Gorji, Ali Ebrahimpoor; Alopaeus, Ville

Prediction of solubility of hydrogen (H_2) in hydrocarbons using QSPR method: MLR data-driven as a simple Machine Learning (ML) algorithm

Published in:
International Journal of Hydrogen Energy

DOI:
10.1016/j.ijhydene.2024.09.433

Published: 11/11/2024

Document Version
Publisher's PDF, also known as Version of record

International Journal of Hydrogen Energy 90 (2024) 803–816

Contents lists available at ScienceDirect

International Journal of Hydrogen Energy

journal homepage: www.elsevier.com/locate/he

Prediction of solubility of hydrogen (H_2) in hydrocarbons using QSPR method: MLR data-driven as a simple Machine Learning (ML) algorithm

Ali Ebrahimpoor Gorji **, Ville Alopaeus *

Aalto University, School of Chemical Technology, Department of Chemical and Metallurgical Engineering, Research Group of Chemical Engineering, P.O. Box 16100, FI-00076 Aalto, Finland

ARTICLE INFO

Handling Editor: Dr M Djukic

Keywords:
H_2 solubility in hydrocarbons
QSPR
Prediction
Machine learning (ML)
Pressure and temperature effects

ABSTRACT

In this study, the 'Quantitative Structure-Property Relationship' (QSPR) method has been applied for the prediction hydrogen (H_2) solubility in different types of hydrocarbons using a new bigger dataset than former studied datasets. The dataset constitutes of 1751 datapoints including 32 unique hydrocarbons at the wide ranges of pressures and temperatures. The simple Machine Learning (ML) algorithm, called '*Multilinear Regression (MLR)*' has been applied for the model development for the first time which has not been studied for this application, yet. The two suggested MLR-QSPR models including novel molecular descriptors, called 'PaDEL' and 'sigma profile' descriptors, have been developed for the first time. The dataset was divided to a training set for the development of models, and to a validation set for external validation. The advantages of this study were discussed and compared with other available models which were developed with other ML algorithms. In these comparisons, some deficiencies of former models have been shown and discussed. Unlike former models, internal validation using Leave One/Multi Out- Cross Validations (LOO-CV/LMO-CV) and Y-scrambling methods were performed on the both MLR-QSPR models using statistical parameters for further assessment. According to the obtained results of statistical parameters ($R^2 = 0.98$ and $Q^2_{LOO\text{-}CV} = 0.98$), the predictive capability of the suggested MLR-QSPR models was acceptable for training set. Regarding the external validation, another statistical parameter like AARD% = 9.79 was also satisfactory for validation set.

1. Introduction

The urbanization and the increasing population rates lead to a massive increase of the fossil fuels and energy consumptions. These large consumptions cause to create some problems like high rate of CO_2-gas emission into atmosphere [1]. To remove this problem and to reduce the rate of global warming, the finding and introducing the new clean energy resources [2] like hydrogen gas (H_2) can play as a key role. For this reason, H_2 has recently received remarkable attractions by researchers [1–4] in comparisons with other resources.

In one hand, H_2 is a carbon-free source which plays prominent role in the energy production and generation [1]. On the other hand, the accurate knowledge of the solubility of H_2 is crucial for the process development of hydroprocesses [5] and industrial processes [3]. Accurate information of the phase equilibria in the geological reservoir is vital for the study of hydrogen reactivity and mobility, as well as the monitoring, control, and optimization of the storage [6]. However, the availability of data of solubility of H_2 in heavy oils is scarce at elevated temperatures and pressures, the solubility of H_2 in different types of pure light hydrocarbon is available [7–16] and has been modeled [1,2] at wide ranges of temperature and pressures.

Regarding the performed experimental studies, a comprehensive literature review of the solubility of H_2 in the different hydrocarbons has been carried out. As can be seen in Table 1, the details of studied hydrocarbons and their properties can be found. The most recent experimental studies for the measurement of solubility of H_2 in hydrocarbons have been performed by our research group (i.e., 2014) [5] and Aslam et al. [17] (i.e., 2015). Regarding to our former study, a continuous flow apparatus with a camera system was modified for gas–liquid equilibrium measurements to achieve the high accuracy of the solubility of H_2 in toluene, hexadecane, and octadecane. Benefits of the modified apparatus were a short time of residence of a sample in the heated zone and no sampling is needed [5]. The benchmark was done by measuring the solubility of H_2 in hexadecane and toluene at high pressures and

* Corresponding author.
** Corresponding author.
E-mail addresses: ali.ebrahimpoorgorji@aalto.fi, ali.ebrahimpoor.chemeng@gmail.com (A.E. Gorji), ville.alopaeus@aalto.fi (V. Alopaeus).

https://doi.org/10.1016/j.ijhydene.2024.09.433
Received 14 June 2024; Received in revised form 20 September 2024; Accepted 29 September 2024
Available online 9 October 2024

temperatures (P = 5–10 MPa, T = 461–575 K) and comparing the results with the literature values. After that, the solubility of H_2 in octadecane which would be a good model for heavy oil systems, was measured for the first time. In contrast, Aslam et al. [17] measured the solubility of H_2 in toluene and methylcyclohexane at low ranges of pressures (lower than 1 MPa) and temperatures. In their studies, the solubility values were measured using the static isochoric saturation method. More experimental studies of solubility of H_2 in different types of hydrocarbons can be found in Table 1.

Regarding the modeling studies, it was tried to apply equations of state (EOSs) for the prediction of solubility of H_2 in hydrocarbons in the preliminary efforts. Since the accuracy of EOSs in the prediction of the solubility of H_2 was restricted, particularly in high-pressure or/and high-temperature conditions [3], different algorithms of Machine Learning (ML) have been used. There are many ML algorithms [45] such as multiple linear regression (MLR), multiple nonlinear regression (MNLR), k-nearest neighbors (kNN), decision trees (DT), random forest (RF), gradient boosting machine (GBM), support vector machine (SVM), artificial neural network (ANN), etc. A comprehensive literature review of the modeling studies using different ML algorithms for the prediction of solubility of H_2 in the different hydrocarbons has been conducted and tabulated in Table 2. The majority of above-mentioned ML algorithms (except MLR) has been studied in past three years. For further comparisons, the most important statistical parameters have been reported for each applied ML algorithm as much as possible.

As can be seen from Table 2, there were two missing points in the performed studies in the literature. The first point was that the molecular variables (predictors or descriptors) of hydrocarbons were almost chosen same (T_c, P_c, and M_w) to distinguish the effect of molecular structures. The second point was a missing of MLR model as a simplest ML algorithm [45]. However, the reported values of statistical parameters (i.e., R^2, RMSE, MSE, and AARD%) were excellent, but all used ML algorithms seem to be a little bit challenging task for general usage by researchers, due to their complexities. Apart from complexity of applied ML algorithms, the high values of these statistical parameters may be attributed to the high numbers of used variables (inputs). For example, the numbers of selected molecular variables (descriptors) for those datasets which had 15 and 11 structural variations of hydrocarbons were quite high. In another word, there were not the rational proportions between the number of applied molecular variables (i.e., 3: (P_c, T_c, and acentric factor (ω))) and number of variations of studied hydrocarbons (15 and 11) which had been studied by Refs. [1,2], and [46]. Moreover, the values of critical properties such T_c and P_c of some hydrocarbons may be unavailable which causes some problems and troubles for the prediction of solubility of H_2 in hydrocarbons. For these reasons, it seems that it is necessary to introduce a simple MLR model including new kind of descriptors (except former suggested descriptors) for more variations of hydrocarbons.

Table 1
The experimental studies of H_2 solubility in different types of hydrocarbons, and hydrocarbons features (critical temperature (T_c), critical pressure (P_c), and molecular weigh (Mw)).

Hydrocarbon Types	Hydrocarbon name	Number of carbons (nC)	T_c (K)	P_c (MPa)	M_w (g/mol)	Refs.
alkane	Methane	1	190.56	4.599	16.042	[7–10]
	Ethane	2	305.32	4.872	30.068	[10,11]
	Propane	3	369.82	4.246	44.095	[12,13]
	Butane	4	425.12	3.786	58.121	[14–16]
	Pentane	5	469.7	3.367	72.148	[18,19]
	Hexane	6	507.49	3.018	86.174	[20,21]
	Heptane	7	540.13	2.727	100.200	[22–24]
	Octane	8	568.88	2.486	114.227	[21–26]
	Decane	10	617.7	2.103	142.279	[19,21,27–31]
	Dodecane	12	658.1	1.817	170.332	[32]
	Hexadecane	16	722.1	1.432	226.438	[5,28,33–35]
	Eicosane	20	771.4	1.198	282.543	[36,37]
	Octacosane	28	844.00	0.954	394.754	[33,36,37]
	Hexatriacontane	36	896.00	0.845	506.965	[28,36,37]
	Hexatetracontane	46	1064.86	0.454	647.229	[28]
	Octadecane	18	747	1.27	254.5	[5]
	2,2,4 trimethylpentane	8	543.9	2.57	114.23	[22]
alkene	1-Octene	8	567	2.68	112.21	[22]
	Ethene	2	282.34	5.041	28.054	[38]
	1-Hexene	6	504	3.143	84.161	[39]
	1-Heptene	7	537.3	2.92	98.188	[39]
cyclic	Cyclohexane	6	553.58	4.073	84.16	[22,40,41]
	Methylcyclohexane	7	572.19	3.471	98.19	[17,22,41]
aromatic	Benzene	6	562.05	4.895	78.11	[22,41,42]
	Toluene	7	591.75	4.108	92.14	[5,17,21,22,41]
	Ethylbenzene	8	617.1	3.61	106.16	[22]
	m-xylene	8	617	3.54	106.16	[22]
	1,2,4-trimethylbenzene	9	649.1	3.232	120.19	[22]
	Cumene	9	631	3.209	120.19	[43]
	Diphenylmethane	13	760	2.71	168.238	[44]
Polycyclic aromatic	Naphthalene	10	748.4	4.05	128.17	[35.42]
	Phenanthrene	14	869	2.87	178.23	[42]
	Pyrene	16	938.2	2.61	202.25	[42]
	1,2,3,4tetrahydronaphthalene	10	720	3.65	132.2	[35]
terpene	Squalane	30	822	0.7	422.8	[26]

Table 2
Numbers of datapoints and studied hydrocarbons at each former dataset, and different used ML algorithms for the prediction of solubility of H_2 in the different hydrocarbons with their independent inputs and their values of statistical parameters.

