A Bayesian neural network methodology to predict the liquid phase diffusion coefficient

A Bayesian neural network methodology to predict the liquid phase diffusion coefficient

Valerio Mariani[a,1], Leonardo Pulga[a], Gian Marco Bianchi[a], Giulio Cazzoli[a]

[a] *Department of Industrial Engineering, Alma Mater Studiorum University of Bologna, Viale Risorgimento 2, 40136 Bologna BO, Italy*

Abstract

This paper deals with the development of an Artificial Neural Network methodology for the prediction of the liquid phase diffusion coefficient between species at infinite dilution in binary mixtures. The proposed methodology was implemented to estimate the diffusion coefficients with improved accuracy with respect to empirical correlations, which are widely used despite their significant errors, especially on organic mixtures. The final aim of the work is to propose a novel methodology to apply in industrial fields where the mass transport by diffusion between organic liquids plays a key role, e.g. the lubricant degradation and scraping due to the dilution with the liquid fuel in the engine combustion chamber and the dilution of heavy oils and bitumens with organic solvents. In spite of the classical use of Artificial Neural Networks, this work is based on the mutual support between a reference empirical correlation and the Neural Network model, where the former is used to directly calculate the diffusion coefficient while the latter is trained to correct the correlation's result depending on the bonds strength in the solvent and the solute. Moreover, since the prediction of the diffusion coefficient for new mixtures (i.e. where no experimental measures are available) with Artificial Neural Network models should not be taken blindly, a Bayesian Neural Network is implemented, since it is capable to provide the degree of uncertainty of its prediction. This Bayesian Neural Network, given a selection of fluid properties as input features, was trained on 263 experimental data collected from literature findings. The proposed methodology achieved the prediction of the 80% of the available data with absolute relative errors lower than 10%, while the most recognized empirical correlations predicted the same data with absolute relative errors from 30 to 45%. A proof of the reliability of this methodology is the fact that the experimental data corresponding to the points predicted with the higher errors however fall in the predicted uncertainty, whose average width is below the ±20% of the predicted mean.

[1] Corresponding author
Email address: valerio.mariani4@unibo.it (Valerio Mariani)

Keywords: Diffusion coefficient, Machine Learning, Infinite dilution, Bayesian Neural Network, Oil dilution

Nomenclature

Variables and symbols	
$\%$w	Mass fraction (%)
A	Association factor (-)
C	Number of carbons
D	Liquid phase diffusion coefficient (m^2/s)
f	Number of input features
L_V	Latent heat of vaporization (kJ/kg)
M	Molecular weight (g/mol)
$MARE$	Mean Absolute Relative Error (%)
MSE	Mean Squared Error (%)
n	Number of points in the dataset
O	Number of oxygens
p	Generic probability distribution
$RMSE$	Root Mean Squared Error (%)
R^2, R_a^2	R-squared, adjusted R-squared
r_j	Paired rank for the jth dataset
r_S	Spearman correlation coefficient
$\overline{r_j}$	Average rank for the jth dataset
T	Temperature (K)
V	Molar volume (cm^3/mol)
V_b	Molar volume at the Normal Boiling Point (cm^3/mol)
w	Weight of the Artificial Neural Network model
X, Y	Generic datasets
Greek letters	
γ	Surface tension coefficient (N/m)
η	Dynamic viscosity (mPa·s)
θ	Machine learning correction factor (-)
μ	Machine learning predicted mean value
ρ	Density (kg/m^3)
σ	Machine learning predicted standard deviation

ϕ	Solvent/solute molar volumes ratio (-)
Subscripts	
1	Solvent
2	Solute
NBT	Normal Boiling Temperature
p	Predicted

1. Introduction

The research topic of this work is the mass diffusion between liquids in multi-fluid systems, with the focus on methods to estimate the diffusion coefficient that is a transport property of great importance for different industrial applications. In petroleum engineering solvents such as light hydrocarbons (HCs) are injected in the oil reservoirs to dilute heavy oils and bitumens that are immobile at reservoir condition because of their high viscosity. Solvent based methods for the oil viscosity reduction are attractive for the production and the secondary recovery of heavy oils and bitumens since they allow both water and energy saving in comparison with thermal methods ([1]) (e.g. Steam Assisted Gravity Drainage (SAGD)), which are affected by water consumption and heat loss, being based on hot steam injection. Since heavy oils and bitumens are a great part of the world wide fossil fuel reserves for the petroleum supply in the next future, the accurate estimation of the diffusion coefficient between the oil and the solvents may help the simulation and the early stage design of the solvent based extraction process. In automotive application dilution between the liquid fuel and the lubricant oil on the cylinder wall is one of the challenges to face for the development of the next low impact internal combustion engines. In order to comply with low green-house gases emission, pollutants emission regulation and fossil fuels saving, downsizing with Direct Injection (DI) is among the most adopted technological solutions in the nowadays gasoline-powered engines. However, despite the many advantages, injecting the liquid fuel directly into the combustion chamber puts some concerns, in particular in downsized engines where the time and the mean free path available for the liquid fuel evaporation are shorter due to the reduced displacement. As a result, the non-evaporated liquid fuel may hit the engine walls, likely resulting in liquid fuel film formation on the piston crown and on the cylinder wall, which is wetted by a thin lubricant oil film. The diffusion between the fuel and the lubricant oil onto the cylinder wall has proven to be source of soot emission ([2]), abnormal destructive combustion events at low speed and high load known as Low Speed Pre-Ignition (LSPI) ([3, 4, 5]), mechanical losses and parts wear increasing ([6]). Reliable methods to estimate the diffusion coefficient between the fuel and the lubricant oil are needed to develop oil-fuel dilution models for the study and assessment of engine configurations in terms of LSPI, emissions increasing and lubricant oil degradation risks.

The experimental measure of the liquid phase diffusion coefficient is known to be annoying, time and cost and care expensive. Moreover, experiments involving oils are challenging due to their own high viscosity and opacity. Currently, the most common approach to estimate the diffusion coefficients is the use of empirical and semi-empirical correlations that were developed decades ago. In this field, a milestone that inspired a number of works is the Einstein-Stokes equation, which describes the diffusion of a dispersed spherical particle that moves with Brownian pattern in a viscous fluid medium depending on molecules frictional resistance and intermolecular forces. The Einstein-Stokes equation was widely used in the study of mass diffusion in dilute solutions where the solute is present in mole concentrations below the 10%, being well represented by a dispersed particle in a fluid medium, i.e. the solvent. Several Authors have based their works on the main assumptions of the Einstein-Stokes equation while improving it by including different terms that positive correlated with the intermolecular forces. Wilke and Chang ([7]) conducted experiments on several different species and introduced the solvent corrected molecular weight with the so called *association factor A*, which depends on the liquid polarity (A = 2.6 for strongly associated liquids (e.g. water), A = 1.5-2 for associated liquids (e.g. alcohols) and A = 1 for non-associated liquids (e.g. HCs)). Siddiqi and Lucas ([8]) assumed that the calculation at the Normal Boiling Point (NBP) of both the solvent and the solute molecular volumes was an adequate measure of the intermolecular forces and used a collection of several hundred literature experimental data to fit their correlation. King ([9]) and Tyn and Calus ([10]) respectively introduced the latent heats of vaporization solvent-solute ratio and the surface tension coefficients solvent-solute ratio as representative terms of the intermolecular forces. Despite the fact that the literature correlations remain helpful and viable methods to estimate the diffusion coefficients, they lack of accuracy, in particular when predicting dilution between HCs. In [8] the Authors reported the average absolute error of different recognized correlations with respect to the experimental data underlining errors about 13-20% for aqueous mixtures and 20-35% for organic mixtures. Thus, since the available correlations do not meet the ever higher accuracy requirements of modelling in industrial applications, accurate, reliable and affordable methods to estimate liquid phase diffusion coefficients are needed.

