



OPEN

Modeling of ionic liquids viscosity via advanced white-box machine learning

Sajad Kiani¹, Fahimeh Hadavimoghaddam^{2,3}, Saeid Atashrouz⁴✉, Dragutin Nedeljkovic⁵, Abdolhossein Hemmati-Sarapardeh^{6,7}✉ & Ahmad Mohadespour⁸✉

Ionic liquids (ILs) are more widely used within the industry than ever before, and accurate models of their physicochemical characteristics are becoming increasingly important during the process optimization. It is especially challenging to simulate the viscosity of ILs since there is no widely agreed explanation of how viscosity is determined in liquids. In this research, genetic programming (GP) and group method of data handling (GMDH) models were used as white-box machine learning approaches to predict the viscosity of pure ILs. These methods were developed based on a large open literature database of 2813 experimental viscosity values from 45 various ILs at different pressures (0.06–298.9 MPa) and temperatures (253.15–573 K). The models were developed based on five, six, and seven inputs, and it was found that all the models with seven inputs provided more accurate results, while the models with five and six inputs had acceptable accuracy and simpler formulas. Based on GMDH and GP proposed approaches, the suggested GMDH model with seven inputs gave the most exact results with an average absolute relative deviation (AARD) of 8.14% and a coefficient of determination (R^2) of 0.98. The proposed techniques were compared with theoretical and empirical models available in the literature, and it was displayed that the GMDH model with seven inputs strongly outperforms the existing approaches. The leverage statistical analysis revealed that most of the experimental data were located within the applicability domains of both GMDH and GP models and were of high quality. Trend analysis also illustrated that the GMDH and GP models could follow the expected trends of viscosity with variations in pressure and temperature. In addition, the relevancy factor portrayed that the temperature had the greatest impact on the ILs viscosity. The findings of this study illustrated that the proposed models represented strong alternatives to time-consuming and costly experimental methods of ILs viscosity measurement.

Keywords Ionic liquids, Viscosity, GMDH, GP, White-box modeling, Leverage method

Ionic Liquids (ILs) are novel, highly tunable, and unique compounds that emerged in response to interest in green chemical technologies¹. In the last few decades, an enormous amount of research has gone into developing ILs for a wide range of uses, from industrial to molecular, such as gas absorption, energy storage, biotechnology, electrochemistry, separation, and fluid flow in porous media^{2–5}. ILs are materials composed only of ions and having a melting point of less than 100 °C. They are created chemically when organic cations (such pyridinium, phosphonium, imidazolium, and ammonium) combine with organic and inorganic anions (like phosphates, halides, and sulfates)^{5,6}. When Paul Walden originally described the IL (ethylammonium nitrate ([NH₃H₂]⁺[NO₃]⁻)) back in 1914, he had no idea that nearly a century later the field of ILs would become very significant⁶.

¹Faculty of Science and Engineering, Swansea University, Swansea SA1 8EN, UK. ²Key Laboratory of Continental Shale Hydrocarbon Accumulation and Efficient Development (Northeast Petroleum University), Ministry of Education, Northeast Petroleum University, Daqing 163318, Heilongjiang, China. ³Institute of Unconventional Oil & Gas, Northeast Petroleum University, Daqing 163318, China. ⁴Department of Chemical Engineering, Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran. ⁵College of Engineering and Technology, American University of the Middle East, Egaila 54200, Kuwait. ⁶Department of Petroleum Engineering, Shahid Bahonar University of Kerman, Kerman, Iran. ⁷College of Construction Engineering, Jilin University, Changchun, China. ⁸Department of Chemical Engineering, McGill University, Montreal, QC H3A 0C5, Canada. ✉email: s.atashrouz@gmail.com; saeid_atashrouz@aut.ac.ir; hemmati@uk.ac.ir; aut.hemmati@gmail.com; ahmad.mohadespour@mail.mcgill.ca

Since 1996, the number of scientific papers on ILs has skyrocketed from just a few to over 8000 by 2020, far outpacing the growth rates of other well-known scientific fields⁷.

The type and arrangement of cations and anions, as well as the quantity of branching chains inside the molecules, are strongly linked to the characteristics of ILs⁸. ILs have several noteworthy characteristics, such as strong ion conductivity, remarkable permittivity, outstanding electrical properties, nonflammability, high heat capacity, and thermal and chemical durability^{8,9}. Decades of research have led to the development of novel ionic liquids that can be synthesized to customize their physical and chemical characteristics for specific applications. As more ion combinations were developed, it became essential to describe their physical and chemical characteristics^{10,11}. The addition of even small quantities of chemical precursors, for instance halides or water, might cause ILs to become very sensitive. Therefore, studying their physicochemical properties is vital. Density, electrical/thermal conductivity, sound speed, surface and interface properties, refractive index, and viscosity are necessary variables that require precise prediction and optimization. Viscosity is one of the key physicochemical characteristics that assist in assessing the purity, fluid dynamics, and intermolecular forces of ILs^{11,12}. ILs have complex thermodynamic and physicochemical properties, therefore, modeling approaches and large datasets are needed to predict their viscosity. ILs have a viscosity range of between 10 and 10,000 mPa s, and their viscosity is much higher than conventional solvents (0.1–100 mPa s), which may be a big issue for applications requiring mass or charge transfer¹³. In this regard, accurate models of ILs viscosity are required for process modeling, which allow to minimize costs/energy and predict physicochemical properties of ILs^{14–16}. Several computational methods, including group contribution methods (GCM), intelligent approaches (IA), and quantitative structure–property relationships (QSPR), can be used to determine the viscosity of ILs^{15,17,18}. For example, Gardas and Coutinho used GCM to estimate the viscosity of ILs across a large temperature range (293–393 K) utilizing 500 data points from 29 distinct ILs (based on imidazolium, pyrrolidinium, and pyridinium). According to the results, 7.7% was the absolute average relative deviation (AARD) for determining the viscosity of ILs¹⁹. Other research was conducted by Gharagheizi et al. In this study, the viscosity of the IL was estimated using a GCM method. The model was based on 443 distinct ILs (1672 data points) with the temperature range from 253.15 to 433.15 K, and the result was an AARD of 6.3%²⁰. Lazzús et al., in turn, developed a GCM-based linear model to predict ILs viscosity at temperatures ranging from 253 to 395 K, with an AARD of approximately 4.5%²¹. At the same time, AARD was about 11.4% in the study of Paduszynski et al. This work detailed the use of feed-forward neural network (FF-NN) based GCM using 13,000 data points (1484 ILs) with temperature and pressure ranges of 253–573 K and 0.06–350 MPa, respectively²². Finally, AARD for linear and nonlinear models were 10.68% and 6.58%, accordingly, as was suggested by the QSPR model of Zhao et al.²³. This paper was based on a data-bank consisting of 1502 experimental points (89 ILs) across a broad range of temperatures (253.15–395.2 K) and pressures (0.1–300 MPa).

The nodes and layers of an artificial neural network (ANN) are controlled by a vast collection of equations. Aside from that, the number of nodes and levels in the network are decided either manually or at random^{24,25}. The use of machine learning methods to model complicated systems has gained popularity recently^{25–31}. Machine learning methods fall into two categories: black and white-box methods. Black-box models such as neural networks or gradient boosting may be quite accurate²⁶. Black-box models (e.g., support vector regression (SVR) and decision tree) rely on a complicated computer-aided process, whereas white-box models (e.g., gene expression programming (GEP) and group method of data handling (GMDH)) clearly provide a simple and explainable approach^{26,32–34}. Because white-box models provide a model that is more like to human language, they are often understandable to experts in practical applications. White-box models are based on patterns, rules, or decision trees^{32,35}. The GMDH methodology, a self-organizing neural network, can not only describe the system's genome using simple polynomials, but it can also employ standard minimization procedures to determine the optimal configuration²⁴. In our previous research, we used several black-box machine learning approaches for modeling the viscosity of ILs. Also, we developed a simple correlation using a trial-and-error procedure. However, the proposed correlation was not accurate enough and could predict the data with an AARD of 28%, which is high for engineering practices²⁵. Thus, developing a more accurate correlation with high accuracy using advanced correlative approaches such as GMDH and GP appears to be a preferable research direction.

This work models a vast set of 2813 experimental viscosity values from 45 distinct IL using GP and GMDH models with diverse inputs. Additionally, empirical and theoretical methods—such as Eyring's theory (ET)—are used to estimate the viscosity of pure ILs. To determine which approach is the most correct, the dependability of the models that are provided is assessed using both graphical and statistical criteria. The sensitivity analysis is also used to determine how different input factors affect viscosity in relation to one another. Lastly, the quality of the experimental data is assessed and the application domain of the suggested models is determined using the leverage technique.

Data collection

A model can be more accurate and widely applicable the more data points it contains. In order to do this, 2813 experimental viscosity data from 45 ILs were gathered from open literature sources at varying pressures (0.06–298.9 MPa), temperatures (253.15–573 K), and viscosities (1.13–9667.6 MPa.s)^{36–48}.

$$\eta = f(T, P, M_w, T_c, T_b, P_c, \omega, V_c) \quad (1)$$

Recognizing the potential risks associated with open literature data, a thorough screening method was implemented. This process evaluates the quality and consistency of experimental data based on specific criteria. Rigorous analysis was applied to any data points that raised questions, with verification achieved through direct contact with the original authors or alternative sources. This scrupulous approach enhances the robustness of

$$f = \frac{KT}{p} \exp\left(-\frac{\Delta G_0^+}{TR}\right) \exp\left(\frac{\pm \gamma T_{yx} \bar{Q}}{2\dot{\alpha} TR}\right) \quad (4)$$

The net velocity (Fig. 1) shows the separation between molecules in layer "A" and layer "B." The computation involves multiplying the net frequency of advancing jumps ($f_+ - f_-$) by the distance travel in each jump ($\dot{\alpha}$). The frequency of forward and backward leaps are denoted by " f_+ " and " f_- ". The following equation is used:

$$f_{xA} - f_{xB} = (f_+ - f_-)\omega \quad (5)$$

Over a fairly small distance " $\dot{\alpha}$ " between the two layers, a linear velocity profile may be observed, allowing:

$$-\frac{dv_x}{dy} = (\gamma/\dot{\alpha})(-f_- + f_+) \quad (6)$$

To sum up, Eqs. (4) and (6) are combined to form the following equation:

$$\begin{aligned} -\frac{dv_x}{dy} &= (\gamma/\dot{\alpha}) \left(\frac{KT}{p} \exp\left(-\frac{\Delta G_0^+}{TR}\right) \right) \left(\exp\left(\frac{+\gamma \tau_{yx} \bar{Q}}{2\dot{\alpha} TR}\right) \right. \\ &\quad \left. - \exp\left(\frac{-\gamma \tau_{yx} \bar{Q}}{2\dot{\alpha} TR}\right) \right) = \left(\frac{\gamma}{\dot{\alpha}}\right) \left(\frac{KT}{p} \exp\left(-\frac{\Delta G_0^+}{TR}\right)\right) \left(2\sinh\frac{\gamma \tau_{yx} \bar{Q}}{2\dot{\alpha} TR}\right) \end{aligned} \quad (7)$$