Research group (year)	Number of datapoints	Studied hydrocarbons in the dataset	Independent variables (inputs)	Applied ML algorithms	Statistical parameters				Ref.
					R^2	*RMSE*	*MSE*	*AARD %*	
Amar et al. (2023)	1484	***15 alkanes***: Methane, ethane, propane, butane, pentane, hexane, heptane, octane, decane, dodecane, hexadecane, eicosane, octacosane, hexatriacontane, hexatetracontane	T, P, Tc, Pc, and acentric factor (ω)	Multilayer Perceptron (ANN)	0.9959	0.004	–	–	[1]
				Cascaded Forward Neural Network (ANN)	0.9969	0.0035	–	–	
				committee machine intelligent system (CMIS)	0.9972	0.0033	–	–	
Tatar et al. (2022)	1845	***15 alkanes***: Methane, ethane, propane, butane, pentane, hexane, heptane, octane, decane, dodecane, hexadecane, eicosane, octacosane, hexatriacontane, hexatetracontane	T, P, Tc, Pc, Mw, boiling point (Tb), and acentric factor (ω)	Decision tree	0.9998[a]	0.0009[a]	0.002[a]	–	[2]
				Random forest	0.9935[a]	0.0051[a]	0.0027[a]	–	
				Gradient boosting	0.9964[a]	0.0038[a]	0.0023[a]	–	
				Extremely randomized trees	0.9944[a]	0.0047[a]	0.0028[a]	–	
Hadavimoghaddam et al. (2022)	1332	***32 hydrocarbons***: Ethane, propane, butane, hexane, heptane, octane, decane, dodecane, hexadecane, eicosane, octacosane, hexatriacontane, hexatetracontane, 2,2,4 trimethylpentane, ethene, 1-hexene, 1-heptene, 1-octene, benzene, toluene, ethylbenzene, m-xylene, cumene, 1,2,4 trimethylbenzene, diphenylmethane, cyclohexane, methylcyclohexane, naphthalene, 1,2,3,4 tetrahydronaphthalene, phenanthrene, pyrene, squalane	T, P, Tc, Pc, and Mw	Genetic programming	0.9861	0.0132	–	–	[3]
				Group method of data handling	0.9687	0.0198	–	–	
Mohammadi et al. (2021)	919	***26 hydrocarbons***: butane, hexane, heptane, octane, decane, dodecane, hexadecane, eicosane, octacosane, hexatriacontane, hexatetracontane, 2,2,4 trimethylpentane, 1-octene, benzene, toluene, ethylbenzene, m-xylene, cumene, 1,2,4 trimethylbenzene, cyclohexane, methylcyclohexane, naphthalene, 1,2,3,4 tetrahydronaphthalene, phenanthrene, pyrene, squalane	T, P, Tc, Pc, and Mw	Extreme gradient boosting	0.9998	0.0007	–	1.81	[4]
				Adaptive boosting support vector regression	0.9995	0.0012	–	3.40	
				Gradient boosting with categorical features support	0.9993	0.0015	–	4.7	
				Light gradient boosting machine	0.9940	0.0045	–	3.51	
				Multi-layer perceptron	0.9917	0.0053	–	6.01	
Jiang et al. (2021)	278	11 aromatic/cyclic hydrocarbons: cyclohexane, toluene, tetrahydrofuran, 1,4-Dioxane, 1-methyl-2-pyrrolidone, benzene, naphthalene, phenanthrene, pyrene, methylcyclohexane, quinoline,	T, P, Tc, Pc, and acentric factor (ω)	Adaptive neuro-fuzzy inference systems	0.9966	0.0052	0.00002	7.88	[46]

(*continued on next page*)

Table 2 (*continued*)

Research group (year)	Number of datapoints	Studied hydrocarbons in the dataset	Independent variables (inputs)	Applied ML algorithms	Statistical parameters				Ref.
					R^2	*RMSE*	*MSE*	*AARD %*	
				Artificial neural networks	0.9953	0.0062	0.00003	8.77	
				Least-squares support vector machines	0.9950	0.0063	0.00004	13.7	

[a] These values have been reported for training set.

The 'Quantitative Structure-Property Relationship' (QSPR) method is commonly used to provide quantitative and qualitative descriptions between macroscopic properties and involved structures [47–50]. Unlike former studies (see Table 2) where T_c, P_c, and ω were assumed as distinct molecular descriptors, QSPR study will reveal the relationship between hydrocarbon structures and H_2 solubility with finding proper descriptors. Therefore, it is computationally possible to reach the high-accurate solubility of H_2 in many hydrocarbons using the QSPR method which is important for the development of thermodynamic models (PC-SAFT and Peng-Robinson) as well as many chemical processes. Up to now, no researchers have applied the QSPR method for the prediction of solubility of H_2 in hydrocarbons at the wide ranges of temperatures and pressures. QSPR method has been applied in the well-known software called 'QSARINS' [51–53] which features several methods of internal and external validations in its environment.

The main aim of this study is to show the advantages of MLR-QSPR model-based in comparison with other applied non-linear and complicated ML algorithms. In particular, the results of this study make comparison with genetic programming (GP) which was developed using 32 hydrocarbons and 1332 datapoints by Ref. [3]. Among of applied ML algorithms (see Table 2), GP ML algorithm can be considered as a white box approach [3]. Most applied ML algorithms were black box. Finding the most proper descriptors and structural variables of hydrocarbons is another main aim of this study. In this study, a bigger dataset (i.e., 32 hydrocarbons and 1751 datapoints) will be applied for the development of the predictive MLR-QSPR model for the prediction of H_2 solubility in hydrocarbons. The developed predictive and reliable MLR-QSPR model aids the chemical engineering community to optimize and design the process for the specific applications.

2. Method

2.1. Basic theory

In this study, the values of the solubility of H_2 in hydrocarbons (i.e., x) are first converted to logarithms-based (i.e., ln(x)). The effects of pressure, temperature, and molecular structures on x were included as the independent variables as below (Eq. (1)):

$$\ln(x) = a_1 \ln(P) + a_2 \ln(T) + F(\text{descriptors}) + a_3 \quad (1)$$

where 'a_1', 'a_2', and 'a_3' are adjustable parameters. In this study, the most important descriptors (i.e., F (descriptors) in Eq. (1)) were applied to distinguish the effect of hydrocarbon structural variations and the effects of pressure and temperature are considered using 'ln P' and 'ln T', respectively.

2.2. Dataset

According to the gathered experimental data [7–44], a bigger dataset with very high structural variations of hydrocarbons and datapoints has been created for QSPR studies in comparison with former datasets which were proposed (listed) in Table 2. The detail of this dataset can be found in Table 3 and Table S1. In total, 1751 datapoints including 32 unique hydrocarbons at wide ranges of pressure (0.1–96.18 MPa) and temperature (92.3–701.5 K) are listed in Table 3.

2.3. Former available models

Among of different ML algorithms in Table 2, there is only one study by Hadavimoghaddam et al. [3] which has been done on the enough variation of hydrocarbons. Also, there were some black-box approaches which were carried out on low variations of hydrocarbons as well as homologous series of alkanes [1,2] using high number of descriptors. Since the MLR-QSPR model is a white-box approach, it is rational to make comparison the obtained results of this study with the most qualitative and quantitative white-box model with enough variation of hydrocarbons (i.e., GP model) [3]. As already shown in Table 2, that model was developed using 32 different structures simultaneously consider the effects of pressure temperature, structural variations of hydrocarbons. The used molecular descriptors were T_c, P_c, and M_w. Hadavimoghaddam et al. [3] evaluated the prediction capability of their predictive model using RMSE and R^2 statistical parameters. They reported R^2 and RMSE values for total data of their dataset 0.986 and 0.013, respectively. It is necessary that their proposed model to be applied on our dataset (see Table 3), for further investigations.

2.4. QSPR method

2.4.1. Calculation of descriptors

In this study, it was attempted to calculate 2D, 1D, and 0D descriptors which were independent of the optimization of molecular structures of hydrocarbon. Before calculation of such descriptors, each molecular structure of hydrocarbons was drawn in ChemBioDraw-Ultra [54]. Then, the drawn structures in 'MDL mol' format fed to PaDEL-Descriptor software [55] for the calculation of descriptors. Descriptors with constant or almost constant values for each hydrocarbon were eliminated. Finally, 1000 descriptors were calculated. Also, the sigma profiles for each involved hydrocarbon in the dataset (see Table 3) have been used for the calculation of molecular descriptors in this study. There is a comprehensive databank of sigma profiles in COSMO-RS software which was developed by Klamt and Eckert [56]. To plot the sigma profile, two factors must be available: 1) the 61 points of specific charge density (e/ Å^2) which typically vary from −0.03 to +0.03 with 0.001 step (interval), and 2) the probability distribution of a molecular surface segment having a specific charge density. In this study, the calculated values (peak, height, the values of second factor of sigma profile) at each point of specific charge density (from −0.03 to +0.03 with 0.001 step) have been considered as the molecular descriptors in the variables (descriptors) selection.