Due to the recent advances in computing power, Artificial Neural Networks (ANNs) have shown their potential to approach several issues of engineering interest, including the estimation of fluid properties such as the thermal diffusivity ([11]), the void fraction ([12]), the liquid hold-up ([13]), the laminar flame speed ([14]), the mixtures Liquid-Liquid Equilibrium ([15]) and Vapour-Liquid Equilibrium ([16]) and the diffusion coefficient itself ([17]). However, the prediction of targets that highly differ from the training points leads to increase the sensitivity of the ANN model to the so

called *epistemic uncertainty*, which is related to the lack of sufficient number of data. Therefore, since the experimental data on mixtures comprising heavy HCs (i.e. number of carbons over *C*9 that behave similarly to mixtures of petroleum and engine interest) are very few, their prediction with plain ANNs (i.e. based on a deterministic approach) may suffer from epistemic uncertainty. Currently, Authors rely on Bayesian Neural Networks (BNNs) ([18, 19]) for the prediction of both the target variable and its confidence with respect to different type of uncertainties (e.g. epistemic and *aleatoric*) ([20]). The present paper is focused on the implementation of a novel BNN methodology for the estimation of the liquid phase diffusion coefficient for dilute binary mixtures once commonly available fluid properties are given. The methodology is developed into two steps: a) implementation of a novel plain ANN workflow called *hybrid mode*; b) implementation of the BNN according to the hybrid workflow. The results are compared with some of the most recognized empirical correlations and validated over experimental data selected from literature findings. Finally, the prediction of the liquid phase diffusion coefficient and the corresponding confidence is performed for some sample mixtures of industrial interest (gasoline-lubricant oil and heavy oil-light HC combinations) that are created on purpose by the present Authors to test the methodology.

2. Methodology

2.1 Database creation

Since the ANN learning relies on numerous and reliable data, the creation of an extensive and consistent database was performed. After a deep literature review, the experimental diffusion coefficient at room conditions (298 K, 101325 Pa) was collected for 263 mixtures given by the combination of 73 different liquids comprising water, alcohols, aromatics, paraffins and other organic compounds. For each liquid, the following properties were collected to play the role of input for the ANN: chemical formula, molecular weight, dynamic viscosity, density at room conditions from [21]; latent heat of vaporization at the saturation point from [21, 22]; molar volume calculated as in Eq. (1) at room conditions and as in Eq. (2) at NBP. In Eq. (2) the density is calculated by using the temperature-dependent correlations in [21] with the Normal Boiling Temperature (NBT) and zero vapor fraction. Despite the fact that several correlations based on the molecule structure are available in literature to estimate the NBP molar volume ([22, 23, 24]), Eq. (2) was adopted once its relative error was checked against the experimental data provided in [24] and compared with the relative error of two recognized correlations. As visible in Table 1, Eq. (2) shows small errors (< 1%) with very regular distribution over different sample liquids.

$$V = \frac{M}{\rho} \tag{1}$$

$$V_b = \frac{M}{\rho_{NBT}} \tag{2}$$

The four empirical correlations by Wilke and Chang (Eq. (3)), Siddiqi and Lucas (Eq. (4, 5)), King (Eq. (6)), Tyn and Calus (Eq. (7)) are selected as benchmark for the comparison with the ANN methodology.

$$D_{12} = 7.4 \cdot 10^{-12} \left[\frac{T}{\eta} \cdot \frac{(A_1\, M_1)^{0.5}}{V_{b2}^{0.6}} \right] \tag{3}$$

$$D_{12} = 9.89 \cdot 10^{-12} \left[\frac{T}{\eta^{0.907}} \cdot \frac{{V_{b1}}^{0.265}}{V_{b2}^{0.45}} \right] \tag{4}$$

$$D_{12} = 2.98 \cdot 10^{-11} \left[\frac{T}{\eta^{1.026}} \cdot \frac{1}{V_{b2}^{0.5473}} \right] \tag{5}$$

$$D_{12} = 4.4 \cdot 10^{-12} \left[\frac{T}{\eta} \cdot \left(\frac{V_{b1}}{V_{b2}} \right)^{1/6} \cdot \left(\frac{L_{V1}\, M_1}{L_{V2}\, M_2} \right)^{1/2} \right] \tag{6}$$

$$D_{12} = 8.93 \cdot 10^{-12} \left[\frac{T}{\eta} \cdot \frac{V_{b2}^{1/6}}{V_{b1}^{1/3}} \cdot \left(\frac{V_1\, \gamma_1^{1/4}}{V_2\, \gamma_2^{1/4}} \right)^{0.6} \right] \tag{7}$$

The created database is summarized in the maps shown in Fig. 1, where the points are placed at the pairs property corresponding to collected solvent-solute combinations, whilst the surrounding bars represent how frequently the liquid corresponding to that property is present in the database as a solvent or as a solute. In Fig. 1 dense data are visible in the properties region of the light HCs (i.e. *C*1-*C*9) with particular focus on the groups *C*6-*C*8, whose properties are respectively in the ranges: 0.1-0.4 mPa·s for viscosity (Fig. 1a); 110-140 cm^3/mol for NBP molar volume (Fig. 1b); 70-110 g/mol for molecular weight (Fig. 1c); 300-350 kJ/kg for latent heat of vaporization (Fig. 1d). The low frequency sparse data visible in Fig. 1a, 1b, 1c are mainly due to heavy HCs points (η = 3-35 mPa·s, V_b = 340-480 cm^3/mol, M = 200-400 g/mol). Those points are not highlighted in Fig. 1d since they are included in the light HCs dense region due to the comparable latent heat of vaporization (≈ 200 kJ/kg). The straights data with high frequency visible in Fig. 1d represent mixtures comprising water, whose latent heat of vaporization (2270 kJ/kg) is significantly higher than that of HCs.

2.2 *Plain hybrid neural network methodology*

ANNs are multi-layer networks of base elements, called *neurons*, that are organized in layers (of any width) connected each other. In this work a Feed-Forward ANN (FFANN) was implemented, where information are transferred one-way without any recurrence or matrix manipulation step from the input layer, whose neurons receive the so called *input features* (variables that positive correlate with the prediction target), to the output layer, whose neurons return the regression task. The layers (two or more) in-between the input and the output layer, which perform the non-linear transformations, are called hidden layers.