If $\frac{\gamma \tau_{yx} \bar{Q}}{2\dot{\alpha} TR} \ll 1$, the Taylor series can also be applied. Finally, the viscosity is derived using the following equation:

$$\eta = \left(\frac{\gamma}{\dot{\alpha}}\right)^2 N_A h / \bar{Q} \exp\left(\frac{\Delta G_0^+}{TR}\right) \quad (8)$$

The unity factor, $\frac{\gamma}{\dot{\alpha}}$, makes the equation without compromising accuracy, since ΔG_0^+ is acquired empirically to ensure that the equation's values match the experimental results. However, it is demonstrated that, for a given fluid, the estimated ΔG_0^+ (free activation energies) are almost constant when fitting Eq. (8) to experimental viscosity values. This translates to the boiling point internal energy of vaporization ($\Delta \hat{U}_{vap} = \Delta H_{vap} - RT\Delta Z_{vap}$), which is given by Eq. (9) as follows⁶³:

$$\Delta G_0^+ \approx 0.408 \Delta \hat{U}_{vap} \quad (9)$$

By using this empiricism and setting $\frac{\dot{\alpha}}{\gamma} = 1$, Eq. (8) becomes as follows when empiricism is set at $\frac{\dot{\alpha}}{\gamma} = 1$:

$$\eta = N_A p / \bar{Q} \exp\left(\frac{0.408 \Delta \hat{U}_{vap}}{TR}\right) \quad (10)$$

The following is an accurate estimate of the vaporization energy provided by the Trouton's rule at the typical boiling point:

$$\Delta \hat{U}_{vap} \approx \Delta \hat{H}_{vap} - T_b R \cong 9.4 T_b R \quad (11)$$

Equation (10), when approximated, reads as follows:

$$\eta = N_A p / \bar{Q} \exp\left(\frac{\lambda T_b}{T}\right) \quad (12)$$

where η indicates the expected viscosity (mPa·s) of pure ILs. N_A and p , respectively, are the Avogadro number (mole^{-1}) and the Plank constant (J·s). The \bar{Q} represents the volume of a mole of liquid ($\text{m}^3 \text{mole}^{-1}$), T_b and T stands for the boiling temperature (K) and temperature (K), respectively. To promote the performance of Eq. (12), a " λ " term was added to Eq. (12) in Excel program for each IL in this study. This term is not constant; rather, it varies depending on ionic liquid. Empiricism $\eta = A \exp(B/T)$ is compatible with eqs. (10) and (12) and appears to be a popular and useful approach. Viscosity decreases with temperature, according to the theory.

Group method of data handling (GMDH)

Ivakhnenko's data-management approach for groups matches Darwin's natural choice concept⁵². By merging two independent variables, the system chooses the optimal polynomial terms. The approach generates a generic multinomial term at each stage. The vast relationship multinomial Volterra–Kolmogorov–Gabor (VKG) analyzes the entire network⁵²:

$$y_i = a + \sum_{i=1}^{N_v} b_i x_i + \sum_{i=1}^{N_v} c_{ij} x_i x_j + \dots + \sum_{i=1}^{N_v} \sum_{j=1}^{N_v} \dots \sum_{k=1}^{N_v} d_{ij\dots k} x_i x_j \dots x_k \quad (13)$$

In the above equation, the count of independent variables in the experiment is denoted by N_v . From a set of measured data with N data points, a matrix can be generated. The measured results $\vec{V}_y = (y_1, y_2, \dots, y_n)$ are represented on the left-hand side of the matrix, while the independent variables $\vec{V}_n = (x_1, x_2, \dots, x_n)$ are represented on the right-hand side of the matrix. Both sides of the matrix are produced from the same set of data. When two independent variables are coupled, a quadratic polynomial $\binom{N_v}{2}$ can be used to estimate the actual data. Using N_v parameters, here is a formula for $\binom{N_v}{2}$:

$$z_i^{\text{GMDH}} = aA_i + bB_i + cA_iB_i + dA_i^2 + eB_i^2 + f \quad (14)$$

The matrix of independent variables can here be built using the vector of new variables $\vec{V}_z = (z_1, z_2, \dots, z_n)$. To modify the parameters of equations, the least squares method is utilized (15). The objective is to maintain the square of the deviation from the actual data as small as possible in each column:

$$\delta_j^2 = \sum_{i=1}^{N_i} [y_i - z_i^{\text{GMDH}}]^2 \quad j = 1, 2, \dots, \binom{N_v}{2} \quad (15)$$

In the above equation, N_i denotes the count of datasets used. The measured data is used to construct training and testing subsets. The proportion of training and testing subsets is chosen at random. Equations are derived using the training set of data (15). The ideal set of parameters (z_i). Variations from planned results must fulfill the following criteria, based on the predefined requirement:

$$\delta_j^2 = \sum_{i=N_i+1}^N [y_i - z_i^{\text{GMDH}}]^2 < \varepsilon \quad j = 1, 2, \dots, \binom{N_v}{2} \quad (16)$$

here, ε is an optional/random value. Just the z columns that meet the criteria are kept, whereas the ones that do not are deleted. The entire variation is preserved after each repetition and compared to the prior repetitions to check if the least variation has been achieved.

Genetic programming (GP)

GP is a breakthrough in optimization computing that combines traditional genetic methods with symbolic improvement⁵³⁻⁵⁵. It is predicated on an approach called "tree representation." This form is incredibly flexible since trees may represent full models of industrial systems, mathematical formulae, or computer programs. Creating model structures like differential equations, kinetic ordering, and steady-state models is best accomplished with this approach^{56,57}. To achieve great variation, GP first creates an initial population, which consists of randomly selected individuals (trees). A new generation is finally formed by the software, which evaluates the individuals, selects individuals for reproduction, creates new individuals by mutation, crossover, and direct reproduction⁵⁷. Unlike other optimization techniques, symbolic improvement uses the architectural arrangement of many symbols to convey workable solutions (that is, vectors of real values).

Model assessment

Statistical criteria

The models' validity was tested using the determination coefficient (R^2), standard deviation (SD), average absolute relative deviation (AARD%), average relative deviation percent (ARD%), and root mean square error (RMSE). Below are the statistical parameters:

Determination Coefficient (R^2): R^2 is a regression coefficient that shows the model's accuracy. The model fits the data better if it is close to 1. R^2 's mathematical formula is as follows:

$$R^2 = \frac{\sum_{i=1}^{N_p} (\eta_i^{\text{exp}} - \bar{\eta})^2 - \sum_{i=1}^{N_p} (\eta_i^{\text{cal}} - \eta_i^{\text{exp}})^2}{\sum_{i=1}^{N_p} (\eta_i^{\text{exp}} - \bar{\eta})^2} \quad (17)$$

Average Relative Deviation (ARD%): The relative deviation of the anticipated outcomes from the experimental data is determined using the ARD%:

$$\text{ARD}\% = \frac{100}{N_p} \sum_{j=1} \left(\frac{\eta_j^{\text{exp}} - \eta_j^{\text{est}}}{\eta_j^{\text{exp}}} \right) \quad (18)$$

Positive and negative ARD (%) represents a model's underestimate and overestimate, respectively.

Standard Deviation (SD): SD is a metric used to quantify the degree of dispersion of data around the central point. This has the following definition:

$$\text{SD} = \left(\frac{1}{N-1} \sum_{j=1}^{N_p} \left(\frac{\eta_j^{\text{exp}} - \eta_j^{\text{est}}}{\eta_j^{\text{exp}}} \right)^2 \right)^{\frac{1}{2}} \quad (19)$$

Average Absolute Relative Deviation (AARD%): The relative absolute deviation is used to quantify the difference between the actual or real data and the projected or represented data. It is shown by the equation that follows:

$$\text{AARD}(\%) : 100 \times \frac{\sum_{j=1}^{N_p} \left| \frac{\eta_j^{nep} - \eta_j^{est}}{\eta_j} \right|}{N_p} \quad (20)$$

Root Mean Square Error (RMSE): The RMSE is a frequently used statistical analysis approach for estimating the discrepancies between experimental and expected values. It goes by the name:

$$\text{RMSE} = \sqrt{\frac{1}{N_p} \sum_{j=1}^{N_p} (\eta_j^{exp} - \eta_j^{est})^2} \quad (21)$$

When calculating the average IL viscosity using experimental/real data, the experimental/real viscosity (η) and the number of data points N_p are represented by the variables “*est*”, and “*exp*”, respectively.

Graphical assessment of the models

Several graphical plots were used in this research to further evaluate the suggested models and measure their predicted performances. Among the visualization plots are diagrams showing the cumulative frequency and error distribution. In order to measure the distribution of error around the zero line and to indicate whether the model has a tendency to make mistakes, the percentage of relative deviation is displayed against target/real values in the error distribution. The cross-plot displays the estimated/represented value of the model in relation to the experimental data. After that, a slope line with a 45° unit is constructed to connect the experimental and represented/predicted values. A more accurate model is indicated by more data points that are shown along this line. The bulk of approximations will be inside a standard error range if the cumulative frequency is calculated from the absolute relative error.

Results and discussion

Development of models

Using 2813 points of data (45 ionic liquids) collected from the literature, models were developed. Table 2’s “Total” refers to the whole set of data (2813 data points) that were used for analysis and modeling in the current research. The database was split into training sets (which made up 80% of the overall dataset) and test sets (20% of the total dataset) at random. The 563 data points in the “testing” set were used to track over-fitting errors and the reliability of the built models. The “training” subset (2250 data sets) caused changes to the model’s structure and tuning parameters. T , P , M_w , V_c , T_b , T_c , P_c and w were the input parameters, and IL viscosity was the output (Table 1).

