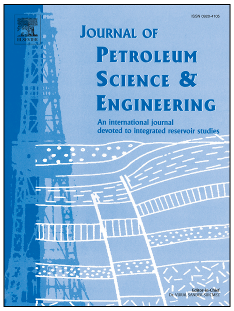

# Prediction of $CO_2$ diffusivity in brine using white-box machine learning

Menad Nait Amar, Ashkan Jahanbani Ghahfarokhi

PII: S0920-4105(20)30131-5

DOI: https://doi.org/10.1016/j.petrol.2020.107037

Reference: PETROL 107037

To appear in: *Journal of Petroleum Science and Engineering*

Received Date: 9 October 2019

Revised Date: 1 February 2020

Accepted Date: 5 February 2020

Please cite this article as: Amar, M.N., Ghahfarokhi, A.J., Prediction of $CO_2$ diffusivity in brine using white-box machine learning, *Journal of Petroleum Science and Engineering* (2020), doi: https://doi.org/10.1016/j.petrol.2020.107037.

# Prediction of $CO_2$ Diffusivity in Brine Using White-Box Machine Learning

Menad Nait Amar [a] and Ashkan Jahanbani Ghahfarokhi [b] *

[a] *Département Etudes Thermodynamiques, Division Laboratoires, Sonatrach, Avenue 1er Novembre, 35000,Boumerdes, Algeria*

[b] *Department of Geoscience and Petroleum, Norwegian University of Science and Technology (NTNU), S. P. Andersens veg 15b, 7031 Trondheim, Norway*

## Abstract

Accurate knowledge of the diffusivity coefficient of $CO_2$ in brine has a significant effect on the design success and monitoring of $CO_2$ storage in saline aquifers, which is a part of carbon capture and sequestration (CCS). Frequently applied experimental approaches for determining this parameter are expensive and time-consuming, and empirical models cannot ensure accurate predictions. Therefore, there is a need to establish cutting-edge correlations for prediction of the diffusivity coefficient of $CO_2$ in brine under various operating conditions. In this work, two white-box machine learning techniques, namely group method of data handling (GMDH) and gene expression programming (GEP) were implemented for correlating the diffusivity coefficient of $CO_2$ in brine with pressure, temperature and the viscosity of the solvent. The obtained results demonstrated the accuracy of the proposed correlations. In addition, statistical and graphical analysis of the performances revealed that GEP correlation outperforms the GMDH correlation, decision trees (DTs), random forest (RF) and all the previous predictive models. GEP correlation exhibited an overall average absolute relative deviation (AARD) of 4.3014% and coefficient of determination ($R^2$) of 0.9979. Finally, by performing the outliers detection, the validity of the GEP correlation was confirmed and only two experimental data points were identified as outliers.

**Keywords** – $CO_2$-brine; diffusivity coefficient; empirical correlations; GEP; GMDH.

*Corresponding author: ashkan.jahanbani@ntnu.no

## 1. Introduction

The increased amount of $CO_2$ in the atmosphere is considered the most important concern of the 21st century around the globe (Boot-Handford et al., 2014). This issue results mainly from fossil fuels being used continuously in different sectors of industry (Azzolina et al., 2015; Boot-Handford et al., 2014; Nait Amar et al., 2019a; Venkatraman and Alsberg, 2017). Therefore, some new approaches have emerged as promising means for reducing the $CO_2$ levels in the atmosphere, among which, carbon capture and sequestration (CCS) in geological formations is still the most attractive one (Amini et al., 2012; Grude et al., 2014; Shahkarami et al., 2014).

CCS has gained much interest within many fields of engineering, environment and energy (Bhakta et al., 2015; Davarazar et al., 2019; Gibbins and Chalmers, 2008; Lee et al., 2010; Mohagheghian et al., 2019; Riahi et al., 2004). The sequestration of the captured $CO_2$ in saline aquifers is the most frequently applied strategy while implementing CCS (Gershenzon et al., 2015, 2014).  In addition, some other applications of $CO_2$ such as enhanced oil recovery methods based on $CO_2$ injection, have served as useful and vital means in reduction of $CO_2$ levels in the atmosphere (Bachu et al., 2004; Ettehadtavakkol et al., 2014; Gozalpour et al., 2005; Holtz et al., 2001; Nait Amar and Zeraibi, 2019).

The process of $CO_2$ injection in saline aquifers is subjected to some mechanisms, namely the contact of $CO_2$ with the in-situ water and its dissolution in brine through molecular diffusion. Therefore, accurate knowledge and determination of the parameters that play important role while monitoring CCS are of vital interest. Diffusivity coefficient which characterizes the diffusivity of fluid is one of these parameters (Cadogan et al., 2014a; Guzmán and Garrido, 2012). Indeed, this parameter has a noticeable effect on the chemical reactions and the interfacial mass transfer occurring deep underground, in addition to impacting the flow path, transport behavior and the quantitative description of diffusion during $CO_2$ injection (Farajzadeh et al., 2009; Guzmán and Garrido, 2012; Mutoru et al., 2011; Trevisan et al., 2014).

Determination of diffusivity coefficient of the $CO_2$ in brine can be done by two distinguished approaches. The first one consists of performing experimental measurements and the second approach is by means of the empirical models. The lab measurements are divided into direct and indirect tests (Lu et al., 2013). The direct procedures are based on the accurate knowledge of the gas concentration in the solvent (Cadogan et al., 2014b; Frank et al., 1996), while the indirect tests exploit the gained information related to the diffusivity of gas such as interfacial tension, volumes of gas and liquids, and the operational conditions (Jang et al., 2018; Li et al., 2016). The experimental approaches are known to deliver accurate results. However, their implementation is time-consuming and demands sophisticated equipment. As a result, several researchers have developed empirical models for estimating the diffusivity coefficient of CO2 in brine. Wilke and Chang (1955) established a predictive correlation for estimating the diffusivity coefficient in numerous dilute solutions. The model employs viscosity and temperature as input parameters.  Lu et al. (2013) developed a model for estimating the diffusivity coefficient of $CO_2$ in water without considering the pressure effect. There model is applicable to cases with temperatures between 268K and 473K. An extended version of  Lu et al. (2013) model at high pressure and temperature conditions, was proposed by Moultos et al. (2016) by applying the concept of molecular dynamics simulations. Although the correlations developed by Lu et al. (2013) and Moultos et al. (2016) are accurate for the $CO_2$-pure water system, they cannot be applied to cases where brine is the solvent. Cadogan et al. (2014a) utilized experimental results for $CO_2$ diffusivity coefficients against brine viscosity at temperature of 298K for establishing a modified Stokes-Einstein relation. Table 1 reports the mathematical formulations of the above models for predicting the diffusivity coefficient of $CO_2$ in water and brine.

An in-depth review of the available correlations for predicting the diffusivity coefficient of $CO_2$ in brine reveals the limitations of these techniques from the applicability and accuracy perspectives (Feng et al., 2019).

In recent years, researchers have shown an increased interest in the application of machine learning techniques for modeling complex systems (Jeong et al., 2018; Nait Amar et al., 2019b; Nait Amar and Zeraibi, 2019; Nomeli and Riaz, 2017; Piotrowski and Napiorkowski, 2012). Machine learning techniques can be divided into computer-aided methods such as support vector regression (SVR) and decision tree, and explicit methods such as gene expression programming (GEP) and group method of data handling (GMDH). The first category is known as black-box approaches, and this means that their paradigms are dependent on a computer-aided technique, while the second category is recognized as white-box methods which means that they deliver explicit expressions (Nait Amar et al., 2019c). Recently, Feng et al. (2019) developed a predictive model for estimating the diffusivity coefficient of $CO_2$ in brine by coupling genetic algorithm with mixed kernel SVR and they obtained satisfactory performances. However, the application and accuracy of their established model depend on calculability efforts, and this presents an issue in terms of flexibility for further utilization.

The main contribution and novelty of this study consist of establishing two distinct explicit and simple-to-use correlations for accurate prediction of diffusivity coefficient of $CO_2$ in brine under various operational conditions. To do so, group method of data handling (GMDH) and gene expression programming (GEP) were implemented with three input parameters, namely pressure, temperature and viscosity of the solvent, using a representative experimental database. Besides, decision trees (DTs) and random forest (RF) were considered for comparison with the best-result explicit correlation. Statistical and graphical assessment criteria were applied for evaluating the newly proposed correlations and compared their performances with prior paradigms. Lastly, Leverage approach was performed to verify the quality of the employed experimental data points and define the realm of application for the best fit established correlation.

The rest of the paper includes 4 sections. Section 2 describes the database which was utilized for developing the correlations. Section 3 briefs the two applied white-box machine learning techniques, namely GMDH and GEP, and the procedure of their implementation in our study. Results are given and discussed in Section 4. The paper ends with Section 5 which summarizes the main findings.

## 2. Data collection and preparation

In the present work, a representative experimental database was collected from the published literature (Cadogan et al., 2015; Cadogan et al., 2014b; Choudhari and Doraiswamy, 1972; Frank, Marco J W and Swaaij, 1996; Lu et al., 2013; Maharajh, 1975; Maharajh and Walkley, 1972; Nijsing et al., 1959; Reddy and Doraiswamy, 1967; Tamimi et al., 1994; Tan and Thorpe, 1992; Thomas and Adams, 1965; Versteeg and van Swaal, 1988; Vivian and Peaceman, 1956; Yang et al., 2006) to develop accurate explicit correlations for estimating the diffusivity coefficient of $CO_2$ in brine at different operating conditions. The database englobes 92 experimental data points with different operating conditions, namely pressure, temperature and viscosity of the solvent. In the context of the affecting variables, salinity affects the solubility, interfacial tension and phase equilibria, thus influencing the diffusivity. In addition, salinity of the solvents affects brine's viscosities (Cadogan, 2015; Feng et al., 2019); therefore, the salinity effect on diffusivity of $CO_2$ in brine is emulated by considering the brine's viscosities as an input parameter while establishing the correlations and the paradigms. The data points collected from previous experimental studies were obtained using various techniques and equipments such as Taylor dispersion, a modified version of Ringborm's apparatus, laminar jet apparatus, laminar falling film, laser-induced fluorescence (LIF), 13C pulsed-field gradient NMR, physical absorption experiments in a stirred vessel operated with a horizontal gas-liquid interface, optical capillary cell via time-dependent Raman spectroscopy, wetted sphere apparatus, Taylor−Aris dispersion method and see-through windowed high-pressure cell. Table 2 reports a detailed statistical insight about the collected data points. In addition, to provide an insightful description of the database, Fig. 1 illustrates frequency histograms of the collected dataset and Fig. 2 demonstrates the correlation between diffusivity coefficient and the considered independent variables through cross plots. According to the histograms shown in Fig. 1, it can be seen that the pressure and temperature are mainly distributed in the medium and low ranges, while viscosity shows a symmetric distribution near its mean. According to Fig. 2 (a), pressure exhibits moderate direct relation with diffusivity coefficient. This analogy is more significant for temperature as can be seen from Fig. 2 (b). From Fig. 2 (c), it can be noted that viscosity of solvent has a negative direct relation with diffusivity coefficient.

# 3. Methodology

## 3.1. Group Method of Data Handling (GMDH)

Group Method of Data Handling (GMDH) is one of the artificial neural network (ANN) types, which is known to generate explicit correlation between input and output parameters of a given system. The resulting correlation by applying a GMDH model takes the form of a polynomial (Dargahi-Zarandi et al., 2017). As an ANN type, GMDH structure involves nodes as basic elements for processing the information. These nodes are arranged in different layers from the input layer to the output layer, with or without intermediate layers (Nait Amar et al., 2019c). The GMDH hybrid version (HGMDH) allowed the improvement of the predictability which was somehow insufficient in the first version developed by (Ivakhnenko, 1971). In HGMDH, the interactions among nodes from different layers are allowed. This procedure brings more robustness when modeling complex cases (Rostami et al., 2019). The mathematical form of HGMDH is expressed as shown below:

$$y_i = a + \sum_{i=1}^{d} \sum_{j=1}^{d} \dots \sum_{k=1}^{d} \vartheta_{ij\dots k} x_i^m x_j^m \dots x_k^m \quad m = 1,2,\dots,2^p \tag{1}$$

where $x$ and $y$ stand for the input and the output parameters, respectively; $\vartheta_{ij\dots k}$ correspond to the polynomial coefficients; $p$ is the number of layers and $d$ is the number of variables.

The following points summarize the calculation procedure in HGMDH with a second order:

- The following equation defines the expression of a node $N_i$ covering two inputs:

$$N_i^{GMDH} = \alpha_0 + \alpha_1 x_i + \alpha_2 x_j + \alpha_3 x_i x_j + \alpha_4 x_i^2 + \alpha_5 x_j^2 \tag{2}$$

- Calculation of polynomial coefficients: least square method is applied to calculate the resulting coefficients in the expressions of the different nodes. The following formula is adapted:

$$\Delta_j^2 = \sum_{i=1}^{N} \left(N_i^{GMDH} - y_i\right)^2 \qquad j = 1,2,\dots,\binom{d}{2} \tag{3}$$

where $d$ and $N$ are the number of variables and data points, respectively.

- Matrix transformation: in order to achieve the final expression, the above equation is transformed to a matrix form (Dargahi-Zarandi et al., 2017; Hemmati-Sarapardeh and Mohagheghian, 2017):

$$Y = A^T X \tag{4}$$

- The final solution is obtained as follows:

$$A^T = yX^T (XX^T)^{-1} \tag{5}$$

where $y = \{y_1, y_2, \dots, y_d\}$ $and$ $A = \{\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5\}$.

## 3.2. Gene Expression Programming (GEP)

Gene expression programming (GEP) is a prevalent evolutionary technique which aims at modeling the systems with accurate explicit expressions. GEP, introduced by Ferreira (Ferreira, 2001), can be regarded as the improved version of genetic programming (GP) which was proposed by Koza (Koza, 1992). GEP employs the standard genetic operators, namely selection, crossover, elitism and mutation, in addition to new implemented actions such as insertion and transposition to search for the reliable correlations. Besides, the codification of the individuals in GEP is performed in the form of chromosomes. In addition, Expression Tree (ET) is introduced for transforming the individuals to real expressions. It is worth mentioning that the genes have a fixed length with terminals which show the variables, such as $\{x_1, x_2, x_3\}$, and operators such as $\{+, /, \times, -, \sqrt{.}, ln\}$ (Teodorescu and Sherwood, 2008).

The main steps of GEP are briefed below:

- Initial population: an initial population of correlations codified in the form of chromosomes is generated randomly. The prediction quality of the correlations differs according to a fitness function.
- Standard operators: the well-known genetic operator including elitism, selection, crossover and mutation are applied. Elitism consists of safeguarding the best correlations for the next generations. Selection allows identification of the correlations to be subjected to reproduction for giving new correlations. Crossover is summarized in the process of exchanging parts between two or more correlations, while mutation is done by means of modifying parts of correlations.
- Transposition and insertion: it consists of jumping and activating parts of the genome in the chromosome (Ferreira, 2001).

The resulting population is assessed and the operators are reiterated until satisfying a stopping condition.

### 3.3. Implementation procedure

As previously highlighted, the aim behind applying GMDH and GEP as white-box machine learning method is to develop an explicit correlation for accurate prediction of $CO_2$ diffusivity in brine under various conditions including pressure, temperature and viscosity. Therefore, the following form is admitted for the two correlations:

$$D_c = f(P, T, \mu) \tag{6}$$

In the above equation, $D_c$ points out the $CO_2$ diffusivity coefficient in brine expressed in $m^2/s$, and $P, T$ and $\mu$ represent the input parameters of the correlations, viz. pressure, temperature and viscosity, respectively. The input parameters are expressed in MPa, K and mPa.s, respectively.

The collected experimental data points that describe these conditions and the obtained diffusivity coefficient of $CO_2$ in brine were prepared for the development of these correlations. The database was divided randomly into a training set with 80% of the whole experimental data points and a testing set which covers the remaining 20%. This dataset partitioning exhibits usually very satisfactory results (Aminu et al., 2019; Benamara et al., 2019; Dargahi-Zarandi et al., 2017; Hemmati-Sarapardeh et al., 2018; Mirjalili, 2015; Yan et al., 2006). Besides, in order to substantiate the better performance and robustness of applied techniques, sensitive analysis of the latter on database were performed.

During the development of the GEP correlation, mean square error (MSE) was the considered fitness function for assessing the chromosomes. MSE is expressed as:

$$MSE = \frac{\sum_{i=1}^{n}(t_i - o_i)^2}{n} \tag{7}$$

where $t_i$ and $o_i$ stand for the measured and the predicted diffusivity coefficient of the $CO_2$ in brine, respectively, and $n$ represents the number of data points.

While developing the GMDH-based correlation, the number of inputs that can be introduced in the hidden and output nodes was specified to three, while the best highest order of the model was investigated by performing a sensitivity analysis.

As indicated, the control parameters of GEP affect the prediction capability. During the establishment of the GEP-based correlation for the prediction of the diffusivity coefficient of $CO_2$ in brine, it was noticed that an increase in the number of chromosomes in the population, the numbers of genes as well as the maximum depth of ET, influence the run-time, the accuracy and the complexity of the generated correlations. Accordingly, these control parameters were tuned. Table 3 reports the final setting of GEP. Consequently, we applied tree encoding, 100 chromosomes, 12 genes, MSE as fitness functions, a function set including $+, -, \times, /$ , exp.$, \sqrt{\ }, INV, ln$, and two point mutation, while the stopping criterion was the maximum number of generations (420).