2.4.2. Model development

As can be seen in Eq. (1), ln (x) is simultaneously a function of hydrocarbon descriptors alongside pressure and temperature. For model construction, the suitable descriptors must be selected from two pools of descriptor (1: 1000 PaDEL-descriptors and 2: 61 points from −0.03 to +0.03 with 0.001 step). There are well-known methods of variables selection such as genetic algorithm (GA) method [57], artificial neural network (ANN) [58], replacement method (RM) [59]. In this study, GA

Table 3
The solubility values (mole fraction[a]) of H_2 in studied hydrocarbons of our dataset with their ranges of pressure, temperature, and experimental values as well as statistical parameter.

hydrocarbons	Temperature range (K)	Pressure range (MPa)	Range of x (***mole fraction***)	AARD%[c]
Methane	116–172	3.37–27.57	0.031–0.349	16.03
	108–183	2.44–27.58	0.011–0.378	
	92–180	0.22–27.63	0.0018–0.392	
	103–173	1.02–10.83	0.0073–0.225	
Ethane	92–255	3.18–28.43	0.0072–0.202	11.62
	148–223	2.02–8.10	0.006–0.055	
Propane	98–348	1.03–20.68	0.0021–0.244	11.51
	277–360	3.44–27.57	0.024–0.319	
Butane	327–394	2.77–16.84	0.019–0.217	11.27
	297–355	2.24–10.72	0.02–0.111	
	172–297	2.06–19.37	0.009–0.111	
Pentane[b]	**273–373**	**0.34–20.68**	**0.001–0.146**	**9.41**
	308–463	**2.86–14.08**	**0.025–0.119**	
Hexane	344–410	1.24–15.11	0.010–0.143	3.34
	298–373	1.38–9.81	0.010–0.093	
Heptane	295	6.99–17.33	0.045–0.112	1.9
	298–323	0.10	0.0006–0.0007	
	238–308	0.10	0.0004–0.0007	
Octane[b]	**295**	**10.44**	**0.066**	**17.49**
	463–543	**1.01–12.53**	**0.008–0.20**	
	298–373	**2.4–15.27**	**0.018–0.137**	
	295	**0.68–1.37**	**0.004–0.008**	
	298–323	**0.10**	**0.0006–0.0007**	
	248–308	**0.10**	**0.0004–0.0007**	
Decane	283–449	1.23–14.21	0.016–0.088	4.35
	503	1.48–10.10	0.017–0.150	
	344–423	3.71–15.04	0.036–0.128	
	293–373	2.04–10.35	0.015–0.088	
	308–573	1.78–16.74	0.025–0.147	
	462–583	1.92–20.26	0.025–0.240	
Dodecane[b]	**344–410**	**1.42–13.24**	**0.014–0.125**	**2.73**
Hexadecane	298–448	1.15–13.88	0.018–0.113	5.64
	301–508	3.79–6.51	0.033–0.098	
	461–542	2.00–15.14	0.031–0.197	
	453–543	2.04–9.74	0.036–0.135	
Eicosane	373–573	1.00–5.08	0.011–0.096	3.6
	323–423	2.23–12.91	0.027–0.124	
Octacosane	342–447	1.46–14.00	0.03–0.178	7.68
	373–573	0.98–5.06	0.014–0.110	
	419–516	4.14–5.10	0.081–0.096	
	348–423	2.86–12.43	0.045–0.172	
Hexatriacontane	357–447	1.37–14.34	0.033–0.210	6.39
	373–573	1.02–5.06	0.015–0.123	
	373–423	3.56–14.32	0.067–0.208	
Cyclohexane[b]	**304–332**	**0.13–4.49**	**0.0006–0.029**	**8.39**
	303	**0.88–4.74**	**0.003–0.019**	
	295	**6.99–17.33**	**0.028–0.068**	
Toluene[b]	**298–373**	**0.87–10.12**	**0.002–0.047**	**6.56**
	303	**1.22–4.41**	**0.004–0.014**	
	293–333	**0.51–0.89**	**0.001–0.003**	

(*continued on next page*)

Table 3 (*continued*)

hydrocarbons	Temperature range (K)	Pressure range (MPa)	Range of x (***mole fraction***)	AARD%[c]
	295	**6.99–17.33**	**0.021–0.050**	
Benzene	303 295 323–423	0.98–4.60 6.99–17.33 2.55–15.73	0.002–0.012 0.017–0.042 0.010–0.058	13.93
Methylcyclohexane	303 293–333 295	1.23–4.32 0.50–0.89 6.99–20.78	0.006–0.021 0.002–0.004 0.033–0.094	6.42
1,2,3,4 tetrahydronaphthalene	453–623	1.53–9.19	0.014–0.085	8.97
Naphthalene	503–623 373–423	1.42–8.67 4.29–19.39	0.012–0.080 0.015–0.056	18.58
1-octene[b]	**295**	**6.99–20.78**	**0.043–0.120**	**3.88**
Ethylbenzene	295	10.44–17.33	0.033–0.054	16.82
m-xylene	295	10.44–17.33	0.034–0.056	1.06
1,2,4 trimethylbenzene	295	6.99–17.33	0.024–0.057	2.47
Squalane	295	0.68–1.37	0.006–0.013	3.5
Phenanthrene	383–423	5.89–21.69	0.016–0.055	9.89
Pyrene	433	5.17–19.73	0.015–0.057	6.33
Cumene	323	1.02–11.70	0.004–0.048	6.48
Diphenylmethane	462–701	2.00–25.00	0.012–0.305	5.9
Octadecane[b]	**462–540**	**5.00–9.97**	**0.087–0.184**	**7.59**
2,2,4 trimethylpentane[b]	**295**	**6.99–20.78**	**0.052–0.145**	**14.69**
Ethene	114–247	4.63–96.18	0.024–0.576	30

[a] Solubility (x) $= \frac{n_{H2}^{Gas}}{n_{Hyd}^{L}}$ (mole fraction).

[b] Bold hydrocarbon means validation set.

[c] Calculated using Eq. (13).

Table 4
Applied statistical parameters in this study.

Introduced parameters	Introduced parameters equations	Eqs. No
Coefficient of determination	$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(Y_i^{exp} - Y_i^{pre} \right)^2}{\sum_{i=1}^{n} \left(Y_i^{exp} - \overline{Y}_i \right)^2}$	(2)
Adjustable coefficient of determination	$R^2_{Adj} = 1 - (1\text{-}R^2) \times (\frac{n-1}{n-p-1})$	(3)
Leave-one-out cross-validated coefficient of determination	$Q^2_{LOO\text{-}CV} = 1 - \frac{\sum_{i=1}^{n} \left(Y_i^{exp} - Y_i^{pre-CV} \right)^2}{\sum_{i=1}^{n} \left(Y_i^{exp} - \overline{Y}_i \right)^2}$	(4)
Fisher function	$F = \frac{\sum_{i=1}^{n} \left(Y_i^{pre} - \overline{Y}_i \right)^2 / p}{\sum_{i=1}^{n} \left(Y_i^{exp} - Y_i^{pre} \right)^2 / n-p-1}$	(5)
Standard residual	$S = \sqrt{\frac{\sum_{i=1}^{n} \left(Y_i^{exp} - Y_i^{pre} \right)^2}{n-p-1}}$	(6)
root mean square error (RMSE)	$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(Y_i^{exp} - Y_i^{pre} \right)^2}{n}}$	(7)
Average absolute deviation	$AAD = \frac{\sum_{i=1}^{n} \left(\lvert Y_i^{exp} - Y_i^{pre} \rvert \right)}{n}$	(8)
Average Absolute relative deviation %	$AARD\% = \frac{\sum_{i=1}^{n} \left(\lvert Y_i^{exp} - Y_i^{pre} \rvert \right) / Y_i^{exp}}{n} \times 100$	(9)
maximum Leverage	$h^* = 3\,(p+1)/n$	(10)

Y_i^{exp}, Y_i^{pre}, $\overline{Y}_i$, n, and p demonstrate experimental values, predicted values, average experimental values, the number of the experimental dataset, and the number of employed descriptors, respectively.

was used to build MLR QSPR models-based COSMO descriptors. The details of the GA-MLR algorithm can be found elsewhere [60,61]. It should be mentioned that the QSARINS software [51–53] was applied to develop the GA-MLR models.

2.4.3. Statistical parameters

The goodness-of-fit of QSPR model should be carefully checked using the standard statistical parameters, including root mean square error (RMSE), coefficient of determination (R^2), leave-one-out cross-validated coefficient of determination ($Q^2_{LOO\text{-}CV}$), adjustable coefficient of determination (R^2_{Adj}), average absolute relative deviation (%AARD), average absolute deviations (AAD), Fisher function (F), standard residual (S), and maximum (or critical) leverage (h*). More detailed information regarding the statistical parameters used in this study can be found in Table 4 (Eqs. (2)–(10)).

Applicability of Domain (AD) analysis as a vital concept of QSPR approach should be considered. It allows [62]: 1) the uncertainty in prediction 2) the extent of extrapolation of QSPR models [63,64]. In order to predict solubility values of H_2 in new hydrocarbons, it is essential that new hydrocarbon lie within the same AD space. In another words, it means that new hydrocarbons are physicochemically, biologically, or structurally similar with molecules used for model development (i.e., training set). The more space of AD, the more reliable predictions of new hydrocarbons. To carry out the external validation using validation set, it is essential to ensure that the validations set of molecules is inside of QSPR model's AD [65].