To begin with, the GMDH method was used to build a new empirical correlation. The viscosity of ILs with 5, 6, and 7 inputs was found to be:

5 Inputs:

$$\begin{aligned} \eta &= -2.87789 + T * 0.0143915 - T * N_1 * 0.00131975 - T^2 * 1.78556e - 05 + N_1 * 1.51847 - N_2^2 * 0.0215748 \\ N_1 &= -0.0340483 + P * 0.00199063 - P * N_2 * 0.000467152 - P^2 * 4.36465e - 06 + N_2 * 1.00943 \\ N_2 &= -0.00484172 + N_9 * 0.535395 + N_8 * N_3 * 1.92525 - N_8^2 * 1.11778 + N_2 * 0.4892 - N_2^2 * 0.814881 \\ N_3 &= 0.00291747 + N_5 * 0.258276 + N_5 * N_4 * 9.67431 - N_4^2 * 5.13243 + N_4 * 0.759826 - N_4^2 * 4.54534 \\ N_4 &= -0.265189 + T_c * 0.00125932 + T_c * N_8 * 0.000298289 - T_c^2 * 8.18089e - 07 + N_6 * 0.53522 + N_6^2 * 0.0574476 \\ N_5 &= 0.408599 + T_c * 0.000100432 + T_c * N_7 * 0.000381021 - T_c^2 * 2.89505e - 07 + N_7 * 0.350992 + N_7^2 * 0.086781 \\ N_6 &= 0.40796 - P_c * 0.0250195 + P_c * N_7 * 0.00115096 + P_c^2 * 0.000437153 + N_7 * 0.854595 + N_7^2 * 0.0297431 \\ N_7 &= 0.274614 + P * N_{11} * 0.0015094 + P^2 * 6.77239e - 06 + N_9 * 0.605881 + N_8^2 * 0.100009 \\ N_8 &= 7.44079 - T * 0.0412914 - T * N_{10} * 0.00645033 + T^2 * 5.49978e - 05 + N_{10} * 3.16623 \\ N_9 &= 10.5667 - T * 0.0467936 - T * w * 0.00582271 + T^2 * 5.09471e - 05 + w * 4.83785 - (w)^2 * 1.60314 \\ N_{10} &= 0.865852 + P * 0.00233429 + P^2 * 4.66924e - 06 + w * 2.3876 - (w)^2 * 1.18929 \end{aligned} \quad (22)$$

	T (K)	P (mPa)	Tc (K)	Pc (bar)	Tb (K)	W	Mw (g/mole)	Exp.Viscosity (MPa s)
Mean	325.62	24.45	1005.87	22.30	723.94	0.60	346.66	191.97
Std	30.92	37.94	280.58	10.05	173.84	0.28	88.09	442.93
Min	253.15	0.06	520.06	2.63	410.77	0.22	201.23	1.13
25%	303.15	0.10	736.99	16.02	586.74	0.34	279.08	25.81
50%	323.15	6.00	1038.70	20.98	712.68	0.53	340.29	64.00
75%	343.15	40.00	1269.93	27.65	862.44	0.87	419.37	177.00
Max	573.00	298.90	1534.60	57.61	1130.30	1.10	515.13	9667.59

Table 1. Dataset statistics acquired in this work.

Statistical criteria	RMSE	SD	R ²	AARPE, %	
7 Input: T, P, T _c , P _c , T _b , w, Mw					
GP	Train	115.62	0.317	0.864	24.378
	Test	246.78	0.331	0.723	24.833
	Total	226.66	0.328	0.798	24.742
GMDH	Train	53.54	0.116	0.987	9.171
	Test	83.46	0.198	0.947	13.269
	Total	65.69	0.136	0.979	8.144
6 Input: T, P, T _c , P _c , w, Mw					
GP	Train	161.96	0.319	0.884	24.593
	Test	284.10	0.375	0.588	27.674
	Total	224.16	0.362	0.699	26.056
GMDH	Train	94.490	0.152	0.959	10.581
	Test	93.975	0.205	0.909	15.977
	Total	94.095	0.164	0.955	11.662
5 Input: T, P, T _c , P _c , w					
GP	Train	214.27	0.309	0.663	23.824
	Test	272.36	0.481	0.5358	32.101
	Total	261.18	0.353	0.650	27.824
GMDH	Train	99.374	0.1146	0.955	8.0798
	Test	121.34	0.236	0.855	16.911
	Total	104.178	0.146	0.944	10.951

Table 2. Calculated the statistical requirements for the developed correlations.

The 6 Inputs:

$$\begin{aligned}
 \eta &= -3.49184 + T * 0.0171992 - T * N_1 * 0.0016858 - T^2 * 2.10848e - 05 + N_1 * 1.68321 - N_1^2 * 0.033313 \\
 N_1 &= 0.0365592 - N_5 * N_3 * 5.74908 + N_2^2 * 2.80849 + N_2 * 0.958807 + N_2^2 * 2.94393 \\
 N_2 &= -0.895957 + M * 0.00539033 + M * N_3 * 0.000164955 - (M)^2 * 7.8437e - 06 + N_3 * 0.960834 \\
 N_3 &= 0.00291747 + N_5 * 0.258276 + N_5 * N_4 * 9.67431 - N_5^2 * 5.13243 + N_4 * 0.759826 - N_4^2 * 4.54534 \\
 N_4 &= -0.265189 + T_c * 0.00125932 + T_c * N_6 * 0.000298289 - T_c^2 * 8.18089e - 07 + N_6 * 0.53522 + N_6^2 * 0.0574476 \\
 N_5 &= 0.408599 + T_c * 0.000100432 + T_c * N_7 * 0.000381021 - T_c^2 * 2.89505e - 07 + N_7 * 0.350992 + N_7^2 * 0.086781 \\
 N_6 &= 0.40796 - P_c * 0.0250195 + P_c * N_8 * 0.00115096 + (P_c^2 * 0.000437153 + N_7 * 0.854595 + N_7^2 * 0.0297431 \\
 N_6 &= 0.40796 - P_c * 0.0250195 + P_c * N_8 * 0.00115096 + (P_c^2 * 0.000437153 + N_7 * 0.854595 + N_7^2 * 0.0297431 \\
 N_7 &= 0.274614 + P * N_8 * 0.0015094 + P^2 * 6.77239e - 06 + N_8 * 0.605881 + N_8^2 * 0.100009 \\
 N_8 &= 10.5667 - T * 0.0467936 - T * w * 0.00582271 + T^2 * 5.09471e - 05 + w * 4.83785 - (w)^2 * 1.60314
 \end{aligned}
 \tag{23}$$

7 Inputs:

$$\begin{aligned}
 \eta &= -0.0100846 - N_7 * N_1 * 2.82949 + N_7^2 * 1.37207 + N_1 * 1.01023 + N_1^2 * 1.45152. \\
 N_1 &= -0.0134572 + P * 0.0021813 - P * N_2 * 0.000587268 - P^2 * 5.26484e - 06 + N_2 * 0.982197 + N_2^2 * 0.00878339; \\
 N_2 &= 0.17471 - P_c * 0.0126706 + P_c * N_3 * 0.00161699 + P_c^2 * 0.00019748 + N_3 * 0.959782; \\
 N_3 &= -0.502724 + M * 0.00270637 - M * N_4 * 0.000181006 - M^2 * 3.44871e - 06 + N_5 * 1.06462; \\
 N_4 &= 0.0538836 - w * 0.420022 - w * N_5 * 0.0630102 + w^2 * 0.347037 + N_5 * 1.06542; \\
 N_5 &= -0.159217 + N_11 * 0.558586 + N_11 * N_6 * 4.28637 - N_11^2 * 2.45863 + N_6 * 0.681851 - N_6^2 * 1.90027; \\
 N_6 &= -0.629478 + N_18 * 0.82026 + N_18 * N_7 * 0.147292 - N_18^2 * 0.232333 + N_7 * 0.683362; \\
 N_7 &= 0.646958 - T_b * 0.000378377 + T_b * N_8 * 0.000677083 - T_b^2 * 3.17028e - 07 + N_8 * 0.236567 + N_8^2 * 0.089335; \\
 N_8 &= 0.274614 + P * N_10 * 0.0015094 + P^2 * 6.77239e - 06 + N_10 * 0.605881 + N_10^2 * 0.100009; \\
 N_9 &= 7.44079 - T * 0.0412914 - T * N_16 * 0.00645033 + T^2 * 5.49978e - 05 + N_16 * 3.16623; \\
 N_10 &= 10.5667 - T * 0.0467936 - T * w * 0.00582271 + T^2 * 5.09471e - 05 + w * 4.83785 - w^2 * 1.60314; \\
 N_11 &= 0.865852 + P * 0.00233429 + P^2 * 4.66924e - 06 + w * 2.3876 - w^2 * 1.18929; \\
 N_12 &= 0.162458 - P_c * T_b * 7.76995e - 05 + P_c^2 * 0.000712479 + T_b * 0.00695937 - T_b^2 * 4.56726e - 06;
 \end{aligned}
 \tag{24}$$

Furthermore, the equations below proposed for 5, 6, and 7 inputs in GP model:

5 Input:

$$\text{Log}(\eta) = \left(c_{17} \frac{\left(\frac{c_0}{\left(\frac{c_1}{\ln(c_2 T_c)} - \left(c_3 T - \frac{c_4}{c_5 w} \right) \right) - (c_6 P + (c_7 + c_8 P))} \right)}{\left(\frac{c_9}{\ln(c_{10} P_c)} - \left(\frac{c_{11} + c_{12} P}{c_{13}} - \left(c_{14} T - \frac{c_{15}}{c_{16} w} \right) \right) \right)} + c_{18} \right)
 \tag{25}$$

$$\begin{aligned}
 c_0 &= 12.339; c_1 = 0.352; c_2 = 0.25504; c_3 = -0.46065; c_4 = 17.271; \\
 c_5 &= 2.5053; c_6 = 0.47496; c_7 = 12.343; c_8 = -0.34241; \\
 c_9 &= 12.343; c_{10} = 0.80501; c_{11} = -8.0693; c_{12} = -0.34241; c_{13} = 8.3702; \\
 c_{14} &= -0.46065; c_{15} = 13.934; c_{16} = 2.5053; c_{17} = -5409.2; c_{18} = -1.09.
 \end{aligned}$$

6 Input:

$$\text{Log}(\eta) = \left(c_{11} \frac{(\exp(\exp(c_0 w)) + c_1)}{(c_2 T + c_3 P_c) \left(\frac{\exp(c_9 M)}{c_{10} w} \right) + c_{12}} \right) \quad (26)$$

$$c_0 = 0.26945; c_1 = 15.569; c_2 = 0.7941; c_3 = 0.5728; c_4 = -12.046; c_5 = 14.544; c_6 = 0.2556; c_7 = 0.3401; c_8 = 1.3188; c_9 = 0.26945; c_{10} = -0.043257; c_{11} = 15270; c_{12} = -1.1226.$$

7 Input:

$$\text{Log}(\eta) = \left(\frac{c_0}{\left(\ln \left(\left(\left(c_1 T_c + \frac{c_2 T_c}{c_3} \right) c_4 w + (c_5 T + (c_6 T_b - c_7 T_b)) \right) \right) \right) \left(\ln \left(\left(c_8 T_c + c_9 T_c c_{10} M + (c_{12} T + (c_{13} T_c + c_{14} P_c)) \right) \right) \right) c_{15} - c_{16} T + (\ln(c_{17} T) c_{18} + (\exp(\ln(c_{19} T c_{20}) - c_{21} P)) \right) \right) * c_{22} + c_{23} \right) \quad (27)$$

$$\begin{aligned}
 c_0 &= 14.019; c_1 = 2.0204; c_2 = 0.25903; c_3 = 2.8184; c_4 = 0.88752; c_5 = 1.5553; c_6 = 0.46073; \\
 c_7 &= 1.2408; c_8 = 2.0204; c_9 = 0.25903; c_{10} = 2.8184; c_{11} = 0.88752; c_{12} = 0.91528; c_{13} = 2.02; c_{14} = 1.8657; \\
 c_{15} &= -3.4975; c_{16} = 1.0567; c_{17} = 1.3443; c_{18} = -3.4975; c_{19} = 1.553; c_{20} = 11.329; c_{21} = 2.3349; c_{22} = 1213.7; c_{23} = -3.6512.
 \end{aligned}$$

The critical temperature and pressure values of the IL are denoted T_c and P_c , respectively. There is also an acentric factor (w), temperature (T), and pressure (P), as well as IL molecular weight (M_w), critical volume (V_c), and IL boiling temperature (T_b). The other parameters are the adjustable correlation coefficients (Table 2).