The FFANN is implemented in *Python* 3.7.4 with the Keras ([25]) and Tensorflow ([26]) open-source software libraries. A novel workflow called *hybrid mode*, which is described in the following, was developed and applied to the plain FFANN. Considering the diffusion coefficient as the prediction target, a standard used ANN is trained to return the regression task, which is the prediction target itself, hence, the diffusion coefficient is directly predicted by the ANN model. In the proposed hybrid mode, the regression task is an intermediate parameter that aims to correct the diffusion coefficient returned by a reference empirical correlation. In this work the classical empirical correlation by Wilke and Chang (Eq. (1)) was considered as the reference after the implementation of slight modifications aimed to integrate the correlation with the ANN in a physical manner. As a result, Eq. (8) was achieved by following these steps: a) due to the dilute solution assumption, the mixture viscosity and temperature were replaced with the ones of the solvent; b) the molecular weight was expressed as function of the density and the molar volume both at room conditions by rearranging Eq. (1); c) the solute NBP molar volume was replaced by the solute molar volume at room conditions; d) the fit exponent for the solute NBP molar volume was turned to 0.5 allowing to put the solvent and the solute molar volumes under the same exponent so that the diffusion coefficient directly depends on the solvent-solute molar volume ratio ($\phi = V_1/V_2$); e) the association factor (A) was removed. At this point, the modified correlation would not account for the intermolecular forces since in the original correlation that dependence was represented by the NBP molar volume and the association factor. However, the original idea of Wilke and Chang to represent the intermolecular forces by introducing a sort of effective molecules weight with a correction factor was maintained. The solvent-solute molar volume ratio (ϕ) is multiplied by a θ correction coefficient resulting in the effective molar volumes ratio depending on the bonds strength in the solvent and the solute. This θ correction is the regression task of the hybrid mode ANN, thus, in this novel workflow the output layer provides the θ value.

$$D_{12} = 7.4 \cdot 10^{-12} \left[\frac{T_1}{\eta_1 \, \rho_1^{-0.5}} \cdot (\theta \, \phi)^{0.5} \right] \tag{8}$$

Besides the fluid properties mentioned in 2.1 (density, viscosity, latent heat of vaporization, molar volume, molecular weight, carbons, hydrogens and oxygens number), other input features called *secondary features* were taken into account. The secondary features were features derived from the application of mathematical operations such (e.g. exponentials) and ratios to the solvent and the solute properties. The introduction of the secondary features was decided after that mathematical operations tested on some fluid properties have proven to give stronger relationships with the regression task. The input features were normalized following the traditional approach of transforming each input variable to present mean value = 0 and standard deviation = 1. The scaled parameters were fitted on the train set only, in order to verify with the test and the validation sets the validity of the whole pipeline, and not just of the regression model.

2.3 *Preventing the over-fitting*

Providing to the input layer with the whole set of features, together with the limited number of data against those recommended for the ANN learning process, increased the risk of overfitting, namely the ANN model memorizes the training points with excellent performances on the train set without being able to generalize to new data. In order to limit this risk, different strategies were adopted: a) selection of only the most meaningful features based on the statistic measure of the degree of association between the regression task and every single primary and secondary feature; b) implementation of an early stop strategy of the network weights update; c) addition of noise to the selected input features; d) integration of one of the hidden layers with a L2 kernel regularizer with the aim to improve the robustness of the prediction [27]; e) use of a repeated k-fold cross-validation approach. Focusing on the first point (a), the features are ranked based on the Spearman rank-order analysis that provides the strength and the direction (positive or negative) of the monotonic relationship between two generic data sets *X* and *Y*. The Spearman rank-order analysis is performed with the calculation of the Spearman correlation coefficient (r_S) in Eq. (9), where r_X and r_Y are the paired ranks for the considered input feature (*X*) and the θ correction (*Y*) while $\bar{r}_{X,Y}$ are the averaged ranks over the number of cases (*n*). For both *X* and *Y*, ranks are integer numbers from 1 to *n* that are assigned to each of the cases according to the following criterion: 1 is assigned to the higher value in the set, then the other reals from 2 to *n* are assigned to each value in the set following the descending order. After applying the Spearman rank-order analysis, only the *f* features with r_S greater than 0.1 (Fig. 2) were provided to the input layer.

$$r_S = \frac{\sum_{i=1}^{n}\left(r_{X,i} - \bar{r}_X\right)\cdot\left(r_{Y,i} - \bar{r}_Y\right)}{\sqrt{\sum_{i=1}^{n}\left(r_{X,i} - \bar{r}_X\right)^2 \cdot \sum_{i=1}^{n}\left(r_{Y,i} - \bar{r}_Y\right)^2}} \tag{9}$$

According to the second point (b), during the training step the maximum number of epochs, i.e. the number of times that the database is used to update the weights of the neurons, can be bypassed with an exit control strategy that stops the training if the performance on the train set would not improve after 10 consecutive iterations. As mentioned in the third point (c), the ANN is integrated with an extra input layer that adds Gaussian noise to the input features. In Fig. 3 focusing on the train set (solid lines), regardless to the noise amplitude, as the epoch increases the Mean Squared Error (*MSE*, Eq. (10)) decreases to a minimum asymptotic value. Despite the *MSE* decreasing trend, the persistence of a significant gap between train and test sets when providing the clean features (i.e. without noise addition, orange curve) is a clear sign of overfitting. Focusing on the test set (dashed lines), after an initial transient behaviour that does not optimize the weights on the train set, the *MSE* with noise addition follows the expected performance, reducing the gap between train and test (blue and green curves). Thus, increasing Gaussian noise has proven to be an effective strategy to prevent the overfitting issues even though it results in a bit lower accuracy on the train results. Moreover, Fig. 3 shows that the use of noise addition allows to achieve the minimum asymptotic value faster, thus, to stop the epochs earlier.

$$MSE = \frac{\sum_i \left(D_i - D_{i,p}\right)^2}{n} \tag{10}$$

Concerning the fifth point (e), in this work the repeated k-fold cross-validation is an iterative split of the dataset under review into train set and test set, so that the evaluation of the model's performance is independent from choice of the data-split.

2.4 *Optimization and comparison*

In order to provide a comprehensive evaluation of the performance of the ANN model, the full database (236 points) was divided into two subsets i.e. the reference set and the validation set. Since the focus of the work is to improve the future predictions with particular focus on mixtures of petroleum and engine interest, the solvent/solute combinations listed in Table 2 were selected as validation set. The remaining data are the reference set, which was used in the repeated k-fold cross-validation process mentioned in 2.3. As a consequence, the points in the validation set (Table 2) were not present in any of the train and test sets split. Thus, they were used as new data to perform a further check of both the accuracy and the generalization capability of the ANN model after the training step. The optimization targets were the network architecture, i.e. the number of hidden layers and the number of neurons per layer, and the model hyperparameters, i.e. the activation functions, the noise standard deviation and the lambda value for the L2 regularizer. Different sets of architectures and hyperparameters were generated automatically with a Bayesian

optimization algorithm [28] once it was provided with 5 initial sample sets defined by the present Authors. Each sample set was tested with the aforementioned repeated k-fold cross-validation (2.3), where the training step was performed with the Adam iterative based optimization algorithm [29] for the network weights update. For the training step, the maximum number of epochs was set to 1E5, however the algorithm is able to stop early if the aforementioned (2.3) exit control strategy condition is reached. Therefore, the accuracy of the model achieved with the architecture and hyperparameters set under review, was evaluated by averaging the error metrics on the test and the validation sets of all the k-repetitions. At this point, the Bayesian optimization algorithm, having recorded the performance increase/decrease associated with the initial sample sets, was able to generate the new sets, which were tested following the same procedure applied to the sample sets.