The space of AD can be specified using two main parameters: 1) the leverage values (h_i) 2) the standardized residual (SDR) and. SDR was defined as Eq. (11):

$$SDR = \frac{Y_i^{exp} - Y_i^{pre}}{\sqrt{\frac{\sum_{m=1}^{n} \left(Y_i^{exp} - Y_i^{pre} \right)^2}{n}}} \quad (11)$$

h_i, represents a measure of a molecule's distance from the center of the training set. It is needed to determine whether new hydrocarbons are within the applicability of domain of the developed QSPR model or not. The parameter can be calculated with Eq. (12).

$$h_i \text{ (or Leverage (i))} = z_i.\ (Z_i^T Z_i)^{-1}.\ z_i^T \quad (12)$$

When z_i, Z are the descriptor row vector of point i and a n × p matrix of descriptors for compounds derived from the training set, respectively. AD of developed QSPR models can be obtained in QSARINS software for

Table 5
The suggested MLR-QSPR models for the training and validation status.

Kind of descriptors	Number of datapoints and hydrocarbons in training set	Model	Eq. No
PaDEL	1464 and 24	ln(x) = 1.7487 **ln(T)** + 0.9903 **ln(P)** + 0.0014 **ATSC0m** – 2.5452 **MATS1e** + 0.1242 **minssCH2** – 0.1686 **ETA_Beta_s** –15.7177	(13)
Sigma profile	1464 and 24	ln(x) = 1.5386 **ln(T)** + 1.0131 **ln(P)** – 1.5564 **(SCD-0.008)** – 0.0751 **(SCD-0.002)** – 0.4608 **(SCD-0.001)** + 0.6626 **(SCD0)** – 0.2904 **(SCD0.001)** + 0.3904 **(SCD0.004)** – 13.1838	(14)

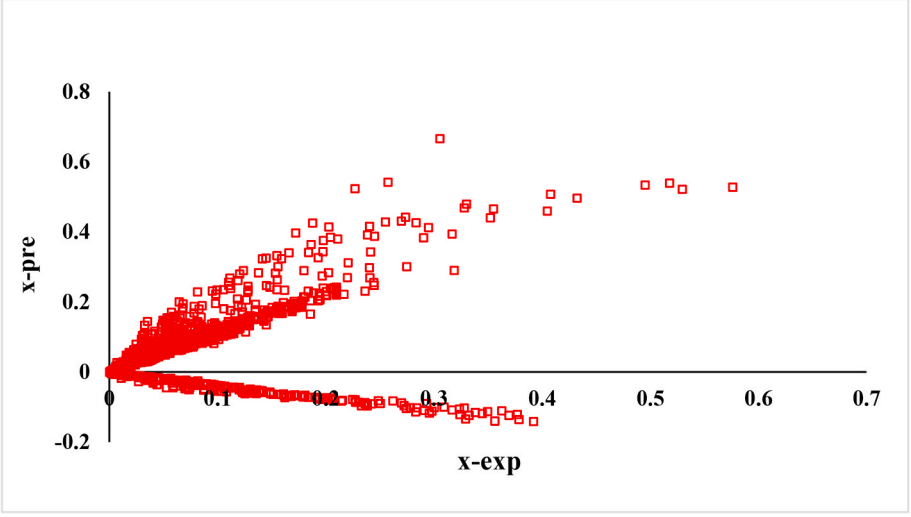

Fig. 1. The predicted values of H_2 solubility in studied hydrocarbons of our dataset (this study) by GP-model versus experimental values.

Table 6
The values of statistical parameters of suggested MLR-QSPR models (ln-based) for both of training and validation sets.

Eqs. No	Sets	Number of datapoints and hydrocarbons	R^2	R2-Adj	Q2-LOO	Q2-LMO	F	S	RMSE
(13)	Training	1464 and 24	0.98	0.98	0.98	0.98	11161	0.1492	0.1488
	Validation	287 and 8	–	–	–		–	–	0.1543
(14)	Training	1464 and 24	0.96	0.96	0.96	0.96	4313	0.2058	0.2051
	Validation	287 and 8	–	–	–	–	–	–	0.2190

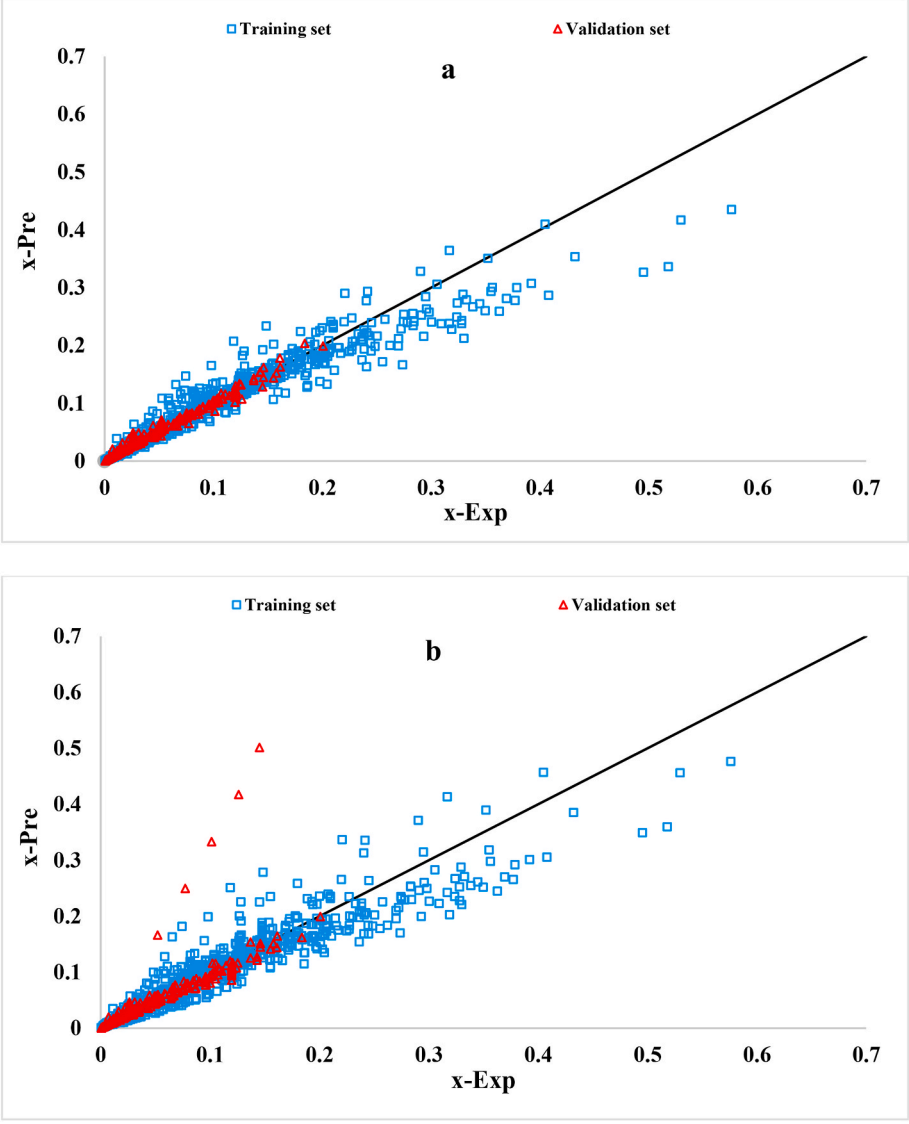

Fig. 2. The predicted versus experimental values for both of training and validation sets using a) PaDEL descriptors (Eq. (13)) and b) sigma profile descriptors (Eq. (14)).

each model and maximum leverage (i.e., h*) can be calculated using Eq. (10).

2.4.4. Internal and external validations

After building of QSPR model, it is essential to conduct internal and external validations on the training (approx. 80% of main dataset) and validation (approx. 20% of main dataset) sets, respectively. Regarding the internal validation, Y-Scrambling, leave multi out -cross validation (LMO-CV), and leave one out -cross validation (LOO-CV) methods should be conducted on the developed QSPR model. These methods were performed on the training set only. Regarding the external validation, the prediction capability of developed QSPR model was evaluated using a validation set. Both of internal and external validations of QSPR models can be carried out in QSARINS software one by one, due their high importance.

3. Results and discussion

3.1. A comparison with former model based-genetic programming (GP)

Prediction capability of the proposed GP-model by Hadavimoghaddam et al. [3], was examined on our dataset which was described in

Table 7
The built MLR models without and with former descriptors (i.e., T_c, P_c, and M_w) for the training and validation status.

Without/with descriptors	Number of datapoints and hydrocarbons in training set	Model	Eq. No
Without	1464 and 24	$\ln(x) = 0.4320\ \mathbf{ln(T)} + 0.9329\ \mathbf{ln(P)} - 7.0204$	(15)
With former descriptors	1464 and 24	$\ln(x) = 1.4141\ \mathbf{ln(T)} + 0.9522\ \mathbf{ln(P)} - 0.0062\ \mathbf{T_C} - 0.3319\ \mathbf{P_C} + 0.0045\ \mathbf{M_w} - 8.9884$	(16)

Table 8
The values of statistical parameters of MLR-models (ln-based) without and with former descriptors for both of training and validation sets.

Eqs. No	Sets	Number of datapoints and hydrocarbons	R^2	R2-Adj	Q2-LOO	F	S	RMSE
(15)	Training	1464 and 24	0.73	0.73	0.73	1962	0.5318	0.5312
	Validation	287 and 8	–	–	–	–	–	0.6059
(16)	Training	1464 and 24	0.92	0.92	0.92	3480	0.2841	0.2836
	Validation	287 and 8	–	–	–	–	–	0.2538

Table 3. The more details of calculations and predicted values of H_2 solubility in studied hydrocarbons of our dataset can be found in Table S2. As can be seen in Table S2, GP-model predicted the negative H_2 solubility in methane. Also, the same deficiency was observed for other datapoints. It indicates that the proposed GP-model used some irrelevant descriptors (variables or predictors: T_c, P_c, and M_w) to distinguish the effect of hydrocarbons on the H_2 solubility. The predicted versus experimental values have been plotted in Fig. 1.