The RMSE, SD, R^2 , and AARPE% for the proposed correlation are calculated for the GP and GMDH models in Table 2.

The cross-plots on the results of the experimental viscosity data and the predicted data for the given correlation are displayed in Fig. 2. Around the unit-slope line, this figure shows a medium-uniform distribution of forecasts. The viscosity of the ILs that were taken from the database was estimated using temperature (T) and boiling temperature (T_b), in accordance with Eyring’s theory (Eq. 13). AARD stands for 21.86%. The expected vs experimental IL viscosity is also plotted in a logarithmic cross-plot in Fig. 2. The data points were somewhat near the diagonal line, indicating moderate conformity. But data indicates that Arrhenius reliance does not match the experimental transport characteristics of ILs, which is why Eyring’s theory does not hold up. In fact, ILs viscosity decreased as temperature rose, and this feature has to be changed by new model improvements. In order to define the thermal characteristics of ILs, the Vogel–Tamman–Fulcher (VTF) development is frequently used. This provides the basis for a complex energy landscape with several local potential energy minimums and a broad variety of energy barriers^{58,59}.

The white-box machine learning models were carried out using GP and GMDH and compared to ET and Mousavi’s model²⁵. "White-box" models in machine learning are those that are easy for experts in the application area to understand. These models, in general, provide a fair mix between explainability and accuracy. The numerical assessment of the created methods is presented in Table 3. With the use of a GMDH optimizer, it was shown that the usage of seven inputs was the best design for forecasting the viscosity of ILs since it can anticipate the whole data collection with more accuracy than other approaches (AARD% = 8.14).

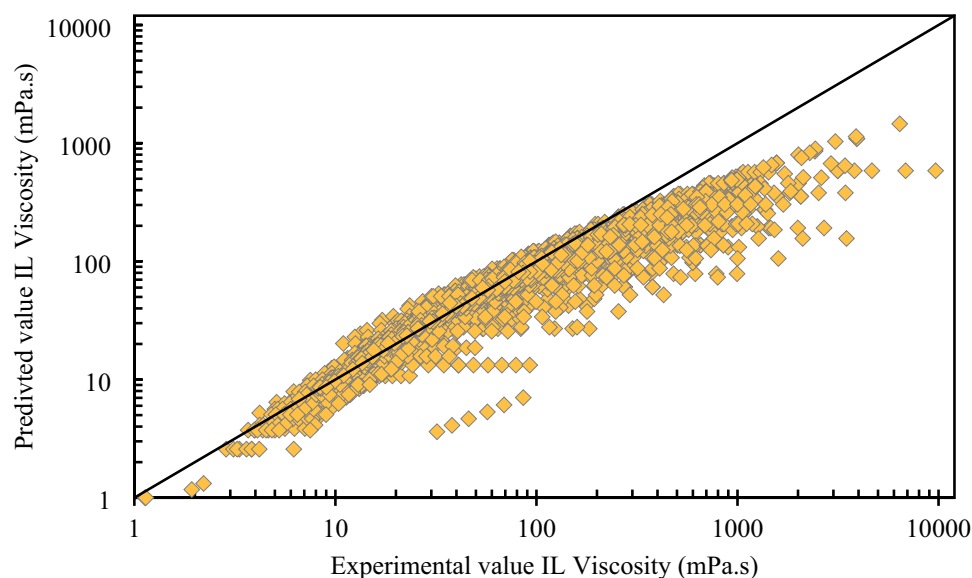


Figure 2. Cross plot of the proposed Eyring’s theory for viscosity of ILs.

	AARD (%)	RMSE	SD	R ²
GP—5 inputs	27.82	261.18	0.35	0.65
GMDH—5 inputs	10.95	104.17	0.14	0.94
GP—6 inputs	26.05	224.16	0.36	0.70
GMDH—6 inputs	11.66	94.09	0.16	0.95
GP—7 inputs	24.74	226.66	0.32	0.79
GMDH—7 inputs	8.14	65.69	0.13	0.98
Eyring modified model	25.76	371.97	0.33	0.29
Mousavi et al. correlation	28.34	394.12	0.36	0.20

Table 3. Statistical comparison between the Eyring modified model correlations and new developed correlation model with various inputs.

Statistical evaluation

To illustrate the error margin, many statistical metrics were computed for both created models with different inputs, including ARD, AARD, RMSE, SD, and R². As more input data was provided, the AARD, ARD, SD, and RMSE values for the test and training datasets decreased, as seen in Table 3 for the GP and GMDH models. The following is a breakdown of the models based on how accurate they are: Mousavi et al. correlation < Eyring theory < GP < GMDH. As a result, the GMDH model may produce more reliable estimates than the other established models.

Graphical error analysis

A number of graphical error evaluations developed from the GP and GMDH models were examined in order to offer a more lucid evaluation of the models' efficacy. To evaluate the models, the predicted viscosity measurements were compared to the experimental measurements shown in Fig. 3(a–c). For the GP and GMDH models, there is a high formation of points around the unit slope in both the test and training datasets. The observed viscosity values are shown to be more accurate by the GMDH correlations than by the GP and Mousavi et al. correlations (Fig. 3c). As indicated in Fig. 3, the data distribution for GMDH correlation with 7 inputs is more on the slope line than the GMDH data with 5 and 6 inputs. The GMDH decreases overall relative deviation, resulting in the smallest error margin.

The AARE% of the white-box machine learning models is shown against the number of input parameters in Fig. 4. Comparing the GP model with experimental data indicated that it was less accurate, less flexible, and less well-suited than the GMDH model. Furthermore, the GMDH model with 7 inputs showed higher accuracy with experimental viscosity data in comparison to the GMDH trained on 5 and 6 inputs.

To show the models' level of competence, comparative graphs are used, such as cumulative frequency plots. The GMDH has the maximum cumulative frequency for a given absolute relative deviation, as seen in Fig. 5. To put it another way, the GMDH model predicted almost 70% of the data points as we got closer to the ARD of less than 4%, but the corresponding values for the GP models were 9%, 11%, and 10%, respectively.

Figure 6 presents a comparison of the created models with respect to their relative deviations. The model's ability to precisely predict the viscosity of ILs is demonstrated by the dense cluster of dots surrounding the zero line. As can be seen, the GMDH model with 5, 6, and 7 inputs estimated viscosity better than GP, Eyring theory, and Mousavi et al. correlations.

Figure 7 compares the acentric factor, molecular weight, boiling temperature, critical temperature / pressure, temperature, and pressure impacts (7 inputs) on AARD (%) for the GMDH and GP models under investigation. According to our findings, the GP model is more sensitive to changes, which leads to greater parameter values than the GMDH model, which was shown to be less susceptible. The GP model, for instance, is very temperature-sensitive (Fig. 7a). Thus, the GMDH model may be applied in a wide range of temperatures with a lower relative error of ARE < 15%, whereas the GP model can only be utilized in a narrow range of temperatures (381–445 K) with a minimum relative error of 14.6%.

AARD values of 8.14% and 25.76% for the 7 inputs are displayed in Table 3 and are thus retained for future analyses since they are among the best responses for the GMDH and GP models. Based on the GMDH and GP correlations, Fig. 8 shows how temperature and pressure affect 1-ethyl-3-methylimidazolium hexafluorophosphate. The anticipated viscosity of ILs using both models is consistent with the experimental dataset, as expected. Viscosity assessments for ILs using GP correlations are, in turn, inconsistent, as seen in Fig. 8, and come with large error margins. As can be observed in Fig. 8b, there is a physical link between the temperature and the GMDH model; but, as the pressure increases, neither model can adequately represent the experimental data.

Identifying outliers in experimental data, GMDH, and GP models

Finding data that significantly differs from the bulk of the data in a database is the aim of outlier (or aberrant) identification^{60,61}. Leverage is a well-known approach for doing this^{60,62}. Standardized residuals (R) and the Hat matrix (H) are used⁶². The R value for each data point can be found using the below equation:

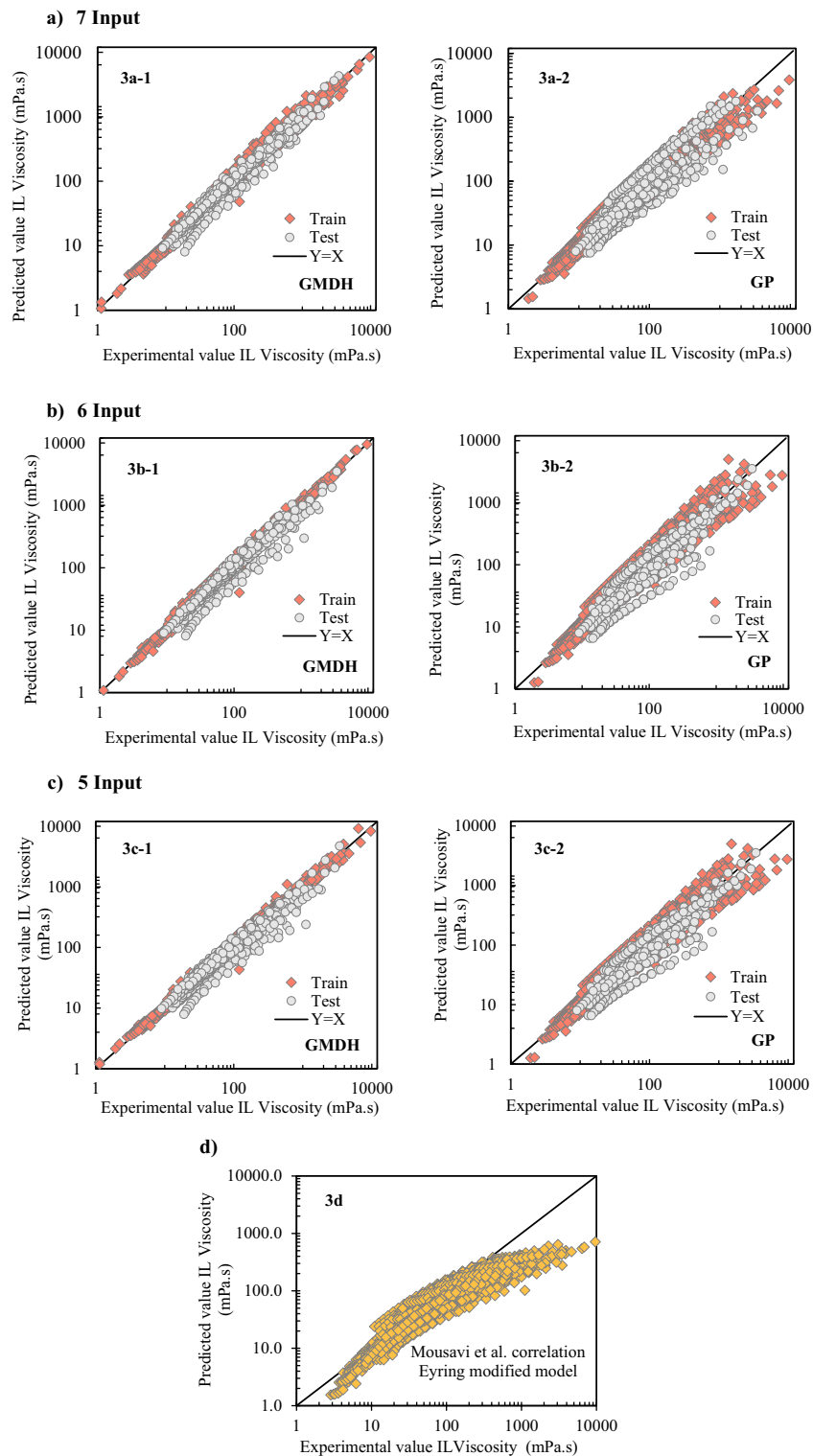


Figure 3. Comparison of the cross plots of the predicted correlations. Subfigures include: Fig 3(a) with 7 inputs (3a-1: GMDH, 3a-2: GP), Fig 3(b) with 5 inputs (3b-1: GMDH, 3b-2: GP), Fig 3(c) with 3 inputs (3c-1: GMDH, 3c-2: GP), and Fig 3(d) (modified Eyring model).

$$R_i = \frac{z_i}{(MSE(1 - H_{ii}))^{\frac{1}{2}}} \quad (28)$$

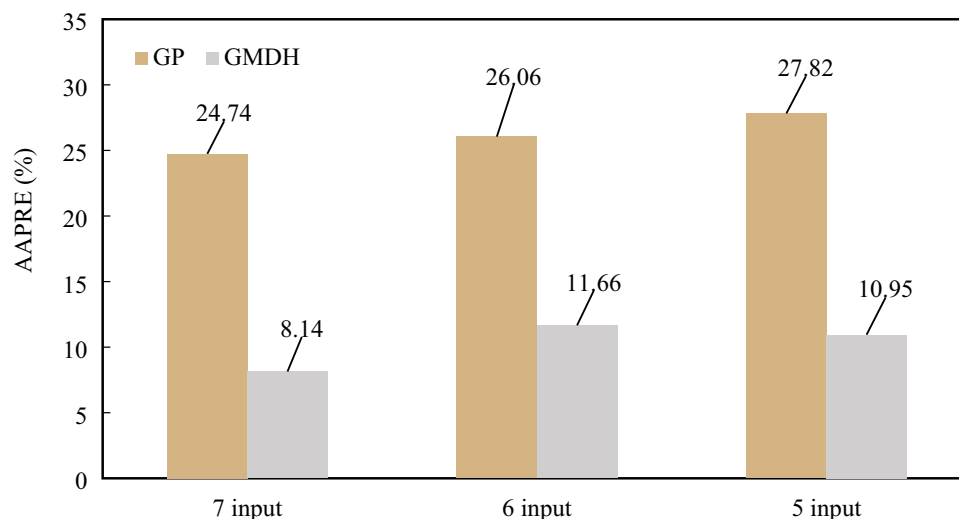


Figure 4. Comparison between the AARE values of the GMDH and GP models.

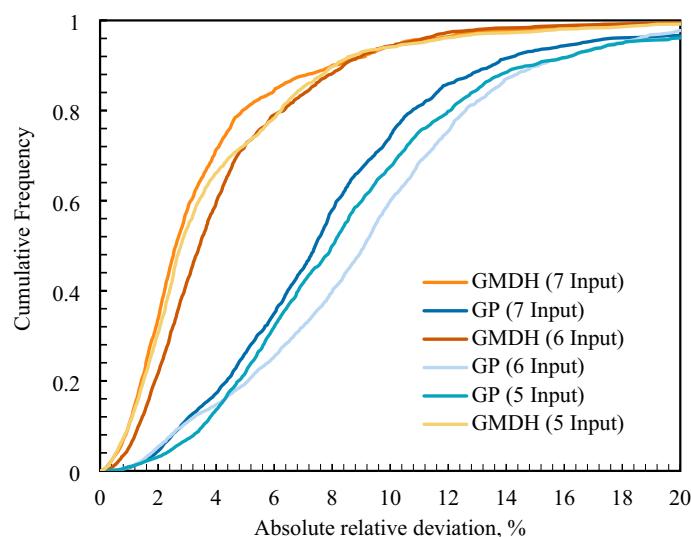


Figure 5. Absolute relative deviation cumulative frequency for various models based on GMDH and GP models.

MSE stands for mean square error (MSE), while the i_{th} data point's error and ii_{th} Hat indices (Leverage) are represented by z_i and H_{ii} ⁶³. In addition, the following formula may be used to calculate Hat index (or Leverage):⁶⁴

$$H = X(X^t X)^{-1} X^t \quad (29)$$

here, X represents a two-dimensional $q \times w$ matrix (where “ q ” shows the number of data and “ w ” is the count of input variables). Also, X^t is transpose of matrix. The outliers were investigated using the Williams plot after the R and H values were measured. In addition, the Leverage limit (H^*), a parameter defined as $3a/b$, where b stands for the count of data points and a is the number of model parameters plus one, is applied in this approach.

The calculated R values must be within $[-3, +3]$ standard deviations in order to encompass 99.7% of the normally distributed data^{17,62}. The model is statistically valid if a significant proportion of data points are in the range of $H^* \geq H \geq 0$ and $3 \geq R \geq -3$ ¹⁷. Since they are highly expected yet outside of the application domain, data points in the range of $-3 \leq R \leq 3$ and $H^* \leq H$ are referred to as “Good High Leverage” points. Conversely, data points with R values larger than or less than -3 are referred to as “Bad High Leverage” data points. These regions are beyond the applicability range of the model and have significant levels of uncertainty. It is clear that reliable data significantly affect the GMDH (7 inputs) model's performance, making it the best model used in this study. The H^* value, as per the suggested model, was 0.0085. The GMDH model's Williams plot is shown in Fig. 9 into the statistically significant range of $0 \leq H \leq 0.0085$ and $-3 \leq R \leq 3$, all data points appear to fit into the established GMDH model. Less residual value normalization leads to an increase in reliability. However,

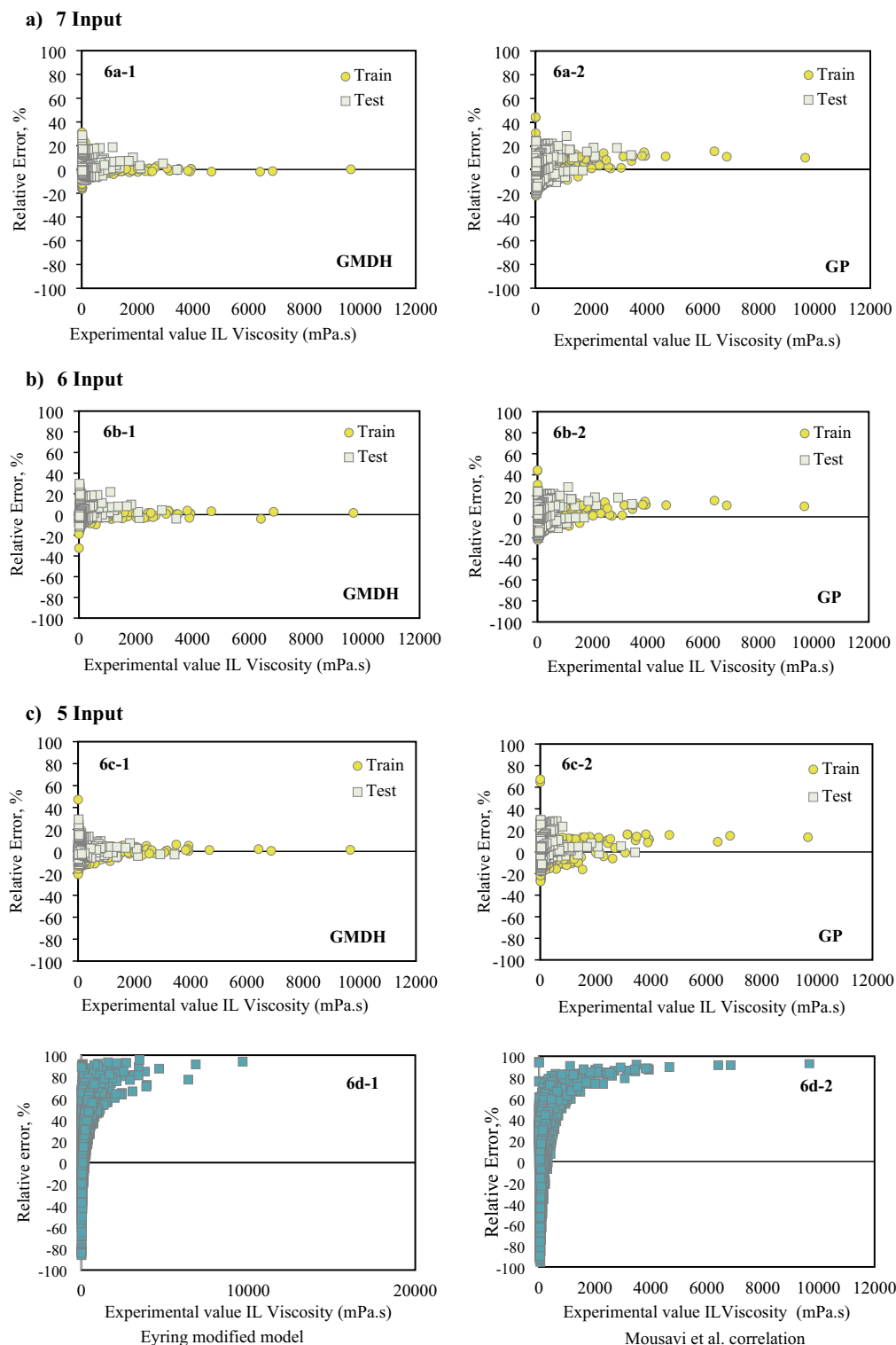


Figure 6. Error distribution plots of the developed correlations compared to the Eyring modified and Mousavi's models. Fig 6(a) with 7 inputs (6a-1: GMDH, 6a-2: GP), Fig 6(b) with 6 inputs (6b-1: GMDH, 6b-2: GP), Fig 6(c) with 5 inputs (5c-1: GMDH, 5c-2: GP), and Fig 6(d) (6d-1: modified Eyring model, 6d-2: modified Mousavi model).