# 4. Results and discussion

## 4.1. Development and evaluation of the correlations

Several statistical indexes including average absolute relative deviation (AARD), coefficient of determination ($R^2$) and root mean square error (RMSE), were considered for assessing the quality of the predictions of the newly proposed correlations. These statistical criteria are defined as follows:

$$AARD\% = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{t_i - o_i}{t_i}\right| \times 100 \qquad (8)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(t_i - o_i)^2}{\sum_{i=1}^{n}(o_i - \bar{t})^2} \qquad (9)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(t_i - o_i)^2} \qquad (10)$$

The graphical representations of the outcomes of the correlations were illustrated for a visual evaluation and a comparison of the performances. The graphical assessment of the performances was done by means of cross plots, relative error distribution, cumulative frequency distribution of the absolute relative error and bar plots. Cross plots give an insight about the reliability of the correlations by representing their predictions against the real values of the output, and then comparing the obtained distribution against the line $Y = X$ which shows the perfect paradigm. The diagram of relative error distribution details the repartition of the relative error through the output values. Satisfactory distribution of the relative error around the zero-error line indicates the reliability of the correlations. The cumulative frequency diagram of the absolute relative error allows the identification of portions of the data points which are predicted with some pre-specified values of the absolute relative error. Predicting a great portion of the data points with a low absolute relative error value ensures the high reliability of the correlations. Bar plots summarize the statistical criteria of the correlations in visual comparative schemes.

As stated in the previous section, sensitivity analyses were conducted to find the best highest order for GMDH and to investigate GEP and GMDH robustness according to samples considered for training and testing phases. To this end, ten realizations were run at each different GMDH highest order, namely two, three and four. The performance comparison through final overall AARD values is depicted in the box plot of Fig. 3. In this figure, the box exhibits specific interquartile range. The red horizontal line denotes the median value, while the lower and higher black horizontal lines represent the best and worst overall AARD values. As can be seen from this figure, considering three as the highest GMDH order results in the best performance with lowest overall AARD value. Therefore, the highest order of the model was set to three. Fig. 4 shows a bar plot comparing AARD values of GEP and GMDH during training and testing phases in the considered ten runs (with different training and testing samples). According to this figure, the best performance of GEP was achieved in the fourth run while GMDH showed its best reliability in the third run. Accordingly, these two best-result models were kept for further comparison.

Fig. 5 schematizes the best resulted GMDH model for predicting the diffusivity coefficient of $CO_2$ in brine.

According to this figure, no middle nodes were resulted in the model. In the same context, the following equation defines the explicit expression of the GMDH model:

$$D_c = 10^{-9} \times [C_0 + C_1 \times \mu + C_2 \times T + C_3 \times P + C_4 \times T \times \mu + C_5 \times P \times \mu + C_6 \times P \times T + C_7 \times \mu^2 + C_8 \times T^2 + C_9 \times P^2 + C_{10} \times P \times T \times \mu + C_{11} \times T \times \mu^2 + C_{12} \times T^2 \times \mu + C_{13} \times P \times \mu^2 + C_{14} \times P \times T^2 + C_{15} \times P^2 \times \mu + C_{16} \times P^2 \times T + C_{17} \times \mu^3 + C_{18} \times T^3 + C_{19} \times P^3] \qquad (11)$$

where $C_0 = -207.739284$; $C_1 = -201.432367$; $C_2 = 1.1500875$; $C_3 = 0.678161$; $C_4 = 1.834310$; $C_5 = -1.309668$; $C_6 = -0.002251$; $C_7 = -25.879322$; $C_8 = -0.00201$; $C_9 = -0.011747$; $C_{10} = 0.004118$; $C_{11} = 0.038415$; $C_{12} = -0.003692$; $C_{13} = 0.082021$; $C_{14} = -2.664159 \times 10^{-7}$; $C_{15} = 0.001794$; $C_{16} = 3.978117 \times 10^{-5}$; $C_{17} = 3.600267$; $C_{18} = 1.477156 \times 10^{-6}$; and $C_{19} = -2.412235 \times 10^{-5}$

After rearrangement, the best resulting correlation based on GEP is expressed as follows:

$$D_c = 10^{-9} \times \left[\frac{A_1}{\mu} + \frac{A_2}{\sqrt{\mu}} + A_3 \times \sqrt{T} + A_4 \times P + A_5\right] \qquad (12)$$

The terms $A_1$, $A_2$, $A_3$, $A_4$, and $A_5$ are defined as shown below:

$$A_1 = -0.0001564 \times P^3 + 0.01113 \times P^2 + 0.02935 \times P - 2.83 \times \sqrt{P} + 8.362$$

$$A_2 = 0.02426 \times (P + T) - 2.466 \times \sqrt{P}$$

$$A_3 = 0.4583 + 0.123 \times P$$

$$A_4 = 6.832 \times 10^{-5} \times P^2 + 0.003955 \times P + 19.34 \times \frac{1}{\sqrt{P}} - 3.801$$

$$A_5 = \frac{30.95 \times \ln(P)}{5.629 \times (P - \mu)} - \frac{0.0006024 \times T \times \sqrt{\mu}}{P} - 4.259 \times \ln(P^2 \times T) - 7.945$$

Fig. 6 illustrates the cross plots of the proposed correlations. The two cross plots demonstrate a promising consistency between predictions of the correlations and the real data, as very satisfactory alignments nearby the unit slope line are noticed for both GMDH and GEP predictions. The fit is surprisingly trustworthy for the two correlations for both training and testing data.

Furthermore, Fig. 7 shows the distribution of the percentage relative error between the real values of the diffusivity coefficient of the $CO_2$ in brine and the values predicted by GMDH and GEP correlations. According to this figure, the distributions of the relative error around the zero-error line for both correlations at the training and testing phases are deemed satisfactory. However, the reported results in this figure reveal that GEP correlation seems to have a better predictive capability compared to GMDH correlation. For a detailed quantification of the performances, Table 4 states the statistical criteria, namely AARD, $R^2$ and RMSE, for the proposed correlations. Moreover, a graphical comparison between the global performances of the two correlations is reported in Fig. 8. It is clear from the evaluation reported in Table 4 that the newly proposed correlations exhibit promising predictive capabilities with overall values of AARD of 8.0404% and 4.3015% for GMDH and GEP, respectively. Besides, the overall determination coefficient of the models indicates the trustworthiness of the correlations fit. The analyses shown in Figs. 6–8 and Table 4 claim the superiority of the GEP-based correlation in the prediction of the diffusivity coefficient of $CO_2$ in brine by considering the whole employed data points. Therefore, GEP correlation is used for the rest of this paper.

The impact of the employed independent variables, namely pressure, temperature and viscosity on the group error of the GEP-based correlation for prediction of the diffusivity coefficient is investigated in Fig. 9 for the whole database. In Fig. 9(a), the GEP correlation shows its worst predictions with an AARD of 5.31% where pressure is in the range of 20–40 MPa, while other intervals of pressure are predicted with an AARD value of less than 4.6%. In Fig. 9(b), the maximum AARD value of the GEP correlation is obtained when temperature is in the range of 298-323 K, while the rest of the intervals are estimated with an AARD value that does not exceed 4.6%.  In Fig. 9(c), the values of the diffusivity coefficient for viscosity of less than 0.3 mPa.s are predicted with an AARD value of about 1%, while the AARD values of other viscosities are between 4.5 and 5.5%. Consequently, the performance of the proposed GEP is deemed again very sufficient in predicting the diffusivity coefficient.

### 4.2. Comparison of the developed GEP correlation with decision trees (DTs), random forest (RF) and the prior models

It would be of interest from the reliability perspective to compare the results of the proposed GEP correlation with other soft computing techniques, namely decision trees (DTs) and random forest (RF) as well as the prior paradigms reported for the prediction of the diffusivity coefficient of $CO_2$ in brine. To keep the work concise, more details about DTs and RF can be found in published literature (Breiman, 2017; Guo et al., 2011; Peters et al., 2007; Wilkinson, 2004). The final tuned control parameters of RF and DTs models are reported as follows:

- RF: number of grown trees: 20; min leaf size: 5; min parent size: 2 × min leaf size; predictor selection: interaction-curvature; splitting criterion: MSE.
- DTs: min leaf size: 1; min parent size: 5; quadratic error tolerance: 1E-6; predictor selection: all-splits; prune criterion: MSE.

While developing DTs and RF models, 80% of the database was used for their training and 20% of testing. The performance evaluation of the best-result DTs and RF paradigms after various runs is reported in Table 5. By comparing the stated results in Tables 4 and 5, it can be deduced that GEP based correlation outperforms both DTs and RF models.

The comparison includes the empirical models, namely those of Othmer and Thakar (1953), Wilke and Chang (1955) and Cadogan et al. (2014a). In addition to the empirical models, the implemented GEP correlation was compared with one of the most recent intelligent paradigms proposed by Feng et al. (2019) based on hybrid genetic algorithm and mixed Kernels-based support vector machine. It is worth mentioning that when performing the comparison with the pre-existing approaches, we included only the points that satisfy the applicability conditions in each correlation. Table 6 and Fig. 10 report the comparison of the proposed GEP correlation and the models based on previously described statistical criteria. According to these statistical analyses, among the existing models, the hybrid model proposed by Feng et al. (2019) outperforms the available empirical models for predicting the diffusivity coefficient with a global predictive AARD of 7.91%. Despite the exhibited accuracy of Feng et al. (2019) model, it is worth mentioning that this model (based on mixed kernel SVR coupled with GA) is resulted through performing some calculability efforts such as the included quadratic programming involved in the establishment of the final solution. Besides, as this paradigm is of the black box type, it is difficult to apply it to other related tasks.

The reported statistical quality measures in Table 6 and Fig. 10 demonstrate the superiority of the newly proposed GEP correlation, as it outperforms both the prior intelligent model and the empirical paradigms.

Being explicit based approaches, the performances of the empirical models and the white-box GEP correlation were compared through the plot of the absolute relative error distribution as shown in Fig. 11. As seen in this figure, 90% of the data points were predicted by GEP correlation with an AARD value of less than 8.5%. The equivalent percentage of the datapoints predicted with this AARD cutoff value by the empirical models are 50%, 25% and 24% for Cadogan et al. (2014a), Othmer and Thakar (1953), and Wilke and Chang (1955), respectively.

As demonstrated in these comparative analyses, the newly implemented GEP correlation was able to predict more reasonable values of the diffusivity coefficient of $CO_2$ in brine. In addition, the GEP-based correlation has an explicit and simple form which can predict the diffusivity coefficient of $CO_2$ in brine more directly than the other intelligent schemes, and hence, it can be applied to other related tasks or implemented in different softwares. The improvement brought by GEP based correlation in the prediction of the coefficient of $CO_2$ diffusivity in brine can be explained by the followed learning strategy during GEP steps which is based on the use of chromosomes, genes, functions, variables, and the traditional and the new genetic operators, which result in more flexibility for capturing the complexity of the modeled phenomenon.

### 4.3. Trend Analysis

To assess the efficiency of the implemented GEP correlation for accurate prediction of diffusivity coefficient as a function of the input parameters, three different sets of experimental measurements included in our study and the values predicted by GEP correlation are depicted as function of the input parameters in Fig. 12. In Fig. 12(a), the comparison is performed with respect to viscosity variation where pressure and temperature are constant. In Fig. 12(b), the comparison is illustrated for different pressures and constant temperature and viscosity. In Fig. 12(c), the comparison is shown for the case where pressure is constant, and temperature and viscosity vary. Furthermore, additional comparison is depicted in Fig. 13 by presenting the real measurements and the predictions of GEP as function of pressure for the whole considered database. As shown in the plots of Figs. 12 and 13, the predicted and real diffusivity coefficient values of $CO_2$ in brine overlap properly regardless of the operating conditions.

### 4.4. Relative importance of input parameters

A sensitivity analysis using the relevancy factor $(r)$ (Chen et al., 2014; Hajirezaie et al., 2015; Shateri et al., 2015), was performed to assess the relative importance of the input variables on the diffusivity coefficient. The relevancy factor is expressed as follows:

$$r(I_j, O) = \frac{\sum_{i=1}^{n}(I_{j,i}-\bar{I}_j)(o_i-\bar{o})}{\sqrt{\sum_{i=1}^{n}(I_{j,i}-\bar{I}_j)^2\sum_{i=1}^{n}(o_i-\bar{o})^2}} \qquad (13)$$

where the subscripts i and j refer to the data index and the variable, respectively; $I$ and $\bar{I}$ represent the input parameter and its average, respectively, while $O$ and $\bar{O}$ refer to the predicted output and its average, respectively. It is worth noting that a high absolute $(r)$ value for an input parameter indicates its noteworthy impact on the output. Furthermore, achieving positive/negative $r$ values for an input suggests a positive/negative effect on the output.

The obtained results regarding the relevancy factor for the diffusivity coefficient of $CO_2$ in brine are exhibited in Fig. 14. According to this figure, temperature has the biggest impact on the outputs. In addition, it can be deduced that viscosity has a negative effect on the output, while pressure and temperature positively affect the diffusivity coefficient.

### 4.5. Outliers detection

In the last part of this study, outliers detection was conducted to assess the quality of the employed experimental data points employed for the establishment of the GEP correlation, and also to define the applicability domain. The well-known Leverage approach was applied (Rousseeuw and Leroy, 2005). The results from the Leverage approach are converted to the famous graphical representation known as William plot (Nait Amar et al., 2019a, 2019b). This plot scatters the standardized residual (R) of the predicted values versus the so-called hat (H) values which corresponds to the diagonal elements of the hat matrix defined as (Gramatica, 2007; Rousseeuw and Leroy, 2005):

$$H = X(X^t X)^{-1} X^t \qquad (14)$$

where $X$ is a matrix with $(n \times d)$ size, with $n$ and $d$ represents the number of samples and the variables, respectively, and $X^t$ is the transpose matrix of $X$. To delineate the applicability in the Williams plot after presenting standardized residual as function of hat values, a Leverage limit value (H*) calculated as $\frac{3(d+1)}{n}$ is utilized. In addition, the data points are selected in the range of ± 3 of standard deviation from the mean, where the cut-off value of 3 covers 99% of the distributed data (Gramatica, 2007; Rousseeuw and Leroy, 2005). The suspected data points known as outliers are defined as the points which are situated in the range of R > 3 or R < − 3 regardless of their hat value in comparison with H*. Hence, existence of great accumulation of the data points in the ranges $0 \leq H \leq H^*$ and $-3 \leq R \leq 3$ indicates the high reliability of the model.

Fig. 15 shows the obtained Williams plot for the newly proposed correlation. This plot reveals that 90 data points are in the intervals of $0 \leq H \leq 0.1304$ and $-3 \leq R \leq 3$, while only two data points are found outside these margins, and hence, they are detected as doubtful data. The Leverage approach confirms the statistical validity of the implemented GEP correlation for predicting the diffusivity coefficient of $CO_2$ in brine.

## 5. Conclusions

In this paper, two new correlations were developed using GMDH and GEP for accurate prediction of the diffusivity coefficient of $CO_2$ in brine. For developing the correlations, a representative experimental database was collected from the published literature, based on pressure, temperature and the viscosity of the solvent, as inputs. According to this study, the following conclusions are drawn:

1. Both GMDH and GEP correlations showed very close prediction capabilities.
2. GEP correlation outperforms the GMDH correlation with an overall AARD value of 4.3014%.
3. The newly implemented GEP correlation exhibited very low AARD values with respect to different intervals of input parameters.
4. The proposed GEP correlation can provide a fast and reasonably-priced estimation of the coefficient of $CO_2$ diffusivity in brine.
5. The developed GEP correlation was compared with DTs, RF, mixed Kernels-based support vector machine coupled with GA and other pre-existing models. The accuracy of the developed correlation was superior to all these models.
6. The trends of the GEP outputs are logical in terms of the independent variables.
7. Temperature was found the most impacting parameter in the prediction of diffusivity coefficient by GEP correlation.
8. The Leverage approach demonstrated the statistical validity of the model and only two data points were detected as outliers.

# References

Amini, S., Mohaghegh, S.D., Gaskari, R., Bromhal, G., others, 2012. Uncertainty analysis of a co2 sequestration project using surrogate reservoir modeling technique, in: SPE Western Regional Meeting.

Aminu, K.T., McGlinchey, D., Chen, Y., 2019. Optimal design for real-time quantitative monitoring of sand in gas flowline using computational intelligence assisted design framework. J. Pet. Sci. Eng. 177, 1059–1071.

Azzolina, N.A., Nakles, D. V, Gorecki, C.D., Peck, W.D., Ayash, S.C., Melzer, L.S., Chatterjee, S., 2015. CO2 storage associated with CO2 enhanced oil recovery: A statistical analysis of historical operations. Int. J. Greenh. Gas Control 37, 384–397.

Bachu, S., Shaw, J.C., Pearson, R.M., others, 2004. Estimation of oil recovery and CO2 storage capacity in CO2 EOR incorporating the effect of underlying aquifers, in: SPE/DOE Symposium on Improved Oil Recovery.

Benamara, C., Nait Amar, M., Gharbi, K., Hamada, B., 2019. Modeling Wax Disappearance Temperature Using Advanced Intelligent Frameworks. Energy & Fuels. https://doi.org/10.1021/acs.energyfuels.9b03296

Bhakta, J.N., Lahiri, S., Pittman, J.K., Jana, B.B., 2015. Carbon dioxide sequestration in wastewater by a consortium of elevated carbon dioxide-tolerant microalgae. J. CO2 Util. 10, 105–112.