The error metrics adopted to evaluate the accuracy of the model are the R-squared (R^2, Eq. (11)), the adjusted R-squared (R_a^2, Eq. (12)), the Mean Absolute Relative Error (*MARE*, Eq. (13)) and the Root Mean Squared Error (*RMSE*, Eq. (14)), which were calculated for the training, the test and the validation sets. In Eq. (11, 12, 13, 14) D is the generic experimental diffusion coefficient, D_p the corresponding predicted value, $\overline{D}$ the average experimental diffusion coefficient calculated on the dataset under analysis, n is the number of points and f is the number of input features.

$$R^2 = 1 - \frac{\sum_{i=1}\left(D_i - D_{i,p}\right)^2}{\sum_{i=1}\left(D_i - \overline{D}\right)^2} \tag{11}$$

$$R_a^2 = 1 - (1 - R^2) \cdot \frac{n-1}{n-f-1} \tag{12}$$

$$MARE = \sum_i \frac{\left|D_i - D_{i,p}\right|}{D_i} \cdot 100 \tag{13}$$

$$RMSE = \sqrt{\frac{\sum_i \left(D_{i,p} - D_i\right)^2}{n}} \cdot 100 \tag{14}$$

For sake of clarity, the optimized hybrid mode ANN is presented while compared with a standard used ANN, where the regression task is the prediction target itself i.e. the diffusion coefficient. The standard ANN was implemented by assuming the same methodologic (overfitting approach) and optimization techniques reported for the hybrid mode. The optimized network architecture for both the standard and hybrid modes is composed by 2 hidden layers of 24 and 32 neurons each. The activation functions are the tanh (hyperbolic tangent) for the first hidden layer and the ReLU (Rectified Linear Unit i.e. a function that set the input to zero if the original input is ≤ 0 and passes the input without any modification if the original input is > 0) for the second. The optimized std of

the Gaussian noise is 0.02 and the lambda value for the L2 regularizer is 0.01. For the output layer, a linear activation function is chosen since no normalization of the target values was performed because not necessary given their almost normal distribution.

Table 3 shows the comparison between the hybrid and the standard modes in terms of accuracy with the considered error metrics. As visible in Table 3, whilst the performance of the hybrid and the standard modes are comparable on the three sets in terms of %*MARE* and %*RMSE*, the hybrid mode behaves significantly better on the test and the validation sets in terms of R^2. As shown in Table 3, the hybrid mode addresses the accuracy target better than the standard mode on both the test set and the validation set. Furthermore, the hybrid mode relies on small R^2 differences between train and test and between train and validation with respect to the standard mode. This is reported since small R^2 differences between train and evaluation (test and validation in this case) are a recognized proof that the model is not overfitting the train set. As a result, one can assume that the developed hybrid mode reasonably prevents the overfitting and that it is capable to generalize to new data better than the standard mode. Moreover, it must be considered that the hybrid mode allows the interpretability of relevant dependencies that are not assigned to the regression task. Fig. 4 shows the target vs predicted regression line for the train, the test and the validation sets. It can be observed that the test set prediction is affected by some random sparse points around the 1:1 line. However, the reliability of the model for the prediction of new mixtures is ensured by the fact that in the region where the gasoline/lubricant oil and the heavy oils/HC solvent mixtures are expected (yellow frame), the prediction of the test set as well as the ones of the validation set are very close to linearity. The R^2, the slope and the intercept associated with Fig. 4 are listed in Table 4. According to the above analysis, the hybrid mode is the winner methodology for the implementation of the Bayesian approach. When the hybrid mode is applied to predict the full database, $R^2 = 0.888$ (slope = 0.922, intercept = 0.117) and $R_a^2 = 0.879$ result. In Fig. 5 a scheme of the hybrid mode workflow with a brief resume of the network topology is shown.

2.5 *Bayesian neural network methodology*

Plain ANNs have the potential to perform extremely accurate predictions based on a sufficiently large dataset and an optimization of the network topology and weights. Their predictions, however, may suffer mostly from two different types of error: aleatoric (due to uncertainty in the training data) and epistemic (due to the distribution of the points in the train set with respect to the prediction domain). Whilst the aleatoric uncertainty is intrinsic in the dataset and it has been addressed by utilizing only accurate and reliable experimental data for the training set, the epistemic uncertainty is traditionally more complicated to calculate and, consequently, to reduce. To this aim,

a new approach called Bayesian has been recently proposed. The Bayesian approach adopts probability distributions in place of single weights for describing the connections between the nodes. The optimization process must therefore identify the optimal distributions rather than the optimal weights, leading to the training of an ensemble of networks which share the same weight distributions, rather than a unique representation of the system.

Given a domain $d(X, Y)$ (our dataset) comprised of n observations, and a probabilistic model $p(Y|X, w)$ that represents the dataset, with w weights of the model, the Bayes' theorem statement (Eq. (15)) says that: given the training data d, the posterior distribution of the weights $p(w|d)$ (i. e. the probability distribution of the weights of the neurons updated after having observed the training data) is proportional to the prior probability of the weights $p(w)$ (i. e. the initially guessed probability distribution of the weights of the neurons) with the proportionality coefficient $p(d|w)$ defined as the likelihood function (Eq. (16)). In Eq. (15) the denominator $p(d)$ is the distribution of the training set and is intended as a normalization term.

$$p(w|d) = \frac{p(d|w) \cdot p(w)}{p(d)} \tag{15}$$

$$p(d|w) = \prod_{i=1}^{n} p(Y_i|X_i, w) \tag{16}$$

The maximization of the product in the numerator of Eq. (15) provides the maximum a-posteriori estimate of the weights. The optimization of the product instead of the maximization of $p(d|w)$ only, prevents the risk of overfitting. The posterior predictive distribution can be rewritten as in Eq. (17), which describes how the final prediction is achieved by performing a weighted average of the predictions of an ensemble of networks, weighted on the posterior probabilities of the parameters w. The numerical implementation of this approach required the computation of the output value a sufficient number of times (> 100, controlled iteratively on the residuals of mean and standard deviation of the predictions) performed with random weights taken from the trained distributions for the connections between nodes.

$$p(Y|X, d) = \int p(Y|X, w) \cdot p(w|d) dw \tag{17}$$

With regards to the training phase, the optimization of the parameters of an ANN is traditionally implemented with the back-propagation algorithm. This requires the computation of the derivatives of the activation functions, which is intractable with the probability distributions of the weights, therefore a variational approximation form has been proposed. The exact derivation is the same as that proposed in the original article ([30]), from which the cost function to minimize is

approximated as in Eq. (18), where $p(w)$ is the prior distribution of the weights, which is called complexity cost and it is assumed as a Gaussian distribution, whilst $q(w|z)$ is the variational form of the distribution $p(w|d)$. The target variational posterior distribution $q(w|z)$ is described by z = (μ, σ), which is parametrized as a Gaussian distribution with mean value vector (μ) and its standard deviation (σ). Therefore, the BNN is parametrized with twice the number of parameters than a plain ANN. The optimization process of the network architecture and hyperparameters has been performed through the same steps as for the plain ANNs, with repeated k-fold cross validation and early stopping during training. The final topology and hyperparameters are the same as those listed in 2.4 for the plain network.

$$F(d,z) \cong \frac{1}{n}\sum_{i=1}^{n}\left[\log\left(q(w_i,z)\right) - \log\left(p(w_i)\right) - \log\left(p(d|w_i)\right)\right] \quad (18)$$

3. Results

3.1 *Validation*

In this section the BNN is validated and discussed against the previously mentioned validation set (Table 2). The final prediction, transformed for calculating the diffusion coefficient, has reported a mean standard deviation of 0.02E-9 m^2/s in the train set and of 0.03734E-9 m^2/s in the test set. The 97% of the training points and the 96% of the test points fall into the 2σ distance from the predicted mean value, that is an appropriate result for a well fitted normal distribution, thus, it is assumed reliable for predicting the epistemic uncertainty on new points. In Fig. 6 the predicted mean value and epistemic uncertainty (mean prediction ± 2σ) are reported with respect to the experimental data. A general good agreement with experiments is obtained for the mean value. Furthermore, for the points with the larger gaps between measurements and predicted mean value, the experimental data is however included in the confidence interval predicted by the BNN. The epistemic uncertainty can be therefore adopted to estimate the applicability of the prediction method for new mixtures of interest for which experimental data are not available.