However, 24 suspicious data points, or fewer than 1% of the total data in Fig. 9, either $R < -3$ or $R > 3$, making them outliers with considerable uncertainty. Furthermore, 77 data points, or 3% of all data had $H > 0.0085$. These points are all in the range of $-3 \leq R \leq 3$, which indicates that they are all Good High Leverage regardless of their Hat (Leverage) values.

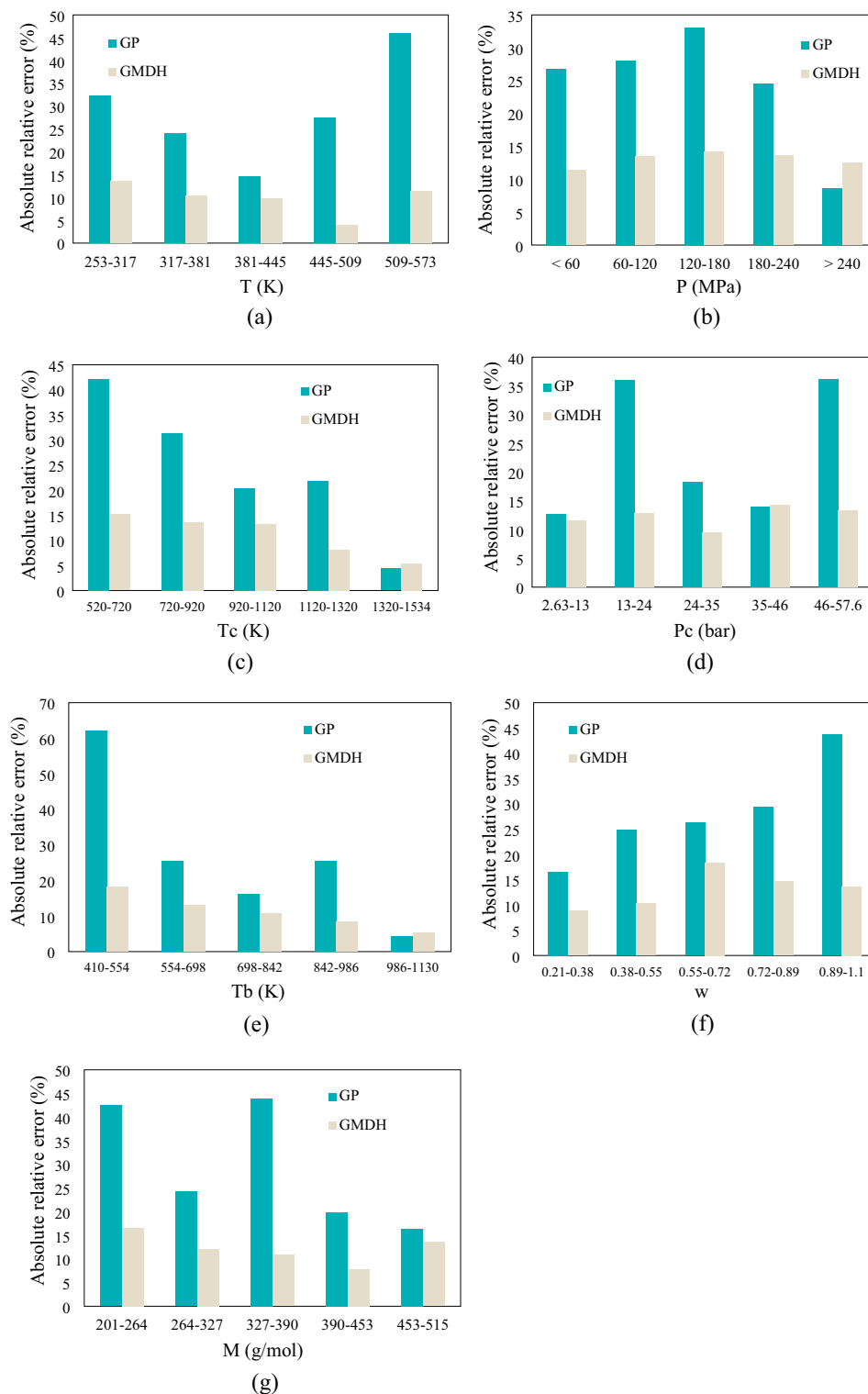


Figure 7. AARE for the correlations between the GMDH and GP correlations with 7 inputs. Temperature; pressure; critical temperature; critical pressure; boiling temperature; acentric factor; and molecular weight are represented by (a–g).

Variables' relative importance

When taking the GMDH model, all input variables were tested to see how much of an influence they had on the viscosity of ILs. The relative significance of the inputs with respect to one another is shown in Fig. 10. One measure used to evaluate each input parameter's impact on the pure viscosity of ILs as a model output is the

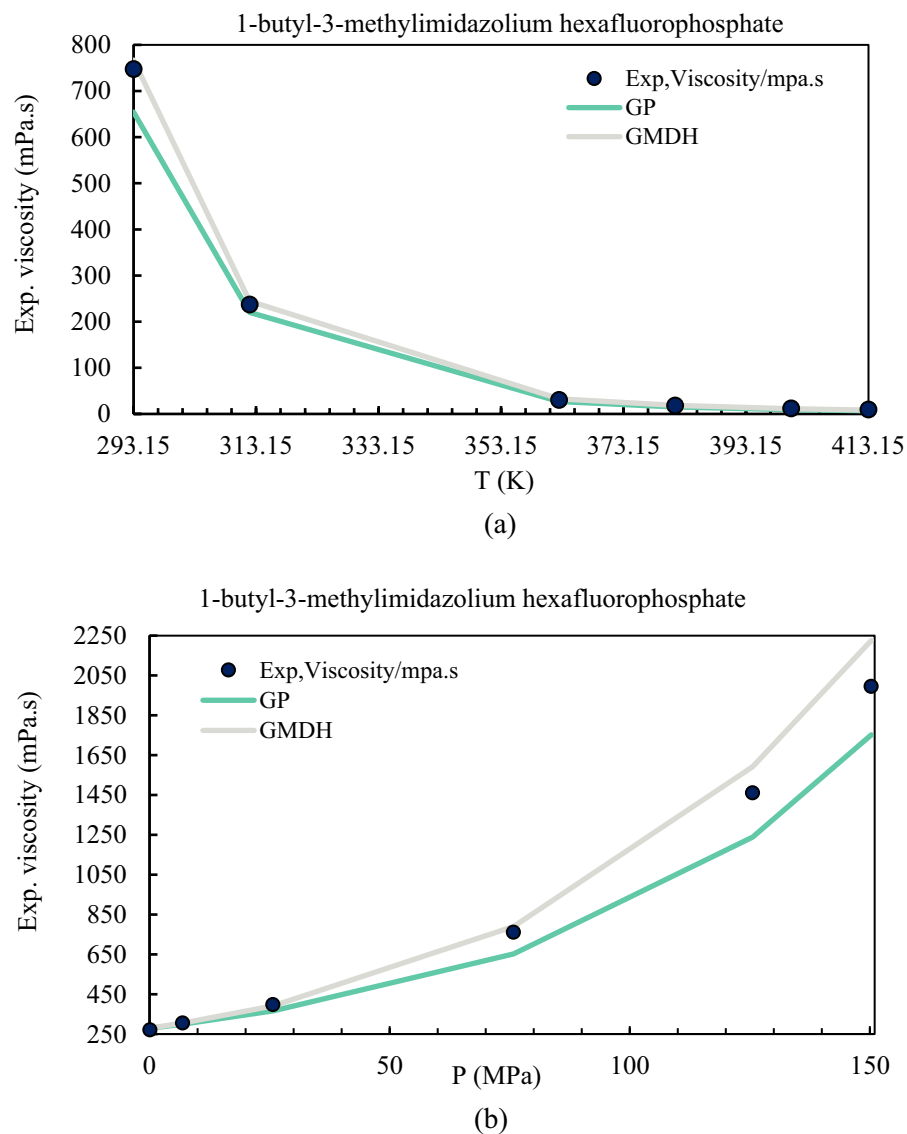


Figure 8. Correlation of the 1-butyl-3-methylimidazolium hexafluorophosphate viscosity for the generated correlation (7 input) with experimental data. (a) Viscosity-temperature; (b) Viscosity-pressure.

relevance factor (r). Negative values indicate an inverse correlation between the input and output parameters, and vice versa. Relevance Factor (r) values are analyzed in accordance with the following equation⁶⁵:

$$r(I_i, \eta) = \frac{\sum_{j=1}^n (I_{ij} - \bar{I}_i)(\eta_j - \bar{\eta})}{\left(\sum_{j=1}^n (I_{i,j} - \bar{I}_i)^2 \sum_{j=1}^n (\eta_j - \bar{\eta})^2\right)^{0.5}} \quad (30)$$

where n represents the number of datasets. Also, the j -th value, and the mean of the I -th input are respectively represented by the variables, $I_{i,j}$, and \bar{I}_i . Whereas $\bar{\eta}$ denotes the average value of the predicted ILs viscosity, while η_j represents the j -th value of the represented/expected viscosity. Based on the GMDH (as the output), Fig. 10 displays the relative effects of each parameter on the pure viscosity of ILs. It is demonstrated that temperature and the acentric factor significantly affect the model's output.

