Despite its broad applications, there is no study conducted on the prediction of speed of sounds in liquid.

Originally, in this communication a robust twelve-parameter QSPR model is introduced to estimate speed of sounds of 73 liquids at wide range of temperatures. GFA is applied for subset variable selection as well linear model derivation. For the sake of more accurate modeling as well as studying the nonlinearity of the speed of sound, LSSVM approach is also practiced to develop a nonlinear model. The results of LSSVM modeling reveal significant improvement of prediction power as well as substantial reduction of predicted values deviation from experimental ones.

For the sake of the investigation of the model reliability, AD of the model is also studied. The presence of the majority of both training and test sets data in the AD generated by Williams graph authenticates the validity of the predictions. Besides, Analysis based on AD of the derived model and LSSVM ($0 \leq h \leq 0.0337$ and $R > 3$ or < -3) implies that reported experimental data for 24 data points are ambiguous and need modification. By the aid of derived model parameters as well as its AD, the reliability of the experimental data could be analyzed to find flawed data points. Moreover, the reliability and predictive capability of the model are adequately scrutinized by various statistical validation techniques. The results of validation techniques pronounced that the model is stable and accurate and is immune of chance correlation. Predicted speed of sounds by both GFA and LSSVM model for studied data points as well as corresponding model descriptors values are provided as supplementary information.

References

1. Goodwin ARH, Trusler JPM. Speed of sound. In: Goodwin ARH, Marsh KN, Wakeham WA, editors. Experimental thermodynamics. Amsterdam: Elsevier; 2003. p. 237–323.
2. Gharagheizi F, Eslamimanesh A, Ilani-Kashkouli P, Mohammadi AH, Richon D. QSPR molecular approach for representation/prediction of very large vapor pressure dataset. Chem Eng Sci. 2012;76:99–107.
3. Gharagheizi F, Eslamimanesh A, Sattari M, Mohammadi AH, Richon D. Corresponding states method for evaluation of the solubility parameters of chemical compounds. Ind Eng Chem Res. 2012;51:3826–31.
4. Gharagheizi F, Eslamimanesh A, Sattari M, Tirandazi B, Mohammadi AH, Richon D. Evaluation of thermal conductivity of gases at atmospheric pressure through a corresponding states method. Ind Eng Chem Res. 2012;51:3844–9.
5. Gharagheizi F, Gohar MRS, Vayeghan MG. A quantitative structure-property relationship for determination of enthalpy of fusion of pure compounds. J Therm Anal Calorim. 2012;109:501–6.
6. Gharagheizi F, Ilani-Kashkouli P, Mohammadi AH. Computation of normal melting temperature of ionic liquids using a group contribution method. Fluid Phase Equilibria. 2012;329:1–7.
7. Mirkhani SA, Gharagheizi F, Ilani-Kashkouli P, Farahani N. Determination of the glass transition temperature of ionic liquids: a molecular approach. Thermochim Acta. 2012;543:88–95.
8. Mirkhani SA, Gharagheizi F, Ilani-Kashkouli P, Farahani N. An accurate model for the prediction of the glass transition temperature of ammonium based ionic liquids: a QSPR approach. Fluid Phase Equilibria. 2012;324:50–63.
9. Frenkel M, Chirico RD, Diky V, Yan X, Dong Q, Muzny C. ThermoData Engine (TDE): software implementation of the dynamic data evaluation concept. J Chem Inf Model. 2005;45:816–38.
10. Gharagheizi F. QSPR analysis for intrinsic viscosity of polymer solutions by means of GA-MLR and RBFNN. Comput Mater Sci. 2007;40:159–67.
11. Mayo SL, Olafson BD, Goddard WA. DREIDING: a generic force field for molecular simulations. J Phys Chem. 1990;94:8897–909.
12. Talete S. Dragon for windows (Software for Molecular Descriptor Calculations), Version 5.5. 2007. http://www.talete.mi.it/.
13. Rogers D, Hopfinger AJ. Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. J Chem Inf Comput Sci. 1994;34:854–66.