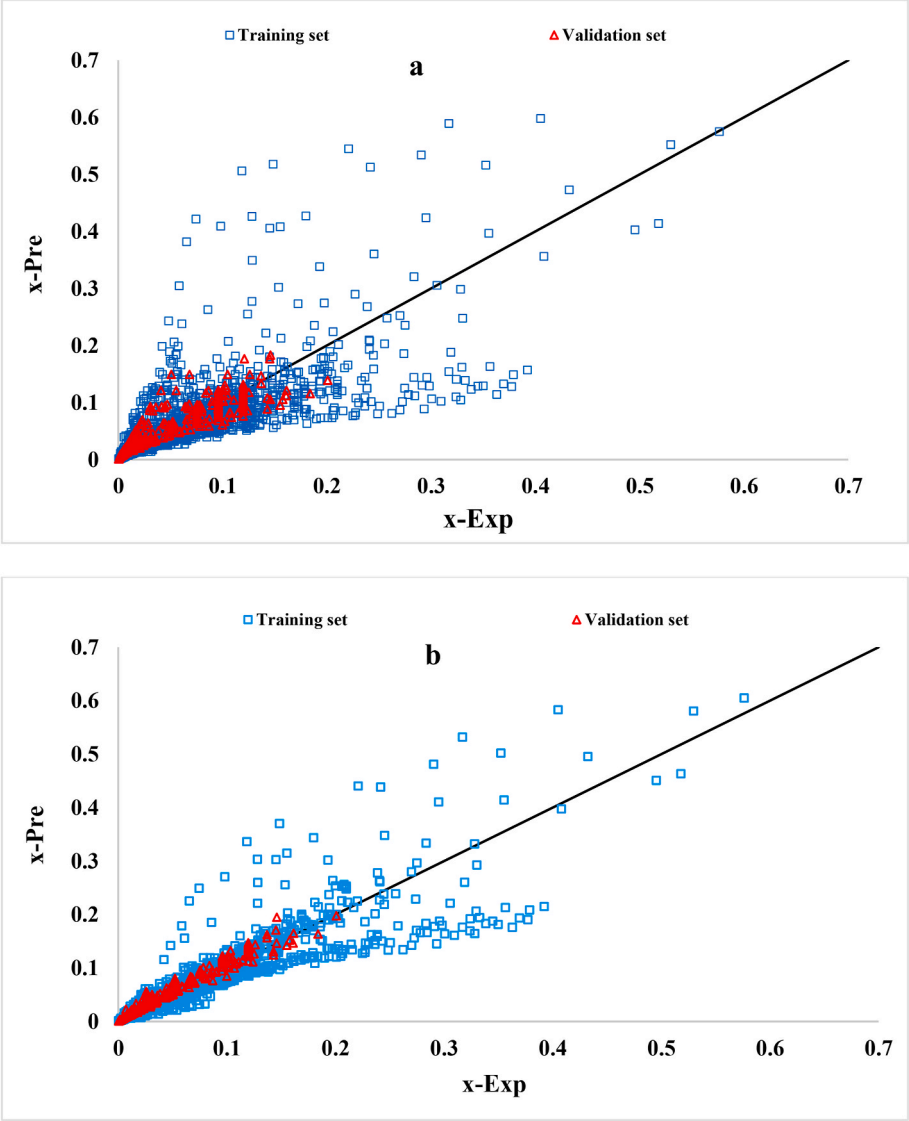

Fig. 3. The predicted versus experimental values for both of training and validation sets using a) without any descriptors (Eq. (15)) and b) with T_c, P_c, and M_w descriptors (Eq. (16)).

Therefore, it seems that it is better to propose a new reliable predictive model including relevant descriptors. Although the proposed GP-model showed a big deficiency for prediction of H_2 in methane, the prediction capability of such model had enough accuracy for prediction of H_2 in other hydrocarbons. For this reason, a comprehensive comparison between the proposed MLR-model in this study and former GP-model will be done in upcoming sections.

3.2. Developed MLR-QSPR models

In this study, an extensive dataset including 32 hydrocarbons from six types of hydrocarbons (see Table 1) and 1751 datapoints was divided to training and validation sets for performing of internal and external validations. All datapoints of eight hydrocarbons (i.e., pentane, octane, dodecane, octadecane, cyclohexane, 1-octene, toluene, and 2,2,4 trimethylpentane) were set aside into validation set. Some types of hydrocarbons (cyclic and branched alkane) were set intentionally aside in the validation set to guarantee presence of hydrocarbons with new types. The main aim of this categorization is to investigate the MLR-QSPR model's prediction capability for new types of hydrocarbons. Since the finding the proper descriptors (except T_c, P_c, and M_w descriptors in Table 2) in MLR-ML algorithms is always interesting, two separate MLR-QSPR models with two different kinds of descriptors (PaDEL and sigma profile descriptors) are suggested here. These two suggested MLR-QSPR models which were built in training and validation status (Eqs. (13) and (14)) are indicated in Table 5.

The values of statistical parameters of each suggested model (ln-based) are shown in Table 6.

As indicated in Table 6, the obtained values of Q^2_{LOO} (as internal validation) of each MLR-QSPR model (either with PaDEL or sigma profile descriptors) were high which are confirming that each model has acceptable capability for prediction of H_2 solubility in studied hydrocarbons of our dataset at the wide ranges of pressures and temperatures (see Table 3). Also, LMO-CV and Y-scrambling techniques have been done on the training set in the QSARINS software for each selected MLR-QSPR model and results confirmed the validity of each model. As external validation, it is also shown that H_2 solubility in hydrocarbons of validation set (**bold** hydrocarbons in Table 3) predicted with enough accuracy based on the obtained values of AARD% (See Table 3 and Table S3). In comparison between these two models, the predicted values by Eq. (13) were in better agreement with experimental data than the predicted values by Eq. (14). The proposed MLR-QSPR model (i.e., Eq. (13)) could take into account the effects of pressure and temperature on the H_2 solubility in majority of studied hydrocarbons (the included hydrocarbons in Table 3) well.

The Williams plot for the training and validation sets which was obtained using MLR-QSPR model (i.e., Eq. (13)), is shown in Fig. S1. The plot confirms that there were no outliers in our dataset as no hydrocarbon at given pressure and temperature has a leverage value higher than the critical leverage (i.e., $h^* = 0.014$) and its SRD is higher than ±3. Although the leverage values of some hydrocarbons were a little bit higher than the critical leverage (h*), the MLR-QSPR models could predict the H_2 solubility in those hydrocarbons well. The plots of predicted versus experimental values for both of training and validation sets which were obtained using MLR-QSPR models (i.e., Eqs. (13) and (14)), are shown in Figure (2).

To show the importance of presence of relevant descriptors in the models, two separate MLR-models 1) without any descriptors (i.e., Eq. (15)) and 2) with T_c, P_c, and M_w descriptors (i.e., Eq. (16)) have been regressed on our dataset, too. These two models have been shown in Table 7.

The values of statistical parameters of each model (ln-based) are shown in Table 8.

The plots of predicted versus experimental values for both training and validation sets obtained using MLR-models (i.e., Eqs. (15) and (16)), are shown in Figure. (3).

A general comparison between all the developed models (i.e., Eqs. (13)–(16)) in this study has been performed using some statistical parameters reported in Table 9.

Table 10
Definition of each descriptor.

Models	Descriptors	Definition	Ref
PaDEL (i.e., Eq. (13))	**ATSC0m**	Centered Broto-Moreau autocorrelation - lag 0/weighted by mass	[66–71]
	MATS1e	Moran autocorrelation - lag 1/weighted by Sanderson electronegativities	[66–71]
	minssCH2	Minimum atom-type E-State: –CH2-	[72]
	ETA_Beta_s	A measure of electronegative atom count of the molecule	[73,74]
Sigma profile (i.e., Eq. (14))	**SCD-0.008**	the probability distribution of a molecular surface segment having a '-0.008 e/ Å^2,	[56]
	SCD-0.002	the probability distribution of a molecular surface segment having a '-0.002 e/ Å^2,	[56]
	SCD-0.001	the probability distribution of a molecular surface segment having a '-0.001 e/ Å^2,	[56]
	SCD0	the probability distribution of a molecular surface segment having a '0 e/ Å^2,	[56]
	SCD0.001	the probability distribution of a molecular surface segment having a '0.001 e/ Å^2,	[56]
	SCD0.004	the probability distribution of a molecular surface segment having a '0.004 e/ Å^2,	[56]

Table 9
Statistical parameters of all developed models in this study for both of training and validation sets (none ln -based).