Viscosity increases with an increase in pressure or acentric factor in pure ionic liquids. As Fig. 10 illustrates, increasing T , M_w , V_c , T_b , T_c , and P_c parameters will result in a decrease in the viscosity of ILs, since they have negative relevance factors. Moreover, the temperature has the most significant effect on the viscosity of ILs compared to other inputs.

We compared our models to a nonlinear artificial neural network (ANN) model using a dataset of 8,523 IL-water mixture viscosity data points¹⁶. The assessment included critical performance indicators such as mean absolute error (MAE) and R-squared (R^2). The results show that the GP and GMDH models have equivalent, if not greater, prediction accuracy, with benefits in simplicity and interpretability. This comparative research not

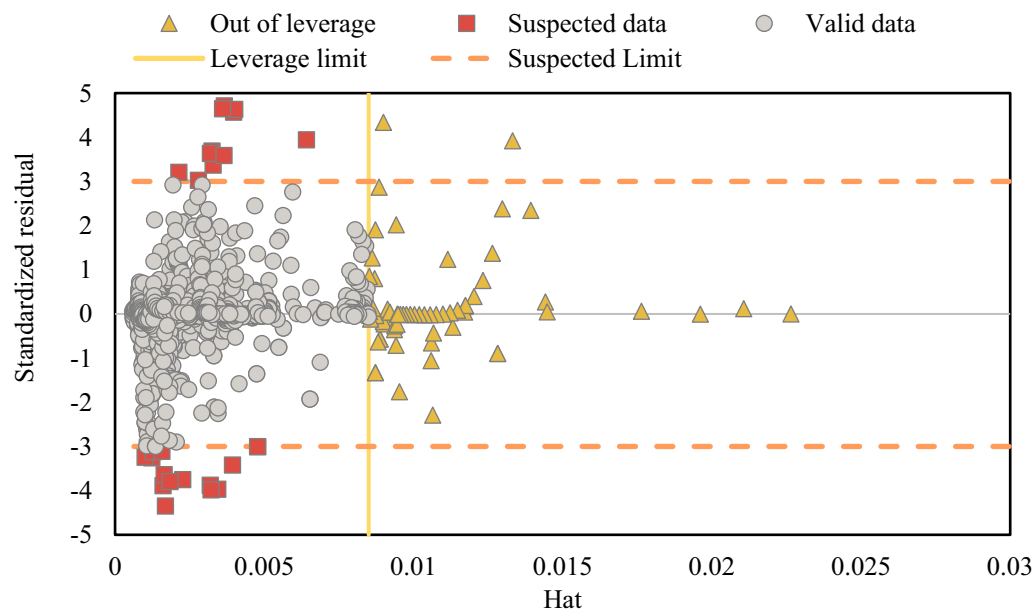


Figure 9. Williams plot for outlier the proposed GMDH.

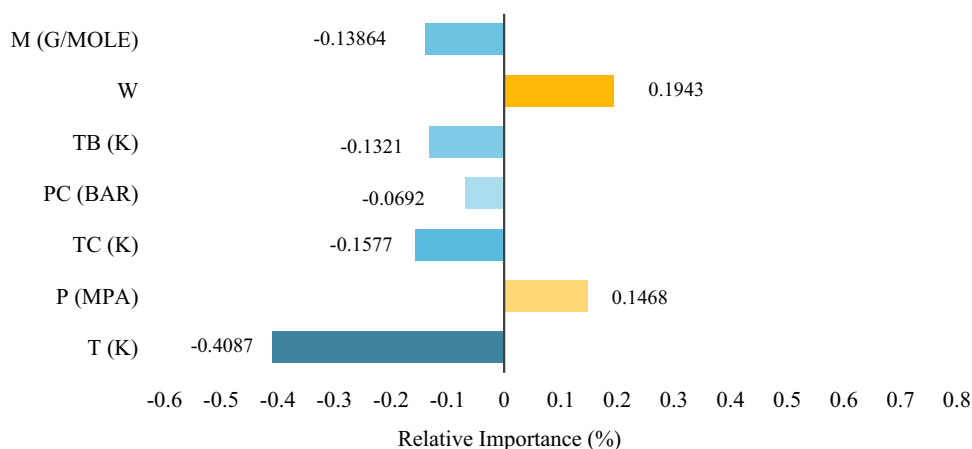


Figure 10. Evaluation of the input parameters' impact on ILs viscosity.

only supports our models' effectiveness but also highlights their potential as reliable methods for forecasting IL viscosity. Our target in this research was in line with the goal of obtaining better predictions about the physical features of ILs¹⁵. The main concern, though, is making accurate predictions regarding the viscosity of pure IL. We used the genetic programming (GP) and group method of data handling (GMDH) techniques to do this. In particular, our study adds custom models with clear benefits, focusing on accuracy and ease of use in determining the viscosity of pure ILs.

Conclusions

The GMDH model was obtained by modeling 2813 experimental findings from 45 ILs based on temperature, pressure, molecular weight, critical volume, and acentric factor. Furthermore, IL viscosity was calculated using temperature and boiling temperature in accordance with Eyring's hypothesis. There were statistical and graphical comparisons between GMDH and experimental data in order to evaluate the model's efficacy. AARD, ARD, RMSE, and R^2 parameters indicated that the GMDH model performed rather well. Using the relevance factor, the impact of input characteristics on the model's target parameter was also investigated. The relevance factor illustrated that the temperature is the most important parameter affecting ILs viscosity. Finally, the employed dataset's reliability and validity were assessed using the leverage statistics. In our case, Williams' plot was applied to study the established paradigm's applicability domain and data collection. Only a small number of data points were found to be outside the realm of applicability. In light of all the above, the developed GMDH model is able to accurately forecast IL viscosity and obtain IL physicochemical parameters in different chemical engineering processes.

Data availability

All data have been gathered from the literature. All references used for extracting the required data have been cited in the text. However, the data will be available from the corresponding author upon reasonable request.

Received: 18 March 2023; Accepted: 20 February 2024

Published online: 15 April 2024

References

- Salgado, J. *et al.* Density and viscosity of three (2, 2, 2-trifluoroethanol+ 1-butyl-3-methylimidazolium) ionic liquid binary systems. *J. Chem. Thermodyn.* **70**, 101–110 (2014).
- Wu, T.-Y., Chen, B.-K., Hao, L., Kuo, C.-W. & Sun, I.-W. Thermophysical properties of binary mixtures {1-methyl-3-pentylimidazolium tetrafluoroborate+ polyethylene glycol methyl ether}. *J. Taiwan Inst. Chem. Eng.* **43**(2), 313–321 (2012).
- Canongia Lopes, J. *et al.* Polarity, viscosity, and ionic conductivity of liquid mixtures containing [C4C1im][Ntf2] and a molecular component. *J. Phys. Chem. B* **115**(19), 6088–6099 (2011).
- Hezave, A. Z., Dorostkar, S., Ayatollahi, S., Nabipour, M. & Hemmateenejad, B. Dynamic interfacial tension behavior between heavy crude oil and ionic liquid solution (1-dodecyl-3-methylimidazolium chloride ([C12mim][Cl]⁺ distilled or saline water/heavy crude oil)) as a new surfactant. *J. Mol. Liq.* **187**, 83–89 (2013).
- Atashrouz, S., Zarghampour, M., Abdollahi, S., Pazuki, G. & Nasernejad, B. Estimation of the viscosity of ionic liquids containing binary mixtures based on the Eyring's theory and a modified Gibbs energy model. *J. Chem. Eng. Data* **59**(11), 3691–3704 (2014).
- Zafarani-Moattar, M. T. & Majdan-Cegincara R. Viscosity, density, speed of sound, and refractive index of binary mixtures of organic solvent+ ionic liquid, 1-butyl-3-methylimidazolium hexafluorophosphate at 298.15 K. *J. Chem. Eng. Data* **52**(6), 2359–2364 (2007).
- Welton, T. Ionic liquids: A brief history. *Biophys. Rev.* **10**(3), 691–706 (2018).
- Freemantle, M. *An Introduction to Ionic Liquids* (Royal Society of Chemistry, 2010).
- Schmidt, H. *et al.* Experimental study of the density and viscosity of 1-ethyl-3-methylimidazolium ethyl sulfate. *J. Chem. Thermodyn.* **47**, 68–75 (2012).
- Torrecilla, J. S., Tortuero, C., Cancilla, J. C. & Díaz-Rodríguez, P. Neural networks to estimate the water content of imidazolium-based ionic liquids using their refractive indices. *Talanta* **116**, 122–126 (2013).
- Torrecilla, J. S., Tortuero, C., Cancilla, J. C. & Díaz-Rodríguez, P. Estimation with neural networks of the water content in imidazolium-based ionic liquids using their experimental density and viscosity values. *Talanta* **113**, 93–98 (2013).
- Zhu, A., Wang, J. & Liu, R. A volumetric and viscosity study for the binary mixtures of 1-hexyl-3-methylimidazolium tetrafluoroborate with some molecular solvents. *J. Chem. Thermodyn.* **43**(5), 796–799 (2011).
- Yu, G., Zhao, D., Wen, L., Yang, S. & Chen, X. Viscosity of ionic liquids: Database, observation, and quantitative structure-property relationship analysis. *AIChE J.* **58**(9), 2885–2899 (2012).
- Burrell, G. L., Burgar, I. M., Separovic, F. & Dunlop, N. F. Preparation of protic ionic liquids with minimal water content and 15N NMR study of proton transfer. *Phys. Chem. Chem. Phys.* **12**(7), 1571–1577 (2010).
- Duong, D. V. *et al.* Machine learning investigation of viscosity and ionic conductivity of protic ionic liquids in water mixtures. *J. Chem. Phys.* **156**(15), 85592 (2022).
- Chen, Y., Peng, B., Kontogeorgis, G. M. & Liang, X. Machine learning for the prediction of viscosity of ionic liquid–water mixtures. *J. Mol. Liq.* **350**, 118546 (2022).
- Hosseinzadeh, M. & Hemmati-Sarapardeh, A. Toward a predictive model for estimating viscosity of ternary mixtures containing ionic liquids. *J. Mol. Liq.* **200**, 340–348 (2014).
- Barycki, M. *et al.* Temperature-dependent structure-property modeling of viscosity for ionic liquids. *Fluid Phase Equilibria* **427**, 9–17 (2016).
- Gardas, R. L. & Coutinho, J. A. A group contribution method for viscosity estimation of ionic liquids. *Fluid Phase Equilibria* **266**(1–2), 195–201 (2008).
- Gharagheizi, F., Ilani-Kashkoui, P., Mohammadi, A. H., Ramjugernath, D. & Richon, D. Development of a group contribution method for determination of viscosity of ionic liquids at atmospheric pressure. *Chem. Eng. Sci.* **80**, 326–333 (2012).
- Lazzús, J. A. & Pulgar-Villarreal, G. A group contribution method to estimate the viscosity of ionic liquids at different temperatures. *J. Mol. Liq.* **209**, 161–168 (2015).
- Paduszynski, K. & Domanska, U. Viscosity of ionic liquids: An extensive database and a new group contribution model based on a feed-forward artificial neural network. *J. Chem. Inf. Model.* **54**(5), 1311–1324 (2014).
- Zhao, Y., Huang, Y., Zhang, X. & Zhang, S. A quantitative prediction of the viscosity of ionic liquids using S σ -profile molecular descriptors. *Phys. Chem. Chem. Phys.* **17**(5), 3761–3767 (2015).
- Atashrouz, S., Pazuki, G. & Alimoradi, Y. Estimation of the viscosity of nine nanofluids using a hybrid GMDH-type neural network system. *Fluid Phase Equilibria* **372**, 43–48 (2014).
- Mousavi, S. P. *et al.* Viscosity of ionic liquids: Application of the Eyring's theory and a committee machine intelligent system. *Molecules* **26**(1), 156 (2021).
- Loyola-Gonzalez, O. Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access* **7**, 154096–154113 (2019).
- Menad, N. A. & Noureddine, Z. An efficient methodology for multi-objective optimization of water alternating CO₂ EOR process. *J. Taiwan Inst. Chem. Eng.* **99**, 154–165 (2019).
- Kang, D., Wang, X., Zheng, X. & Zhao, Y.-P. Predicting the components and types of kerogen in shale by combining machine learning with NMR spectra. *Fuel* **290**, 120006 (2021).
- Mohammadi, M.-R. *et al.* Modeling the solubility of light hydrocarbon gases and their mixture in brine with machine learning and equations of state. *Sci. Rep.* **12**(1), 14943 (2022).
- Ma, J., Kang, D., Wang, X. & Zhao, Y.-P. Defining kerogen maturity from orbital hybridization by machine learning. *Fuel* **310**, 122250 (2022).
- Lv, Q. *et al.* Modelling CO₂ diffusion coefficient in heavy crude oils and bitumen using extreme gradient boosting and Gaussian process regression. *Energy* **275**, 127396 (2023).
- Dong, G. Exploiting the power of group differences: Using patterns to solve data analysis problems. *Synth. Lect. Data Min. Knowl. Discov.* **11**(1), 1–146 (2019).
- Rudin, C. Please stop explaining black box models for high stakes decisions. *Statistics* **1050**, 26 (2018).
- Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**(5), 206–215 (2019).
- Loyola-González, O. *et al.* PBC4cip: A new contrast pattern-based classifier for class imbalance problems. *Knowl.-Based Syst.* **115**, 100–109 (2017).