Comparing the four benchmark empirical correlations and the BNN methodology on the full database, the cumulative distributions of the Absolute Relative Error (*ARE*) shown in Fig. 7 are reported. It must be underlined that the general error reduction achieved by the BNN can be attributed to the database points that were used in the training set. However, the gain in accuracy is visible by the fact that the 70% of the database is predicted with errors below the 5% while the empirical correlations achieve the same error threshold with about the 20% of the database. Moreover, the proposed methodology predicts almost the full database (90%) with errors below 25% against errors over 40% that may be committed with the empirical correlations. The blue area

over the BNN distribution in Fig. 7 represents the cumulative error of the predicted confidence interval, whose trend is assessed as follows: when the experimental measure of the point falls in the predicted confidence interval (± 2σ), the error associated to the point is zero while when the confidence interval is exceeded (positive or negative), the *ARE* is calculated as the difference between the experimental measure and the maximum deviation from the predicted mean (D_{12} + 2σ if positive, D_{12} - 2σ if negative). Fig. 7 shows that for the 80% of the full database, which is predicted with *ARE* in the range 0-10%, the BNN is capable to provide results with a reliable confidence.

Before moving on the prediction of the diffusion coefficients of petroleum and engine mixtures, Fig. 8 shows the trend of the predictive potential of this BNN on new HC mixtures. Fig. 8 is realized assuming constant properties for the solvent, identified as n-heptane for representing a generic fuel, while the number of carbon atoms of the solute is increased from *C*10 (threshold carbons number conventionally used to intend heavy HCs) to *C*34 (representative of the mean carbons number in oils ([31])). As shown in Fig. 8, the higher is the solute carbon number, the lower is the liquid diffusivity (as reported in literature) with an increased uncertainty in the BNN's prediction the more the target is distant from the training domain. However, since the predicted mean values well matches the reference experimental points available from *C*10 to *C*16, one can expect that the real diffusion coefficients for the higher carbon numbers in the solute remain close to the traced mean curve (red) and, at worst, in the yellow band (half the confidence interval) with absolute relative errors below the 20%.

3.2 *Prediction of new mixtures*

Once the validation step has been reported in terms of both improved accuracy against the current standard, i.e. empirical correlations (Fig. 7), with reasonable proofs of reliability (confidence intervals, Fig. 8), the BNN methodology is applied to organic mixtures of industrial interest, which are created on purpose by the present Authors in order to test the methodology against liquids that are very uncommon in diffusivity measurements. With regards to the petroleum application, n-hexane (C_6H_{14}), n-heptane (C_7H_{16}) and naphthalene ($C_{10}H_8$) are considered as a solute, being common proposals in solvent injection methods whilst a cold lake blend crude oil is assigned as a solvent in representation of heavy oils and bitumens. Concerning the automotive application, two different SAE lubricant oils, the single-grade oil SAE 30 and the multi-grade oil SAE 10W-30, are considered as a solutes whilst a four-component (45.95w% i-octane (C_8H_{18}), 12.91w% n-heptane (C_7H_{16}), 37.18w% toluene (C_7H_8), 3.96w% 1-pentene (C_5H_{10})) surrogate representative of the thermo-physical properties (density, viscosity, latent heat of vaporization) of a commercial gasoline

is assigned as a solvent. In the calculation, the cold lake oil, the gasoline surrogate and both the SAE oils are treated as pseudo-pure liquids by averaging their properties.

The first two rows of Table 5 show the gasoline-lubricant oil combinations. It can be noticed that the predicted mean values are consistent with the trend shown in Fig. 8. Indeed, since the combination of n-heptane (*C*7) with *C*34 (representative of the average carbons number for HC lubricant oils) gives diffusion coefficients about 1.0E-9 m^2/s, it is reasonable to think that the combination of commercial gasolines, which usually comprise small fractions of *C*10-*C*12, with real lubricant oils, which may contain fractions with carbons number about *C*50, shows diffusion coefficients lower than 1.0E-9 m^2/s. Another proof of the reasonability and the robustness of the mean predictions is that using the SAE 10W-30 instead of the SAE 30 the diffusion coefficient increases consistently with the experimental evidence that the diffusion is faster the lighter is the solute (SAE 10W-30 in this case) due to the reduced friction between the solute particle and the medium. Furthermore, even though the confidence interval predicts maximum deviations from the mean values of 18% (SAE 30) and 26% (SAE 10W-30), it needs to be remembered that Fig. 8 showed that the real values tend to be included in half the confidence interval, hence, maximum deviations of 9% (SAE 30) and 13% (SAE 10W-30) can be expected.

The last three rows in Table 5 show the results for solvent-solute combinations that are representative of heavy oils dilution with the solvent injection method. As proof of concept, it is remarkable that even though the database lacks of specific information on heavy and high viscosity liquids comparable to those oils, the predicted orders of magnitude and values are very close to experimental findings on topic. In [32] the Authors measured the diffusion coefficient at infinite dilution and at room conditions of n-hexane and naphthalene in both i-octane and different HC oils with increasing viscosity from 3 to 5000 mPa·s. Considering the viscosity-diffusion coefficient curves presented in [32], diffusion coefficients in the range 0.03-0.05 m^2/s can be observed corresponding to the viscosity of the cold lake oil at 293 K (≈ 30 mPa·s). According to the last three rows in Table 5, the three considered solvents are comparable between each other in terms of the predicted mean value even though n-hexane and naphthalene show a bit faster dilution. Moreover, in the case of oil dilution with naphthalene, the implemented BNN guarantees a very small uncertainty.

4. Conclusion

In this work the implementation of a Bayesian Neural Network based methodology is performed for the prediction of the liquid phase diffusion coefficient in binary mixtures at infinite dilution. The need of this implementation is related to the fact that the results returned by the empirical

correlations, that are commonly used for those estimations, may be affected by severe errors. The proposed methodology deals with the coupling (in a physical manner) of the predictive power of Neural Network models with the interpretability of the most relevant dependencies given by the empirical correlations and with the capability of the Bayesian implementation to approach the uncertainties (in particular the epistemic uncertainty) predicting the confidence interval of the solution. This methodology has led to more accurate predictions ($MARE \approx 9\%$) with respect to several experimental data of different species against the empirical correlations of Wilke and Chang, Siddiqi and Lucas, King, Tyn and Calus ($MARE \approx 35\%$) that are the current standard for those estimations. The reliability of the methodology has been shown by checking that almost all the available measures (80%) are included in the confidence interval predicted by the Bayesian Neural Network. The methodology was applied to the simulation of gasoline-lubricant oil and heavy oil-liquid solvent mixtures resulting in predictions that are consistent with experimental evidences both in quantitative terms i.e. order of magnitude and value, and in qualitative terms i.e. capability to capture diffusion trends.