Eqs. No	Sets	R^2	RMSE	AAD	AARD%
(13)	Training	0.94	0.0188	0.0085	10.02
	Validation	0.99	0.0055	0.0031	9.79
(14)	Training	0.90	0.0224	0.0119	14.94
	Validation	0.65	0.0336	0.0091	15.38
(15)	Training	0.47	0.0566	0.0324	49.23
	Validation	0.75	0.0225	0.0155	70.24
(16)	Training	0.78	0.0344	0.0173	21.66
	Validation	0.96	0.0117	0.0085	23.70

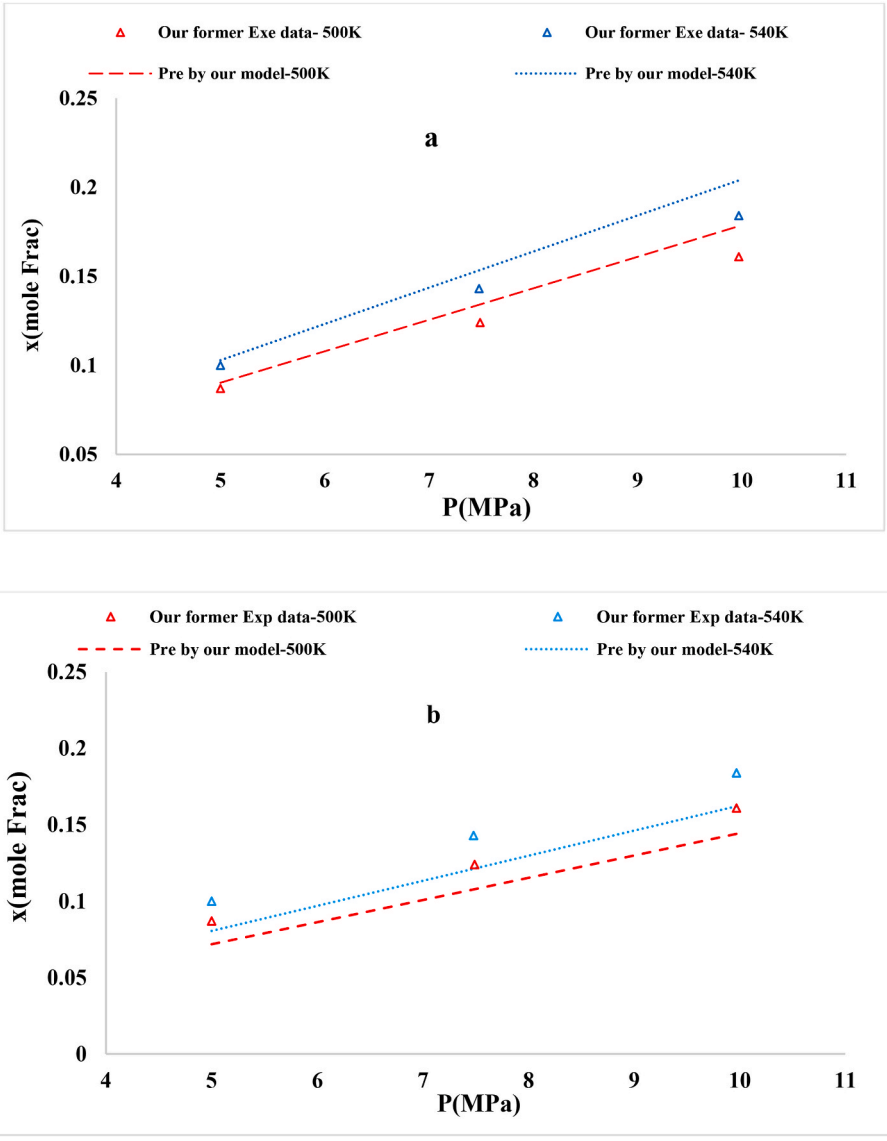

Fig. 4. The predicted values by our MLR-QSPR models (a: Eq. (13) and b: Eq. (14)) and our former experimental values [5] at two temperatures and different pressures.

As can be seen in Table 9 and Fig. 3, both developed MLR-models (i.e., Eqs. (15) and (16)) had not acceptable accuracy for the prediction of H_2 solubility in hydrocarbons. This point verifies that suggested MLR-QSPR models (i.e., Eqs. (13) and (14) in this study, could efficiently predict the target, due to their relevant descriptors (either PaDEL or sigma profile descriptor).

To introduce the appeared descriptors in the suggested MLR-QSPR model with PaDEL descriptors (i.e., Eq. (13)), it should be mentioned that '**ATSC0m**', '**MATS1e**', descriptors are Autocorrelation descriptors [66]. These descriptors are topological and can be calculated using molecular graphs. Detail of these descriptors can be found in Moreau and Broto [66–69] as well as Moran [70] and Geary [71]. '**minssCH2**'is an Electro-Topological-State-Atom descriptor [72]. '**ETA_Beta_s**' is an Extended-Topochemical-Atom descriptor [73,74].

Six points (i.e., ***(−0.008), (−0.002), (−0.001), (0), (0.001), and (0.004)***) of specific charge density (SCD) from sigma profile are the hydrocarbon descriptors. The values of peak (or height) for each hydrocarbon at above points of SCD are used and important for the prediction of H_2 solubility in hydrocarbons. In both suggested MLR-QSPR models, the effects of pressure and temperature on the H_2 solubility in hydrocarbons have been taken into account using 'ln P' and 'ln T', respectively. The values of these two kinds of descriptors can be found in Table S4. The definition of each above descriptor has been listed in Table 10.

In the former studies (see Table 2), no research groups included our former experimental data for H_2-octadecane system which were reported in 2014. For this regard, the prediction capability of our MLR-QSPR models for this system which was one of validation systems has been examined. As can be seen in Fig. 4, the predicted values by our MLR-QSPR models had an excellent consistency with our former experimental data.

To make a comparison with other available models which were developed by different ML algorithms, it should be mentioned that the prediction capability of some models (see Table 2), never examined by new types of hydrocarbons (cyclic and aromatic). For example [1,2], examined the performance of their models only for some homologous series of alkanes. It means that their proposed models have not covered an extensive variation of different hydrocarbons. However, with knowing this point, it has been tried to make comparison our models with CMIS model for some hydrocarbons. This comparison has been

indicated in Fig. 5. As can be seen in Fig. 5, the simple MLR-QSPR model has quite the same capability for the prediction in comparison with most complicated non-linear CMIS model. More details can be found in Table S5. Keep in mind that CMIS model had been developed only for prediction of H_2 solubility in n-alkane hydrocarbons. Unlike our MLR-QSPR models, the prediction capability of such model never examined for other types of hydrocarbons.

A comprehensive comparison of prediction capability between the suggested MLR-QSPR model (i.e., Eq. (13)) and the former GP-model which was the unique and white-box predictive model, has been conducted in this study for the prediction H_2 solubility in hydrocarbons (except methane). This comparison has been demonstrated in Fig. 6. As can be seen in Fig. 6, the simple MLR-QSPR model shows an excellent result for the prediction in comparison with GP-model. More details can be found in Table S6. As it is clear, GP-model has not good efficiency for the prediction of solubility of H_2 in light hydrocarbons and overestimates the prediction of solubility of H_2 in majority of hydrocarbons, due to some irrelevant descriptors (T_c, P_c, and M_w).

As can be seen in Fig. 7, the residual values (i.e., Exp - Pre) seem to reduce as a function of M_W. It shows that suggested MLR-QSPR model (i.e., Eq. (13)) may be used for the prediction of H_2 solubility in heavier hydrocarbons which could be difficult to purify and measure, even if they are outside the present range of studied dataset.

As with all data-driven models, the validity depends on the quality and quantity of the data. The present models could be improved further by experimental solubilities in hydrocarbons which solubility data is not currently available in the literature.

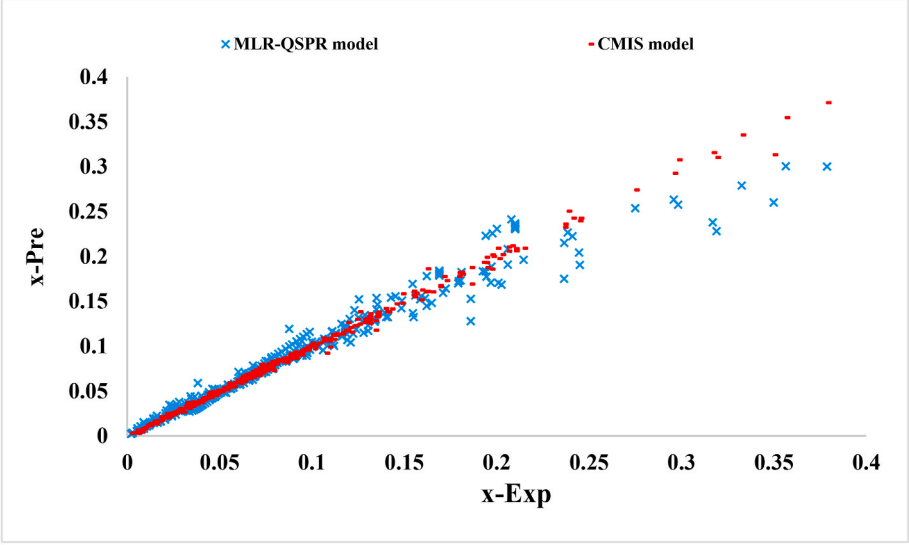

Fig. 5. The comparison between our simple MLR-QSPR model (i.e., Eq. (13)) and most complicated none-linear CMIS model [1].

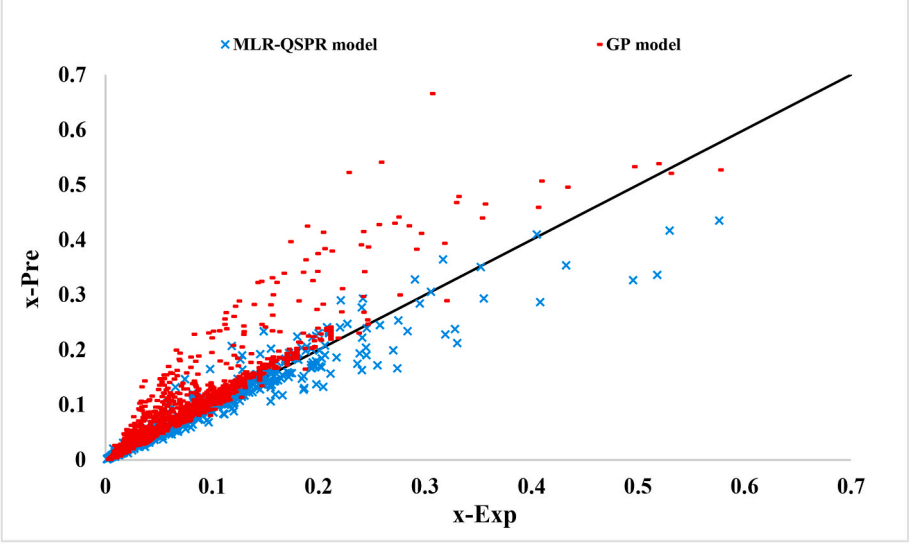

Fig. 6. The comparison between our simple MLR-QSPR model (i.e., Eq. (13)) and GP-model [3].