14. Friedman JH. Multivariate adaptive regression splines. Ann Stat. 1991;19:1–67.
15. Holland JH. Adaptation in natural and artificial systems : an introductory analysis with applications to biology, control, and artificial intelligence. 1st MIT Press ed. ed: MIT Press; 1992.
16. Eslamimanesh A, Gharagheizi F, Illbeigi M, Mohammadi AH, Fazlali A, Richon D. Phase equilibrium modeling of clathrate hydrates of methane, carbon dioxide, nitrogen, and hydrogen + water soluble organic promoters using support vector machine algorithm. Fluid Phase Equilibria. 2012;316:34–45.
17. Gharagheizi F, Eslamimanesh A, Farjood F, Mohammadi AH, Richon D. Solubility parameters of nonelectrolyte organic compounds: determination using quantitative structure-property relationship strategy. Ind Eng Chem Res. 2011;50: 11382–95.
18. Mousavisafavi SM, Gharagheizi F, Mirkhani SA, Akbari J. A predictive quantitative structure-property relationship for glass transition temperature of 1,3-dialkyl imidazolium ionic liquids—Part 2. The nonlinear approach. J Therm Anal Calorim. 2013;111:1639–48.
19. Gramatica P. Modelling chemicals in the environment. In: Livingstone DJ, Davis AM, editors. Drug design strategies: quantitative approaches. London: The Royal Society of Chemistry; 2012. p. 458–78.
20. Gramatica P. Principles of QSAR models validation: internal and external. QSAR Comb Sci. 2007;26:694–701.
21. Gharagheizi F, Ilani-Kashkouli P, Mirkhani SA, Farahani N, Mohammadi AH. QSPR molecular approach for estimating Henry's law constants of pure compounds in water at ambient conditions. Ind Eng Chem Res. 2012;51:4764–7.
22. Mirkhani SA, Gharagheizi F. Predictive quantitative structure-property relationship model for the estimation of ionic liquid viscosity. Ind Eng Chem Res. 2012;51:2470–7.
23. Mirkhani SA, Gharagheizi F, Sattari M. A QSPR model for prediction of diffusion coefficient of non-electrolyte organic compounds in air at ambient condition. Chemosphere. 2012;86:959–66.
24. Mousavisafavi SM, Mirkhani SA, Gharagheizi F, Akbari J. A predictive quantitative structure-property relationship for glass transition temperature of 1,3-dialkyl imidazolium ionic liquids—Part 1. The linear approach. J Therm Anal Calorim. 2013;111:235–46.
25. Bonchev D. Information theoretic indices for characterization of chemical structures. Chichester: Research Studies Press; 1983.
26. Balaban AT, Balaban T-S. New vertex invariants and topological indices of chemical graphs based on information on distances. J Math Chem. 1991;8:383–97.
27. Mekenyan O, Peitchev D, Bonchev D, Trinajstic N. Arzneim Forsch. 1986;36:176–83.
28. Gasteiger JE. Software-Entwicklung in der Chemie 10 = Software development in chemistry: GDCh; 1995.
29. Todeschini R, Bettiol C, Giurin G, Gramatica P, Miana P, Argese E. Modeling and prediction by using WHIM descriptors in QSAR studies: submitochondrial particles (SMP) as toxicity blosensors of chlorophenols. Chemosphere. 1996;33:71–9.
30. Consonni V, Todeschini R, Pavan M. Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. J Chem Inf Comput Sci. 2002;42:682–92.
31. Krzanowski WJ. Principles of multivariate analysis: a user's perspective. Rev ed. Oxford: Oxford University Press; 2000.
32. Todeschini R, Consonni V. Molecular descriptors for chemoinformatics. 2nd ed., rev. and Enl. ed. Weinheim: Wiley-VCH 2009.
33. Gharagheizi F, Eslamimanesh A, Mohammadi AH, Richon D. QSPR approach for determination of parachor of non-electrolyte organic compounds. Chem Eng Sci. 2011;66:2959–67.
34. Gharagheizi F, Eslamimanesh A, Mohammadi AH, Richon D. Representation/prediction of solubilities of pure compounds in water using artificial neural network-group contribution method. J Chem Eng Data. 2011;56:720–6.
35. Gharagheizi F, Eslamimanesh A, Mohammadi AH, Richon D. Use of artificial neural network-group contribution method to determine surface tension of pure compounds. J Chem Eng Data. 2011;56:2587–601.
36. Gharagheizi F, Gohar MRS, Vayeghan MG. A quantitative structure-property relationship for determination of enthalpy of fusion of pure compounds. J Therm Anal Calorim. 2011;27:1–6.
37. Efron B. Better bootstrap confidence intervals. J Am Stat Assoc. 1987;82:171–85.
38. Lindgren F, Hansen B, Karcher W, Sjöström M, Eriksson L. Model validation by permutation tests: applications to variable selection. J Chemom. 1996;10:521–32.
39. Chiou J. Hybrid method of evolutionary algorithms for static and dynamic optimization problems with application to a fed-batch fermentation process. Comput Chem Eng. 1999;23:1277–91.