Boot-Handford, M.E., Abanades, J.C., Anthony, E.J., Blunt, M.J., Brandani, S., Mac Dowell, N., Fernández, J.R., Ferrari, M.-C., Gross, R., Hallett, J.P., others, 2014. Carbon capture and storage update. Energy Environ. Sci. 7, 130–189.

Breiman, L., 2017. Classification and regression trees. Routledge.

Cadogan, S., 2015. Diffusion of CO2 in fluids relevant to carbon capture, utilisation and storage. PhD thesis, Imp. Coll. London.

Cadogan, S.P., Hallett, J.P., Maitland, G.C., Trusler, J.P.M., 2015. Diffusion coefficients of carbon dioxide in brines measured using 13C pulsed-field gradient nuclear magnetic resonance. J. Chem. Eng. Data 60, 181–184. https://doi.org/10.1021/je5009203

Cadogan, Shane P, Hallett, J.P., Maitland, G.C., Trusler, J.P.M., 2014a. Diffusion coefficients of carbon dioxide in brines measured using 13C pulsed-field gradient nuclear magnetic resonance. J. Chem. Eng. Data 60, 181–184.

Cadogan, Shane P, Maitland, G.C., Trusler, J.P.M., 2014b. Diffusion coefficients of CO2 and N2 in water at temperatures between 298.15 K and 423.15 K at pressures up to 45 MPa. J. Chem. Eng. Data 59, 519–525.

Chen, G., Fu, K., Liang, Z., Sema, T., Li, C., Tontiwachwuthikul, P., Idem, R., 2014. The genetic algorithm based back propagation neural network for MMP prediction in CO2-EOR process. Fuel 126, 202–212.

Choudhari, R. V., Doraiswamy, L.K., 1972. Physical Properties in Reaction of Ethylene and Hydrogen Chloride in Liquid Media: Diffusivities and Solubilities. J. Chem. Eng. Data 17, 428–432. https://doi.org/10.1021/je60055a012

Dargahi-Zarandi, A., Hemmati-Sarapardeh, A., Hajirezaie, S., Dabir, B., Atashrouz, S., 2017. Modeling gas/vapor viscosity of hydrocarbon fluids using a hybrid GMDH-type neural network system. J. Mol. Liq. 236, 162–171.

Davarazar, M., Jahanianfard, D., Sheikhnejad, Y., Nemati, B., Mostafaie, A., Zandi, S., Khalaj, M., Kamali, M., Aminabhavi, T.M., 2019. Underground carbon dioxide sequestration for climate change mitigation--A scientometric study. J. CO2 Util. 33, 179–188.

Ettehadtavakkol, A., Lake, L.W., Bryant, S.L., 2014. CO2-EOR and storage design optimization. Int. J. Greenh. Gas Control 25, 79–92.

Farajzadeh, R., Zitha, P.L.J., Bruining, J., 2009. Enhanced mass transfer of CO2 into water: experiment and modeling. Ind. Eng. Chem. Res. 48, 6423–6431.

Feng, Q., Cui, R., Wang, S., Zhang, J., Jiang, Z., 2019. Estimation of CO2 diffusivity in brine by use of the genetic algorithm and mixed kernels-based support vector machine model. J. Energy Resour. Technol. 141, 41001.

Ferreira, C., 2001. Algorithm for solving gene expression programming: a new adaptive problems. Complex Syst. 13, 87–129.

Frank, Marco J W, K.J. a M., Swaaij, W.P.M. Van, 1996. Marco J. W. Frank,* Johannes A. M. Kuipers, and Wim P. M. van Swaaij. J. Chem. Eng. Data 41, 297–302. https://doi.org/10.1021/je950157k

Frank, M.J.W., Kuipers, J.A.M., van Swaaij, W.P.M., 1996. Diffusion coefficients and viscosities of CO2+ H2O, CO2+ CH3OH, NH3+ H2O, and NH3+ CH3OH liquid mixtures. J. Chem. Eng. Data 41, 297–302.

Gershenzon, N.I., Ritzi, R.W., Dominic, D.F., Soltanian, M., Mehnert, E., Okwen, R.T., 2015. Influence of small-scale fluvial architecture on CO2 trapping processes in deep brine reservoirs. Water Resour. Res. 51, 8240–8256.

Gershenzon, N.I., Soltanian, M., Ritzi Jr, R.W., Dominic, D.F., 2014. Influence of small scale heterogeneity on CO2 trapping processes in deep saline aquifers. Energy Procedia 59, 166–173.

Gibbins, J., Chalmers, H., 2008. Carbon capture and storage. Energy Policy 36, 4317–4322.

Gozalpour, F., Ren, S.R., Tohidi, B., 2005. CO2 EOR and storage in oil reservoir. Oil gas Sci. Technol. 60, 537–546.

Gramatica, P., 2007. Principles of QSAR models validation: internal and external. QSAR Comb. Sci. 26, 694–701. https://doi.org/10.1002/qsar.200610151

Grude, S., Landrø, M., Dvorkin, J., 2014. Pressure effects caused by CO2 injection in the Tubåen Fm., the Snøhvit field. Int. J. Greenh. Gas Control 27, 178–187.

Guo, L., Chehata, N., Mallet, C., Boukir, S., 2011. Relevance of airborne lidar and multispectral image data for urban scene classification using Random Forests. ISPRS J. Photogramm. Remote Sens. 66, 56–66.

Guzmán, J., Garrido, L., 2012. Determination of carbon dioxide transport coefficients in liquids and polymers by NMR spectroscopy. J. Phys. Chem. B 116, 6050–6058.

Hajirezaie, S., Hemmati-Sarapardeh, A., Mohammadi, A.H., Pournik, M., Kamari, A., 2015. A smooth model for the estimation of gas/vapor viscosity of hydrocarbon fluids. J. Nat. Gas Sci. Eng. 26, 1452–1459.

Hemmati-Sarapardeh, A., Mohagheghian, E., 2017. Modeling interfacial tension and minimum miscibility pressure in paraffin-nitrogen systems: Application to gas injection processes. Fuel 205, 80–89.

Hemmati-Sarapardeh, A., Varamesh, A., Husein, M.M., Karan, K., 2018. On the evaluation of the viscosity of nanofluid systems: Modeling and data assessment. Renew. Sustain. Energy Rev. 81, 313–329.

Holtz, M.H., Nance, P.K., Finley, R.J., 2001. Reduction of greenhouse gas emissions through CO2 EOR in Texas. Environ. Geosci. 8, 187–199.

Ivakhnenko, A.G., 1971. Polynomial theory of complex systems. IEEE Trans. Syst. Man. Cybern. 364–378.

Jang, H.W., Yang, D., Li, H., 2018. A Power-Law Mixing Rule for Predicting Apparent Diffusion Coefficients of Binary Gas Mixtures in Heavy Oil. J. Energy Resour. Technol. 140, 52904.

Jeong, H., Sun, A.Y., Lee, J., Min, B., 2018. A learning-based data-driven forecast approach for predicting future reservoir performance. Adv. Water Resour. 118, 95–109.

Koza, J.R., 1992. Genetic programming II, automatic discovery of reusable subprograms. MIT Press, Cambridge, MA.

Lee, J.W., Hawkins, B., Day, D.M., Reicosky, D.C., 2010. Sustainability: the capacity of smokeless biomass pyrolysis for energy production, global carbon capture and sequestration. Energy Environ. Sci. 3, 1695–1705.

Li, H., Yang, D., others, 2016. Determination of individual diffusion coefficients of solvent/CO 2 mixture in heavy oil with pressure-decay method. SPE J. 21, 131–143.

Lu, W., Guo, H., Chou, I.-M., Burruss, R.C., Li, L., 2013. Determination of diffusion coefficients of carbon dioxide in water between 268 and 473 K in a high-pressure capillary optical cell with in situ Raman spectroscopic measurements. Geochim. Cosmochim. Acta 115, 183–204.

Maharajh, D.M., 1975. Solubility and Diffusion of Gases in Water.

Maharajh, D.M., Walkley, J., 1972. The Temperature Dependence of the DiEusion Coefficients of Ar, COz, CH4, CH3CI, CH3Br, and CHCB2F in Water. Can. J. Chem. 51, 944–952.

Mirjalili, S., 2015. How effective is the Grey Wolf optimizer in training multi-layer perceptrons. Appl. Intell. 43, 150–161.

Mohagheghian, E., Hassanzadeh, H., Chen, Z., 2019. CO2 sequestration coupled with enhanced gas recovery in shale gas reservoirs. J. CO2 Util. 34, 646–655.

Moultos, O.A., Tsimpanogiannis, I.N., Panagiotopoulos, A.Z., Economou, I.G., 2016. Self-diffusion coefficients of the binary (H2O+ CO2) mixture at high temperatures and pressures. J. Chem. Thermodyn. 93, 424–429.

Mutoru, J.W., Leahy-Dios, A., Firoozabadi, A., 2011. Modeling infinite dilution and Fickian diffusion coefficients of carbon dioxide in water. AIChE J. 57, 1617–1627.

Nait Amar, M., Hemmati-Sarapardeh, A., Varamesh, A., Shamshirband, S., 2019a. Predicting solubility of CO2 in brine by advanced machine learning systems: Application to carbon capture and sequestration. J. CO2 Util. 33, 83–95.

Nait Amar, M., Zeraibi, N., 2019. An efficient methodology for multi-objective optimization of water alternating CO2 EOR process. J. Taiwan Inst. Chem. Eng. 99, 154–165.

Nait Amar, M., Zeraibi, N., Hemmati-Sarapardeh, A., Shamshirband, S., 2019b. Modeling temperature-based oil-water relative permeability by integrating advanced intelligent models with grey wolf optimization: Application to thermal enhanced oil recovery processes. Fuel 242, 649–663.

Nait Amar, M., Zeraibi, N., Hemmati-Sarapardeh, A., Shamshirband, S., Mosavi, A., Chau, K., 2019c. Modeling temperature dependency of oil-water relative permeability in thermal enhanced oil recovery processes using group method of data handling and gene expression programming. Eng. Appl. Comput. Fluid Mech. 13, 724–743.

Nijsing, R.A.T.O., Hendriksz, R.H., Kramers, H., 1959. Absorption of CO2 in jets and falling films of electrolyte solutions, with and without chemical reaction. Chem. Eng. Sci. 10, 88–104. https://doi.org/10.1016/0009-2509(59)80028-2

Nomeli, M.A., Riaz, A., 2017. A data driven model for the impact of IFT and density variations on CO2 storage capacity in geologic formations. Adv. Water Resour. 107, 83–92.

Othmer, D.F., Thakar, M.S., 1953. Correlating diffusion coefficient in liquids. Ind. Eng. Chem. 45, 589–593.

Peters, J., De Baets, B., Verhoest, N.E.C., Samson, R., Degroeve, S., De Becker, P., Huybrechts, W., 2007. Random forests as a tool for ecohydrological distribution modelling. Ecol. Modell. 207, 304–318.

Piotrowski, A.P., Napiorkowski, J.J., 2012. Product-Units neural networks for catchment runoff forecasting. Adv. Water Resour. 49, 97–113.

Reddy, K.A., Doraiswamy, L.K., 1967. Estimating liquid diffusivity. Ind. Eng. Chem. Fundam. 6, 77–79. https://doi.org/10.1021/i160021a012

Riahi, K., Rubin, E.S., Taylor, M.R., Schrattenholzer, L., Hounshell, D., 2004. Technological learning for carbon capture and sequestration technologies. Energy Econ. 26, 539–564.

Rostami, A., Hemmati-Sarapardeh, A., Karkevandi-Talkhooncheh, A., Husein, M.M., Shamshirband, S., Rabczuk, T., 2019. Modeling heat capacity of ionic liquids using group method of data handling: A hybrid and structure-based approach. Int. J. Heat Mass Transf. 129, 7–17.

Rousseeuw, P.J., Leroy, A.M., 2005. Robust regression and outlier detection. John wiley & sons.

Shahkarami, A., Mohaghegh, S., Gholami, V., Haghighat, A., Moreno, D., 2014. Modeling pressure and saturation distribution in a CO2 storage project using a Surrogate Reservoir Model (SRM). Greenh. Gases Sci. Technol. 4, 289–315.

Shateri, M., Ghorbani, S., Hemmati-Sarapardeh, A., Mohammadi, A.H., 2015. Application of Wilcoxon generalized radial basis function network for prediction of natural gas compressibility factor. J. Taiwan Inst. Chem. Eng. 50, 131–141.

Tamimi, A., Rinker, E.B., Sandall, O.C., 1994. Diffusion Coefficients for Hydrogen Sulfide, Carbon Dioxide, and Nitrous Oxide in Water over the Temperature Range 293–368 K. J. Chem. Eng. Data 39, 330–332. https://doi.org/10.1021/je00014a031

Tan, K.K., Thorpe, R.B., 1992. Gas diffusion into viscous and non-Newtonian liquids. Chem. Eng. Sci. 47, 3565–3572. https://doi.org/10.1016/0009-2509(92)85071-I

Teodorescu, L., Sherwood, D., 2008. High energy physics event selection with gene expression programming. Comput. Phys. Commun. 178, 409–419.

Thomas, W.J., Adams, M.J., 1965. Measurement of the diffusion coefficients of carbon dioxide and nitrous oxide in water and aqueous solutions of glycerol. Trans. Faraday Soc. 61, 668–673. https://doi.org/10.1039/tf9656100668

Trevisan, L., Pini, R., Cihan, A., Birkholzer, J.T., Zhou, Q., Illangasekare, T.H., 2014. Experimental investigation of supercritical CO2 trapping mechanisms at the intermediate laboratory scale in well-defined heterogeneous porous media. Energy Procedia 63, 5646–5653.

Venkatraman, V., Alsberg, B.K., 2017. Predicting CO2 capture of ionic liquids using machine learning. J. CO2 Util. 21, 162–168.

Versteeg, G.F., van Swaal, W.P.M., 1988. Solubility and Diffusivity of Acid Gases (CO2, N2O) in Aqueous Alkanolamine Solutions. J. Chem. Eng. Data 33, 29–34. https://doi.org/10.1021/je00051a011

Vivian, J.E., Peaceman, D.W., 1956. Liquid□side resistance in gas absorption. AIChE J. 2, 437–443. https://doi.org/10.1002/aic.690020404

Wilke, C.R., Chang, P., 1955. Correlation of diffusion coefficients in dilute solutions. AIChE J. 1, 264–270.

Wilkinson, L., 2004. Classification and regression trees. Systat 11, 35–56.

Yan, Y., Xu, L., Lee, P., 2006. Mass flow measurement of fine particles in a pneumatic suspension using electrostatic sensing and neural network techniques. IEEE Trans. Instrum. Meas. 55, 2330–2334.

Yang, D., Tontiwachwuthikul, P., Gu, Y., 2006. Dynamic interfacial tension method for measuring gas diffusion coefficient and interface mass transfer coefficient in a liquid. Ind. Eng. Chem. Res. 45, 4999–5008. https://doi.org/10.1021/ie060047e

| Solvent | Model | Expression | Included parameters |
|---|---|---|---|

**Table 1.** Summary of the existing empirical models for predicting the diffusivity coefficient of $CO_2$

| | | | |
|---|---|---|---|
| Brine | (Othmer and Thakar, 1953) | $D_{CO_2} = \frac{14 \times 10^{-9}}{\mu^{1.1} V_m^{0.6}}$ | • Molar volume of the diffusing substance ($V_m$ in cm$^3$/gmol).<br>• Viscosity of the solvent ($\mu$ in $mPa \cdot s$) |
| | (Wilke and Chang, 1955) | $D_{CO_2} = 7.4 \times 10^{-8} \frac{T\sqrt{\emptyset M}}{\mu V_m^{0.6}}$ | • Temperature ($T$ in K).<br>• The association parameter $\phi$.<br>• Molecular weight of solvent (M). |
| | (Cadogan et al., 2014a) | $D_{CO_2} = \frac{k_B T}{n_{SE} \pi \mu a}$ | • $k_B$=1.38065×10$^{-23}$ J/K.<br>• $n_{SE}$ is the Stokes-Einstein number.<br>• The hydrodynamic radius of the solute (a in pm): $a = 168[1 + 2.0 \times 10^{-3}(T - 298)]$ |
| Pure water | (Lu et al., 2013) | $D_{CO_2} = 13.942 \times 10^{-9} \left[\frac{T}{227} - 1\right]^{1.7094}$ | • Temperature (T in K). |
| | (Moultos et al., 2016) | $D_{CO_2} = D_0(P) \left[\frac{T}{Ts} - 1\right]^{m(P)}$ | • $D_0$=$a_1$ln($P$)+$a_2$, m=$b_1$ln($P$)+$b_2$, where $a$1=-2.3097×10$^{-9}$, $a$2=2,1064×10$^{-8}$, $b$1=-0.17812 and $b$2=2.59406; $P$ is the pressure. |

**Table 2.** Summary of the gathered data

| | Max | Avg. | Min | SD |
|---|---|---|---|---|
| **P (MPa)** | 49.30 | 9.64 | 0.1000 | 14.8030 |
| **T (K)** | 473.15 | 317.76 | 273 | 39.8 |
| **Viscosity (mPa.s)** | 1.9500 | 0.9003 | 0.1390 | 0.4720 |
| **Diffusivity coefficient (×$10^{-9}$ $m^2/s$)** | 16.1000 | 3.3522 | 0.3100 | 3.0874 |

**Table 3.** GEP setting parameters used in the study

| **Parameters** | **Value/setting** |
|---|---|
| Chromosome | 100 |
| Gene | 12 |
| Operators used | $+, -, \times, /, \text{exp.}, \sqrt{\ }, INV, ln$ |
| Generations | 420 |
| Mutation rate | 0.45 |
| Inversion rate | 0.12 |