36. Gaciño, F. M., Paredes, X., Comuñas, M. J. & Fernández, J. Effect of the pressure on the viscosities of ionic liquids: Experimental values for 1-ethyl-3-methylimidazolium ethylsulfate and two bis (trifluoromethyl-sulfonyl) imide salts. *J. Chem. Thermodyn.* **54**, 302–309 (2012).
37. Gaciño, F. M., Paredes, X., Comuñas, M. J. & Fernández, J. Pressure dependence on the viscosities of 1-butyl-2, 3-dimethylimidazolium bis (trifluoromethylsulfonyl) imide and two tris (pentafluoroethyl) trifluorophosphate based ionic liquids: New measurements and modelling. *J. Chem. Thermodyn.* **62**, 162–169 (2013).
38. Xu, Y., Chen, B., Qian, W. & Li, H. Properties of pure n-butylammonium nitrate ionic liquid and its binary mixtures of with alcohols at T=(293.15 to 313.15) K. *J. Chem. Thermodyn.* **58**, 449–459 (2013).
39. Yu, Z., Gao, H., Wang, H. & Chen, L. Densities, viscosities, and refractive properties of the binary mixtures of the amino acid Ionic Liquid [bmim][Ala] with methanol or benzylalcohol at T=(298.15 to 313.15) K. *J. Chem. Eng. Data* **56**(6), 2877–2883 (2011).
40. Domańska, U., Zawadzki, M. & Lewandowska, A. Effect of temperature and composition on the density, viscosity, surface tension, and thermodynamic properties of binary mixtures of N-octylisoquinolinium bis (trifluoromethyl) sulfonyl imide with alcohols. *J. Chem. Thermodyn.* **48**, 101–111 (2012).
41. Fendt, S., Padmanabhan, S., Blanch, H. W. & Prausnitz, J. M. Viscosities of acetate or chloride-based ionic liquids and some of their mixtures with water or other common solvents. *J. Chem. Eng. Data* **56**(1), 31–34 (2011).
42. Domańska, U., Skiba, K., Zawadzki, M., Padaszyński, K. & Królikowski, M. Synthesis, physical, and thermodynamic properties of 1-alkyl-cyanopyridinium bis (trifluoromethyl) sulfonyl imide ionic liquids. *J. Chem. Thermodyn.* **56**, 153–161 (2013).
43. Rocha, M. A., Ribeiro, F. M., Ferreira, A. I. L., Coutinho, J. A. & Santos, L. M. Thermophysical properties of [CN– 1C1im][PF6] ionic liquids. *J. Mol. Liq.* **188**, 196–202 (2013).
44. Diogo, J. C., Caetano, F. J., Fareira, J. M. & Wakeham, W. A. Viscosity measurements of three ionic liquids using the vibrating wire technique. *Fluid Phase Equilibria* **353**, 76–86 (2013).
45. Liu, X., Afzal, W. & Prausnitz, J. M. Unusual trend of viscosities and densities for four ionic liquids containing a tetraalkyl phosphonium cation and the anion bis (2, 4, 4-trimethylpentyl) phosphinate. *J. Chem. Thermodyn.* **70**, 122–126 (2014).
46. Qian, W., Xu, Y., Zhu, H. & Yu, C. Properties of pure 1-methylimidazolium acetate ionic liquid and its binary mixtures with alcohols. *J. Chem. Thermodyn.* **49**, 87–94 (2012).
47. Ochędzan-Siodłak, W., Dziubek, K. & Siodłak, D. Densities and viscosities of imidazolium and pyridinium chloroaluminate ionic liquids. *J. Mol. Liq.* **177**, 85–93 (2013).
48. Yan, F. *et al.* Prediction of ionic liquids viscosity at variable temperatures and pressures. *Chem. Eng. Sci.* **184**, 134–140 (2018).
49. Kirkwood, J. G., Buff, F. P. & Green, M. S. The statistical mechanical theory of transport processes. III. The coefficients of shear and bulk viscosity of liquids. *J. Chem. Phys.* **17**(10), 988–994 (1949).
50. Eyring, H. Viscosity, plasticity, and diffusion as examples of absolute reaction rates. *The Journal of chemical physics* **4**(4), 283–291 (1936).
51. Plawsky, J. L. *Transport Phenomena Fundamentals* (CRC Press, 2009).
52. Ivakhnenko, A. G. Polynomial theory of complex systems. *IEEE Trans. Syst. Man Cybern.* **4**, 364–378 (1971).
53. Shokir, E. E. M., Emera, M., Eid, S. & Wally, A. A new optimization model for 3D well design. *Oil Gas Sci. Technol.* **59**(3), 255–266 (2004).
54. McKay, B., Willis, M. & Barton, G. Steady-state modelling of chemical process systems using genetic programming. *Comput. Chem. Eng.* **21**(9), 981–996 (1997).
55. Koza, J. R. & Koza, J. R. *Genetic Programming: on the Programming of Computers by Means of Natural Selection* (MIT Press, 1992).
56. Madar, J., Abonyi, J. & Szeifert, F. Genetic programming for system identification. *Intelligent Systems Design and Applications (ISDA 2004) Conference, Budapest, Hungary* (2004).
57. Shokir, E. M. E.-M. & Dmour, H. N. Genetic programming (GP)-based model for the viscosity of pure and hydrocarbon gas mixtures. *Energy Fuels* **23**(7), 3632–3636 (2009).
58. Garcia-Garabal, S. *et al.* Transport properties for 1-ethyl-3-methylimidazolium n-alkyl sulfates: Possible evidence of grotthuss mechanism. *Electrochim. Acta* **231**, 94–102 (2017).
59. Tammann, G. & Hesse, W. Die Abhängigkeit der Viscosität von der Temperatur bei unterkühlten Flüssigkeiten. *Z. Anorganische Allgemeine Chemie* **156**(1), 245–257 (1926).
60. Shateri, M., Ghorbani, S., Hemmati-Sarapardeh, A. & Mohammadi, A. H. Application of Wilcoxon generalized radial basis function network for prediction of natural gas compressibility factor. *J. Taiwan Inst. Chem. Eng.* **50**, 131–141 (2015).
61. Atashrouz, S., Mirshekar, H. & Hemmati-Sarapardeh, A. A soft-computing technique for prediction of water activity in PEG solutions. *Colloid Polym. Sci.* **295**(3), 421–432 (2017).
62. Hemmati-Sarapardeh, A., Aminshahidi, B., Pajouhandeh, A., Yousefi, S. H. & Hosseini-Kalozakh, S. A. A soft computing approach for the determination of crude oil viscosity: Light and intermediate crude oil systems. *J. Taiwan Inst. Chem. Eng.* **59**, 1–10 (2016).
63. Atashrouz, S., Mirshekar, H. & Mohaddespour, A. A robust modeling approach to predict the surface tension of ionic liquids. *J. Mol. Liq.* **236**, 344–357 (2017).
64. Rousseeuw, P. J. & Leroy, A. M. *Robust Regression and Outlier Detection*. Syria Studies vol. 7 (Wiley, 1987).
65. Mousavi, S.-P. *et al.* Modeling surface tension of ionic liquids by chemical structure-intelligence based models. *J. Mol. Liq.* **342**, 116961 (2021).

Acknowledgements

The study was funded by the Foreign Young Talent Program (Grant No: DL2022011001L) from the Ministry of Science and Technology of China.

Author contributions

S.K.: Investigation, Visualization, Writing-Original Draft, F.H.: Conceptualization, Validation, Modeling, S.A.: Writing-Review & Editing, Methodology, Data curation, Supervision, D.N.: Writing-Review & Editing, Validation, A.H.-S.: Methodology, Validation, Supervision, Writing-Review & Editing, A.M.: Writing-Review & Editing, Validation, Supervision.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.A., A.H.-S. or A.M.

Reprints and permissions information is available at www.nature.com/reprints.

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International

License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024