Authors contributions

Valerio Mariani: Conceptualization, Investigation, Methodology, Writing-Original draft, Writing-Review and editing; Gian Marco Bianchi: Supervision, Writing-Review and editing; Giulio Cazzoli: Conceptualization, Supervision; Leonardo Pulga: Methodology, Formal analysis, Visualization, Writing-Review and editing.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] S. R. Upreti, A. Lohi, R. A. Kapadia, R. El-Haj, Vapor extraction of heavy oil and bitumen: a review, Energy & Fuels 21(3):1562-1574 (2007), doi:10.1021/ef060341j

[2] M. Raza, L. Chen, F. Leach, S. Ding, A review of particulate number (PN) emissions from gasoline direct injection (GDI) engines and their control techniques, Energies 11(6) (2018), doi:10.3390/en11061417

[3] O. Welling, J. Moss, J. Williams, N. Collings, Measuring the impact of engine oils and fuels on low-speed pre-ignition in downsized engines, SAE International Journal of Fuels and Lubricants, 7(1):1-8 (2014), doi:10.4271/2014-01-1219

[4] M. Amman, D. Mehta, T. Alger, Engine operating condition and gasoline fuel composition effects on low-speed pre-ignition in high-performance spark ignited gasoline engines, SAE International Journal of Engines, 4(1):274-285 (2011), doi:10.4271/2011-01-0342
[5] M. Amman, T. Alger, B. Westmoreland, A. Rothmaier, The effect of piston crevices and injection strategy on low-speed pre-ignition in boosted SI engines, SAE International Journal of Engines, 5(3):1216-1228 (2012), doi:10.4272/2012-01-1148
[6] Y. Zhou, W. Li, B. Stump, R. Connatser, S. Lazarevic, J. Qu, Impact of fuel contents on tribological performance of PAO base oil and ZDDP, Lubricants 6(3) (2018), doi:10.3390/lubricants6030079
[7] C. R. Wilke, P. Chang, Correlation of diffusion coefficients in dilute solutions, AIChE Journal 1(2):264-270 (1955), doi:10.1002/aic.690010222
[8] M. A. Siddiqi, K. Lucas, Correlations for prediction of diffusion in liquids, The Canadian Journal of Chemical Engineering 64(5):839-843 (1986), doi:10.1002/cjce.5450640519
[9] C. J. King, L. Hsueh, K. W. Mao, Liquid phase diffusion of non-electrolytes at high dilution, Journal of Chemical & Engineering Data 10(4):348-350 (1965), doi:10.1021/je60027a014
[10] M. T. Tyn, W. F. Calus, Diffusion coefficients in dilute binary liquid mixtures, Journal of Chemical Engineering Data 20(1):106-109 (1975), doi:10.1021/je60064a006
[11] M. Lashkarbolooki, A. Z. Hezave, M. Bayat, Thermal diffusivity of hydrocarbons and aromatics: Artificial neural network predicting model, Journal of Thermophysics and Heat Transfer 31(3):621-627 (2017), doi:10.2514/1.t5041
[12] S. Azizi, E. Ahmadloo, M. M. Awad, Prediction of void fraction for gas-liquid flow in horizontal, upward and downward inclined pipes using artificial neural network, International Journal of Multiphase Flow 87:35-44 (2016), doi:10.1016/j.ijmultiphaseflow.2016.08-004
[13] S. Azizi, M. M. Awad, E. Ahmadloo, Prediction of water holdup in vertical and inclined oil-water two-phase flow using artificial neural network, International Journal of Multiphase Flow 80:181-187 (2016), doi:10.1016/j.ijmultiphaseflow.2015.12.010
[14] L. Pulga, G. M. Bianchi, S. Falfari, C. Forte, A machine learning methodology for improving the accuracy of laminar flame simulations with reduced chemical kinetics mechanism, Combustion and Flame 216:72-81 (2020), doi:10.1016/j.combustflame.2020.02.021
[15] H. Ghanadzadeh, M. Ganji, S. Fallahi, Mathematical model of liquid-liquid equilibrium for a ternary system using the GMDH-type neural network and genetic algorithm, Applied Mathematical Modelling 36(9):4096-4105 (2012), doi:10.1016/j.apm.2011.11.039

[16] S. L. Pandharipane, M. S. Anish, S. Ankit, G. Sagar, Modelling combined VLE of ten binary mixtures using artificial neural networks, Proceedings of the International Conference on Intuitive Systems & Solutions (2012)
[17] M. Mohadesi, G. Moradi, H. S. Mousavi, Estimation of binary infinite dilute diffusion coefficient using artificial neural network, Journal of Chemical and Petroleum Engineering, 48:27-45 (2014)
[18] A. Kendall, Y. Gal, What uncertainties do we need in Bayesian deep learning for computer vision?, in: U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Proceedings of the thirtieth Conference on Neural Information Processing Systems, Curran Associates, Inc., Long Beach, CA, US, 2017, pp. 5574-5584
[19] R. Rohekar, Y. Gurwicz, S. Nisimov, G. Novik, Modelling uncertainty by learning a hierarchy of deep neural connections, in: H. Wallach, A. Beygelzimer, F. d' Alché-Buc, E. Fox, R. Garnett (Eds.), Proceedings of the thirty-third Conference on Neural Information Processing Systems, Curran Associates, Inc., Vancouver, BC, Canada, 2019, 4244-4254
[20] A. D. Kiureghian, O. Ditlevsen, Aleatory or epistemic? does it matter?, Structural Safety 31(2):105-112 (2009), doi:10.1016/j.strusafe.2008.06.020
[21] D. W. Green, M. Z. Southard, Perry's Chemical Engineers' Handbook ninth ed., McGraw Hill, 2018, pp. 93-274
[22] C. L. Yaws, Thermophysical properties of chemicals and hydrocarbons second ed., Gulf Professional Publishing, 2014, pp. 366-389
[23] W. Schotte, Prediction of the molar volume at the normal boiling point, The Chemical Engineering Journal 48(3):167-172 (1992), doi:10.1016/0300-9467(92)80032-6
[24] I. S. Panidi, V. A. Trofimov, N. S. Lepikhina, Calculation of the molar volume of liquid hydrocarbons, Chemistry and technology of Fuels and Oils 42(6):440-445 (2006), doi:10.1007/s10553-006-0104-1
[25] F. Chollet, Keras, https://keras.io (2015)
[26] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moor, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, X. Zheng, Tensorflow: a system for large-scale machine learning, in: Proceedings of the twelfth USENIX Symposium on Operating Systems Design and Implementation, USENIX Association, Savannah, GA, US, 2016, pp. 265-283, arXiv:1605.08695v2
[27] J. Kukačka, V. Golkov, D. Cremers, Regularization for deep learning: a taxonomy, 2017, arXiv:1710.10686

[28] J. Snoek, H. Larochelle, R. P. Adams, Practical Bayesian optimization of machine learning algorithms, in: L. Bottou, F. C. N. Pereira, K. Q. Weinberger, C. J. C. Burges, P. Bartlett (Eds.), Proceedings of the twenty-sixth Conference on Neural Information Processing Systems, Curran Associates, Inc., Lake Tahoe, CA, US, 2012, pp. 2951-2959, arXiv:1206.2944
[29] D. P. Kingma, J. Ba, Adam: a method for stochastic optimization, third International Conference for Learning Representations, 2015, arXiv:1412.6980v9
[30] C. Blundell, J. Cornebise, K. Kavakcuoglu, D. Wierstra, Weight uncertainty in neural networks, in: F. Bach, D. Blei (Eds.), Proceedings of the thirty-second International Conference on Machine Learning, JMLR.org, Lille, FR, 2015, pp. 1613-1622
[31] Z. Liang, L. Chen, M. S. Alam, S. Z. Rezaei, C. Stark, H. Xu, R. M. Harrison, Comprehensive chemical characterization of lubricating oils used in modern vehicular engines utilizing gc x gc-tofms, Fuel 220:792-799 (2018), doi:10.1016/j.fuel.2017.11.142
[32] T. G. Hiss, E. L. Cussler, Diffusion in high viscosity liquids, AIChE Journal 19(4):698-703 (1973), doi:10.1002/aic.690190404

Table 1

Comparison between experimental and calculated NBP molar volumes (cm^3/mol). Relative error in parenthesis calculated as (calc. - exp.)/exp. · 100