**Table 4.** Performance analysis of the implemented models.

| | | **GEP** | **GMDH** |
|---|---|---|---|
| **Training data** | AARD (%) | 3.8584 | 8.6269 |

| | | | |
|---|---|---|---|
| | R$^{2}$ | 0.9980 | 0.9943 |
| | RMSE (×10$^{-9}$ m$^{2}$/s$^{)}$ | 0.1427 | 0.2479 |
| **Test data** | AARD (%) | 6.0035 | 5.6292 |
| | R$^{2}$ | 0.9978 | 0.9874 |
| | RMSE (×10$^{-9}$ m$^{2}$/s$^{)}$ | 0.1245 | 0.2271 |
| **All data** | AARD (%) | 4.3014 | 8.0404 |
| | R$^{2}$ | 0.9979 | 0.9937 |
| | RMSE (×10$^{-9}$ m$^{2}$/s$^{)}$ | 0.1391 | 0.2440 |

**Table 5.** Performance analysis of the implemented decision trees (DTs) and random forest (RF) models.

| | | **DTs** | **RF** |
|---|---|---|---|
| **Training data** | AARD (%) | 4.2969 | 6.3627 |

| | | | |
|---|---|---|---|
| | R$^2$ | 0.9980 | 0.9973 |
| | RMSE (×10$^{-9}$ m$^2$/s$^{)}$ | 0.1598 | 0.1647 |
| **Test data** | AARD (%) | 8.8426 | 9.0015 |
| | R$^2$ | 0.9924 | 0.9940 |
| | RMSE (×10$^{-9}$ m$^2$/s$^{)}$ | 0.2532 | 0.2764 |
| **All data** | AARD (%) | 5.1862 | 6.8790 |
| | R$^2$ | 0.9969 | 0.9966 |
| | RMSE (×10$^{-9}$ m$^2$/s$^{)}$ | 0.1785 | 0.1870 |

**Table 6.** Comparison of the performances with prior models

| | **GEP** | **Feng et al.** | **Othmer** | **Wilke** | **Cadogan** |
|---|---|---|---|---|---|

| | | | and Thakar | and Chang | et al. |
|---|---|---|---|---|---|
| **AARD (%)** | 4.3014 | 7.91 | 12.75 | 12.60 | 13.84 |
| **$R^2$** | 0.9979 | 0.9960 | 0.9661 | 0.9434 | 0.9858 |
| **RMSE** | 0.1391 | 0.1954 | 0.5661 | 0.7311 | 0.3661 |