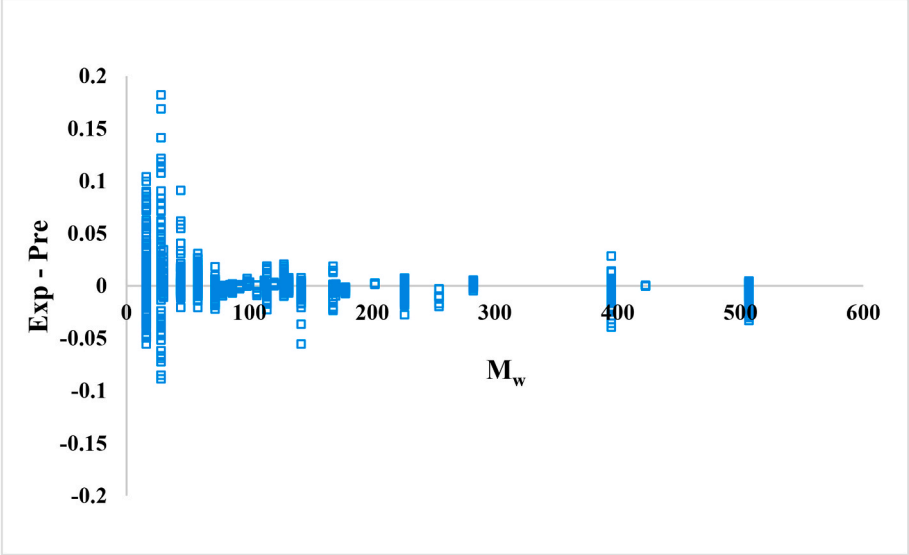

Fig. 7. Residual values of Eq. (13) as a function of molecular weight (M_w).

4. Conclusion

In this study, the strengths, and weaknesses of different applied ML algorithms for the prediction of H_2 solubility in different types of hydrocarbons as well as former molecular variables (i.e., T_c, P_c, and M_w descriptors) have been discussed for the first time. The suggested MLR-QSPR models as the simple ML algorithms, could successfully predict the H_2 solubility in hydrocarbons as functions of pressure, temperature, and suitable molecular descriptors. In this study, H_2 solubility in hydrocarbons has been predicted with very good accuracy using MLR-QSPR models which had two different kinds of descriptors 1) PaDEL and 2) sigma profile. The obtained values of statistical parameters (RMSE, AAD, R^2, and $Q^2_{LOO\text{-}CV}$) of suggested MLR-QSPR models were acceptable for training and validation sets. The obtained values of AARD% parameter were 10.02 and 9.79 for training and validation sets, respectively, expressing the high prediction capability of MLR-QSPR model with PaDEL descriptors. The internal and external validations verified that the prediction of H_2 solubility in different types of hydrocarbons which had not been experimentally studied can be practical with respect to the leverage value of hydrocarbons and AD of MLR-QSPR models.

CRediT authorship contribution statement

Ali Ebrahimpoor Gorji: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Ville Alopaeus:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was partially supported by the 'In-situ equilibrium shifting in CO_2 utilization reactions by novel absorbents (CO2Shift)' Project (351113). The authors gratefully and thankfully appreciate to **Prof. Paola Gramatica** (University of Insubria) for providing the free license of QSARINS software for the development of Multilinear Regression models.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijhydene.2024.09.433.

References

[1] Amar MN, Alqahtani FM, Djema H, Ourabah K, Ghasemi M. Predicting the solubility of hydrogen in hydrocarbon fractions: advanced data-driven machine learning approach and equation of state. J Taiwan Inst Chem Eng 2023;153:105215.
[2] Tatar A, Esmaeili-Jaghdan Z, Shokrollahi A, Zeinijahromi A. Hydrogen solubility in n-alkanes: data mining and modelling with machine learning approach. Int J Hydrogen Energy 2022;47(85):35999–6021.
[3] Hadavimoghaddam F, Mohammadi MR, Atashrouz S, Nedeljkovic D, Hemmati-Sarapardeh A, Mohaddespour A. Data-driven modeling of H2 solubility in hydrocarbons using white-box approaches. Int J Hydrogen Energy 2022;47(78):33224–38.
[4] Mohammadi MR, Hadavimoghaddam F, Pourmahdi M, Atashrouz S, Munir MT, Hemmati-Sarapardeh A, Mohaddespour A. Modeling hydrogen solubility in hydrocarbons using extreme gradient boosting and equations of state. Sci Rep 2021;11(1):17911.
[5] Saajanlehto M, Uusi-Kyyny P, Alopaeus V. A modified continuous flow apparatus for gas solubility measurements at high pressure and temperature with camera system. Fluid Phase Equil 2014;382:150–7.
[6] Chabab S, Théveneau P, Coquelet C, Corvisier J, Paricaud P. Measurements and predictive models of high-pressure H2 solubility in brine (H2O+ NaCl) for underground hydrogen storage application. Int J Hydrogen Energy 2020;45(56):32206–20.
[7] Benham AL, Katz DL. Vapor-liquid equilibria for hydrogen–light-hydrocarbon systems at low temperatures. AIChE J 1957;3(1):33–6.
[8] Hong JH, Kobayashi R. Vapor-liquid equilibrium study of the hydrogen-methane system at low temperatures and elevated pressures. J Chem Eng Data 1981;26(2):127–31.
[9] Tsang CY, Clancy P, Calado JCG, Streett WB. Phase equilibria in the H2/CH4 system at temperatures from 92.3 to 180.0 K and pressures to 140 MPa. Chem Eng Commun 1980;6(6):365–83.
[10] Sagara H, Arai Y, Saito S. Vapor-liquid equilibria of binary and ternary systems containing hydrogen and light hydrocarbons. J Chem Eng Jpn 1972;5(4):339–48.
[11] Heintz A, Streett WB. Phase equilibriums in the hydrogen/ethane system at temperatures from 92.5 to 280.1 K and pressures to 560 MPa. J Chem Eng Data 1982;27(4):465–9.
[12] Trust DB, Kurata F. Vapor-liquid phase behavior of the hydrogen-propane and hydrogen-carbon monoxide-propane systems. AIChE J 1971;17(1):86–91.
[13] Burriss WL, Hsu NT, Reamer HH, Sage BH. Phase behavior of the hydrogen-propane system. Ind Eng Chem 1953;45(1):210–3.
[14] Klink AE, Cheh HY, Amick Jr EH. The vapor-liquid equilibrium of the hydrogen—n-butane system at elevated pressures. AIChE J 1975;21(6):1142–8.
[15] Nelson EE, Bonnell WS. Solubility of hydrogen in n-butane. Ind Eng Chem 1943;35(2):204–6.

[16] Aroyan HJ, Katz DL. Low temperature vapor-liquid equilibria in hydrogen-n-butane system. Ind Eng Chem 1951;43(1):185–9.
[17] Aslam R, Müller K, Müller M, Koch M, Wasserscheid P, Arlt W. Measurement of hydrogen solubility in potential liquid organic hydrogen carriers. J Chem Eng Data 2016;61(1):643–9.
[18] Freitag NP, Robinson DB. Equilibrium phase properties of the hydrogen—methane—carbon dioxide, hydrogen—carbon dioxide—n-pentane and hydrogen—n-pentane systems. Fluid Phase Equil 1986;31(2):183–201.
[19] Connolly JF, Kandalic GA. Gas solubilities, vapor-liquid equilibria, and partial molal volumes in some hydrogen-hydrocarbon systems. J Chem Eng Data 1986;31(4):396–406.
[20] Gao W, Robinson RL, Gasem KA. Solubilities of hydrogen in hexane and of carbon monoxide in cyclohexane at temperatures from 344.3 to 410.9 K and pressures to 15 MPa. J Chem Eng Data 2001;46(3):609–12.
[21] Brunner E. Solubility of hydrogen in 10 organic solvents at 298.15, 323.15, and 373.15 K. J Chem Eng Data 1985;30(3):269–73.
[22] Peramanu S, Pruden BB. Solubility study for the purification of hydrogen from high pressure hydrocracker off-gas by an absorption-stripping process. Can J Chem Eng 1997;75(3):535–43.
[23] Lachowicz SK, Newitt DM, Weale KE. The solubility of hydrogen and deuterium in n-heptane and n-octane at high pressures. Trans Faraday Soc 1955;51:1198–205.
[24] Cook MW, Hanson DN, Alder BJ. Solubility of hydrogen and deuterium in nonpolar solvents. J Chem Phys 1957;26(4):748–51.
[25] Connolly JF, Kandalic GA. Thermodynamic properties of solutions of hydrogen in n-octane. J Chem Therm 1989;21(8):851–8.
[26] Kim KJ, Way TR, Feldman KT, Razani A. Solubility of hydrogen in octane, 1-octanol, and squalane. J Chem Eng Data 1997;42(1):214–5.
[27] Prausnitz JM, Benson PR. Solubility of liquids in compressed hydrogen, nitrogen, and carbon dioxide. AIChE J 1959;5(2):161–4.
[28] Florusse LJ, Peters CJ, Pamies JC, Vega LF, Meijer H. Solubility of hydrogen in heavy n-alkanes: experiments and saft modeling. AIChE J 2003;49(12):3260–9.
[29] Schofield BA, Ring ZE, Missen RW. Solubility of hydrogen in a white oil. Can J Chem Eng 1992;70(4):822–4.
[30] Park J, Robinson RLJ, Gasem KA. Solubilities of hydrogen in heavy normal paraffins at temperatures from 323.2 to 423.2 K and pressures to 17.4 MPa. J Chem Eng Data 1995;40(1):241–4.
[31] Sebastian HM, Simnick JJ, Lin HM, Chao KC. Gas-liquid equilibrium in the hydrogen+ n-decane system at elevated temperatures and pressures. J Chem Eng Data 1980;25(1):68–70.
[32] Gao W, Robinson RL, Gasem KA. High-pressure solubilities of hydrogen, nitrogen, and carbon monoxide in dodecane from 344 to 410 K at pressures to 13.2 MPa. J Chem Eng Data 1999;44(1):130–2.
[33] Breman BB, Beenackers AACM, Rietjens EWJ, Stege RJH. Gas-liquid solubilities of carbon monoxide, carbon dioxide, hydrogen, water, 1-alcohols (1. ltoreq. n. ltoreq. 6), and n-paraffins (2. ltoreq. n. ltoreq. 6) in hexadecane, octacosane, 1-hexadecanol, phenanthrene, and tetraethylene glycol at pressures up to 5.5 MPa and temperatures from 293 to 553 K. J Chem Eng Data 1994;39(4):647–66.
[34] Lin HM, Sebastian HM, Chao KC. Gas-liquid equilibrium in hydrogen+ n-hexadecane and methane+ n-hexadecane at elevated temperatures and pressures. J Chem Eng Data 1980;25(3):252–4.
[35] Luo H, Ling K, Zhang W, Wang Y, Shen J. A model of solubility of hydrogen in hydrocarbons and coal liquid. Energy Sources, Part A Recovery, Util Environ Eff 2010;33(1):38–48.
[36] Huang SH, Lin HM, Tsai FN, Chao KC. Solubility of synthesis gases in heavy n-paraffins and Fischer-Tropsch wax. Ind Eng Chem Res 1988;27(1):162–9.
[37] Park J, Robinson RLJ, Gasem KA. Solubilities of hydrogen in heavy normal paraffins at temperatures from 323.2 to 423.2 K and pressures to 17.4 MPa. J Chem Eng Data 1995;40(1):241–4.
[38] Heintz A, Streett WB. Phase equilibria in the H2/C2H4System at temperatures from 114.1 to 247.1 K and pressures to 600 MPa. Ber Bunsen Ges Phys Chem 1983;87(4):298–303.
[39] Sokolov V, Polyakov A. Solubility of H2 in n-decane, n-tetradecane, 1-hexane, 1-octene, isopropyl benzene, 1-methyl naftalene and decalin. Zh Prikl Khim 1977;50:1403–5.
[40] Ronze D, Fongarland P, Pitault I, Forissier M. Hydrogen solubility in straight run gasoil. Chem Eng Sci 2002;57(4):547–53.
[41] Tsuji T, Shinya Y, Hiaki T, Itoh N. Hydrogen solubility in a chemical hydrogen storage medium, aromatic hydrocarbon, cyclic hydrocarbon, and their mixture for fuel cell systems. Fluid Phase Equil 2005;228:499–503.
[42] Park J, Robinson RL, Gasem KA. Solubilities of hydrogen in aromatic hydrocarbons from 323 to 433 K and pressures to 21.7 MPa. J Chem Eng Data 1996;41(1):70–3.
[43] Phiong HS, Lucien FP. Solubility of hydrogen in α-methylstyrene and cumene at elevated pressure. J Chem Eng Data 2002;47(3):474–7.
[44] Simnick JJ, Liu KD, Lin HM, Chao KC. Gas-liquid equilibrium in mixtures of hydrogen and diphenylmethane. Ind Eng Chem Process Des Dev 1978;17(2):204–8.
[45] Lemaoui T, Eid T, Darwish AS, Arafat HA, Banat F, AlNashef I. Revolutionizing inverse design of ionic liquids through the multi-property prediction of over 300,000 novel variants using ensemble deep learning. Mater Sci Eng R Rep 2024;159:100798.
[46] Jiang Y, Zhang G, Wang J, Vaferi B. Hydrogen solubility in aromatic/cyclic compounds: prediction by different machine learning techniques. Int J Hydrogen Energy 2021;46(46):23591–602.
[47] Gorji AE, Sobati MA, Alopaeus V, Uusi-Kyyny P. Toward solvent screening in the extractive desulfurization using ionic liquids: QSPR modeling and experimental validations. Fuel 2021;302:121159.
[48] Gorji AE, Sobati MA. Toward molecular modeling of thiophene distribution between the ionic liquid and hydrocarbon phases: effect of hydrocarbon structure. J Mol Liq 2019;287:110976.
[49] Gorji AE, Sobati MA. How anion structures can affect the thiophene distribution between imidazolium-based ionic liquid and hydrocarbon phases? A theoretical QSPR study. Energy Fuels 2019;33(9):8576–87.
[50] Gorji AE, Sobati MA. Effect of the cation structure on the thiophene distribution between the ionic liquid with NTf2 anion and the hydrocarbon rich phases: a QSPR study. J Mol Liq 2020;313:113551.
[51] Gramatica P. Principles of QSAR modeling: comments and suggestions from personal experience. Int J Quant Struct-Property Relat (IJQSPR) 2020;5(3):61–97.
[52] Gramatica P, Sangion A. A historical excursus on the statistical validation parameters for QSAR models: a clarification concerning metrics and terminology. J Chem Inf Model 2016;56(6):1127–31.
[53] Gramatica P, Chirico N, Papa E, Cassani S, Kovarich S. QSARINS: a new software for the development, analysis, and validation of QSAR MLR models. 2013.
[54] ChemBioDraw, Ultra level, 12.0.2.1076 version, 2010 CambridgeSoft.
[55] Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem 2011;32(7):1466–74.
[56] Diedenhofen M, Eckert F, Klamt A. Prediction of infinite dilution activity coefficients of organic compounds in ionic liquids using COSMO-RS. J Chem Eng Data 2003;48(3):475–9.
[57] Modarresi H, Modarress H, Dearden JC. QSPR model of Henry's law constant for a diverse set of organic chemicals based on genetic algorithm-radial basis function network approach. Chemosphere 2007;66(11):2067–76.
[58] Shahlaei M. Descriptor selection methods in quantitative structure–activity relationship studies: a review study. Chem Rev 2013;113(10):8093–103.
[59] Khooshechin S, Dashtbozorgi Z, Golmohammadi H, Acree Jr WE. QSPR prediction of gas-to-ionic liquid partition coefficient of organic solutes dissolved in 1-(2-hydroxyethyl)-1-methylimidazolium tris (pentafluoroethyl) trifluorophosphate using the replacement method and support vector regression. J Mol Liq 2014;196:43–51.
[60] Holland JH. Adaption in natural and artificial systems. Ann Arbor MI: The University of MichiganPress; 1975.
[61] Haupt RL, Haupt SE. Practical genetic algorithms. second ed. Hoboken, New Jersey: John Wiley & Sons, Inc.; 2004.
[62] Lemaoui T, Darwish AS, Hammoudi NEH, Abu Hatab F, Attoui A, Alnashef IM, Benguerba Y. Prediction of electrical conductivity of deep eutectic solvents using COSMO-RS sigma profiles as molecular descriptors: a quantitative structure–property relationship study. Ind Eng Chem Res 2020;59(29):13343–54.
[63] Ojha PK, Roy K. Comparative QSARs for antimalarial endochins: importance of descriptor-thinning and noise reduction prior to feature selection. Chemometr Intell Lab Syst 2011;109(2):146–61.
[64] He W, Yan F, Jia Q, Xia S, Wang Q. Description of the thermal conductivity λ (T, P) of ionic liquids using the structure–property relationship method. J Chem Eng Data 2017;62(8):2466–72.
[65] Amini Z, Fatemi MH, Gharaghani S. Hybrid docking-QSAR studies of DPP-IV inhibition activities of a series of aminomethyl-piperidones. Comput Biol Chem 2016;64:335–45.
[66] Todeschini R, Consonni V. Molecular descriptors for chemoinformatics. Weinheim: Wiley VCH; 2009. p. 27–37.
[67] Moreau G, Broto P. The autocorrelation of a topological structure: a new molecular descriptor. Nouv J Chim 1980;4:359–60.
[68] Moreau G, Broto P. Autocorrelation of molecular structures, application to SAR studies. Nouv J Chim 1980;4:757–64.
[69] Broto P, Moreau G, Vandycke C. Molecular structures: perception, autocorrelation descriptor and SAR studies. Autocorrelation descriptor. Eur J Med Chem 1984;19:66–70.
[70] Moran PAP. Notes on continuous stochastic phenomena. Biometrika 1950;37:17–23.
[71] Geary RC. The contiguity ratio and statistical mapping. Inc Statistician 1954;5:115–45.
[72] Hall LH, Kier LB. Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information. J Chem Inf Comput Sci 1995;35:1039–45.
[73] Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem 2011;32(7):1466–74.
[74] Roy K, Ghosh G. QSTR with extended topochemical atom indices. 2. Fish toxicity of substituted benzenes. J Chem Inf Comput Sci 2004;44:559–67.