Liquid	Experimental	Le Bas [24]	Tyn & Calus [24]	Eq. (2)
Water (H_2O)	18.70	14.80 (-20.85%)	19.34 (3.44%)	18.78 (0.45%)
Acetone (C_3H_6O)	77.50	74.00 (-4.52%)	78.43 (1.20%)	77.47 (-0.03%)
Methanol (CH_4O)	42.50	37.00 (-12.94%)	42.05 (-1.06%)	42.74 (0.56%)
Cyclohexane (C_6H_{12})	117.00	118.20 (1.03%)	116.43 (-0.49%)	116.66 (-0.29%)
n-Heptane (C_7H_{16})	162.00	162.80 (0.49%)	164.71 (1.67%)	162.98 (0.60%)

Single column fitting

Table 2

Mixtures selected for the validation set and their diffusion coefficient

Solvent/solute	**D_{12} x1E9 (m^2/s)**
Butanol ($C_4H_{10}O$) /Oleic-acid ($C_{18}H_{34}O_2$)	0.25
n-Hexadecane ($C_{16}H_{34}$) /n-Decane ($C_{10}H_{22}$)	0.57
n-Hexadecane /n-Octane (C_8H_{18})	0.68
n-Hexadecane /n-Heptane (C_7H_{16})	0.74
n-Hexadecane /n-Hexane (C_6H_{14})	0.85
n-Tetradecane ($C_{14}H_{30}$) /n-Octane	0.84
n-Tetradecane /n-Heptane	0.93
Kerosene /Carbon-tetrachloride (CCl_4)	0.96
n-Tetradecane /Toluene (C_7H_8)	1.02
n-Hexadecane /n-Hexane	1.42

2-column fitting

Table 3

Comparison between standard and hybrid mode plain Neural Networks on different datasets

Dataset	Error metric	Standard mode	Hybrid mode
Train	R^2	0.996	0.987
	MARE (%)	2.77	3.52
	RMSE (%)	6.06	8.10
Test	R^2	0.860	0.953
	MARE (%)	21.53	22.46
	RMSE (%)	23.00	21.00
Validation	R^2	0.914	0.999
	MARE (%)	7.62	5.94
	RMSE (%)	7.90	7.60

Single column fitting

Table 4

Regression line results for the prediction of train, test and validation sets with the hybrid mode

Dataset	R^2	Slope	Intercept
Train	0.987	0.986	0.019
Test	0.953	0.916	0.128
Validation	0.999	1.090	-0.073

2-column fitting

Table 5

Bayesian neural network predictions (diffusion coefficient and its confidence) for mixtures of industrial interest

Solvent/Solute	**Mean x 1E-9 (m^2/s)**	**± 2σ**
Gasoline surr/SAE 30	0.6885	0.1263
Gasoline surr/SAE 10W-30	0.8864	0.2317
Cold lake oil/n-Hexane (C_6H_{14})	0.0255	0.0177
Cold lake oil/n-Heptane (C_7H_{16})	0.0230	0.0166
Cold lake oil/Naphthalene ($C_{10}H_8$)	0.0256	0.0095

Single column fitting

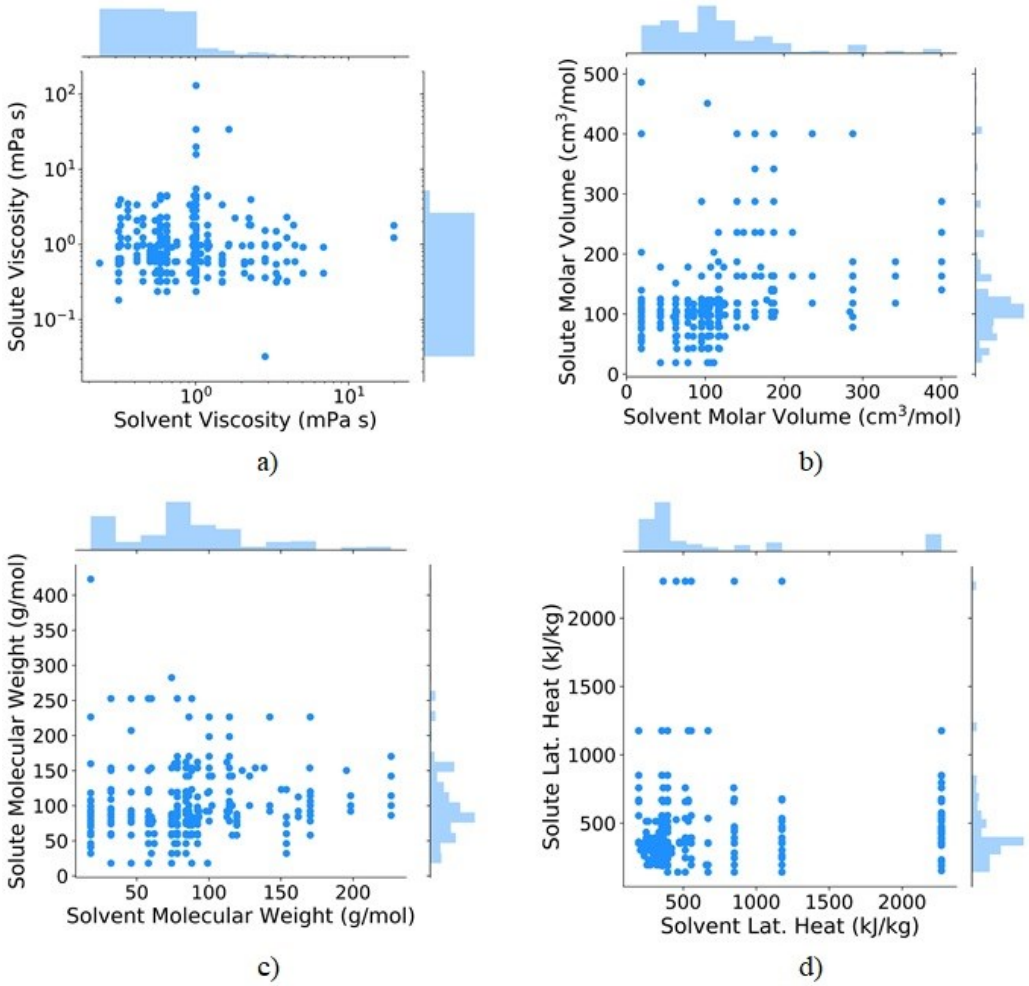

Fig. 1. Dynamic viscosity (a), molar volume (b), molecular weight (c), latent heat of vaporization (d) distribution and frequency in the database

Single column fitting

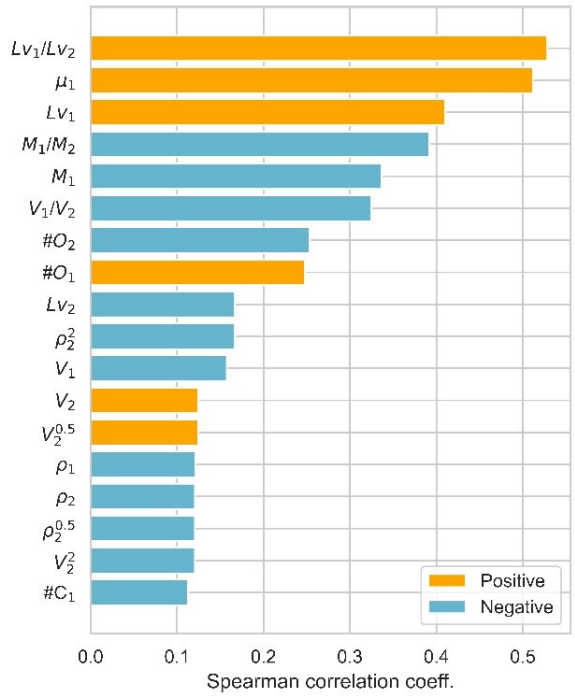

Fig. 2. Results of the Spearman rank-order analysis on the *f* top ranked features