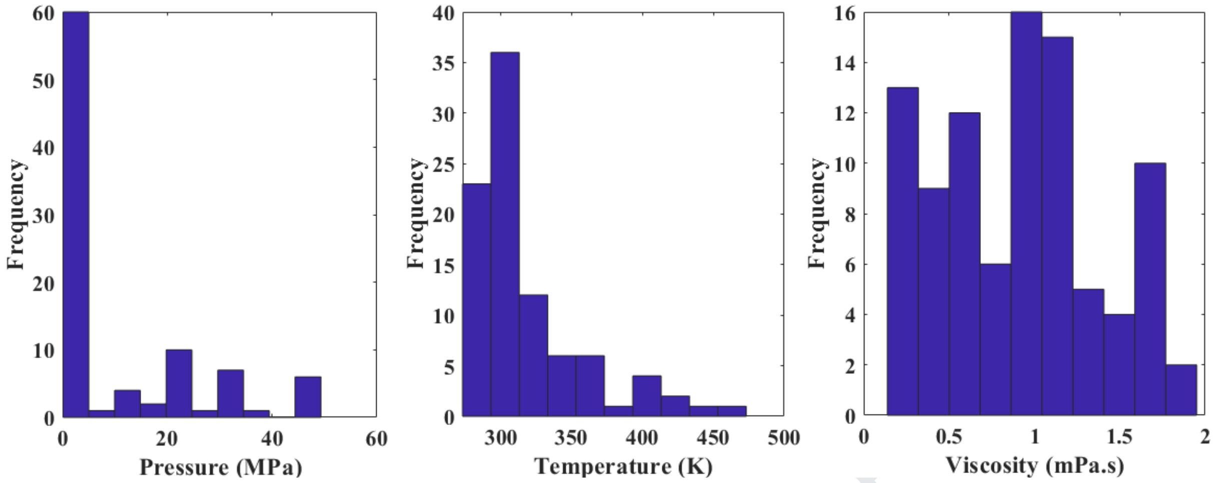

**Fig. 1.** Frequency histograms of the collected dataset

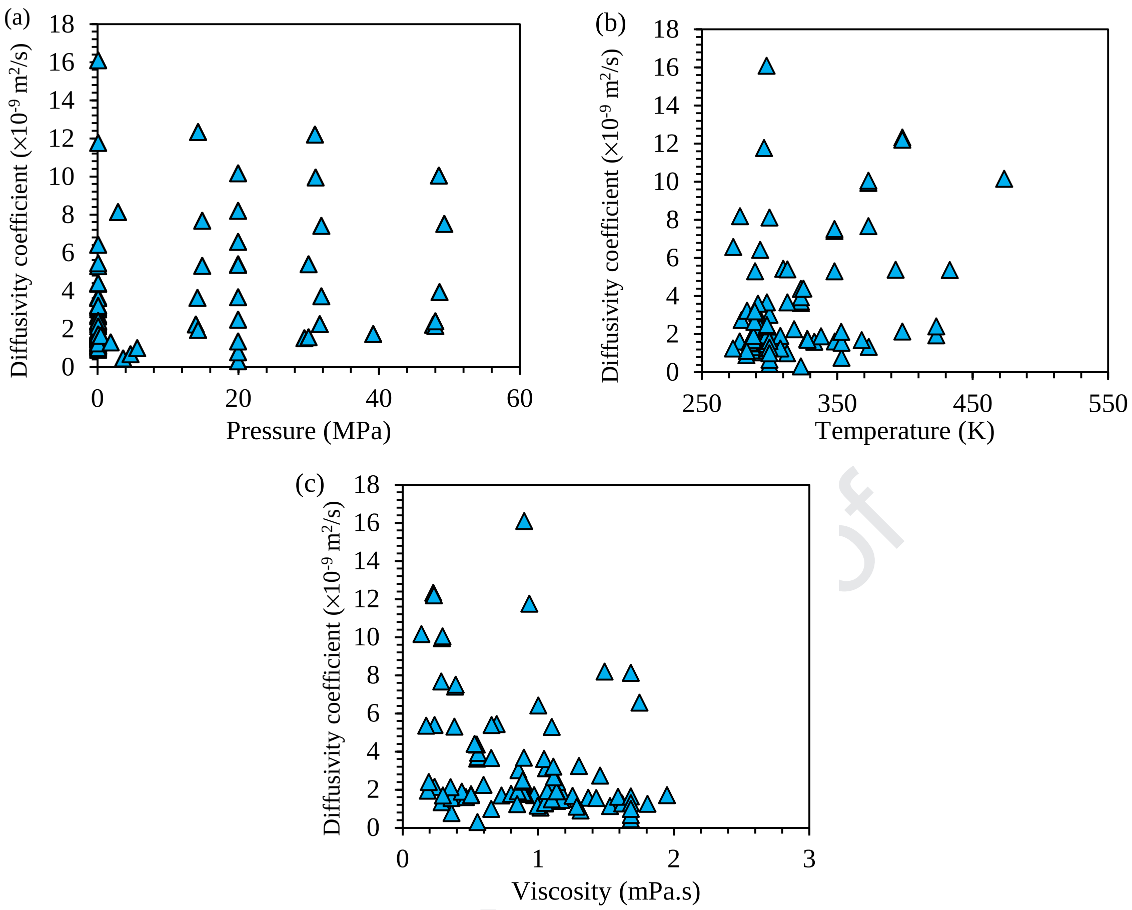

**Fig. 2.** Variation of diffusivity coefficient versus the independent variables

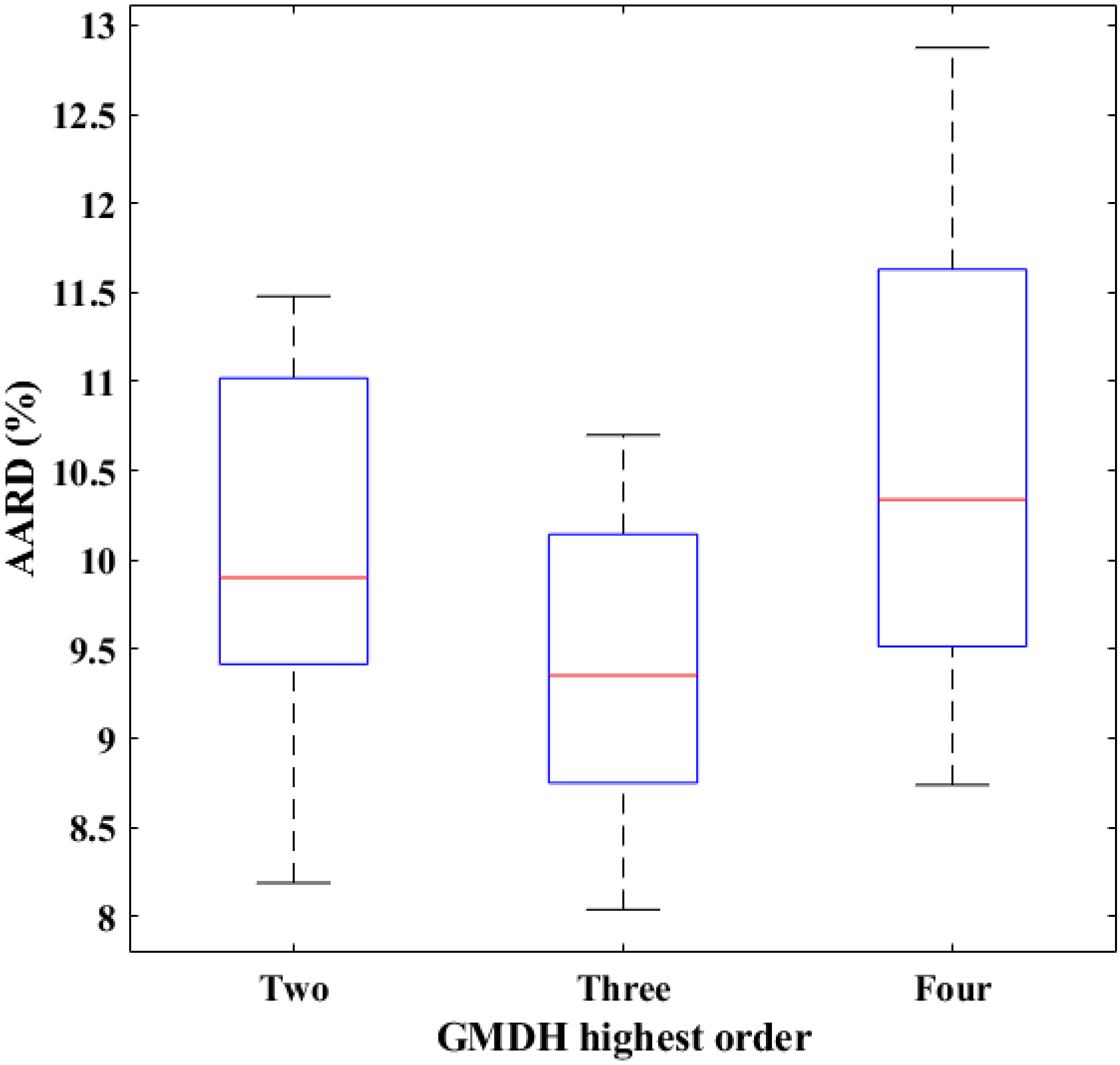

**Fig. 3.** Results obtained for ten realizations with different GMDH highest orders

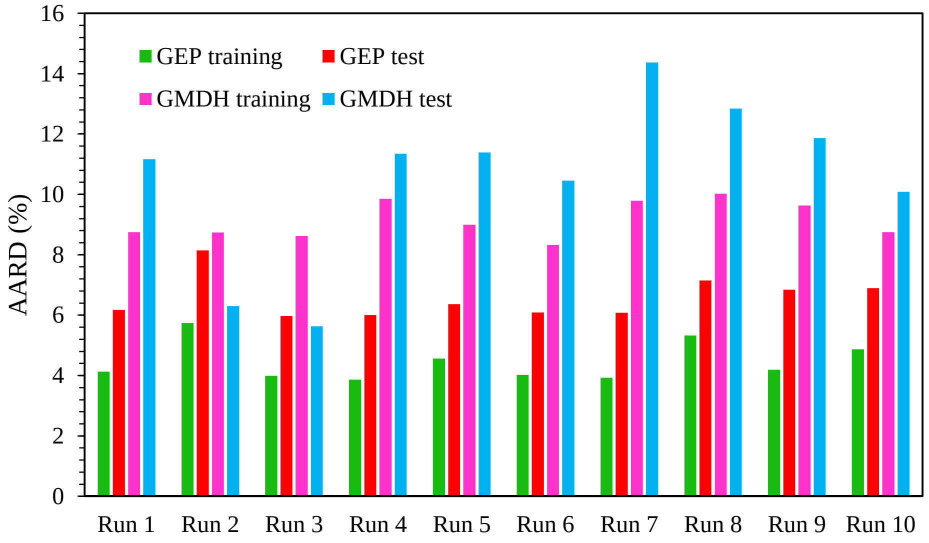

**Fig. 4.** Results of the GEP and GMDH sensitivity analysis on training and test sets

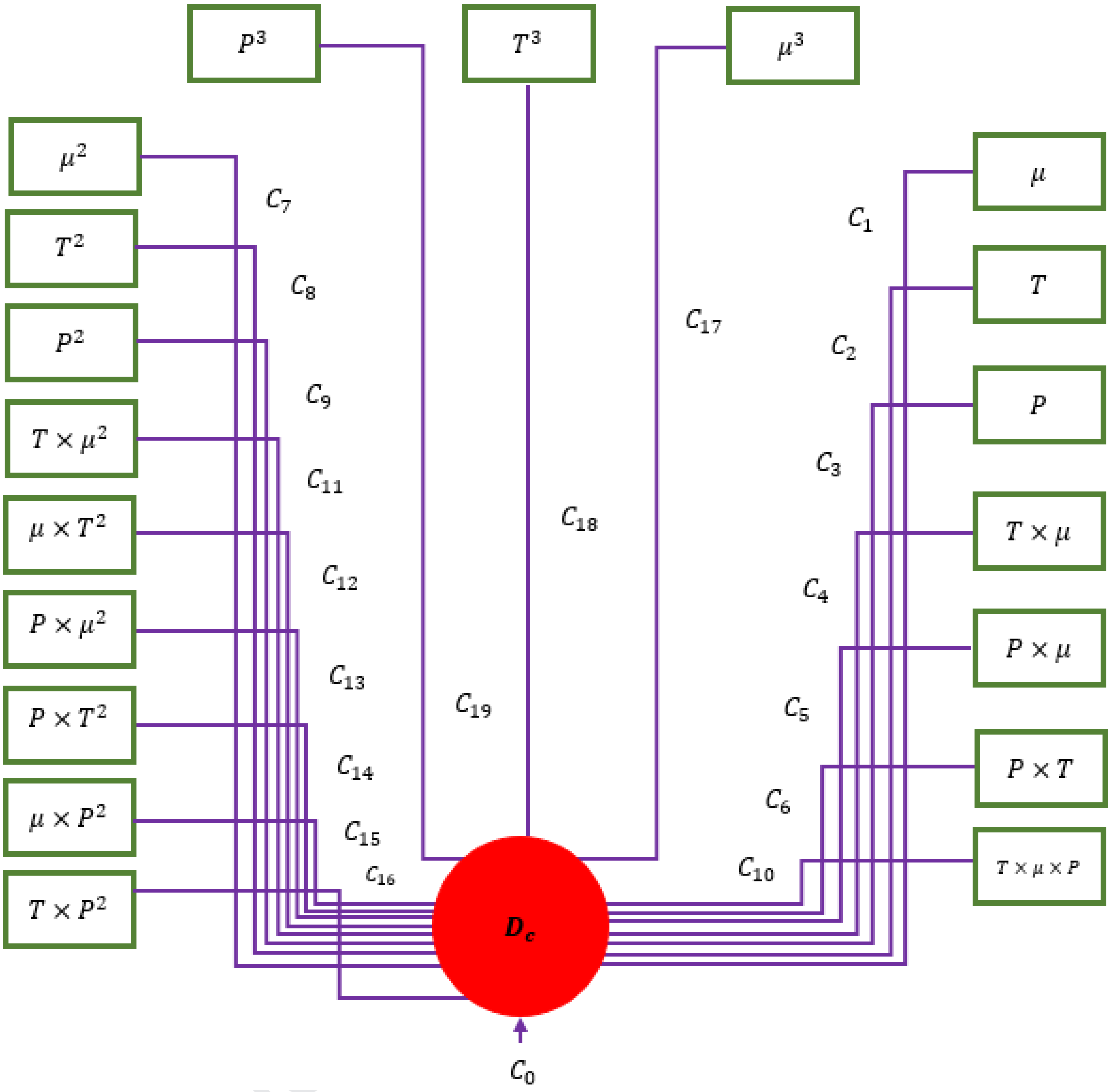

**Fig. 5.** A schematic structure of the implemented GMDH for predicting diffusivity coefficient

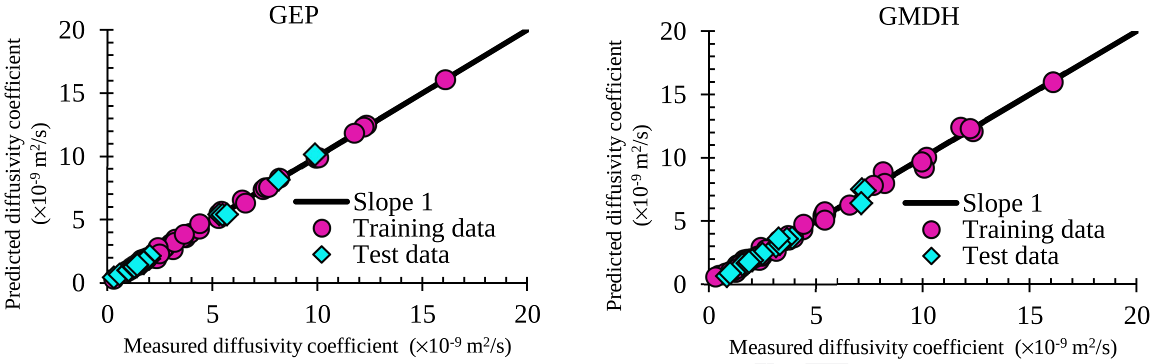

**Fig. 6.** Cross plots of the established GEP and GMDH correlations for diffusivity coefficient prediction

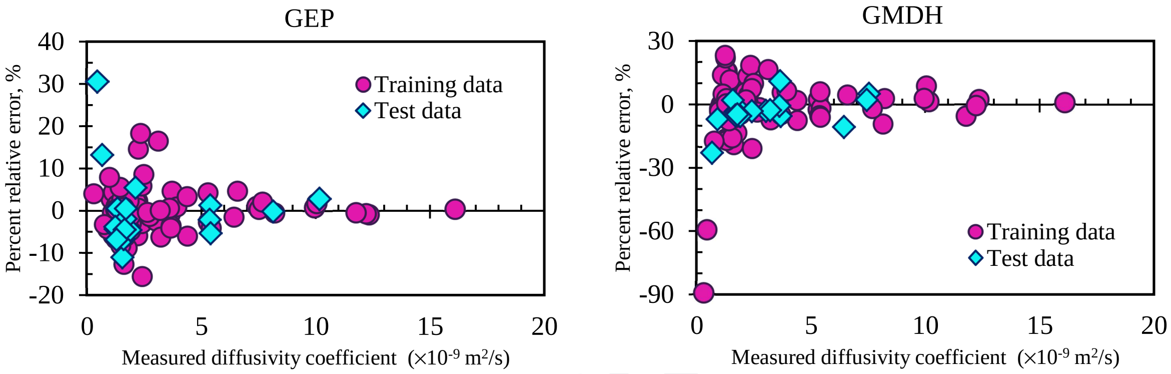

**Fig. 7.** Error distribution for the developed correlations

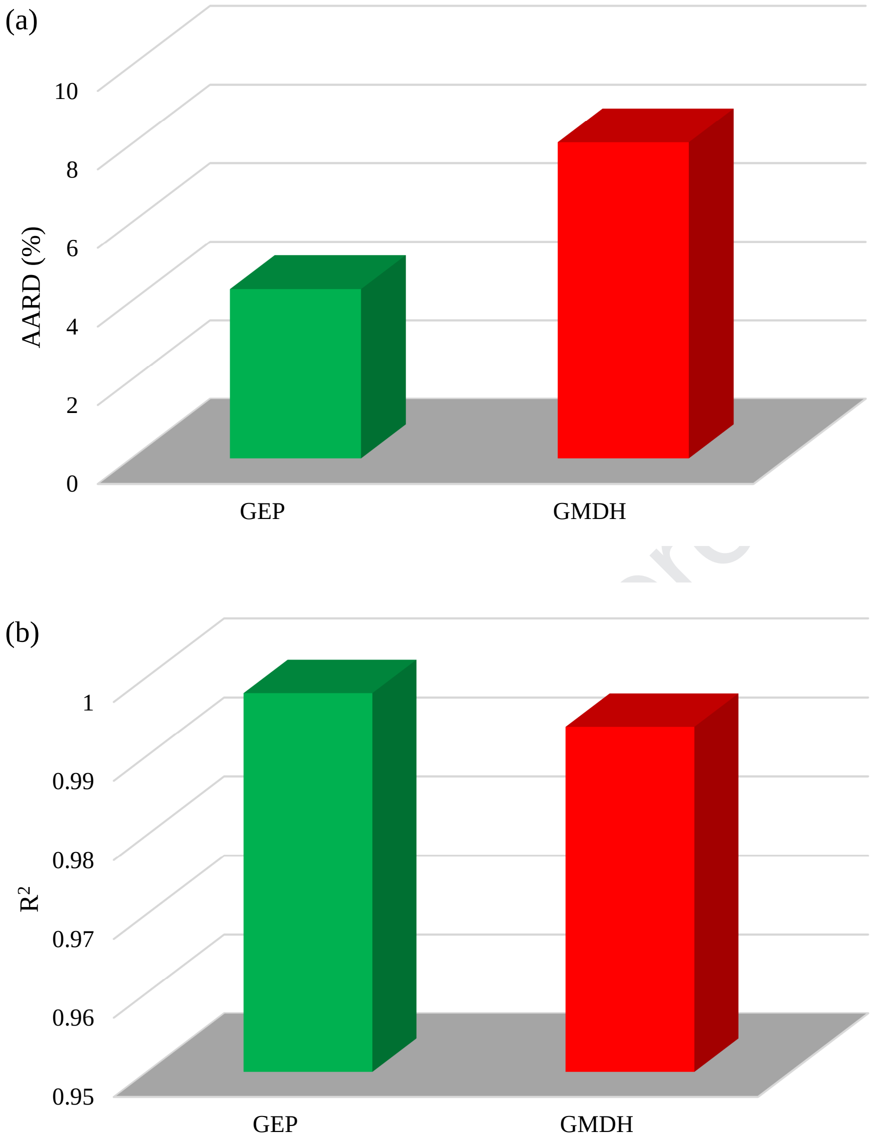

**Fig. 8.** Comparison between the performances of the correlations: (a) AARD and (b) $R^2$

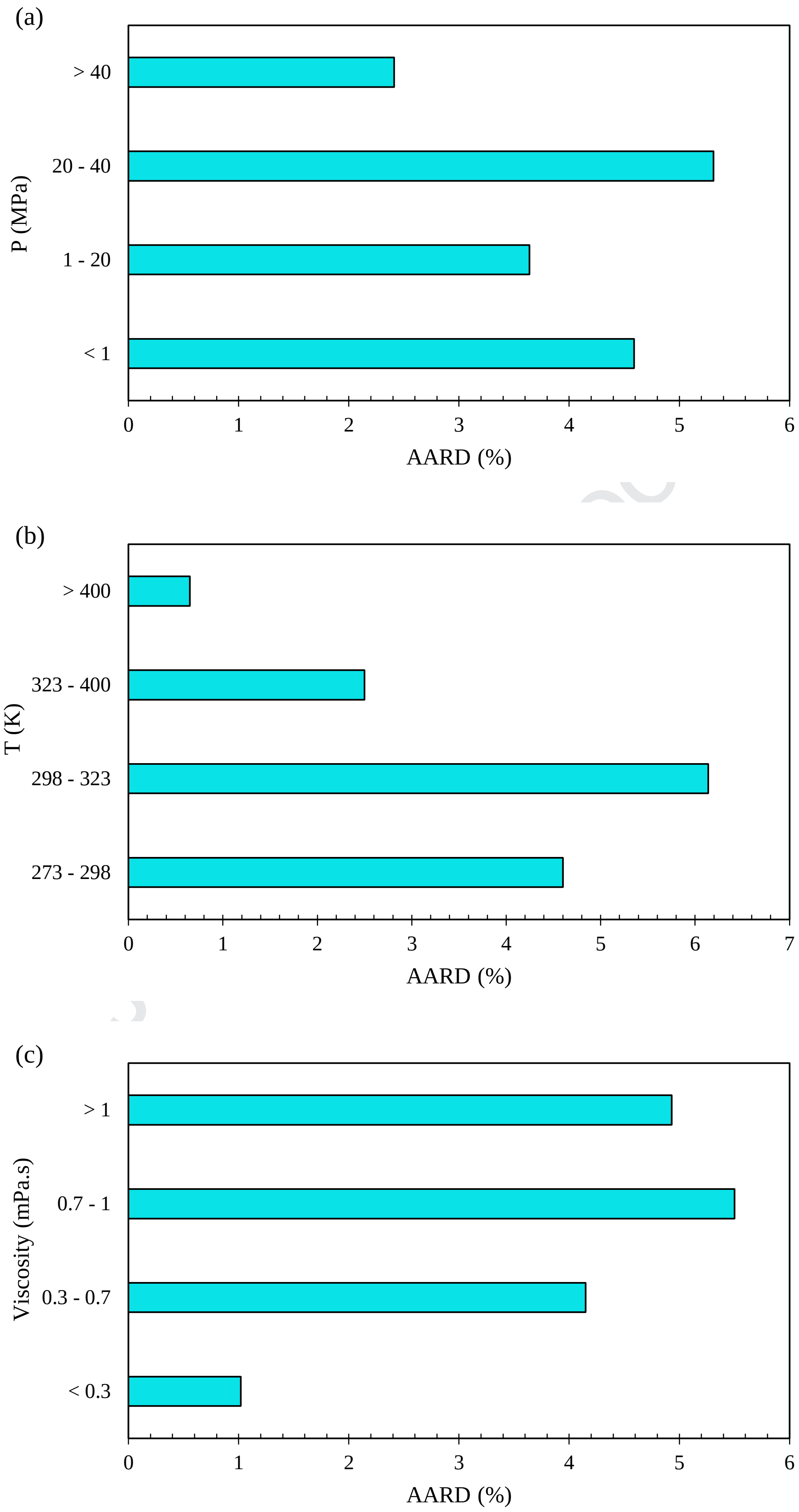

**Fig. 9.** Comparison of the error distribution of the models with respect to input parameters

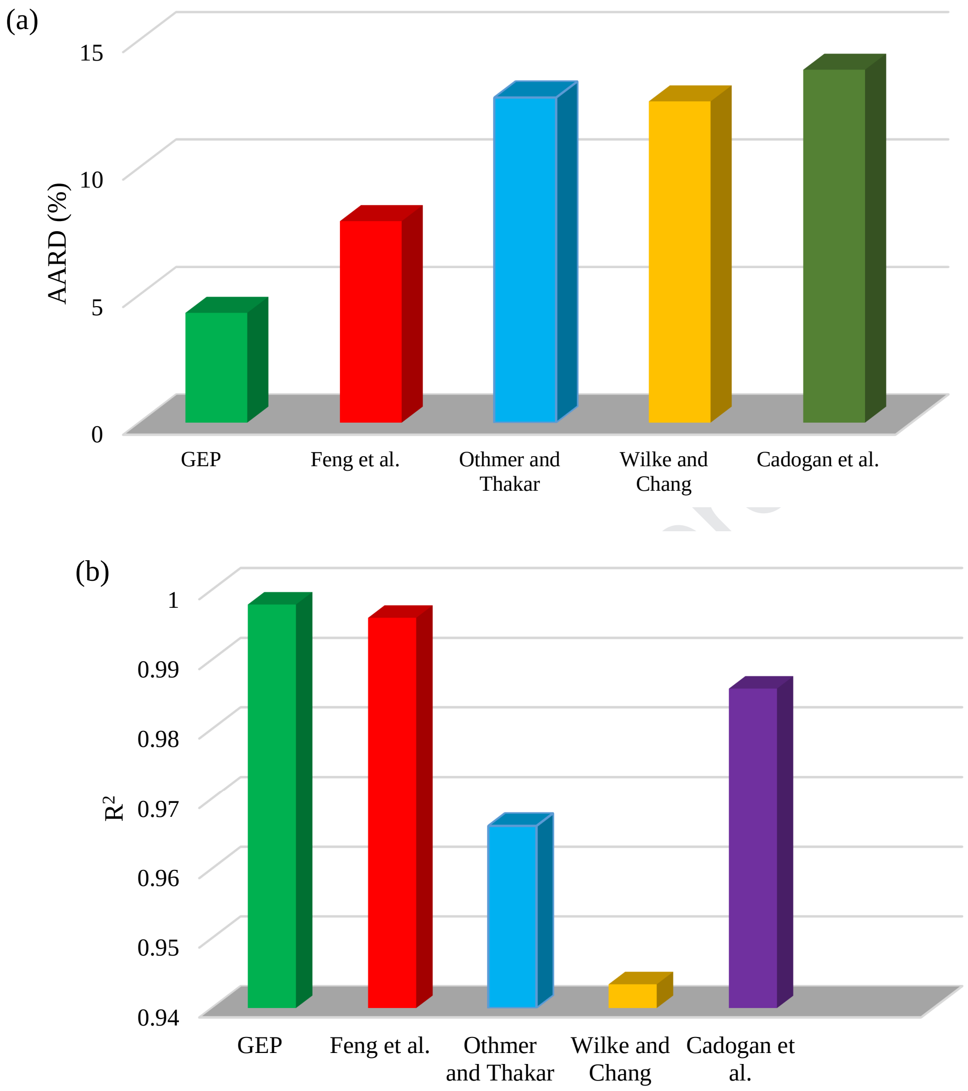

**Fig. 10.** Comparison between the performances of GEP correlation and the prior models: (a) AARD and (b) $R^2$

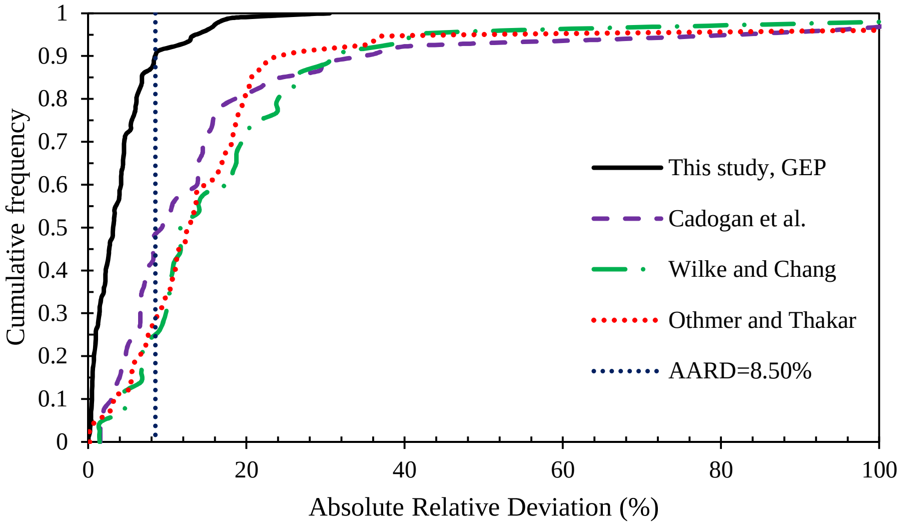

**Fig. 11.** Cumulative frequency vs. absolute percent relative deviation of GEP correlation and the prior empirical models.

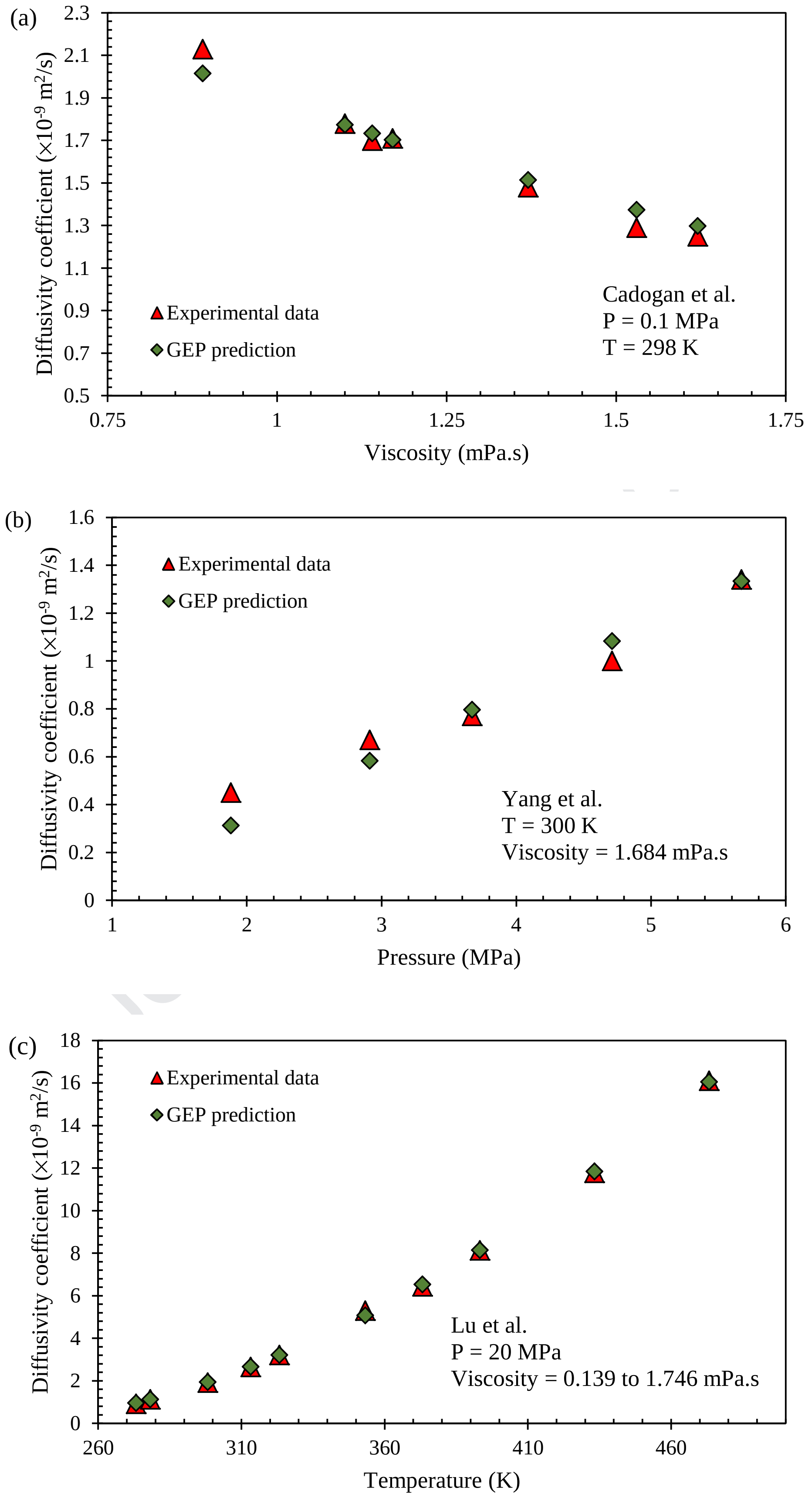

**Fig. 12.** Comparison of the diffusivity coefficient obtained from measurements and generated by GEP correlation, as function of the input parameters.