2-column fitting

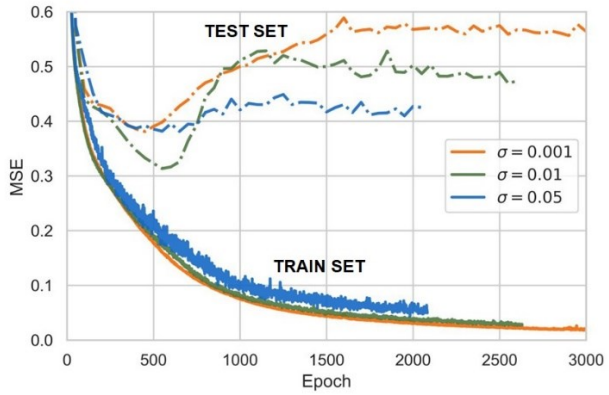

Fig. 3. Mean Squared Error for increasing level of input noise (σ) on train (solid lines) and test (dashed lines) sets

2-column fitting

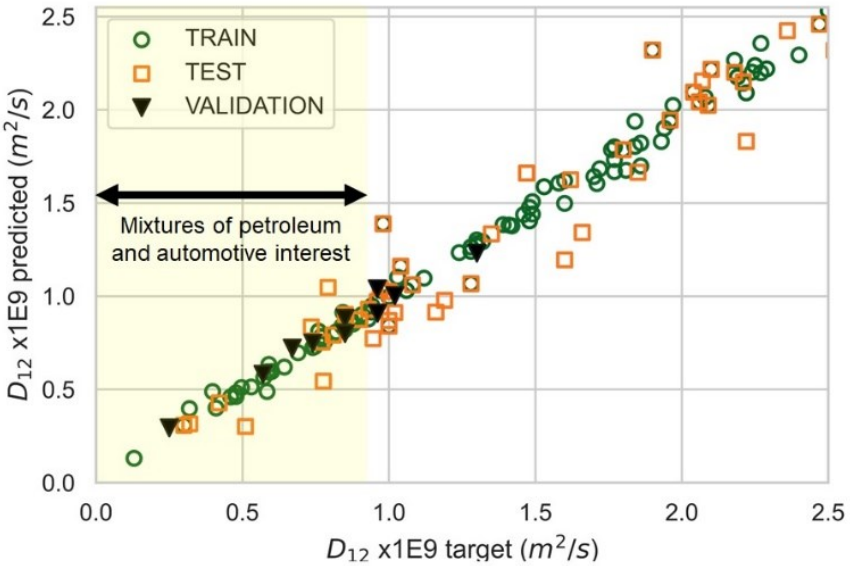

Fig. 4. Linear regression test on the predicted diffusion coefficients on train, test and validation sets

Single column fitting

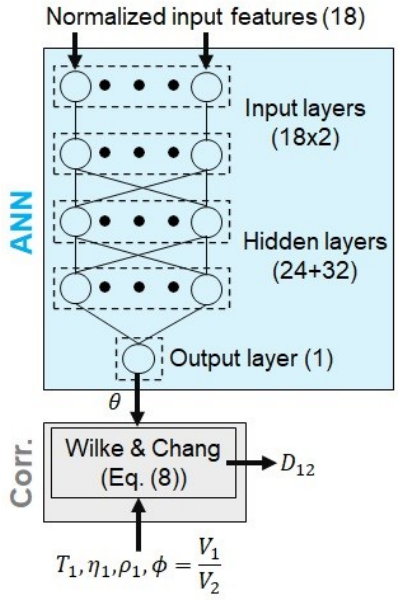

Fig. 5. Scheme of the architecture and the workflow of the hybrid mode Artificial Neural Network

2-column fitting

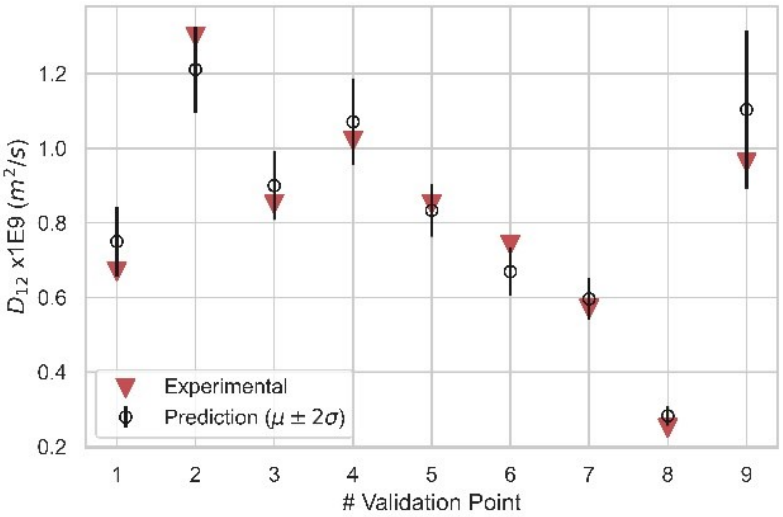

Fig. 6. Validation of the Bayesian Neural Network on the validation set

Single column fitting

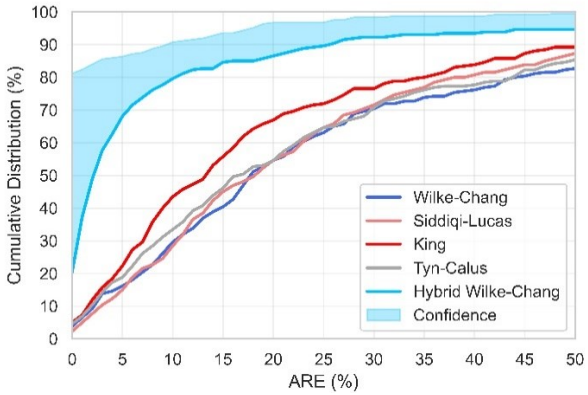

Fig. 7. Comparison of the Absolute Relative Error distribution on the full database with four literature empirical correlations and the Bayesian Neural Network

2-column fitting

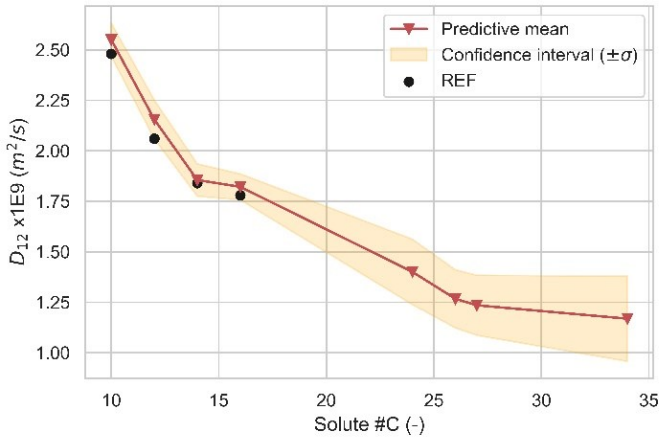

Fig. 8. Bayesian Neural Network prediction and its confidence for different combinations of n-heptane (solvent) with increasing carbons number n-alkanes (solutes)

2-column fitting