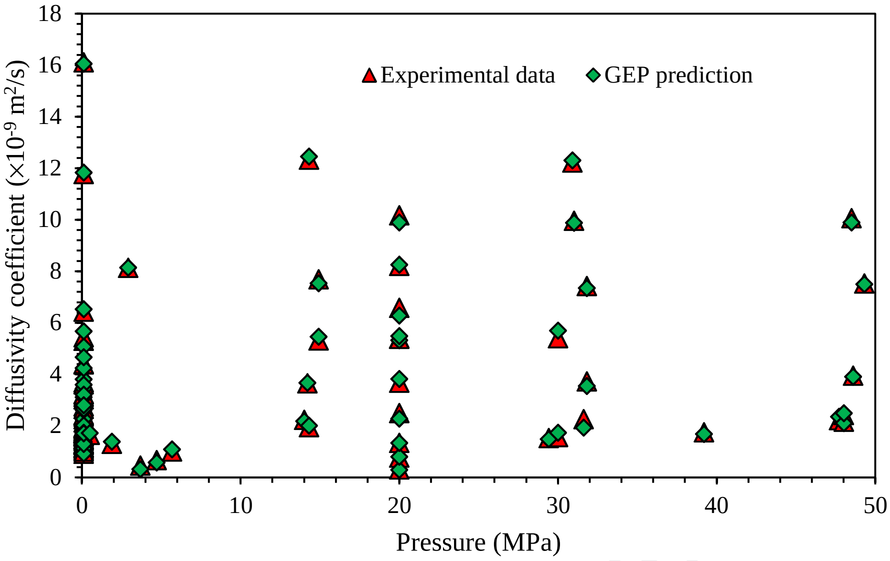

**Fig. 13.** Comparison of the diffusivity coefficient obtained from measurements and generated by GEP correlation, for the whole employed pressure values.

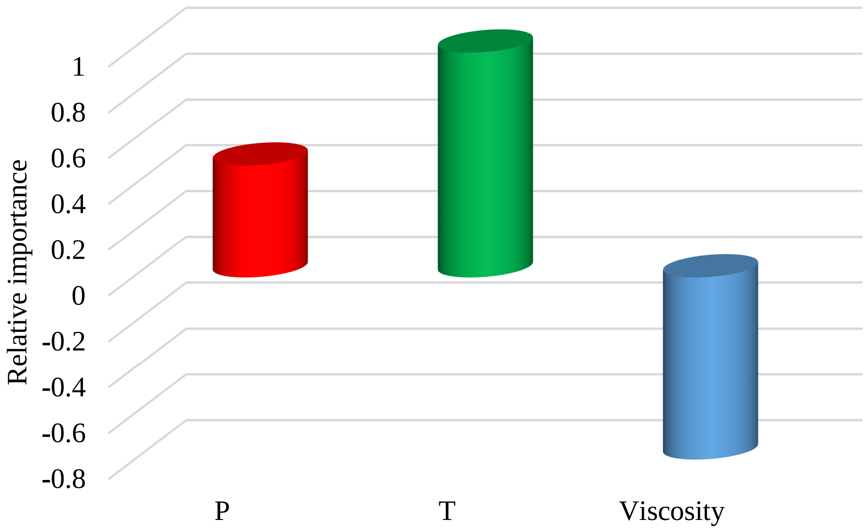

**Fig. 14.** Relevancy factor

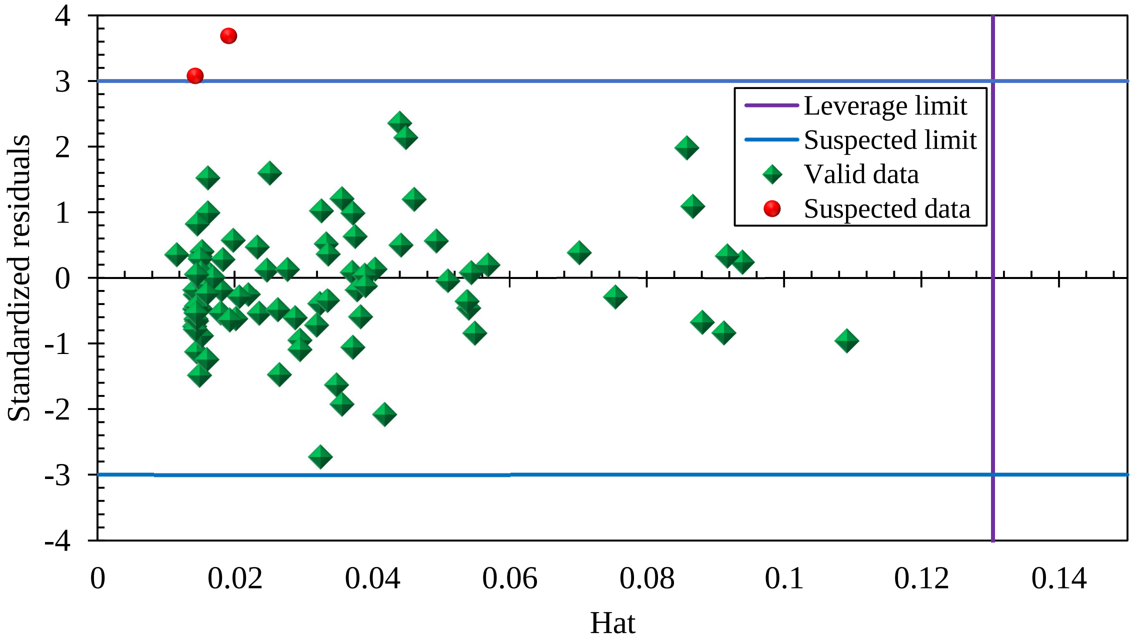

**Fig. 15.** The Williams plot of GEP correlation.

**Table 1.** Summary of the existing empirical models for predicting the diffusivity coefficient of $CO_2$

| Solvent | Model | Expression | Included parameters |
|---|---|---|---|
| Brine | (Othmer and Thakar, 1953) | $D_{CO_2} = \frac{14 \times 10^{-9}}{\mu^{1.1} V_m^{0.6}}$ | • Molar volume of the diffusing substance ($V_m$ in cm$^3$/gmol).<br>• Viscosity of the solvent ($\mu$ in $mPa \cdot s$) |
|  | (Wilke and Chang, 1955) | $D_{CO_2} = 7.4 \times 10^{-8} \frac{T\sqrt{\emptyset M}}{\mu V_m^{0.6}}$ | • Temperature ($T$ in K).<br>• The association parameter $\phi$.<br>• Molecular weight of solvent (M). |
|  | (Cadogan et al., 2014a) | $D_{CO_2} = \frac{k_B T}{n_{SE} \pi \mu a}$ | • $k_B$=1.38065×10$^{-23}$ J/K.<br>• $n_{SE}$ is the Stokes-Einstein number.<br>• The hydrodynamic radius of the solute (a in pm): $a = 168[1 + 2.0 \times 10^{-3}(T - 298)]$ |
| Pure water | (Lu et al., 2013) | $D_{CO_2} = 13.942 \times 10^{-9} \left[\frac{T}{227} - 1\right]^{1.7094}$ | • Temperature (T in K). |
|  | (Moultos et al., 2016) | $D_{CO_2} = D_0(P) \left[\frac{T}{Ts} - 1\right]^{m(P)}$ | • $D_0$=$a_1$ln($P$)+$a_2$, m=$b_1$ln($P$)+$b_2$, where $a$1=-2.3097×10$^{-9}$, $a$2=2,1064×10$^{-8}$, $b$1=-0.17812 and $b$2=2.59406; $P$ is the pressure. |

**Table 2.** Summary of the gathered data

| | **Max** | **Avg.** | **Min** | **SD** |
|---|---|---|---|---|
| **P (MPa)** | 49.30 | 9.64 | 0.1000 | 14.8030 |
| **T (K)** | 473.15 | 317.76 | 273 | 39.8 |
| **Viscosity (mPa.s)** | 1.9500 | 0.9003 | 0.1390 | 0.4720 |
| **Diffusivity coefficient (×10$^{-9}$ m$^2$/s)** | 16.1000 | 3.3522 | 0.3100 | 3.0874 |

**Table 3.** GEP setting parameters used in the study

| **Parameters** | **Value/setting** |
|---|---|
| Chromosome | 100 |
| Gene | 12 |
| Operators used | $+, -, \times, /,$ exp.$, \sqrt{\ }, INV, ln$ |
| Generations | 420 |
| Mutation rate | 0.45 |
| Inversion rate | 0.12 |

**Table 4.** Performance analysis of the implemented models.

| | | **GEP** | **GMDH** |
|---|---|---|---|
| **Training data** | AARD (%) | 3.8584 | 8.6269 |
| | R$^{2}$ | 0.9980 | 0.9943 |
| | RMSE (×10$^{-9}$ m$^{2}$/s$^{)}$ | 0.1427 | 0.2479 |
| **Test data** | AARD (%) | 6.0035 | 5.6292 |
| | R$^{2}$ | 0.9978 | 0.9874 |
| | RMSE (×10$^{-9}$ m$^{2}$/s$^{)}$ | 0.1245 | 0.2271 |
| **All data** | AARD (%) | 4.3014 | 8.0404 |
| | R$^{2}$ | 0.9979 | 0.9937 |
| | RMSE (×10$^{-9}$ m$^{2}$/s$^{)}$ | 0.1391 | 0.2440 |

**Table 5.** Performance analysis of the implemented decision trees (DTs) and random forest (RF) models.

| | | **DTs** | **RF** |
|---|---|---|---|
| **Training data** | AARD (%) | 4.2969 | 6.3627 |
| | R$^{2}$ | 0.9980 | 0.9973 |
| | RMSE (×10$^{-9}$ m$^{2}$/s$^{)}$ | 0.1598 | 0.1647 |
| **Test data** | AARD (%) | 8.8426 | 9.0015 |
| | R$^{2}$ | 0.9924 | 0.9940 |
| | RMSE (×10$^{-9}$ m$^{2}$/s$^{)}$ | 0.2532 | 0.2764 |
| **All data** | AARD (%) | 5.1862 | 6.8790 |
| | R$^{2}$ | 0.9969 | 0.9966 |
| | RMSE (×10$^{-9}$ m$^{2}$/s$^{)}$ | 0.1785 | 0.1870 |

**Table 6.** Comparison of the performances with prior models

| | GEP | Feng et al. | Othmer and Thakar | Wilke and Chang | Cadogan et al. |
|---|---|---|---|---|---|
| **AARD (%)** | 4.3014 | 7.91 | 12.75 | 12.60 | 13.84 |
| **$R^2$** | 0.9979 | 0.9960 | 0.9661 | 0.9434 | 0.9858 |
| **RMSE** | 0.1391 | 0.1954 | 0.5661 | 0.7311 | 0.3661 |

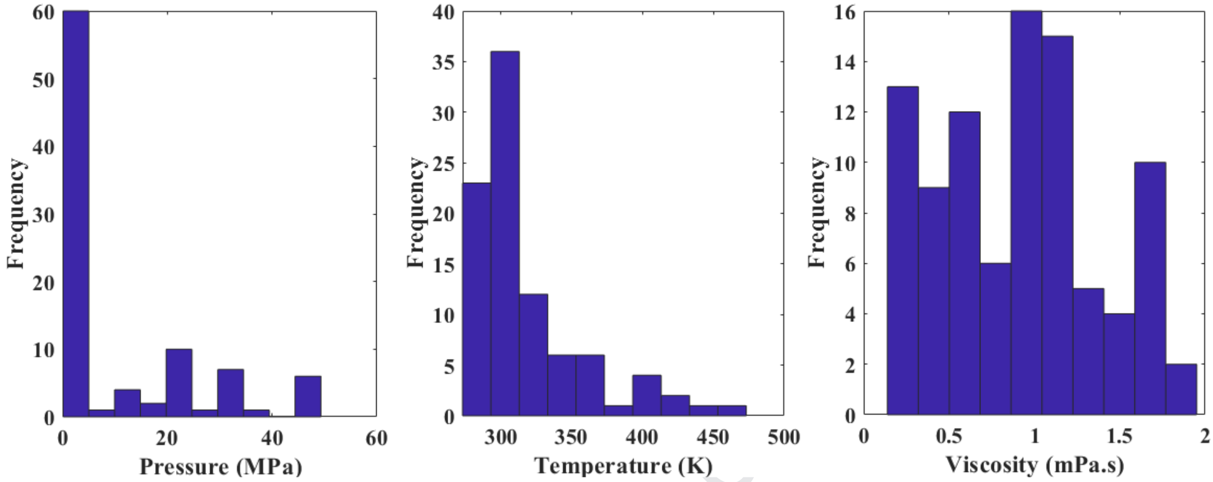

**Fig. 1.** Frequency histograms of the collected dataset

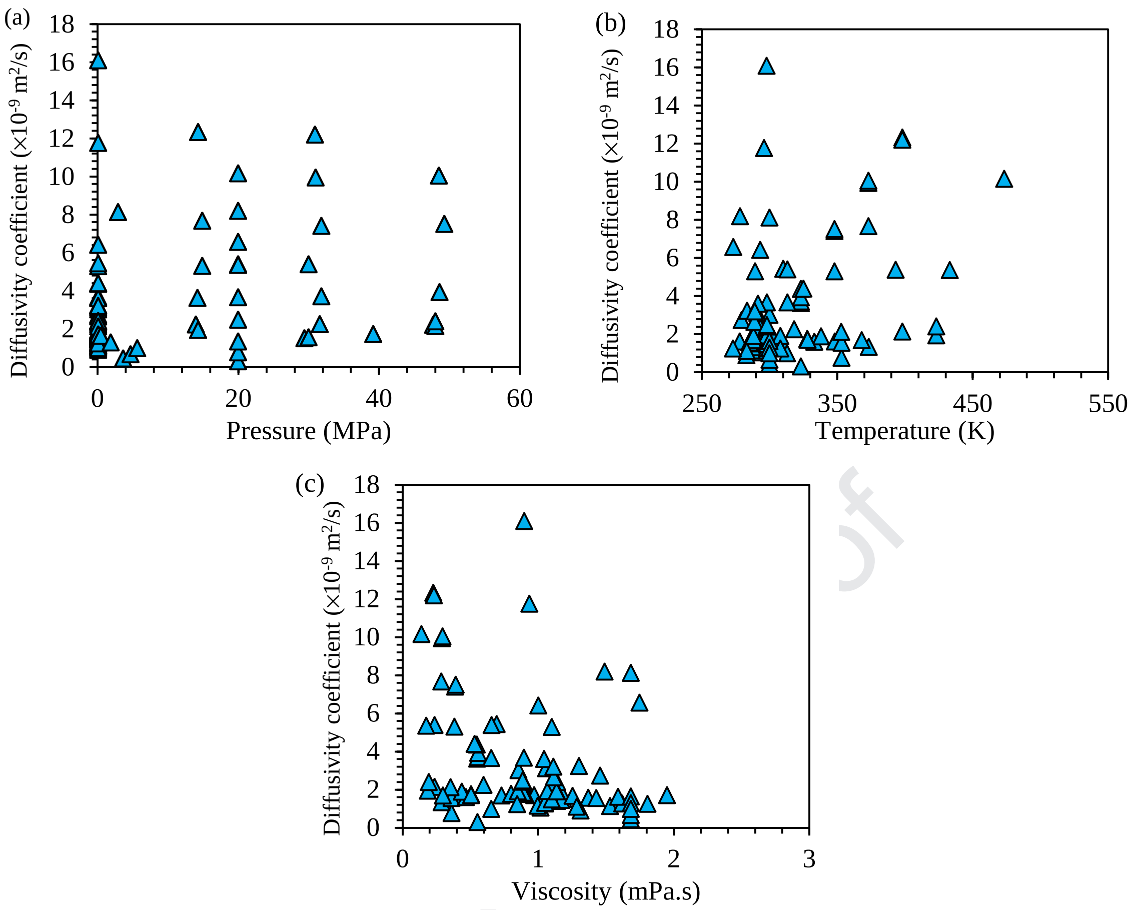

**Fig. 2.** Variation of diffusivity coefficient versus the independent variables

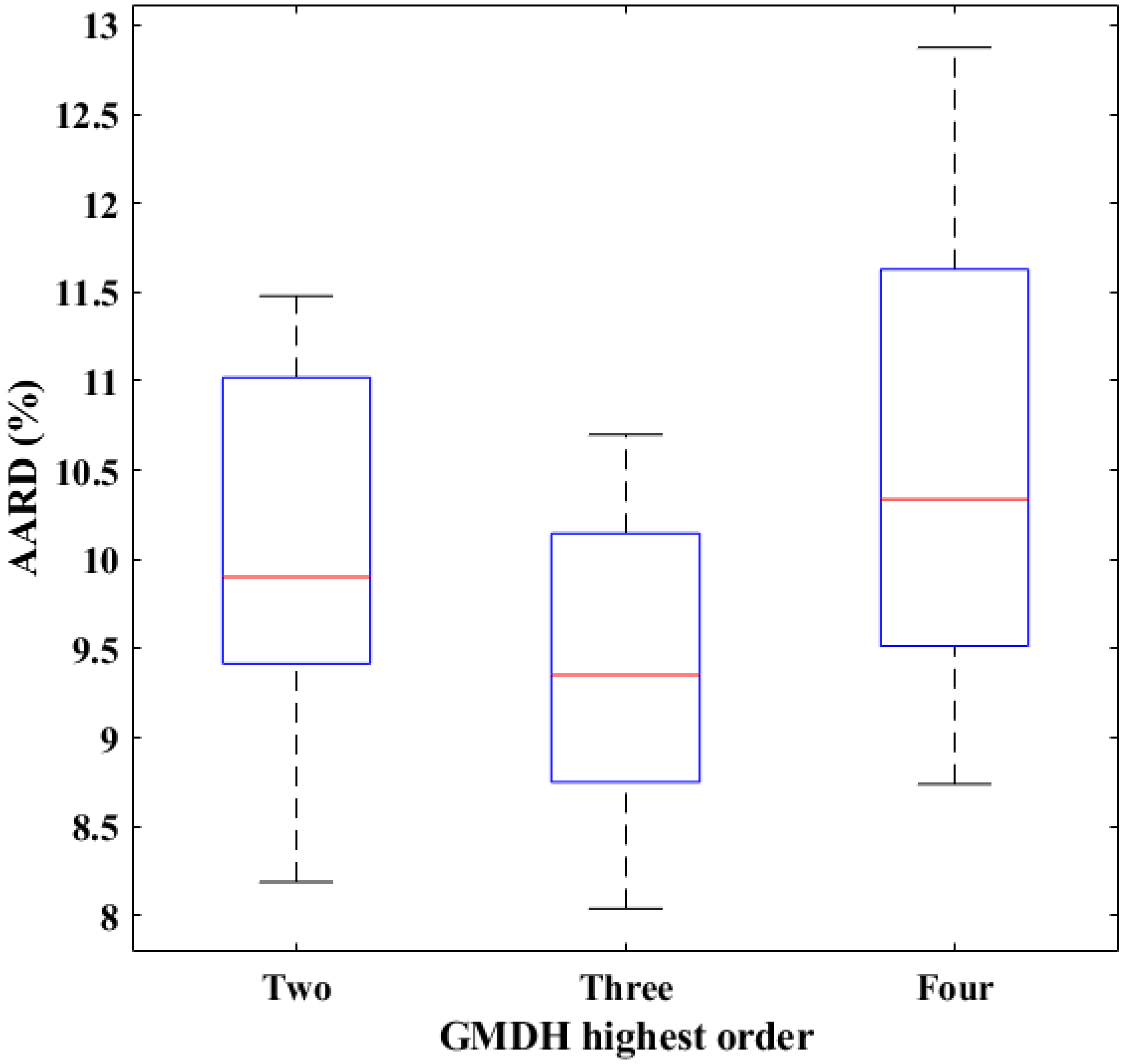

**Fig. 3.** Results obtained for ten realizations with different GMDH highest orders

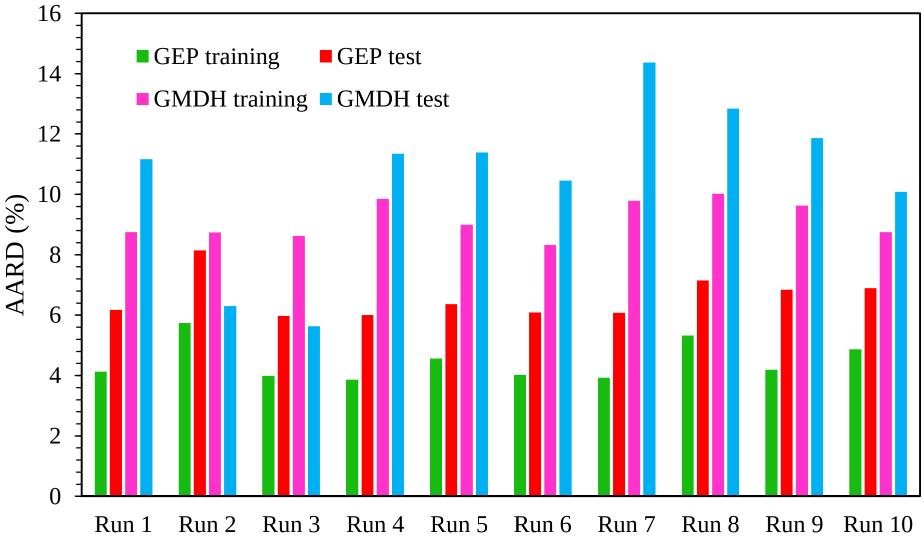

**Fig. 4.** Results of the GEP and GMDH sensitivity analysis on training and test sets

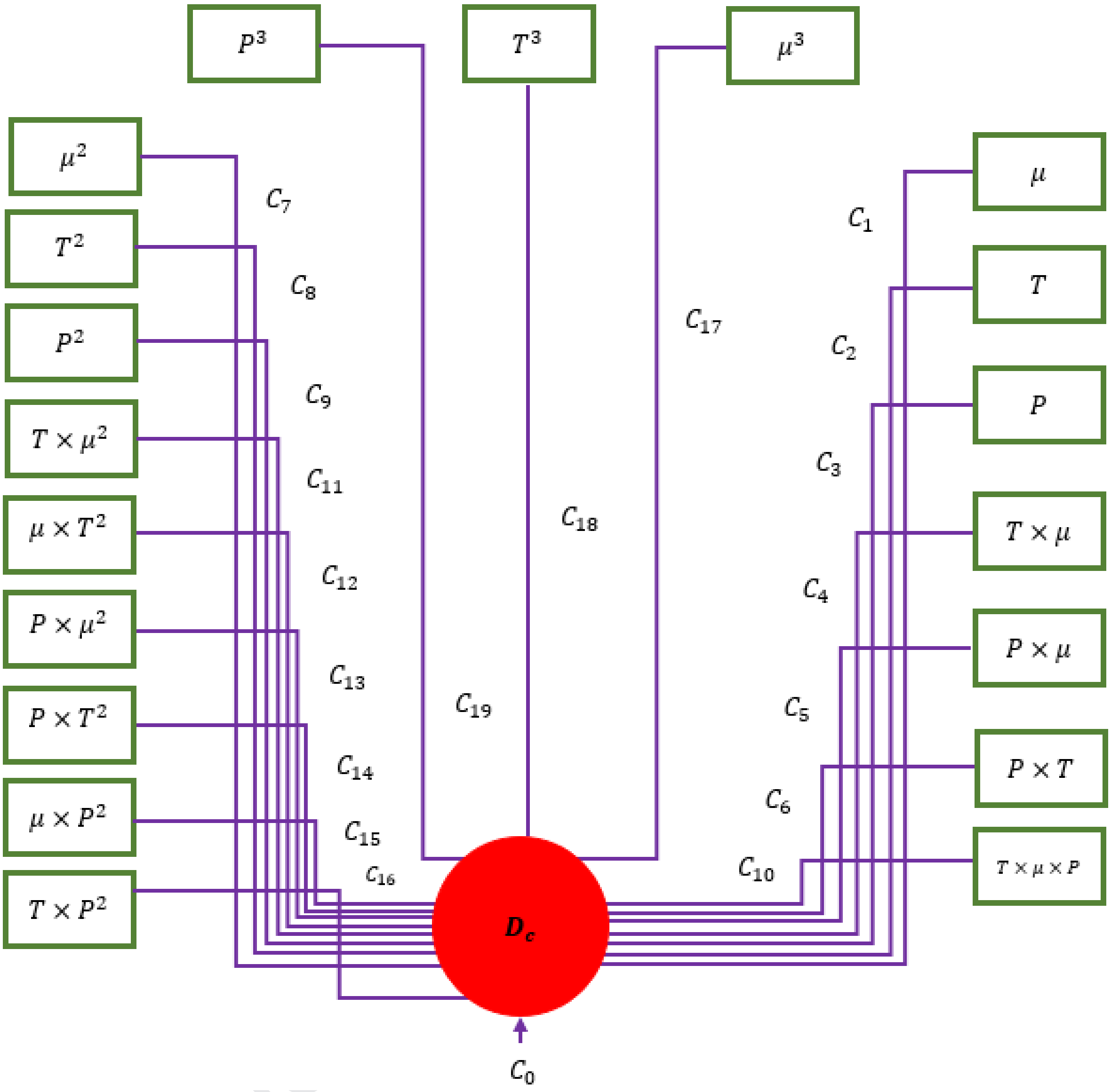

**Fig. 5.** A schematic structure of the implemented GMDH for predicting diffusivity coefficient

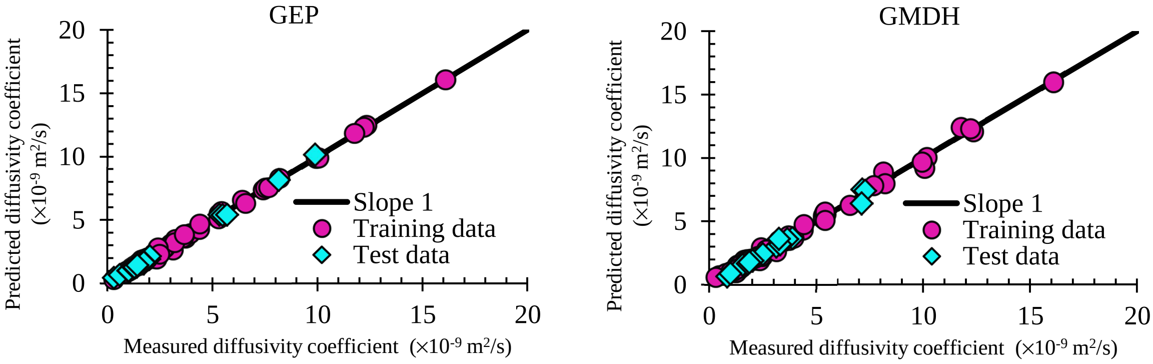

**Fig. 6.** Cross plots of the established GEP and GMDH correlations for diffusivity coefficient prediction

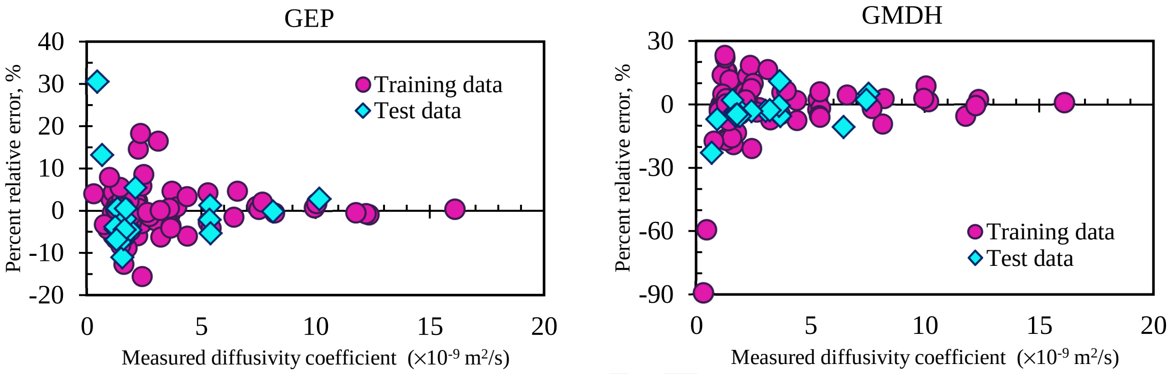

**Fig. 7.** Error distribution for the developed correlations

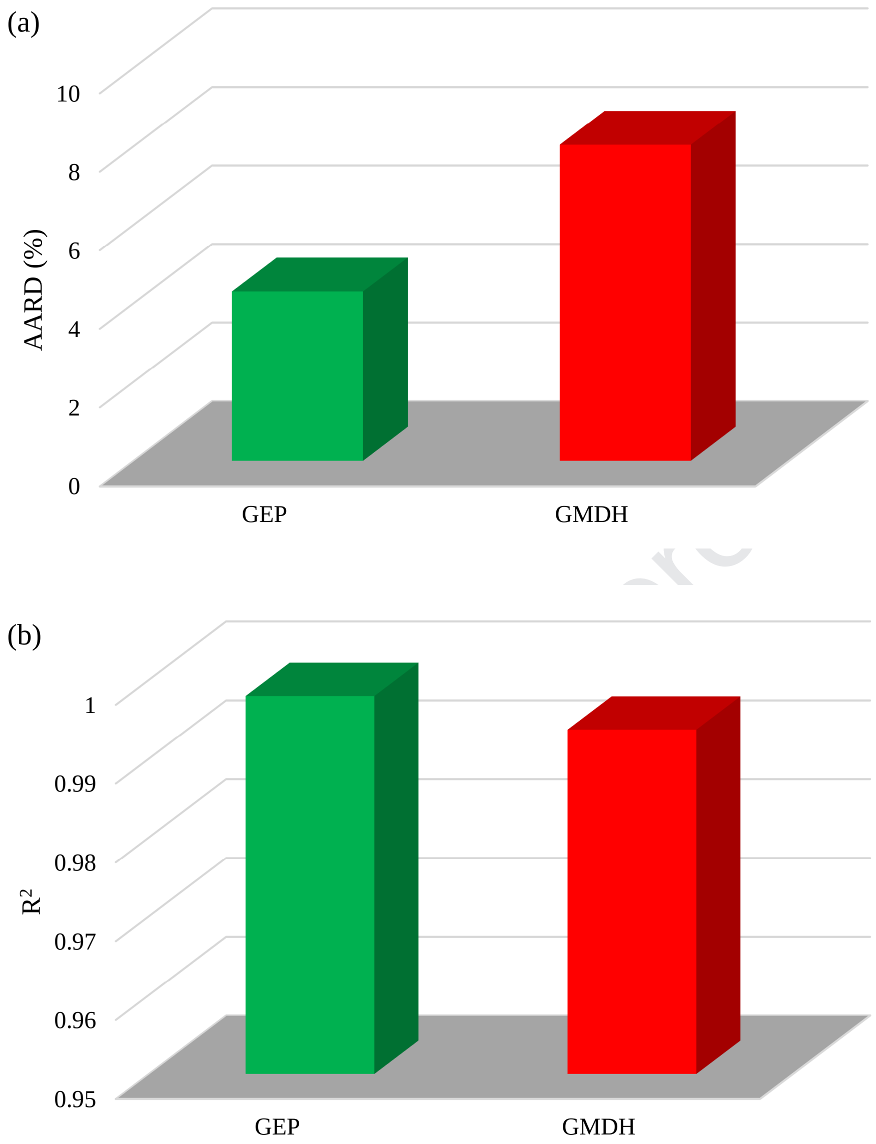

**Fig. 8.** Comparison between the performances of the correlations: (a) AARD and (b) $R^2$

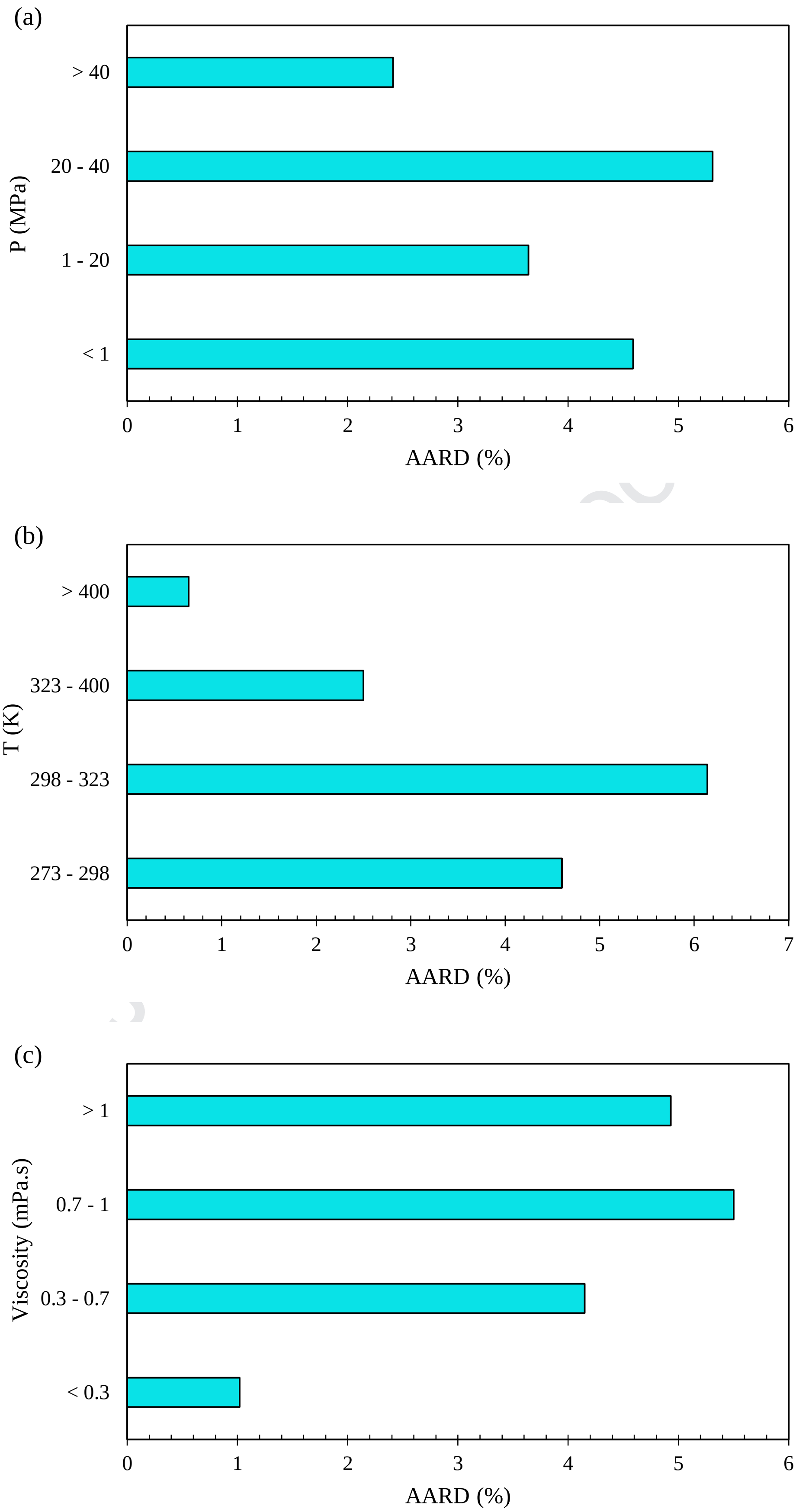

**Fig. 9.** Comparison of the error distribution of the models with respect to input parameters

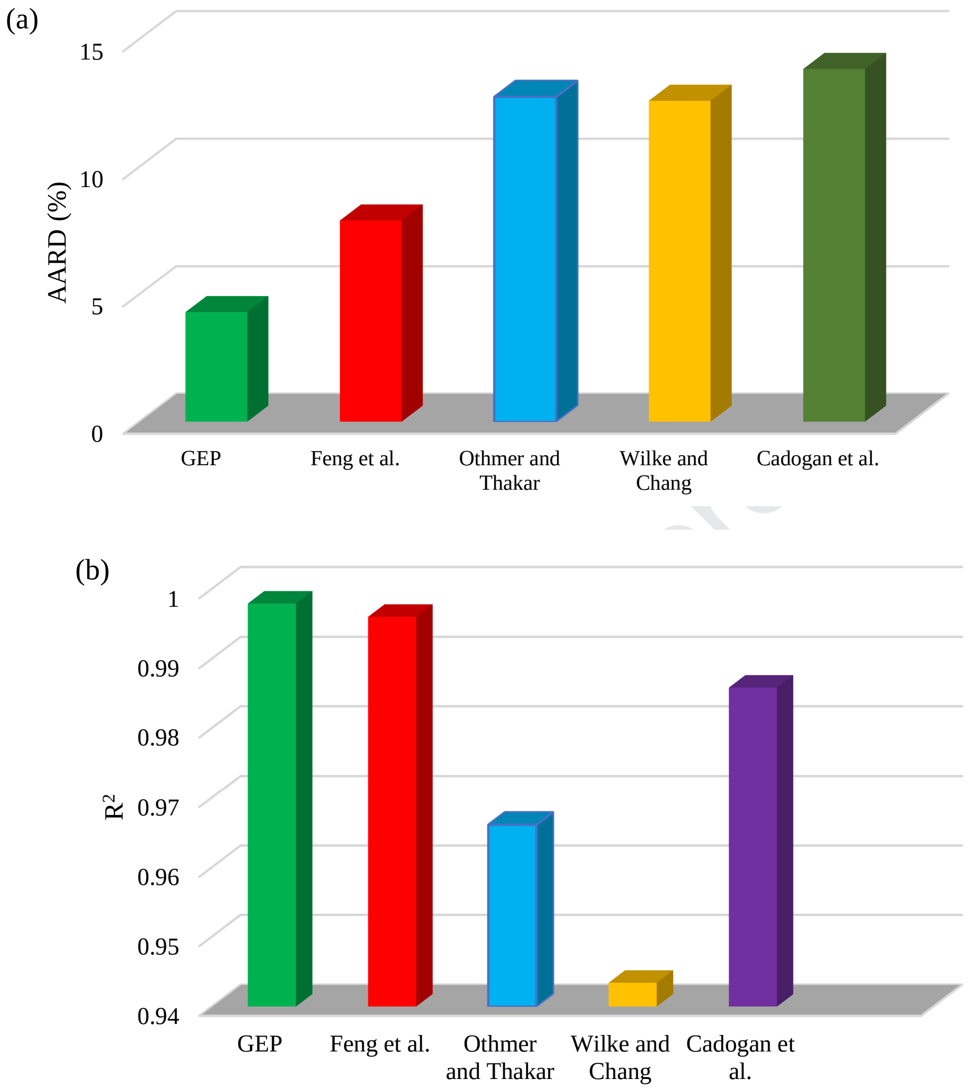

**Fig. 10.** Comparison between the performances of GEP correlation and the prior models: (a) AARD and (b) $R^2$

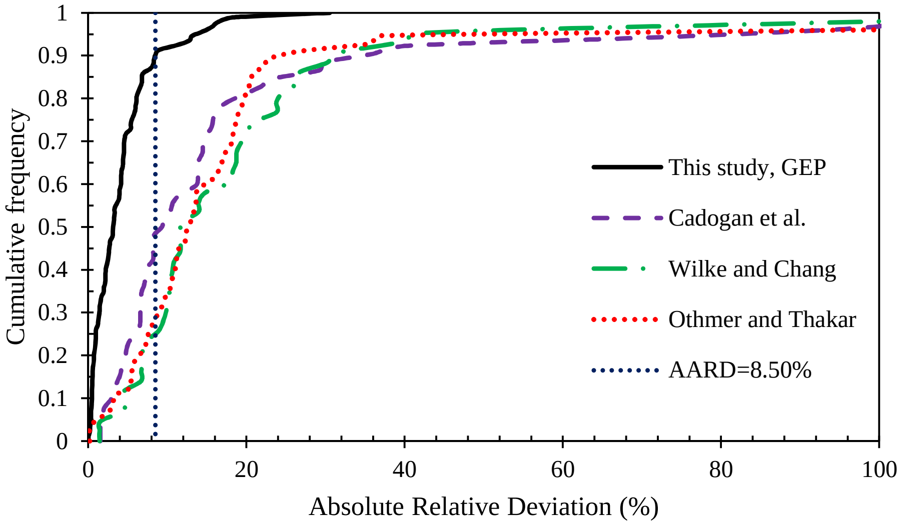

**Fig. 11.** Cumulative frequency vs. absolute percent relative deviation of GEP correlation and the prior empirical models.

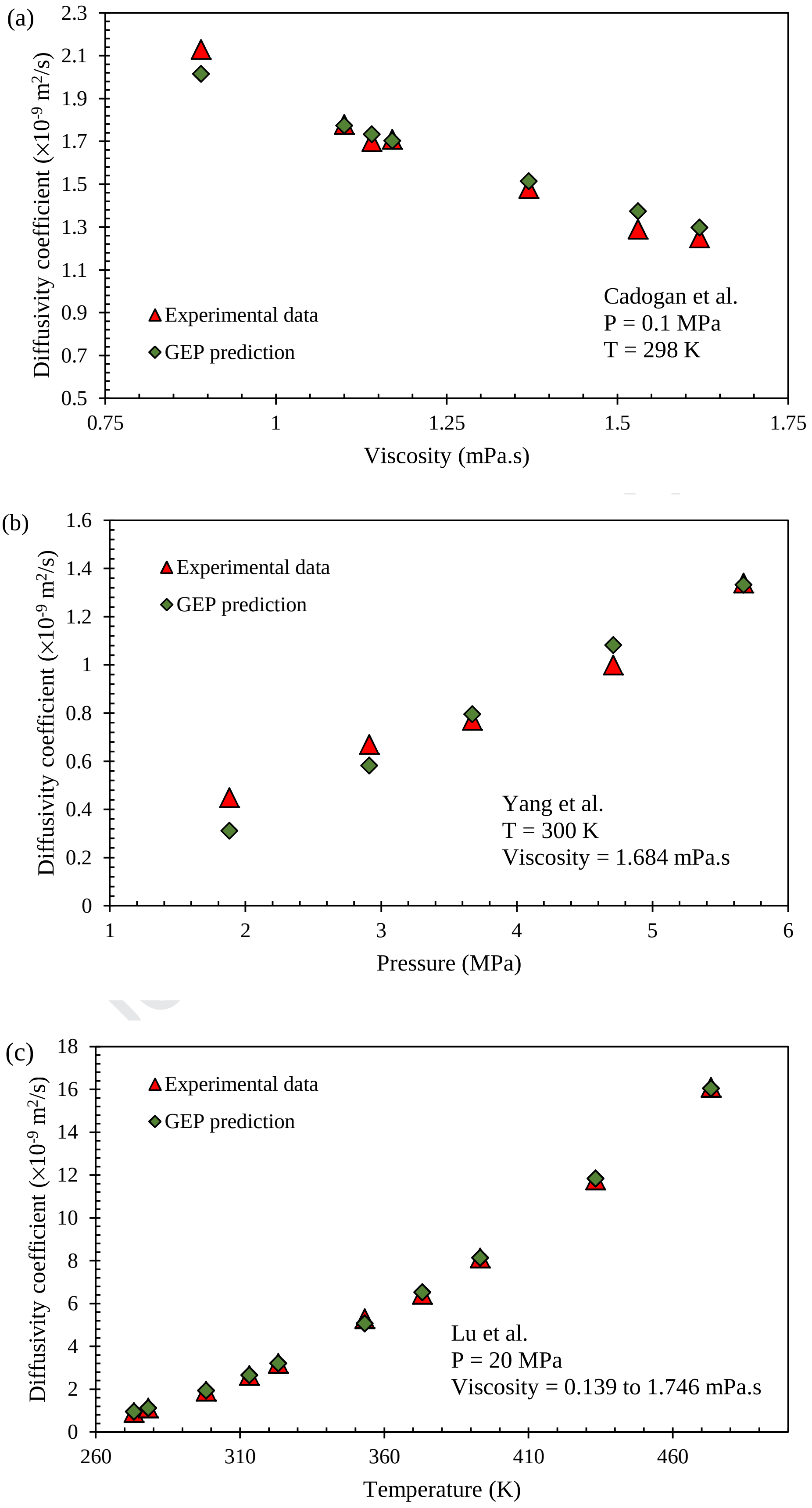

**Fig. 12.** Comparison of the diffusivity coefficient obtained from measurements and generated by GEP correlation, as function of the input parameters.

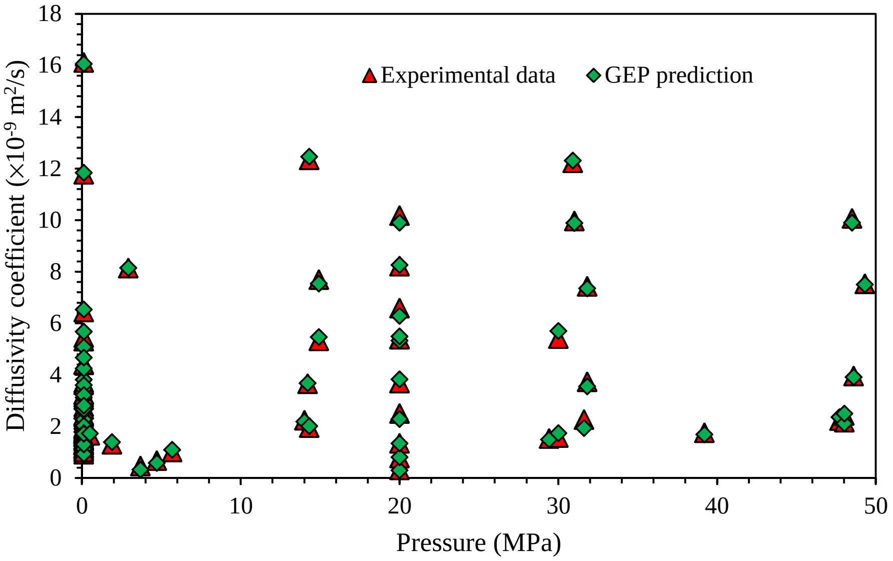

**Fig. 13.** Comparison of the diffusivity coefficient obtained from measurements and generated by GEP correlation, for the whole employed pressure values.

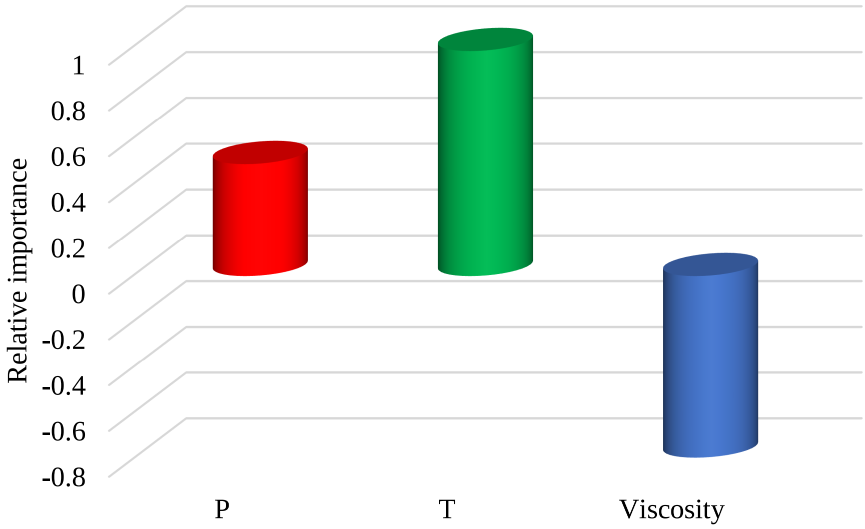

**Fig. 14.** Relevancy factor

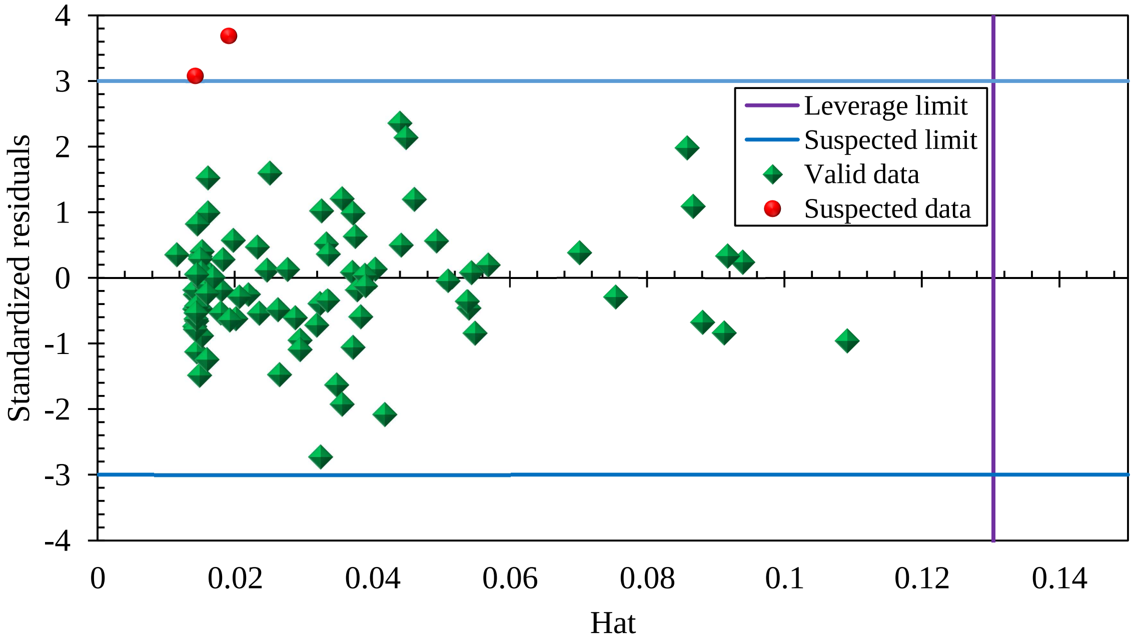

**Fig. 15.** The Williams plot of GEP correlation.

# Highlights

- Two white-box machine learning techniques were implemented for predicting the diffusivity of $CO_2$ in brine.
- GEP is the best developed correlation.
- GEP correlation outperforms the prior paradigms.

# Authors' contributions

Menad Nait Amar: Data curation, Formal analysis, Methodology, Investigation and Modeling, Software, Writing.

Ashkan Jahanbani Ghahfarokhi: Supervision, Methodology, Writing, Reviewing and Editing.

**Declaration of interests**

☒ 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.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: